



**POLITECHNIKA POZNAŃSKA**  
WYDZIAŁ INŻYNIERII ZARZĄDZANIA



Mgr inż. Maciej Niemir  
**ROZPRAWA DOKTORSKA**

**DOSKONALENIE JAKOŚCI  
ELEKTRONICZNEGO KATALOGU PRODUKTÓW  
Z WYKORZYSTANIEM SZTUCZNEJ INTELIGENCJI**

**Promotor:**

Dr hab. inż. Beata MRUGALSKA, prof. PP

Poznań, 2024



*Serdeczne podziękowania składam  
promotorowi pracy  
Pani dr hab. inż. Beacie Mrugalskiej, prof. PP  
oraz opiekunowi doktoratu wdrożeniowego  
Panu dr inż. Krzysztofowi Muszyńskiemu*

*za nieocenioną pomoc,  
ukierunkowanie badań,  
wsparcie przy realizacji rozprawy doktorskiej  
i ogromną życzliwość,*

*a także mojej Żonie  
za motywację, cierpliwość  
i niegasnącą wiarę we mnie.*



## Spis treści

Streszczenie.....	7
Summary .....	8
1. Wprowadzenie .....	9
1.1. Uzasadnienie podjęcia tematu.....	9
1.2. Przegląd literatury tematu w bazach bibliograficznych.....	11
1.3. Przedmiot i podmiot badań .....	16
1.4. Cel pracy .....	17
1.5. Problem badawczy i zadania badawcze .....	17
1.6. Pytania badawcze.....	18
1.7. Metodyka badań.....	19
2. Przegląd literatury i podstaw teoretycznych.....	20
2.1. Pojęcie jakości.....	20
2.2. Jakość w kontekście danych .....	24
2.3. Wymiary jakości danych.....	26
2.4. Znaczenie jakości danych w ujęciu danych produktowych .....	33
2.5. Przegląd metod pozyskiwania danych o produktach .....	37
2.6. Podstawowe dane produktowe opisujące i identyfikujące produkt .....	41
2.7. Główne źródła błędów w danych produktowych.....	47
2.8. Systemy zarządzania danymi produktowymi i wyzwania w kontekście problematyki jakości .....	52
2.9. Taksonomia sposobów oczyszczania danych produktowych .....	55
2.10. Rola walidacji danych produktowych.....	58
2.11. Przegląd metod walidacji danych .....	60
2.12. Przegląd koncepcji i rozwiązań walidacji danych .....	67
3. Eksploracja standardów i wytycznych w kontekście podstawowych danych produktowych – badania własne .....	86
3.1. Materiały i metoda badania.....	86
3.2. Rezultaty badań własnych.....	89
3.3. Wnioski .....	105
4. Analiza problemów jakościowych w danych produktowych dostępnych online – badania własne.....	106
4.1. Cel badań.....	106

4.2.	Poprawność numerów GTIN .....	107
4.3.	Nazwa produktu w wybranych sklepach elektronicznych .....	109
4.4.	Nazwa produktu/oferty na platformie zakupowej.....	114
4.5.	Nazwy produktów w różnych katalogach produktowych.....	115
4.6.	Zgodność nazwy ze zdjęciem produktu .....	117
4.7.	Wnioski .....	118
5.	Analiza problemu jakości danych produktowych .....	120
5.1.	Audyt danych i wywiad ekspercki .....	120
5.2.	Analiza i ocena możliwości rozwiązania wykrytych błędów w danych.....	123
5.3.	Wnioski .....	140
6.	Opracowanie modelu walidacji danych produktowych .....	142
6.1.	Idea segmentacji względem podejmowanych działań .....	142
6.2.	Omówienie proponowanego modelu .....	145
6.3.	Metodyka oceny wpływu walidatorów na dane w zaproponowanym modelu .	148
7.	Weryfikacja koncepcji modelu w podziale na 3 warstwy .....	153
7.1.	Poziom 1: Normalizacja danych .....	153
7.2.	Poziom 2: Walidacja techniczno - logiczna.....	155
7.3.	Poziom 3: Sugestie poprawy jakości .....	161
7.4.	Ocena wpływu walidatorów na dane w zaproponowanym modelu.....	227
8.	Podsumowanie i kierunki dalszych badań.....	231
8.1.	Zrealizowane zadania badawcze .....	233
8.2.	Odpowiedzi na pytania badawcze.....	235
8.3.	Kierunki dalszych badań.....	237
8.4.	Wdrożenie wyników prac .....	238
	Bibliografia .....	240
	Spis rysunków .....	253
	Spis tabel .....	254

## Streszczenie

Rozprawa doktorska o charakterze teoretyczno-empirycznym dotyczy doskonalenia jakości elektronicznego katalogu produktów z wykorzystaniem sztucznej inteligencji. Celem pracy było opracowanie oraz adaptacja modelu walidacji do zarządzania danymi produktowymi w katalogach elektronicznych, integrującego nowoczesne rozwiązania z zakresu sztucznej inteligencji. Realizację rozpoczęto od analizy literatury przedmiotu oraz eksploracji standardów i praktyk stosowanych w sektorze e-commerce. W ramach eksploracji zbadano wytyczne dotyczące wprowadzania danych produktowych na platformach e-commerce oraz przeanalizowano bazy produktowe katalogów i sklepów internetowych. Na podstawie analizy największej bazy unikalnych produktów oznaczonych polskimi numerami GTIN zidentyfikowano kluczowe wyzwania związane z walidacją podstawowych danych o produktach. Opracowano strukturę wielopoziomowego modelu walidacji, którego celem była integracja tradycyjnych metod walidacji regułowych z technikami bazującymi na sztucznej inteligencji. Wyselekcjonowano i zaadaptowano 45 algorytmów walidacyjnych, odpowiadających na zidentyfikowane problemy oraz mieszczących się w opracowanej strukturze, szczególną uwagę poświęcając możliwościom wykorzystania sztucznej inteligencji do ich budowy. Rozwiązania przetestowano w warunkach rzeczywistych, ostatecznie potwierdzając ich możliwy wpływ na doskonalenie jakości danych w katalogach produktowych.

Opracowany model walidacji danych produktowych wraz z przedstawionymi przykładami stanowi krok w kierunku implementacji zaawansowanych technologii w procesach biznesowych, zgodnie z założeniami Jakości 4.0. W efekcie, rozwiązanie wspiera innowacyjność oraz konkurencyjność przedsiębiorstw w erze cyfrowej, umożliwiając im lepsze dostosowanie się do dynamicznie zmieniających się wymagań rynku.

## Summary

The doctoral dissertation, theoretical and empirical in nature, focuses on improving the quality of electronic product catalogs using artificial intelligence. The objective of the work was to develop and adapt a validation model for managing product data in electronic catalogs, integrating modern AI solutions. The research began with a literature review and exploration of standards and practices used in the e-commerce sector. As part of the exploration, guidelines for entering product data on e-commerce platforms were examined, and product databases from catalogs and online stores were analyzed. Based on the analysis of the largest database of unique products marked with Polish GTIN numbers, key challenges related to the validation of basic product data were identified. A multi-level validation model structure was developed, aimed at integrating traditional rule-based validation methods with AI-based techniques. A selection of forty-five validation algorithms was adapted, addressing the identified problems, and fitting within the developed structure, with particular attention paid to the potential use of AI in their development. The solutions were evaluated in real-world conditions, confirming their potential impact on improving data quality in product catalogs.

The developed product data validation model, along with the provided examples, represents a step toward the implementation of advanced technologies in business processes, in line with the principles of Quality 4.0. As a result, the solution supports innovation and competitiveness of enterprises in the digital age, enabling them to better adapt to the dynamically changing market demands.



# 1. Wprowadzenie

## 1.1. Uzasadnienie podjęcia tematu

We współczesnej erze cyfrowej, gdzie globalny rynek e-commerce przechodzi przez fazę intensywnego rozwoju, wymiana informacji cyfrowych staje się nieodłącznym elementem współczesnego handlu. Autorzy badań naukowych i przedstawiciele przemysłu zgadzają się, że zachowanie dobrej jakości danych stanowi kluczowy problem w organizacjach a ich brak może mieć znaczący negatywny wpływ na biznes (Wang i Strong, 1996; Ballou i in., 2004; Haug i in., 2011). Przez jakość danych rozumiemy przede wszystkim dokładność, kompletność, aktualność i spójność informacji o produktach prezentowanych w e-katalogach (Wang i Strong, 1996). Jakość danych stała się kluczowym elementem determinującym sukces lub porażkę wielu przedsiębiorstw (Cao i Zhang, 2011), mając znaczący wpływ na efektywność transakcji biznesowych (Hole i in., 2018). Zaniedbania w tym zakresie, niespójne, niejasne, niepełne lub błędne informacje mogą prowadzić do problemów w zakresie satysfakcji klienta, logistyki i ostatecznej sprzedaży produktów online, a przede wszystkim negatywnie wpływać na wiarygodność firmy (Qalati i in., 2021) i generować niepotrzebne koszty (Appelbaum i in., 2017). Deficyt w zarządzaniu jakością danych może mieć więc poważne konsekwencje finansowe dla przedsiębiorstw (Haug i in., 2011). W poprzedniej dekadzie badania dotyczyły głównie wymiany i jakości danych produktowych w wewnętrznych i międzyorganizacyjnych scenariuszach łańcucha dostaw (Schemm i Legner, 2008; Hüner i in., 2011; Otto, 2015). Zarządzanie danymi już wtedy było niepraktyczne i szukano zautomatyzowanych rozwiązań (Russom, 2011). Pandemia COVID-19 przyspieszyła proces wdrażania nowoczesnych technologii informacyjnych, w tym aplikacji e-commerce, co nie tylko umożliwiło firmom przetrwanie kryzysu, ale również zdobycie przewagi konkurencyjnej (Jelonek, 2023). To jednak znacznie zwiększyło zapotrzebowanie na dane. W obecnych warunkach, przy gwałtownym wzroście skali handlu elektronicznego i zbiorach danych często przekraczających miliony rekordów, ręczne zarządzanie nimi stało się po prostu niemożliwe (Cao, 2023; Russom, 2011). Złożoność tego problemu potęgowana jest dodatkowo brakiem jednolitej standaryzacji w obszarze katalogowania danych produktowych a także brakiem wiarygodnego źródła danych o produktach konsumenckich, które byłoby dostępne, kompletne i przydatne dla zastosowań B2C (ang. business to consumer). Dostępne źródła są ukierunkowane na wymagania łańcucha dostaw, obejmują tylko podzbiór dostępnych produktów, a dostęp do danych wymaga indywidualnych umów z każdą firmą, co nie zmienia się od dwóch dekad (Nakatani i in., 2006; Schemm i in., 2007; Schemm i Legner, 2008). Dostawcy aplikacji skierowanych do

konsumentów zmuszeni są agregować dane z wielu źródeł o różnych formatach, schematach i jakości, lub wprowadzać dane manualnie z etykiet produktów co namnaża problemów z utrzymaniem jakości (Karpischek i in., 2014).

W odpowiedzi na rosnące wyzwania na światowym rynku e-commerce, globalna organizacja GS1, twórca i propagator globalnych standardów identyfikacji i kodowania produktów, podjęła ambitny krok w kierunku opracowania uniwersalnego, wzorcowego katalogu wszystkich produktów wytwarzanych na świecie w formie rejestru podstawowych danych, tzw. Global Registry Platform (GS1.pl, 2023). Kluczowym założeniem tego przedsięwzięcia jest, aby to producenci byli bezpośrednio odpowiedzialni za dostarczanie oraz za jakość danych dotyczących oferowanych przez nich produktów. Jednakże, zebranie i utrzymanie aktualnego katalogu o tak ogromnej skali, gdzie liczne atrybuty produktów są rejestrowane w formie cyfrowej, tekstowej lub obrazowej, napotkało ponownie na wyzwania związane z jakością, z identycznego powodu jak już opisany. Rozproszona forma i wysoka różnorodność zbieranych danych, otwarta dla szerokiego grona użytkowników powoduje, że obecnie każdy z przedstawicieli firm wprowadza dane we własny, nieuporządkowany globalnie sposób. Ze względu na skalę (np. w Polsce jest to kilkadziesiąt milionów produktów) dane nie mogą zostać poddane manualnej weryfikacji.

W świetle powyższego, zaistniała pilna potrzeba poszukiwania innowacyjnych rozwiązań, które, z wykorzystaniem nowoczesnych technik takich jak sztuczna inteligencja, zwiększą wykrywanie błędów i potencjalnie zautomatyzują proces doskonalenia jakości danych. Szczególnie generatywna sztuczna inteligencja, dzięki swoim zdolnościom do realizacji skomplikowanych zadań, może stanowić kluczowy element tego procesu, umożliwiając tworzenie precyzyjnych sugestii poprawek oraz usprawnienie procesów walidacji. To wyzwanie stało się bezpośrednią motywacją do rozpoczęcia badań, które doprowadziły do powstania niniejszej rozprawy doktorskiej o charakterze wdrożeniowym.

Warto podkreślić, że przedstawione w niej rozwiązania nie zostały ograniczone jedynie do teoretycznej analizy - są one proponowane jako praktyczne narzędzia (tzw. walidatory), które mogą znaleźć zastosowanie w doskonaleniu jakości dowolnego katalogu. Dzięki temu niniejsze badania, nie tylko mają wpływ na rozwój wiedzy naukowej w zakresie poprawy jakości danych, ale przyczyniają się także do rzeczywistych zmian na rynku e-commerce.

## 1.2. Przegląd literatury tematu w bazach bibliograficznych

Jakość danych produktowych oraz metody jej poprawy z użyciem różnych metod walidacji w e-katalogach stanowi wąski obszar zagadnień badawczych. W ramach przeglądu literaturowego podjęto próbę rozpoznania tej tematyki, koncentrując się na problematyce szeroko pojętej walidacji danych produktowych w katalogach elektronicznych. Szczególnie interesujące były nowoczesne metody oparte o sztuczną inteligencję, które mogą oferować skuteczniejsze rozwiązania w zakresie walidacji, w porównaniu do klasycznego podejścia bazującego na prostych regułach deterministycznych. W celu dokładnego zrozumienia istniejącej problematyki oraz stanu wiedzy na ten temat, przeprowadzono przeszukiwanie w czterech bazach bibliograficznych: **Google Scholar**, **IEEE Xplore**, **Scopus** i **Web of Science**. Pierwsza z nich, **Google Scholar**, to wszechstronne źródłem literatury naukowej, które obejmuje publikacje z różnych dziedzin i pozwala na szybkie zlokalizowanie kluczowych prac w omawianym zakresie. Oprócz renomowanych czasopism i materiałów konferencyjnych, indeksuje także prace dyplomowe i raporty techniczne. Drugie źródło, **IEEE Xplore**, jest skoncentrowane na literaturze z zakresu informatyki i inżynierii. Bazując na tej bazie, zwrócono szczególną uwagę na prace dotyczące technologii i metod związanych z przetwarzaniem danych oraz sztuczną inteligencją. Kolejna przeszukiwana baza, **Scopus**, jest jednym z najważniejszych międzynarodowych źródeł literatury naukowej, która dostarcza kompleksowy przegląd prac z różnych dyscyplin, w tym zarządzania, informatyki i technologii. Ostatnia w omawianym zestawieniu, **Web of Science**, to cenione narzędzie bibliograficzne, które integruje literaturę z różnych dziedzin naukowych, w tym nauk przyrodniczych, społecznych i humanistycznych. Indeksuje przede wszystkim renomowane czasopisma naukowe i konferencje. Zapewnia dokładną kategoryzację oraz narzędzia do analizy cytowań. Dzięki takiemu połączeniu, możliwe było uzyskanie dokładnego obrazu aktualnych badań i trendów w omawianym obszarze. W celu rzetelnego odwzorowania zakresu poszukiwanych treści, w zapytaniach kierowanych do baz bibliograficznych połączono terminologię z dwóch kluczowych obszarów:

- A. Terminy odnoszące się do aspektów przetwarzania i zarządzania danymi, gdzie wspólnym mianownikiem jest zapewnienie jakości danych:
  - a. **"Data validation", "validation rules" / "walidacja danych", "reguły walidacji"**, odnoszą się do procesu sprawdzania czy dane są poprawne i spełniają określone kryteria.

- b. **"Data cleaning" / "czyszczenie danych"** obejmują usuwanie, poprawianie lub wypełnianie brakujących lub nieprawidłowych danych.
  - c. **"Data verification" / "weryfikacja danych"** dotyczą sprawdzania poprawności i spójności danych.
  - d. **"Data normalization" / "normalizacja danych"** to proces przekształcania danych do wspólnej skali lub formatu, co jest kluczowe dla wielu zastosowań, takich jak analiza danych czy uczenie maszynowe.
  - e. **"Data anomalies" / "anomalie w danych"** dotyczą wykrywania niespójności, błędów lub nietypowych wartości w danych, które mogą wskazywać na błąd lub oszustwo.
- B. Słowa kluczowe skoncentrowane na specyficznym zastosowaniu danych - w kontekście katalogów produktów:
- a. **"Electronic product catalogs" / "elektroniczne katalogi produktów"** odnoszą się do baz danych zawierających informacje o produktach dostępnych online.
  - b. **"Product data" / "dane produktów"** to ogólny termin odnoszący się do wszelkich danych związanych z produktami, takich jak ich nazwy, opisy, ceny itp.
  - c. **"Product master data" / „główne dane produktu”** odnoszą się do centralnej bazy danych zawierającej wszystkie kluczowe informacje o produktach, które są konsekwentnie i spójnie używane w całej organizacji.

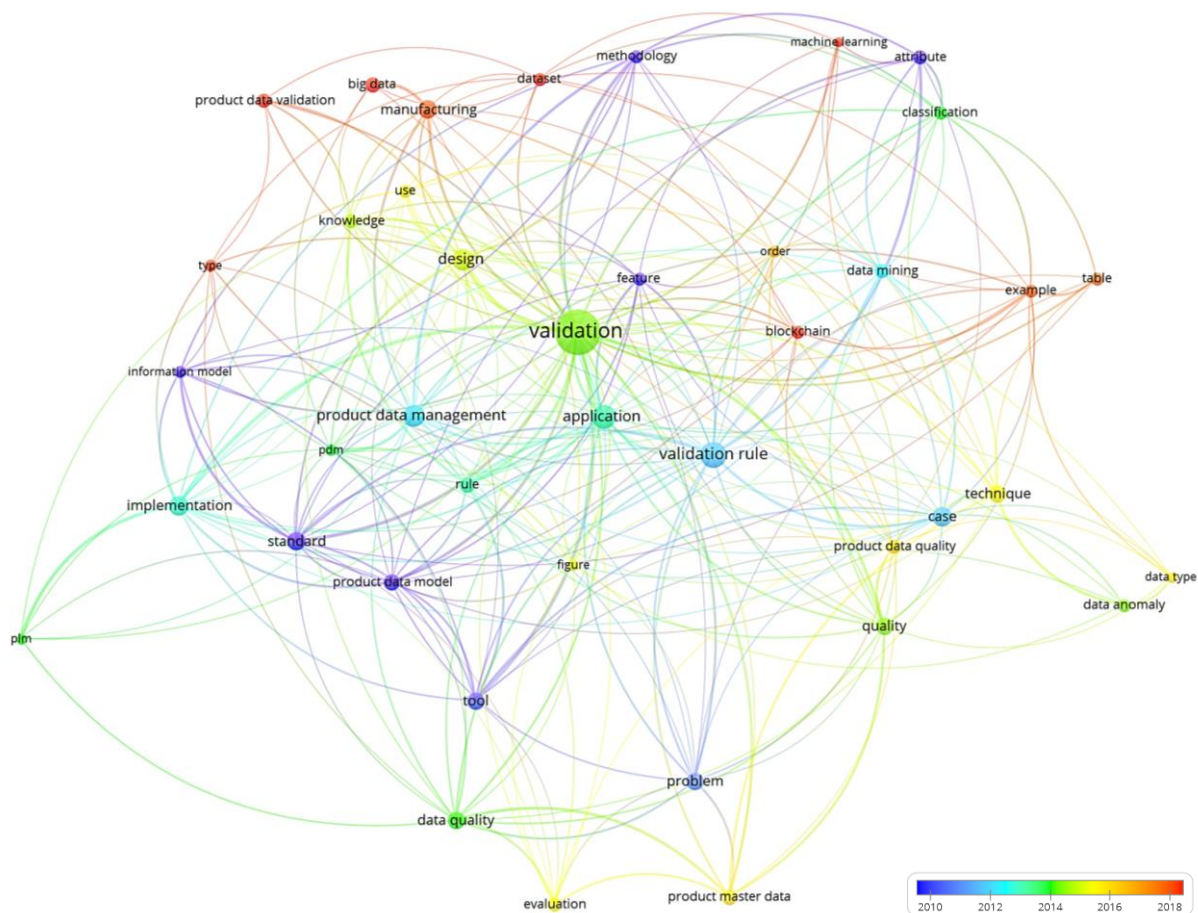
Kryteria oraz wyniki wyszukiwania odpowiednio złączonych haseł w wytypowanych bazach przedstawiono w Tabeli 1. Składnia zapytania umieszczona w tabeli ma charakter poglądowy, samo zapytanie do bazy skonstruowano dostosowując je do konkretnych wymogów wyszukiwawczych baz danych. Przeszukiwanie baz bibliograficznych zostało ograniczone do publikacji z ostatnich 10 lat, ponieważ technologie i metodyki w dziedzinie zarządzania danymi produktowymi ewoluują szybko, a starsze źródła mogą nie odzwierciedlać aktualnego stanu wiedzy i praktyk w tej dziedzinie. W analizie bibliograficznej zwrócono uwagę na najnowsze trendy i osiągnięcia w tej dziedzinie. Ostatnia dekada przyniosła znaczący postęp w zakresie wykorzystania uczenia maszynowego w różnych obszarach, oferując znacznie bardziej zaawansowane rozwiązania niż tradycyjne metody.

Tabela 1. Wyniki wyszukiwania poszczególnych haseł w bazach (stan na dzień 2022-01-02)

Kryteria wyszukiwania (hasła)	Scopus	Google Scholar	Web of Science	IEEE Xplore
("data validation" OR "data cleaning" OR "data verification" OR "data normalization" OR "data anomalies") AND ("product catalog" OR "product data" OR "product master data")	21	999	89	19
("validacja danych" OR "czyszczenie danych" OR "weryfikacja danych" OR "normalizacja danych" OR "anomalie w danych") AND ("katalogi produktów" OR "dane produktu" OR "główne dane produktu")	0	1	0	0

Źródło: opracowanie własne.

Wyniki wyszukiwania poszczególnych haseł w bazach zaprezentowano na Rysunku 1 w postaci wizualizacji sieci powiązań, z oznaczeniem najważniejszych klastrów. Kolorami oznaczono daty publikacji od niebieskiego - rok 2012, do czerwonego - rok 2022.

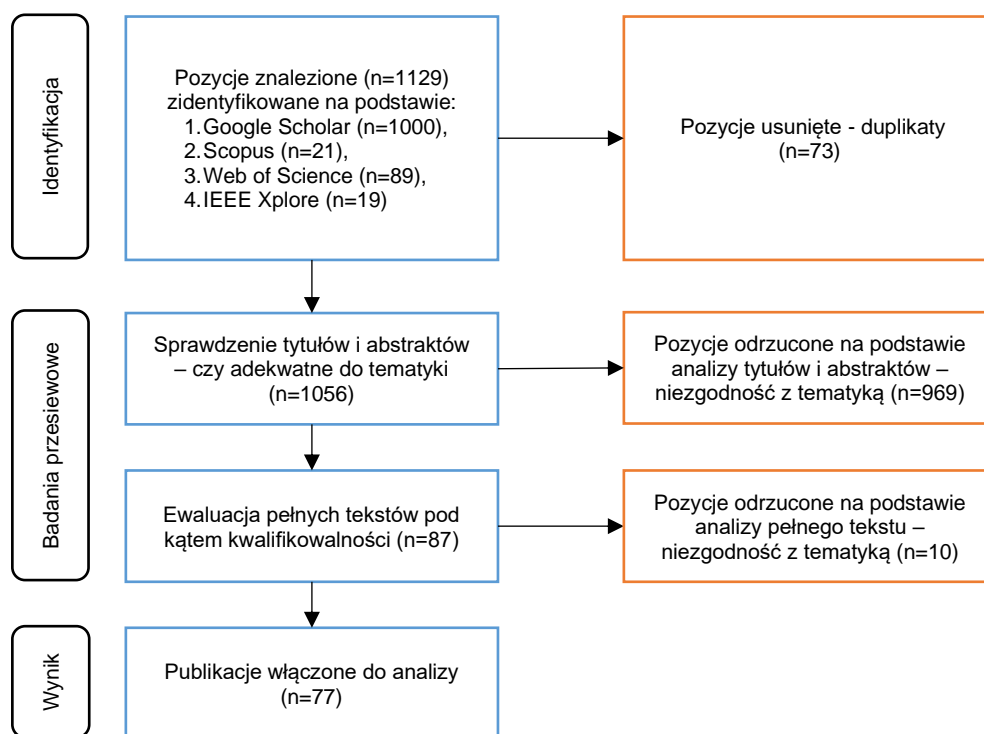


Rysunek 1. Wizualizacja słów kluczowych i powiązań publikacji  
Źródło: opracowanie własne

Słowa kluczowe do budowy wizualizacji zostały wyekstrahowane z tytułów i abstraktów publikacji przeszukiwanych publikacji. Na ich podstawie można dokonać kilku obserwacji i wniosków:

- A. Słowo "validation" (walidacja) znajduje się w centralnej części sieci, co sugeruje, że jest to główny temat lub kluczowe pojęcie, wokół którego koncentrują się inne tematy. Jest ono połączone z wieloma innymi pojęciami, co wskazuje na jego znaczenie w kontekście przedstawionych badań.
- B. Istnieją pewne kluczowe kategorie powiązane z walidacją, takie jak "product data management", "data quality", "validation rule" i "product data model". Te kategorie mają wiele połączeń, co wskazuje na ich znaczenie w kontekście walidacji danych produktowych.
- C. Kolorowe linie wskazują na różne lata, od 2011 do 2021. Można zauważyć, że niektóre tematy stały się bardziej popularne w późniejszych latach, co może sugerować ewolucję w badaniach i zainteresowaniach naukowców.
- D. Pojęcia takie jak "machine learning" (uczenie maszynowe), "big data" i "blockchain" wskazują, że badania w tej dziedzinie ewoluowały w kierunku nowych technologii i metodologii w ciągu ostatniej dekady.
- E. Sieć pokazuje złożoność relacji między różnymi kategoriami i jak są one ze sobą powiązane. Na przykład, "product data quality" jest powiązane zarówno z "validation" jak i "data anomaly", co wskazuje na złożoność zarządzania jakością danych produktowych.
- F. Nie występują słowa kluczowe związane ze społecznościowym charakterem bazy produktów, czy też z crowdsourcingowym modelem zarządzania danymi.

Przegląd publikacji (Rysunek 2) przeprowadzono przy wykorzystaniu metody PRISMA (Page i in., 2021).



Rysunek 2. Proces selekcji publikacji zgodnie z metodą PRISMA  
Źródło: opracowanie własne

Zidentyfikowano szereg kluczowych publikacji, które koncentrują się na zagadnieniach walidacji, weryfikacji oraz normalizacji danych w kontekście katalogów elektronicznych i zarządzania danymi produktowymi. Wiele z tych prac podkreśla znaczenie jakości danych w biznesie, szczególnie w sektorze e-commerce, wskazując na potencjalne negatywne skutki niewłaściwie zarządzanych danych, takie jak błędy w decyzjach biznesowych, straty finansowe, czy obniżenie zaufania klientów. Niektóre z tych publikacji prezentują innowacyjne techniki i narzędzia do automatycznego wykrywania anomalii w danych, wykorzystując zaawansowane metody analizy, takie jak uczenie maszynowe czy analiza wzorców. Inne skoncentrowane są na znaczeniu spójnej klasyfikacji produktów i wyzwaniach związanych z integracją danych pochodzących z różnych źródeł. Niewiele uwagi poświęcono jednak ocenie jakości danych w rozumieniu np. semantycznej analizy tekstu czy określenia przydatności zdjęć multimedialnych za pomocą sztucznej inteligencji, czyli walidacji pól, których interpretacja wymaga zaawansowanego zrozumienia kontekstu oraz analizy treści. Tego rodzaju pola, jak nazwy i opisy produktów czy zdjęcia produktów, stanowią bardzo istotny element każdego katalogu elektronicznego. Wśród zidentyfikowanych prac nie znaleziono także tych, które analizują specyfikę polskiego rynku e-commerce.

Przegląd literatury pozwolił na głębsze zrozumienie istniejących rozwiązań, narzędzi i praktyk w dziedzinie zarządzania danymi produktowymi. Na podstawie przeprowadzonej analizy literatury oraz wyników wyszukiwania publikacji naukowych, zaobserwowano **lukę poznawczą** dotyczącą kompleksowego podejścia do automatycznej walidacji danych w katalogach elektronicznych. Choć wielu Autorów analizowanych badań koncentruje się na poszczególnych aspektach czyszczenia danych czy identyfikacji anomalii, niewielu z nich oferuje zintegrowane podejście uwzględniające zarówno metody regułowe, jak i techniki bazujące na sztucznej inteligencji, istotne w podejściu do Jakości 4.0 (Wawak, 2022; Zonnenshain i Kenett, 2020). Ponadto, brakuje prac omawiających adaptację metod walidacyjnych do danych produktowych opisywanych w języku polskim oraz prac, które obejmowałyby wszystkie rodzaje produktów produkowanych na rynki bez względu na branżę. W związku z tym istnieje potrzeba opracowania struktury modelu walidacji danych produktów w elektronicznych katalogach, który uwzględniłby możliwość zastosowania sztucznej inteligencji w procesie walidacji danych, a także zaprojektowania i weryfikacji konkretnych propozycji rozwiązań walidacyjnych, które pomogą w udoskonalaniu jakości danych produktów w katalogach.

### **1.3. Przedmiot i podmiot badań**

**Przedmiotem badań** podjętych w niniejszej rozprawie doktorskiej było opracowanie modelu automatycznej walidacji danych produktowych uwzględniającego sztuczną inteligencję, wspierającego doskonalenie jakości danych zawartych w katalogach elektronicznych. Model ten bazował na zróżnicowanych algorytmach walidacyjnych, które wspierały weryfikację podstawowych atrybutów danych produktowych na różnych poziomach walidacji. Zakres badań obejmował identyfikację najczęstszych problemów związanych z zarządzaniem danymi produktowymi, opracowanie schematu modelu walidacji oraz zaprojektowanie i analizę wybranych algorytmów walidacyjnych. Główną ideą, oprócz implementacji niezbędnych algorytmów deterministycznych będących podstawą walidacji, było zbadanie możliwości wykorzystania sztucznej inteligencji (SI) jako uzupełnienia funkcji walidacyjnej w katalogach elektronicznych.

Prace zrealizowano w ramach programu doktoratu wdrożeniowego Ministerstwa Nauki i Szkolnictwa Wyższego. Celem programu było rozwiązanie rzeczywistego problemu przedsiębiorstwa. Zaproponowany w niniejszej dysertacji model walidacyjny jak i wybrane



walidatory zostały wdrożone. W rozdziale 8.4 przedstawiono działania mające na celu dalszą komercjalizację wyników badań, wskazując na potencjalne korzyści biznesowe oraz kolejne możliwości wdrożenia opracowanych rozwiązań.

**Podmiotem badań** był elektroniczny katalog danych produktowych Fundacji GS1, w którym dane pochodzą bezpośrednio od producentów i właścicieli marek, korzystających z kodów kreskowych do znakowania swoich produktów. Katalog zawiera kilkadziesiąt milionów unikalnych danych produktów, jest uzupełniany i aktualizowany przez około 60 tysięcy firm, w tym producentów i bezpośrednich dystrybutorów produktów na rynek polski. Struktura danych produktowych obejmuje elementy takie jak: nazwa produktu, marka, zdjęcie, unikalny kod GTIN (Global Trade Item Number), zawartość netto oraz klasyfikacja produktu - GPC (Global Product Classification). Ze względu na skalę zgromadzonych danych oraz miesięczny przyrost nowych wpisów, ich manualna weryfikacja przez administratorów katalogu jest niemożliwa.

#### **1.4. Cel pracy**

**Celem rozprawy doktorskiej** było udoskonalenie jakości danych produktowych poprzez opracowanie i adaptację modelu walidacji, który uwzględnił możliwość korzystania z nowoczesnych rozwiązań z zakresu sztucznej inteligencji. Oczekiwany **rezultatem** było dokładniejsze odwzorowanie rzeczywistości we wprowadzanych danych produktowych, przyczyniające się do spełnienia oczekiwań rynku w zakresie precyzji informacji w katalogu.

#### **1.5. Problem badawczy i zadania badawcze**

Analiza literatury przedmiotu ujawniła lukę poznawczą w zakresie kompleksowego podejścia do automatycznej walidacji danych produktowych w elektronicznych katalogach, które łączyłoby metody regułowe z technikami bazującymi na sztucznej inteligencji. Choć sztuczna inteligencja dynamicznie się rozwija, to jej zastosowanie w kontekście walidacji danych niesie za sobą ryzyko związane z popełnianiem błędów i odrzucaniem poprawnych danych. Z drugiej strony, może jednak znacząco wspomóc proces identyfikacji problematycznych obszarów pełniąc rolę asystenta, tzw. „co-pilota” dla osób wprowadzających informacje. Tym samym sztuczna inteligencja może stanowić narzędzie wspierające doskonalenie jakości danych produktowych, umożliwiając identyfikację potencjalnych

nieprawidłowości i sugerowanie poprawek. W związku z tym sformułowano następujący problem badawczy: **Jak można udoskonalić jakość danych produktowych bazując na modelu walidacji, który wykorzysta potencjał sztucznej inteligencji?**

W trakcie prac nad dysertacją sformułowano następujące **zadania badawcze**:

**Zadanie 1.** Analiza literatury przedmiotu.

**Zadanie 2.** Eksploracja standardów i analiza sektora e-commerce w kontekście przyjętych norm i praktyk w doniesieniu do procesu wprowadzania danych o produktach.

**Zadanie 3.** Identyfikacja wyzwań związanych z walidacją danych na podstawie analizy bazy danych, będącej przedmiotem badań.

**Zadanie 4.** Opracowanie struktury wielopoziomowego modelu walidacji.

**Zadanie 5.** Selekcja, adaptacja i ocena algorytmów walidacyjnych do zastosowania w zaproponowanym modelu walidacji.

Rozważania przedstawione w rozprawie doktorskiej wpisują się w dziedzinę nauk społecznych oraz dyscyplinę nauki o zarządzaniu i jakości. W dysertacji zapoznano się i skorzystano z 229 pozycji literaturowych, które posłużyły jako podstawa dla rozważań teoretycznych oraz praktycznych. Wykazano także potrzebę dalszej kontynuacji badań w obszarze doskonalenia jakości danych w e-katalogach produktów za pomocą technik walidacyjnych.

## **1.6. Pytania badawcze**

W realizacji celu rozprawy doktorskiej pomocne były następujące **pytania badawcze**:

**Pytanie 1.** Jakie standardy i praktyki są stosowane w sektorze e-commerce w kontekście procesu wprowadzania danych o produktach?

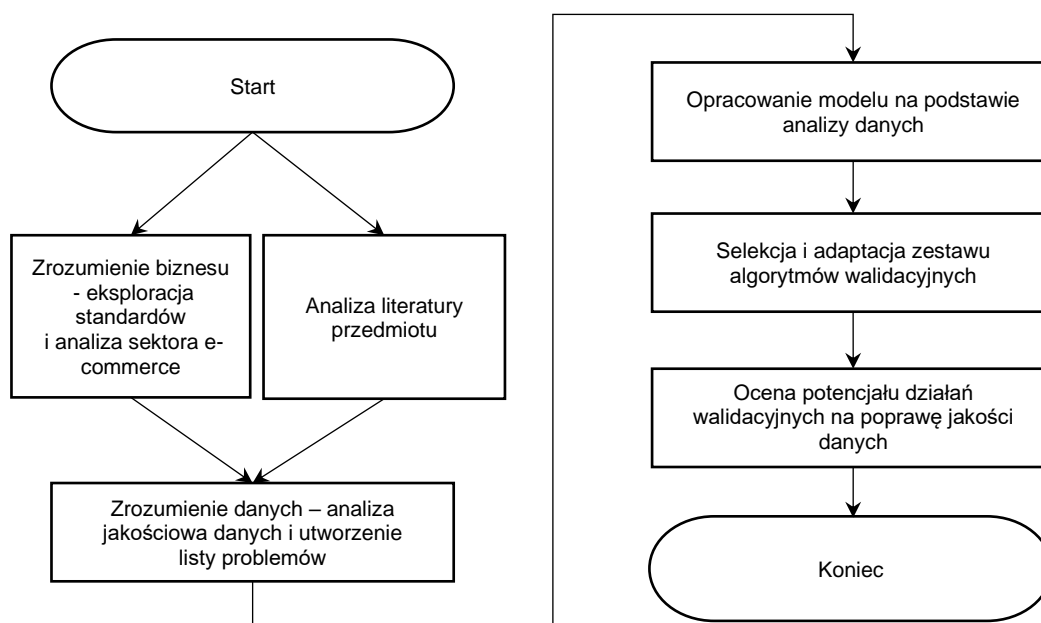
**Pytanie 2.** Jakie problemy jakościowe można znaleźć w danych produktowych w elektronicznych katalogach?

**Pytanie 3.** Jaka struktura modelu walidacji danych produktowych zapewni skorzystanie z metod sztucznej inteligencji w elektronicznych katalogach?

**Pytanie 4.** Jakie algorytmy walidacyjne mogą pozytywnie wpłynąć na doskonalenie jakości danych produktów w elektronicznych katalogach?

## 1.7. Metodyka badań

Jako typ badań wybrano badania stosowane w sekwencji **od jakościowej do ilościowej**, co umożliwiło głębokie zrozumienie problemu poprzez poznanie jego natury (**jakościowe**) a następnie kwantyfikację wyników, oraz ewaluację rozwiązań (**ilościowe**). Każde z pytań badawczych było przedmiotem odrębnej analizy. Schemat realizacji pracy przedstawiono na Rysunku 3. Tworząc go, skorzystano z kluczowych elementów ramy modelu CRISP-DM (Cross Industry Standard Process for Data Mining) służącego do wydobywania wiedzy z danych (Chapman, 2000).



Rysunek 3. Schemat realizacji pracy  
Źródło: opracowanie własne

Schemat składa się z kilku kluczowych etapów. Pierwszym było „Zrozumienie biznesu”. Na tym etapie przeprowadzono eksplorację standardów oraz analizę sektora e-commerce, aby zrozumieć kontekst biznesowy oraz istniejące standardy dotyczące danych produktowych. Analiza ta pozwoliła na identyfikację kluczowych wymagań i problemów, które mają bezpośredni wpływ na jakość danych produktowych. Równocześnie przeprowadzono analizę literatury przedmiotu. Celem tego etapu było zbadanie istniejących prac naukowych i technicznych, które są związane z problematyką jakości danych produktowych. Analiza literatury pozwoliła na identyfikację luk badawczych oraz najlepszych praktyk stosowanych w dziedzinie zarządzania jakością danych. Kolejnym etapem była analiza jakościowa danych zakończona utworzeniem listy zidentyfikowanych problemów związanych z jakością danych produktowych (etap tzw. „zrozumienia danych”). Na podstawie zebranych informacji

opracowano własny model walidacji danych produktowych. Następnie przeprowadzono „Selekcję i adaptację zestawu algorytmów walidacyjnych”. W tym zakresie, na podstawie wcześniej zidentyfikowanych problemów wybrano najbardziej odpowiednie algorytmy do walidacji danych produktowych. Zostały one dostosowane do specyficznych potrzeb walidacji, a ich działanie było iteracyjnie testowane i optymalizowane. Ostatnim etapem była ocena potencjału działań walidacyjnych na poprawę jakości danych. Na tym etapie zbadano realny lub potencjalny wpływ każdego z walidatorów na udoskonalenie danych produktów podmiotu badań.

## **2. Przegląd literatury i podstaw teoretycznych**

### **2.1. Pojęcie jakości**

Jakość to koncepcja złożona i wielowymiarowa. Jest ona interpretowana na wiele sposobów zarówno w życiu codziennym, w środowisku akademickim jak i w przemyśle (Charantimath, 2011; Zhang, 2001). Trudności w jej definiowaniu wynikają z wielu czynników, takich jak złożoność pojęcia, różnorodność potrzeb klientów, dynamiczne zmiany w technologii, a także ewolucja świadomości i praktyk zarządzania jakością (Kindlarski, 1988; Martin i in., 2020; Mroczko, 2012). Kindlarski podkreśla, że jedna uniwersalna definicja jakości jest niemożliwa do osiągnięcia, dopóki postęp technologiczny będzie się rozwijał, a różnorodność perspektyw będzie wzbogacać jej rozumienie.

Począwszy od lat 30. XX wieku, kiedy Walter A. Shewhart (1931) położył fundamenty pod statystyczną kontrolę jakości, podkreślając istotę redukcji zmienności w procesach produkcyjnych, pojęcie jakości zaczęło przybierać coraz bardziej złożone formy. Shewhart rozpoznał, że jakość nie jest stała i wymaga ciągłego monitorowania i dostosowywania. W latach 50. i 60. Joseph M. Juran i Philip B. Crosby rozwinęli koncepcję jakości, skupiając się odpowiednio na "przydatności do użytku" (Juran, 2003) i "zgodności ze specyfikacjami" (Crosby, 1967). Juran podkreślał, że produkt musi służyć potrzebom klienta, a jakość powinna być planowana i zarządzana. Crosby natomiast zdefiniował jakość jako spełnienie wytyczonych specyfikacji i promował ideę "zero defektów", która miała prowadzić do obniżenia kosztów i zwiększenia zadowolenia klienta.

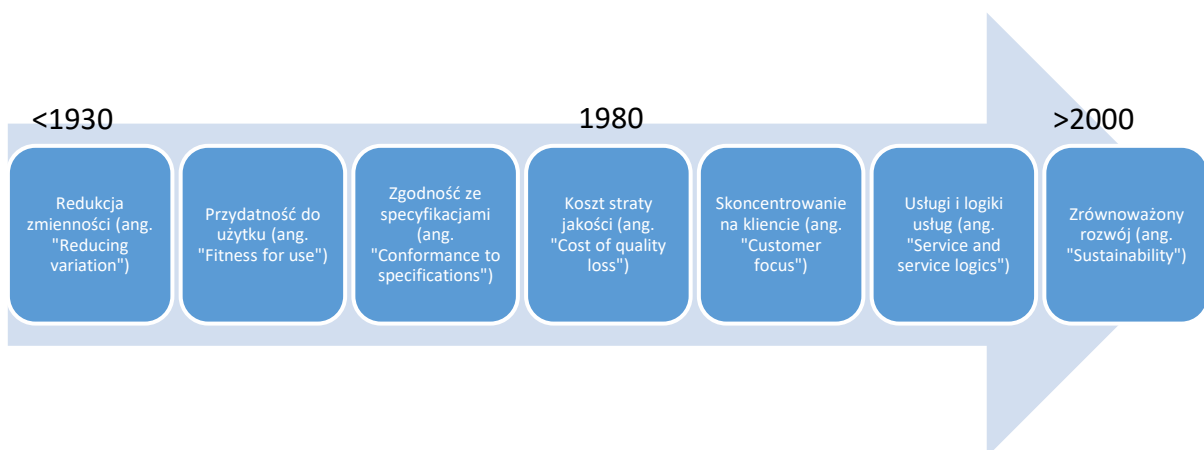
W latach 70. i 80. Genichi Taguchi (1986) wprowadził innowacyjną perspektywę na jakość, skupiając się na "koszcie straty jakości" i wpływie, jaki produkt ma na użytkownika i społeczeństwo po zakupie. Taguchi promował ideę, że jakość powinna być mierzona przez

funkcję strat, która kwantyfikuje koszty nawet niewielkich odchyień od optymalnych parametrów działania produktu. W podobnym czasie, W. Edwards Deming (2018) wprowadził pojęcie skoncentrowane na kliencie, twierdząc, że jakość powinna odpowiadać zarówno obecnym, jak i przyszłym jego potrzebom. Deming, podobnie jak Juran, był zwolennikiem ciągłego doskonalenia i planowania jakości.

W latach 80. i 90. Christian Grönroos (1984) oraz Parasuraman, Zeithaml i Berry (1985) poszerzyli tę perspektywę, wprowadzając do debaty pojęcie jakości w usługach, podkreślając, że jakość ma zarówno wymiary techniczne, jak i funkcjonalne. Grönroos w szczególności zwrócił uwagę na to, jak usługi są doświadczane przez klientów, a jakość usług została zdefiniowana jako różnica między oczekiwaniami klientów a ich percepcją usługi.

Na początku XXI wieku, wraz z rosnącą świadomością ekologiczną i społeczną, pojawiło się nowe rozumienie jakości, włączające zrównoważony rozwój. Autorzy publikacji podkreślają, że jakość obejmuje również odpowiedzialność społeczną i środowiskową, poszerzając koncepcję jakości o aspekty zrównoważonego rozwoju (Corbett i Cutler, 2000; Craig i Lemon, 2008; Isaksson i Garvare, 2003).

W ten sposób, przez ostatnie dziesięciolecia, pojęcie jakości przekształciło się z prostego spełnienia specyfikacji do kompleksowego procesu zarządzania, który uwzględnia potrzeby klientów, długoterminowe wpływy na społeczeństwo i środowisko, a także ciągłe dążenie do doskonalenia w każdym aspekcie działalności organizacji.



Rysunek 4. Oś czasu kluczowych zmian w definicji jakości  
 Źródło: opracowanie własne na podstawie (Martin i in., 2020)

Analizując rozwój pojęcia jakości na przestrzeni lat, warto zauważyć, że sama definicja, początkowo odnosząca się do produktów fizycznych, zaczęła obejmować dane jako istotny element jakościowy w latach 80. XX wieku. To był czas, kiedy rozwój technologii

informacyjnych i pojawienie się systemów zarządzania bazami danych zaczęło podkreślać znaczenie danych jako kluczowego zasobu w organizacjach. W tym okresie, z uwagi na rosnącą zależność od systemów informatycznych i informatyzacji procesów biznesowych, jakość danych stała się równie ważna jak jakość produktów fizycznych. Rozwój systemów IT i coraz większe poleganie na danych cyfrowych w zarządzaniu i podejmowaniu decyzji biznesowych sprawiło, że jakość danych stała się kluczowym aspektem efektywności organizacyjnej, zarządzania ryzykiem i zadowolenia klienta. Mimo to, nawet wczesne definicje jakości mogą być użyteczne i adaptowalne w kontekście danych. Choć Shewhart koncentrował się na produkcji, jego koncepcje można zaadaptować do zarządzania jakością danych, argumentując, że redukcja błędów jest równie ważna w cyfrowym środowisku. Podobnie, Juran i Crosby skupili się na „przydatności do użytku” i „zgodności ze specyfikacjami”. W kontekście danych, przydatność do użytku można rozumieć jako zapewnienie, że dane są wartościowe dla użytkowników końcowych katalogów. Zgoda Crosbiego na ideę „zero defektów” może być ambitna w środowisku, gdzie dane są dynamiczne i często zmienne, ale samo dążenie do minimalizacji błędów jest całkowicie zasadne. Z kolei perspektywa kosztu straty jakości (Taguchi) - jest niezwykle istotna w kontekście danych, co wielokrotnie podkreślają badania dotyczące wpływu jakości danych na sprzedaż i kondycje firm. Zatem dbałość o aspekty techniczne i funkcjonalne jakości danych jest niezbędna. Na koniec, rozszerzenie pojęcia jakości o zrównoważony rozwój w XXI wieku ma szczególne znaczenie. Odpowiedzialne zarządzanie danymi wymaga uwzględnienia ich wpływu na społeczność i środowisko, co jest kluczowe zważywszy choćby na problem ostatniej mili i niepotrzebnych zwrotów źle zamówionych towarów, których bezpośrednią przyczyną są błędy w danych produktowych.

W kontekście globalnego rozwoju koncepcji jakości, warto przybliżyć także wkład polskich naukowców. Podobnie jak u zagranicznych autorów, definicje jakości również zaczęły się od fizycznego produktu, by w późniejszym czasie pojęcie stało się bardziej uniwersalne. Podejście do jakości wg Oyrzanowskiego (1984) koncentruje się na cechach i właściwościach produktu, podkreślając, że jakość powinna być rozpatrywana w kontekście całego cyklu jego życia. Oyrzanowski zwracał uwagę na to, że wszystkie etapy, od projektowania po użytkowanie, mają wpływ na ostateczną jakość produktu. Kindlarski (1988) oraz Wawak (1989) podkreślają, że jakość powinna być postrzegana jako stopień zaspokojenia potrzeb użytkowników, przy czym uwzględniają ich możliwości ekonomiczne. Ich podejście wskazuje na konieczność równoważenia jakości z dostępnością finansową, co ma kluczowe znaczenie dla zadowolenia klientów. Skrzypek (2003) rozszerzyła definicję jakości, uwzględniając

techniczne, ekonomiczne i środowiskowe aspekty. Jej podejście podkreśla, że lojalność klientów i sukces konkurencyjny firmy zależą od zintegrowanego podejścia do jakości, które obejmuje różne wymiary produktu i jego wpływ na otoczenie. Zymonik (2003) skoncentrowała się na tworzeniu wartości dla konsumenta, rozumiejąc jakość nie tylko w kategoriach właściwości produktu, ale przede wszystkim jako wartość dostarczaną klientowi. Jej koncepcja jakości łączy obiektywne charakterystyki produktu z subiektywnymi odczuciami klienta, uznając oba te aspekty za kluczowe w dostarczaniu wartości. Blikle (2009) definiuje jakość jako miarę braku wad w produkcji, skupiając się na negatywnych aspektach z punktu widzenia klienta. Według tej definicji, wada to każda nieoczekiwana i negatywna cecha produktu z perspektywy klienta, co przesuwając nacisk na subiektywne doświadczenie i zadowolenie klienta. Kolman i współautorzy (1996) przedstawili jakość jako wielowymiarową koncepcję, obejmującą aspekty fizyczne, umowne i użytkowe produktu. W ich ujęciu, jakość to nie tylko właściwości fizyczne, ale także zgodność z wymaganiami i oczekiwaniami oraz praktyczna użyteczność i funkcjonalność produktu. Mantura (2012) zwrócił uwagę na praktyczny aspekt jakości, wskazując na rosnące znaczenie badań jakościowych w różnych organizacjach. Jego podejście podkreśla, że jakość jest kluczowym elementem w zarządzaniu i rozwoju firm, wymagającym ciągłej oceny i dostosowywania. Hamrol (2005) rozszerzył definicję jakości, twierdząc, że obejmuje ona nie tylko zgodność z wymaganiami, ale także dążenie do doskonałości i przydatność użytkową. Jego podejście traktuje jakość jako dynamiczny proces, w którym ciągłe doskonalenie i innowacje są niezbędne do spełnienia rzeczywistych potrzeb i oczekiwań użytkowników. Mroczko (2012) skoncentrował się na idei, że jakość jest nie tylko wynikiem spełniania wymagań technicznych czy funkcjonalnych, ale również odzwierciedleniem oczekiwań i potrzeb klienta. Podkreśla, że jakość jest procesem dynamicznym, obejmującym nie tylko aspekty produktu, ale również sposób, w jaki jest on postrzegany i doświadczany przez klienta. W swojej koncepcji, Mroczko wykroczył poza tradycyjne kryteria jakości, takie jak wydajność czy niezawodność, zwracając uwagę na znaczenie budowania trwałej wartości przez skupienie się na całkowitym doświadczeniu klienta, od pierwszego kontaktu z produktem aż po obsługę posprzedażową. Jego podejście sugeruje, że jakość to nie tylko cecha produktu, ale również kluczowy element strategii biznesowej, mający wpływ na lojalność klienta, reputację firmy i jej konkurencyjność na rynku.

Przyszłość definicji jakości może ewoluować w związku z wpływem zjawiska powszechnej transformacji cyfrowej. Zauważono, że cyfryzacja, obejmująca społeczeństwo i organizacje na wszystkich poziomach, wprowadza nowe wymiary do pojęcia jakości

(Parviainen i in., 2017). Dzięki dostępowi do dużych zbiorów danych, firmy są w stanie dokładniej analizować i optymalizować swoje procesy, co wskazywane jest jako kluczowe dla podnoszenia jakości (Gölzer i Fritzsche, 2017). Ponsignon i współautorzy (2019) podkreślają, że transformacja cyfrowa wymusza rozwój nowych kompetencji, zwłaszcza w zakresie współpracy z działami IT, co jest niezbędne do efektywnego wykorzystania narzędzi cyfrowych w zarządzaniu jakością. Dodatkowo, Birch-Jensen i współautorzy (2020) wskazują na nowe rodzaje kanałów informacji zwrotnej od klientów, które umożliwiają organizacjom szybsze i bardziej precyzyjne reagowanie na potrzeby rynku oraz oczekiwania konsumentów. Rozwijana ostatnio koncepcja Jakości 4.0 (Küpper i in., 2019; Sony i in., 2020; Wawak, 2022) łączy tradycyjne metody zarządzania jakością z nowoczesnymi technologiami cyfrowymi, takimi jak sztuczna inteligencja, analiza big data czy Internet Rzeczy (IoT). To podejście może zrewolucjonizować sposób, w jaki jakość jest definiowana, mierzona i ciągle doskonalona, czyniąc ją bardziej predyktywną, adaptacyjną i zintegrowaną z doświadczeniem klienta w czasie rzeczywistym. W efekcie, przyszłość definicji jakości będzie prawdopodobnie obejmować większy nacisk na dane i analitykę, adaptacyjność do zmieniających się warunków rynkowych i preferencji klientów, a także integrację z bardziej zaawansowanymi technologicznie procesami zarządzania i produkcji.

## **2.2. Jakość w kontekście danych**

Postępy w definiowaniu jakości omówione w rozdziale 2.1 miały bezpośredni wpływ na sposób, w jaki obecnie postrzega się jakość danych. Od prostego spełnienia specyfikacji do kompleksowego procesu zarządzania, który uwzględnia potrzeby użytkowników, swoisty „klientocentryzm”, wpływy społeczne i środowiskowe oraz ciągle doskonalenie – wszystkie te aspekty są kluczowe dla zapewnienia ich wysokiej jakości.

W ogólnych terminach, jakość danych jest opisywana jako zdolność danych do zaspokojenia wyrażonych i domyślnych potrzeb, gdy dane są używane w określonych warunkach (International Organization for Standardization, 2008). Ta definicja jest kluczowa, ponieważ podkreśla praktyczne wykorzystanie danych, co jest zgodne z perspektywą tematyki pracy doktorskiej. Warto zauważyć, że definicja ta akcentuje "używanie danych w określonych warunkach", co podkreśla znaczenie kontekstu, w którym dane są wykorzystywane.

W literaturze, jakość danych wskazywana jest jako pojęcie, którego istota wiąże się z ogólną domeną jaką są dane, ale jest na tyle obszerne, że sprowadzenie go do jednoznacznej



definicji jest dużym wyzwaniem (Oliveira i in., 2005). Termin, podobnie jak dla ogólnej definicji jakości, ma wiele znaczeń związanych z kontekstem danych, z domeną danych, obszarem oraz dziedziną, w której są wykorzystywane (Caballero i Piattini, 2003; Sidi i in., 2012). Taleb i współautorzy (2021) podkreślają różnice interpretacyjne definicji pomiędzy środowiskiem naukowym a biznesem. Chen i współautorzy (2012) jakość danych sprowadzają do „zdolności danych do zaspokajania wyrażonych i domniemych potrzeb przy użyciu w określonych warunkach”, natomiast Sidi i współautorzy (2012) definiują jakość danych jako właściwość odpowiadającą zarządzaniu jakością, która jest odpowiednia do użycia lub spełnia potrzeby użytkownika. To pokazuje, że niezależnie od definicji, centralnym punktem jest zawsze użytkownik i jego wymagania względem danych. W raporcie firmy Gartner (2021) termin "jakość danych" odnosi się do procesów i technologii identyfikacji, zrozumienia oraz korygowania wad w danych, które wspierają skuteczne zarządzanie danymi i analityką w ramach operacyjnych procesów biznesowych oraz zarządzanie danymi i analityką w procesach decyzyjnych. To podejście jest szczególnie ważne w kontekście jakości danych produktowych, ponieważ podkreśla aktywny charakter zarządzania jakością – nie tylko jako stan do osiągnięcia, ale jako ciągły proces poprawy przy użyciu najróżniejszych metod i technologii.

Wiele różnic w definicji jakości danych ma swoją genezę w różnicach jej konceptualizacji poprzez wymiary (np. dokładność, kompletność, obiektywność, spójność, aktualność, ważność i wiarygodność) na podstawie badań empirycznych, ram ontologicznych i semiotycznych, np. syntaktyczne (struktura danych), semantyczne (znaczenie danych), pragmatyczne (użycie danych) i społeczne (wspólne rozumienie znaczenia symboli) lub doświadczeń praktyków. Istnieje jednak zgoda co do tego, że jakość danych zależy zawsze od jakości źródła danych (Maier, 2013). To podkreśla między innymi znaczenie selektywnego wyboru i oceny źródeł danych na wstępnym etapie procesu walidacji.

Bardzo często jakość danych definiowana jest jako "przydatność do użytku" (Ballou i in., 2004; Batini i in., 2009; Lederman i in., 2003; Wang i Strong, 1996; Watts i in., 2009) i rozumiana jako zdolność zbioru danych do spełnienia wymagań użytkowników jednocześnie wskazując, że jakość danego obiektu danych może różnić się w zależności od okoliczności (Heinrich i Klier, 2009; Hüner i in., 2011; Silvola i in., 2019). Jeżeli zbiór danych, używany w określonych warunkach, spełnia wszystkie wyrażone i domyślne wymagania użytkowników, można wtedy powiedzieć, że jest wysokiej jakości. Niewystarczająca jakość danych może mieć znaczący negatywny wpływ na działalność biznesową (Haug i in., 2011). Przytaczana definicja

"przydatności do użytku", jest szczególnie ważna w kontekście katalogów produktowych. Oznacza to, że jakość danych musi być oceniana w kontekście ich praktycznego wykorzystania przez użytkowników. W katalogach, gdzie różni użytkownicy mogą mieć różne wymagania i oczekiwania, elastyczność i dostosowanie danych do tych potrzeb są kluczowe.

Pojęcie "przydatności do użytku" w kontekście jakości danych może być jednak źródłem nieporozumień, zwłaszcza gdy rozważamy je w kontekście globalizacji i standaryzacji. W kontekście globalnym, gdzie dane są udostępniane i wykorzystywane przez różnorodne grupy użytkowników z różnych kultur i kontekstów biznesowych, to co jest uznawane za przydatne w jednym środowisku, może nie być w innym. W takim rozumieniu pojęcie przydatności do użytku może być sprzeczne z potrzebą standaryzacji. Standaryzacja danych ma na celu zapewnienie spójności, interoperacyjności i wymienności danych na poziomie międzynarodowym. Jednak koncentrowanie się wyłącznie na przydatności danych dla konkretnego użytkownika lub grupy użytkowników może prowadzić do utraty szerszej perspektywy, niezbędnej do tworzenia ogólnie akceptowalnych standardów. Nie oznacza to, że przydatność do użytku jest całkowicie niezgodna ze standaryzacją. Można ją traktować jako jeden z aspektów jakości danych, który powinien być zrównoważony z innymi kryteriami, takimi jak dokładność, kompletność, aktualność, spójność czy interoperacyjność. W idealnym przypadku, dane powinny być zarówno przydatne dla konkretnych użytkowników, jak i zgodne ze standardami globalnymi, co zapewnia ich użyteczność w szerszym kontekście.

W kontekście jakości danych w społecznościowych katalogach elektronicznych, istotne jest zatem znalezienie równowagi między zapewnieniem przydatności danych dla poszczególnych użytkowników a ich standaryzacją, aby dane były użyteczne i akceptowalne na szerszym, globalnym rynku. To wymaga holistycznego podejścia do jakości danych, uwzględniającego różnorodne wymiary i potrzeby.

### **2.3. Wymiary jakości danych**

Wymiary jakości danych stosowane są głównie do mierzenia, kwantyfikowania i zarządzania jakością danych (Batini i in., 2009; Glowalla i in., 2014; Sidi i in., 2012). Wykorzystanie wymiarów jakości danych staje się szczególnie istotne w procesach zarządzania danymi, gdyż pozwalają one na ocenę czy dane są odpowiednie do zastosowań, dla których zostały zebrane. Już w 1968 roku zauważono, że pożądane atrybuty definiujące jakość danych to: istotność, terminowość i dokładność (Feltham, 1968). Jeżeli rozważamy wysoką jakość

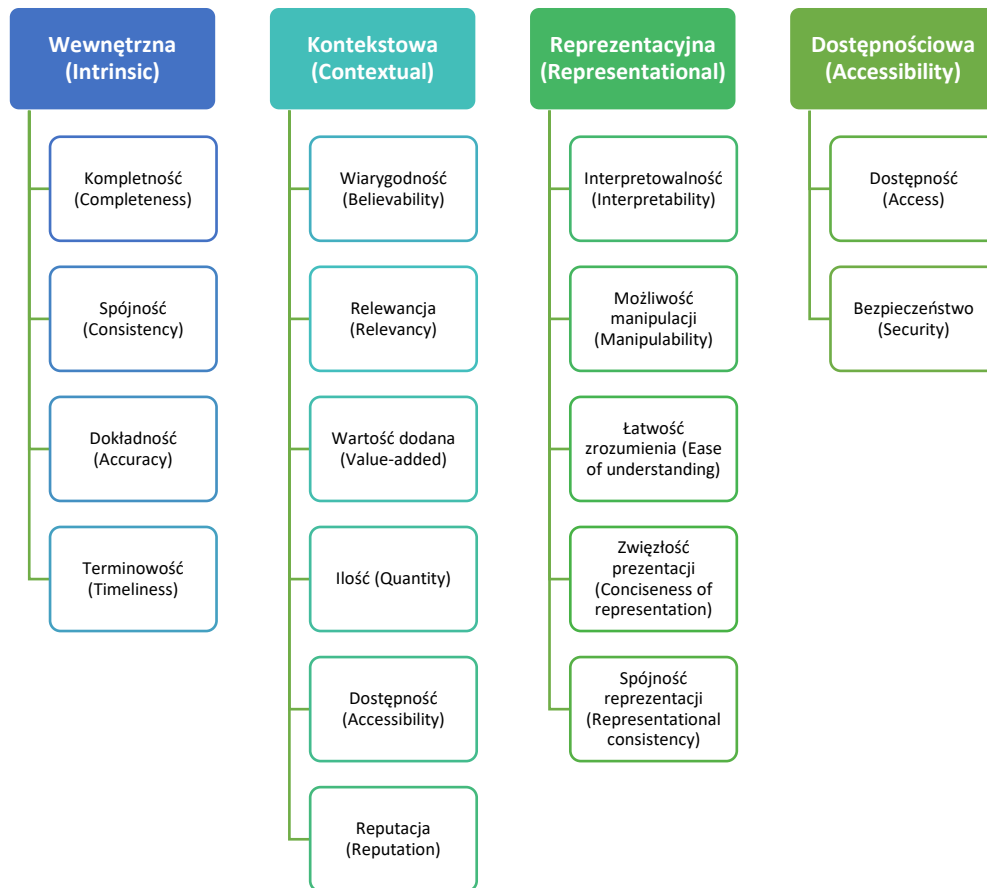
danych, to dane/informacje muszą być istotne, dokładne, oparte na faktach, kompletne, niezawodne, ustrukturyzowane, precyzyjne, czytelne i rozsądne (Zmud, 1978). Reeves i Bednar (1994) zauważyli, że istnieje pięć definicji związanych z pojęciem jakości danych: wartość, zgodność z specyfikacjami, zgodność z wymaganiami, przydatność do użytku oraz spełnienie i/lub przekroczenie oczekiwań klienta. Często cytowaną definicję przedstawili Ballou i Pazer (1985), którzy podzielili jakość danych na: dokładność, aktualność, kompletność i spójność. Argumentują, że wymiar dokładności jest najłatwiejszy do oceny, ponieważ jest to tylko kwestia analizy różnicy między wartością prawidłową a faktycznie używaną. Twierdzą również, że ocena aktualności może być przeprowadzona w podobnie nieskomplikowany sposób. Ocena kompletności danych może być również stosunkowo prosta, pod warunkiem sprawdzania czy dane są kompletne, zamiast definiowania poziomu kompletności, na przykład procentu kompletności danych. Natomiast ocena spójności jest bardziej skomplikowana, ponieważ wymaga co najmniej dwóch schematów reprezentacji do dokonania porównania.

Wymiary jakości danych można zorganizować też w czterech obszarach, co zaproponował Wand i Wang (1996), tj. wewnętrznym, kontekstowym, dostępnościowym i reprezentacyjnym (Batini i in., 2009; Cappiello i in., 2013; Gandomi i Haider, 2015; Saha i Srivastava, 2014; Wang i Strong, 1996).

- A. Wewnętrzna jakość danych, która koncentruje się na własnościach danych samej w sobie, niezależnie od kontekstu ich użycia. Wymiary w tym obszarze mogą obejmować dokładność, spójność i unikalność. Dane powinny być poprawne, logicznie i nie zawierać zbędnych duplikatów.
- B. Kontekstowa jakość danych, która dotyczy przydatności danych w kontekście konkretnego zadania lub procesu. Wymiary mogą tutaj obejmować kompletność, relewancję i aktualność. Kluczowe jest, aby dane były kompletne, istotne i aktualne w odniesieniu do konkretnego zastosowania.
- C. Dostępnościowa jakość danych, która dotyczy łatwości dostępu do danych oraz ich bezpieczeństwa. Wymiary w tym obszarze mogą zawierać dostępność, bezpieczeństwo i zgodność z przepisami. Dane powinny być łatwo dostępne dla uprawnionych użytkowników, ale jednocześnie chronione przed nieuprawnionym dostępem i zgodne z obowiązującymi regulacjami.
- D. Reprezentacyjna jakość danych, która odnosi się do sposobu prezentacji danych. Wymiary mogą tu obejmować czytelność, interpretowalność i spójność

reprezentacji. Ważne jest, aby dane były prezentowane w sposób zrozumiały, łatwy do interpretacji i spójny w różnych systemach.

Na Rysunku 5 przedstawiono przykład takiego podziału wg 4 obszarów (Taleb i in., 2021).



Rysunek 5. Wymiary jakości danych – podział na 4 kategorie  
 Źródło: opracowanie własne na podstawie (Taleb i in., 2021)

W opracowaniu niniejszej dysertacji, główny nacisk położony został na jakość samych danych i informacji w nich zawartych, nie na dostęp do nich. Szczególnie jest tutaj istotna jakość wewnętrzna, która jest odpowiedzialna za dokładność i spójność danych, oraz jakość reprezentacji, która poprzez czytelność i spójność, ułatwia ich interpretację i analizę. Należy zaznaczyć, że liczba wymiarów we wspomnianych obszarach może być różna. Przykładowo, Wand i Wang (1996) dla każdego obszaru definiują zestaw łącznie osiemnastu wymiarów (Tabela 2), a w wewnętrznym zaproponowali dekompozycję jakości danych na cztery wymiary: kompletny (ang. complete), jednoznaczny (ang. unambiguous), znaczący (ang. meaningful), poprawny (ang. correct), tymczasem Cichy i Rass (2019) zidentyfikowali ich aż dwadzieścia. Definicję Wang i Stronga omawiają Haug i inni (2009), którzy argumentują, że

"jakość danych reprezentacyjnych" może być postrzegana jako forma dostępności, zamiast oddzielnego obszaru. W związku z tym definiują trzy obszary jakości danych: wewnętrzną, dostępności i użyteczności. Levitin i Redman (1998) przedstawiają inną perspektywę, argumentując, że ponieważ procesy produkcyjne danych mają wiele podobieństw do procesów wytwarzania produktów fizycznych, procesy produkcyjne danych można by postrzegać jako tworzące produkty danych dla konsumentów danych. Na podstawie tego spojrzenia na dane jako zasoby, Levitin i Redman dyskutują, jak trzynastce podstawowych właściwości zasobów organizacyjnych może być przetłumaczone na właściwości dla danych.

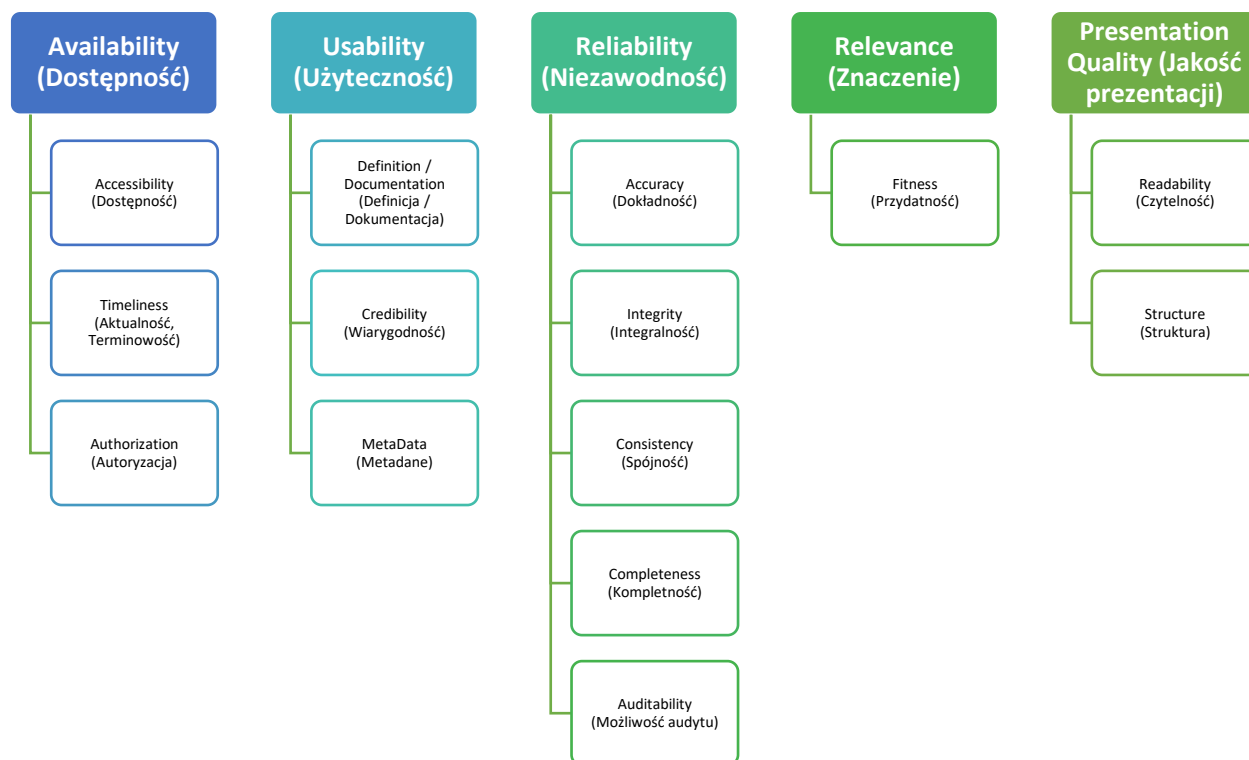
Tabela 2. Wymiary wraz z tłumaczeniem i omówieniem ich znaczenia

<b>Termin oryginalny</b>	<b>Tłumaczenie na język polski</b>	<b>Opis</b>
Accuracy	Dokładność	Dane są wolne od błędów i precyzyjnie odzwierciedlają rzeczywistość.
Reliability	Niezawodność	Dane są spójne i stabilne w czasie.
Timeliness	Aktualność	Dane są dostępne i użyteczne w odpowiednim czasie.
Relevance	Znaczenie	Dane są istotne i przydatne dla określonego kontekstu czy zadania.
Completeness	Kompletność	Dane zawierają wszystkie niezbędne elementy.
Currency	Aktualizacja	Dane są regularnie aktualizowane, aby odzwierciedlały najnowsze informacje.
Consistency	Spójność	Dane są logicznie zgodne i nie zawierają sprzeczności.
Flexibility	Elastyczność	Dane mają zdolność adaptacji do nowych wymagań lub warunków.
Precision	Precyzja	Dane charakteryzują się wysokim stopniem szczegółowości i dokładności.
Format	Format	Dane są odpowiednio zorganizowane i prezentowane.
Interpretability	Interpretowalność	Dane można łatwo zrozumieć i zinterpretować.
Sufficiency	Wystarczalność	Dane są wystarczające do zaspokojenia potrzeb użytkowników.
Usableness	Użyteczność	Dane są łatwe w użyciu dla użytkowników.
Usefulness	Przydatność	Dane przyczyniają się do osiągnięcia określonych celów lub zadań.
Content	Zawartość	Dane zawierają istotne treści informacyjne.
Efficiency	Efektywność	Dane są generowane, przetwarzane i wykorzystywane w sposób ekonomiczny.
Importance	Ważność	Dane mają znaczenie dla użytkowników lub procesów.
Informativeness	Informatywność	Dane są bogate w użyteczne informacje.
Level of Detail	Poziom szczegółowości	Dane charakteryzują się wysokim poziomem detali i głębi informacji.
Quantitativeness	Ilościowość	Dane prezentują informacje w formie liczbowej.
Clarity	Jasność	Dane są łatwe do zrozumienia, proste i jednoznaczne.

Termin oryginalny	Tłumaczenie na język polski	Opis
Scope	Zakres	Dane obejmują określony obszar tematyczny lub zakres.
Comparability	Porównywalność	Dane można sensownie porównywać z innymi danymi.
Understandability	Zrozumiałość	Dane są łatwe do zrozumienia przez użytkowników.
Conciseness	Zwiężłość	Dane są przedstawione w skondensowanej formie, bez zbędnych szczegółów.
Freedom from Bias	Wolność od uprzedzeń	Dane są obiektywne i wolne od subiektywnych opinii lub stronniczości.

Źródło: Opracowanie własne na podstawie Wand i Wang (1996)

Odminną strukturę podziałową zaproponowali Cai i Zhu (2015), dzieląc jakość danych na 5 obszarów i ich poszczególne elementy. Obszary te to: dostępność, użyteczność, wiarygodność, trafność oraz jakość prezentacji. Ostatni wymiar został dodany, ponieważ ustalono, że zwiększa on zadowolenie klientów. Na Rysunku 6 przedstawiono diagram podziału, a następnie opisano poszczególne wymiary w Tabeli 3.



Rysunek 6. Diagram wymiarów jakości

Źródło: opracowanie własne na podstawie (Cai i Zhu, 2015)

Tabela 3. Omówienie diagramu wymiarów

Obszar	Wymiar	Opis
Availability (Dostępność)	Accessibility (Dostępność)	Łatwość, z jaką użytkownicy mogą uzyskać dostęp do danych. Dotyczy również dostępności danych w odpowiednich formatach.
	Timeliness (Terminowość)	Stopień, w jakim dane są aktualne i dostępne w odpowiednim czasie.
	Authorization (Autoryzacja)	Procesy i mechanizmy kontrolujące dostęp do danych, zapewniające, że tylko upoważnione osoby mają do nich dostęp.
Usability (Użyteczność)	Definition/Documentation (Definicja/Dokumentacja)	Dostępność i jakość dokumentacji opisującej dane, ich strukturę, znaczenie i sposób użycia.
	Credibility (Wiarygodność)	Stopień, w jakim użytkownicy uważają dane za wiarygodne i godne zaufania.
	MetaData (Metadane)	Dane opisowe, które dostarczają informacji o danych, ułatwiając ich zrozumienie, zarządzanie i użycie.
Reliability (Niezawodność)	Accuracy (Dokładność)	Precyzja i poprawność danych; odnosi się do tego, jak blisko dane odpowiadają rzeczywistości.
	Integrity (Integralność)	Zachowanie spójności i kompletności danych w całym ich cyklu życia.
	Consistency (Spójność)	Brak sprzeczności w danych, zarówno wewnątrz jednego zbioru, jak i pomiędzy różnymi zbiorami danych.
	Completeness (Kompletność)	Pełność danych, w tym obecność wszystkich niezbędnych wartości danych.
	Auditability (Możliwość audytu)	Możliwość przeprowadzenia audytu lub przeglądu danych w celu weryfikacji ich dokładności i integralności.
Relevance (Znaczenie)	Fitness (Przydatność)	Odnosi się do tego, na ile dane są odpowiednie do celów, dla których są wykorzystywane.
Presentation Quality (Jakość prezentacji)	Readability (Czytelność)	Łatwość, z jaką dane mogą być przeczytane i zrozumiane przez użytkowników.
	Structure (Struktura)	Organizacja danych w sposób, który ułatwia ich zrozumienie i analizę.

Źródło: opracowanie własne na podstawie Cai i Zhu (2015)

W kontekście przeprowadzonej analizy dotyczącej wymiarów jakości danych, brakuje jasnego ujęcia standaryzacji jako kluczowego elementu dla zapewnienia użyteczności i wiarygodności. Modele jakości danych, jak te zaproponowane przez Wand i Wanga czy Cai i Zhu, choć różnią się w zakresie skupienia na różnych wymiarach jakości, mogą być trudne do zastosowania na skalę globalną bez pewnego stopnia standaryzacji. W społecznościowych katalogach elektronicznych, gdzie dane pochodzą z różnych źródeł i są używane przez

użytkowników na całym świecie, standaryzacja danych jest niezbędna. Z drugiej strony, globalizacja wpływa na jakość danych poprzez zwiększanie różnorodności i skali, w jakiej dane są generowane i wykorzystywane. W kontekście globalnym, dane muszą być zrozumiałe i użyteczne dla użytkowników z różnych kultur i systemów prawnych, co stanowi wyzwanie dla zarządzania jakością danych. Na przykład, to co jest uznawane za „dane wysokiej jakości” w jednym kontekście kulturowym lub biznesowym, może nie być takie w innym. To rodzi pytania dotyczące uniwersalności wymiarów jakości danych i potrzeby ich dostosowania do różnych środowisk.

Podsumowując, różnorodność podejść i modeli wymiarów jakości danych w literaturze naukowej podkreśla złożoność i wielowymiarowość tego zagadnienia. Różne modele, takie jak te zaproponowane przez Wanga i Stronga, czy Cai i Zhu, wskazują na istotność różnych aspektów jakości danych, od wewnętrznych cech danych po ich użyteczność i reprezentację. Każdy z tych modeli podkreśla różne, ale ważne elementy, które powinny być brane pod uwagę podczas oceny jakości danych. Model jakości danych może być wybrany lub dostosowany w zależności od specyficznych wymagań i kontekstu, w jakim dane są używane. Na przykład, w kontekście naukowym, kluczowe może być uwzględnienie dokładności i kompletności danych, podczas gdy w biznesie większe znaczenie mogą mieć aspekty takie jak użyteczność, dostępność czy zgodność z przepisami. W obszarze crowdsourcingu, czy uogólniając – społecznościowej wymianie danych, szczególnie istotne mogą być takie aspekty jak wiarygodność, interpretowalność, spójność, czy aktualność, szczególnie w sytuacji, gdy dane są ciągle aktualizowane przez różne źródła. Jeśli katalog ma zasięg globalny, powinien być dostosowany do różnorodności kulturowej, językowej i regulacyjnej, co oznacza, że elastyczność i adaptacyjność modelu są kluczowe. Model powinien ponadto promować standaryzację i być zgodny z międzynarodowymi standardami, aby zapewnić szeroką użyteczność danych. Jednocześnie, model wymiarów jakości danych powinien uwzględniać użyteczność i przydatność danych dla użytkowników końcowych. Jest to istotne, zwłaszcza jeśli katalog ma służyć szerokiemu spektrum użytkowników z różnymi potrzebami. Ostateczny wybór modelu powinien także uwzględniać równowagę między kompleksowością a praktycznością. Należy rozważyć, czy bardziej złożony model będzie efektywny w implementacji i utrzymaniu, czy prostszy model będzie wystarczający do spełnienia potrzeb użytkowników. Ostatecznie, wybierając model wymiarów jakości danych, kluczowe jest znalezienie takiego, który najlepiej odpowiada specyficznym wymaganiom społecznościowych



katalogów elektronicznych, jest skuteczny w implementacji i może być dostosowany do zmieniających się potrzeb i warunków.

#### **2.4. Znaczenie jakości danych w ujęciu danych produktowych**

W ostatnich dekadach, postęp w dziedzinie informatyki znacząco przyczynił się do rozwoju możliwości organizacji w zakresie gromadzenia dużych ilości danych produktowych. Jednak wraz ze wzrostem objętości danych, rośnie również złożoność ich zarządzania. Haug i inni (2011) zwracają uwagę na to, że im organizacje gromadzą i zarządzają coraz większymi i bardziej skomplikowanymi zasobami informacyjnymi, tym ryzyko pogorszenia jakości danych rośnie. Dodatkowo, istnieje problem lokalnego zarządzania danymi przez firmy, co prowadzi do tworzenia "wysp informacyjnych". Te wyspy charakteryzują się redundantnym przechowywaniem, zarządzaniem i przetwarzaniem danych, co sprawia, że kluczowe dane stają się trudne do zrozumienia i wykorzystania, powstają różnice i nieścisłości.

Rozpoczynając analizę znaczenia jakości danych warto przybliżyć funkcjonujące w publikacjach naukowych pojęcie „brudnych danych”. Jest to określenie na dane nieprecyzyjne, niekompletne i niespójne (Marsh, 2005), czyli w ogólnym rozumieniu dane niskiej jakości (Guoling i Qinyun, 2008). Podobnie definiuje je Zhou i inni (2011) i podkreśla, że odnoszą się do danych, które są przestarzałe, niespójne lub niekompletne, a przez to trudne do właściwego przeszukiwania, agregacji i analizy. Marsh (2005), w kontekście "brudnych danych", stwierdza dodatkowo, że nie mogą one już być traktowane jako stosunkowo nieszkodliwy problem. Podkreśla, że brudne dane mogą mieć i faktycznie mają bezpośredni wpływ na różne aspekty działalności biznesowej, w tym na poziomy zapasów, zamówienia sprzedażowe, percepcję klientów, ich lojalność oraz rentowność. Ta definicja wskazuje na znaczącą rolę, jaką jakość danych odgrywa w efektywnym zarządzaniu przedsiębiorstwem i podejmowaniu decyzji biznesowych. Guoling i Qinyun (2008) wskazują dodatkowo na znaczną częstotliwość występowania „brudnych danych” w handlu elektronicznym, z kolei Liu i inni oraz Putri i Pujani (Liu i in., 2019; Wilda Kurnia Putri i Vera Pujani, 2019) zbadali oddziaływanie błędów w tym obszarze i ich znaczący wpływ na wyniki finansowe firm, zwłaszcza w kontekście sprzedaży online. Niemir i Mrugalska (2022a) wskazują, że dane produktowe pierwotnie wykorzystywane online, w dłuższej perspektywie mogą być wykorzystywane szerzej, np. do synchronizacji i integracji z porównywarkami cenowymi, platformami e-marketplace, do zwiększenia widoczności e-sklepu w wyszukiwarkach poprzez wykorzystanie danych strukturalnych, do wymiany informacji na poziomie logistycznym. Z tego powodu kluczowe jest zachowanie standaryzacji i wysokiej jakości już w momencie

tworzenia katalogu danych produktowych. Należy podkreślić znaczenie tego aspektu, nawet jeżeli obecnie firma nie dostrzega takiej konieczności, a obecny zakres danych wydaje się wystarczający dla ich bieżącej działalności. W sytuacji, gdy w przyszłości okaże się, że dane są niekompletne – a więc nie spełniają minimalnych wymagań integracyjnych partnerów biznesowych, jak zauważają Niemir i Mrugalska (2021), na przykład brak numerów GTIN, niekompletność danych o marce produktu lub nieodpowiednie strukturyzowanie danych (niewłaściwy format nazwy produktu lub zdjęcia głównego), to może to stanowić znaczną przeszkodę, a nawet uniemożliwić współpracę z powodu kosztów związanych ze zmianami w bazie danych produktowych. Brak standaryzacji w tym zakresie całkowicie zablokuje możliwość wykorzystania nowoczesnych technologii, co podkreśla Żuchowski (2022).

Haug i współpracownicy (2011) zwracają uwagę, że niska jakość danych może prowadzić do szeregu negatywnych konsekwencji w firmie. Do takich należą obniżona satysfakcja klientów, nieefektywność procesów decyzyjnych, spadek wydajności oraz zmniejszona satysfakcja pracowników z pracy. Ponadto, niska jakość danych zwiększa koszty operacyjne przez konieczność wykrywania i korygowania błędów. Dane są kluczowym elementem codziennych operacji i decyzji, a także definiują wspólne terminy w przedsiębiorstwie, co oznacza, że ich jakość ma istotny wpływ na kulturę organizacyjną. Słabej jakości dane utrudniają budowanie zaufania do informacji firmy, co może prowadzić do braku akceptacji dla inicjatyw opartych na tych danych. Z kolei operacje na nieprecyzyjnych danych obniżają wiarygodność firmy w zakresie oferowanych produktów, jak podają Qalati i inni (2021). Dodatkowo, niespójności danych są jedną z przyczyn zwrotów produktów, co negatywnie wpływa na środowisko ze względu na konieczność zorganizowania dodatkowego transportu, jak zauważają Kawa i Pierański (2021).

Wg Salem i Boufares (2019) jakość danych jest kluczowa dla zapewnienia inteligentnego wydobycia wiedzy z baz danych. Podejmowanie decyzji na podstawie błędnych danych negatywnie wpływa na działania organizacji. Prasad i in. (2011) stwierdzają, że niska jakość danych prowadzi do błędnych raportów, braku możliwości analizy kompletnego obrazu klientów z różnych segmentów, a także do słabej obsługi klienta. Brak jakości danych skutkuje ubytkiem klientów oraz niską satysfakcją konsumentów.

Drd'gan i Metz (2017) przeanalizowali koszty związane z jakością, zarówno pośrednie, jak i bezpośrednie. Koszty pośrednie objęły niezadowolenie pracowników, duplikację danych, skupienie na niewłaściwych segmentach rynku czy złe planowanie produkcji. Koszty bezpośrednie - błędne dostawy, błędy w płatnościach, problemy z efektywnością, długie czasy

dostawy. Według autorów, zła jakość danych wpływa zarówno na operacyjne jak i strategiczne zadania w firmie. Prowadzi również do wzrostu kosztów operacyjnych, ponieważ czas i inne zasoby są zaangażowane w proces wykrywania i korygowania błędów. Dane, które są tworzone i wykorzystywane we wszystkich codziennych operacjach firmy, stanowią krytyczny wkład we wszystkie decyzje i niejawnie definiują wspólne terminy w przedsiębiorstwie, co potwierdza Cai i Zhu (2015).

Powszechnym założeniem jest, że dokładność informacji lub warstwa prezentacji danych ma znaczący wpływ na czas potrzebny do podjęcia decyzji, jak również na efektywność podejmowania decyzji. Według Drd'gan i Metz (2017) jednym z najważniejszych czynników niskiej jakości danych jest ryzyko złej analizy. Jeśli dane są pełne błędów, oznacza to, że wszelkie przeprowadzane analizy mogą prowadzić do zupełnie błędnych prognoz. Innym czynnikiem, który jest konsekwencją poprzedniego, jest brak danych. Niska jakość danych może również oznaczać, że firma nie posiada odpowiednich danych do analizy na wstępie. Kolejną konsekwencją niskiej jakości danych jest brak widoczności. Posiadając błędne wyniki analiz niskiej jakości danych, firma nie jest w stanie właściwie wizualizować trendów gospodarczych, co w większości przypadków prowadzi do strat. Z kolei Redman (1998) wykazał, że znacząca część danych w realnym świecie jest wadliwa, przy czym współczynnik błędów waha się zwykle od 0,5% do 30%, a wartości od 1% do 5% są bardzo powszechne.

Podczas analizy skutków niskiej jakości danych, staje się oczywiste, że wiele firm ponosi znaczące koszty związane z tym problemem, mimo że dokładna ich wysokość jest trudna do oszacowania. Jak zauważają Haug i współautorzy (2011), w badaniach w zakresie jakości danych nie opracowano jeszcze standardowych metod ich pomiaru. Z drugiej strony, istnieją przypadki, które dowodzą, że nawet marginalne niedokładności danych mogą mieć istotny wpływ ekonomiczny. Drobne błędy w danych mogą nie być znaczącym problemem w procesach produkcyjnych, ale mogą prowadzić do utraty sprzedaży i zakłóceń operacyjnych w organizacjach zajmujących się działalnością po sprzedażową. W kontraście do niewielkiej liczby dużych badań dotyczących jakości danych w artykułach akademickich, wielu ekspertów branżowych dostarcza takie studia. Do tych ekspertów branżowych należą grupy takie jak Gartner Group, Price Waterhouse Coopers oraz The Data Warehousing Institute, które wskazują na kryzys w zarządzaniu jakością danych oraz na niechęć wśród wysokiej rangi decydentów do podjęcia odpowiednich działań (Marsh, 2005). Marsh podsumowuje wyniki tych badań w następujących punktach:

- A. 88% wszystkich projektów integracji danych albo całkowicie się nie powiodło, albo znacznie przekroczyło swoje budżety.
- B. 75% organizacji zidentyfikowało koszty wynikające z brudnych danych
- C. 33% organizacji opóźniło lub anulowało nowe systemy IT z powodu złej jakości danych.
- D. 611 miliardów dolarów rocznie traci się w USA na źle skierowane wysyłki pocztowe i samą obsługę.
- E. Według Gartnera, złe dane to główna przyczyna awarii systemów CRM.
- F. Mniej niż 50% firm twierdzi, że ma pełne zaufanie do jakości swoich danych.
- G. Projekty z zakresu business intelligence (BI) często kończą się niepowodzeniem z powodu brudnych danych.
- H. Tylko 15% firm jest pewnych jakości zewnętrznych danych.
- I. Dane klientów zwykle degenerują się o 2% miesięcznie, czyli o 25% rocznie.
- J. Organizacje zwykle przeceniają jakość swoich danych i nie doceniają kosztu błędów.
- K. Procesy biznesowe, oczekiwania klientów, systemy źródłowe i reguły zgodności ciągle się zmieniają. Systemy zarządzania jakością danych muszą odzwierciedlać te zmiany.
- L. Ogromne ilości czasu i pieniędzy są wydawane zazwyczaj na walkę w celu stłumienia natychmiastowego kryzysu, zamiast radzenia sobie z długofalowym problemem.

Kolejne badania ujawniają, że słabej jakości dane kosztują amerykańskie przedsiębiorstwa miliardy dolarów, jak wskazują w raportach Eckerson (2002b), Redman (2004), English (1998), oraz ponownie Eckerson w "Achieving Business Success through a Commitment to High Quality Data" (2002a). Dodatkowo, aż 75% firm doświadcza spadku zysków z powodu niskiej jakości danych (Kenyon i in., 2001). W związku z tym, wysoka jakość danych staje się kluczowym czynnikiem dla organizacji w osiągnięciu ich strategicznych celów.

Wg Zahari i in. (2015) jakość danych jest kluczowym wskaźnikiem, decydującym o powodzeniu projektów migracji danych. Nieadekwatna jakość podstawowych danych uniemożliwia wsparcie dla reguł biznesowych lub procesów. Projekty migracji danych wiążą się z wysokim ryzykiem, zwłaszcza przy przenoszeniu dużych zbiorów danych. Największym wyzwaniem tych projektów jest zachowanie metadanych w ich pierwotnym stanie. Migrowanie danych o niskiej jakości, nieprecyzyjnych lub niewystarczających może prowadzić do nieodpowiednich założeń, mylnych wyników i błędnych decyzji. Zapewnienie wysokiego

poziomu jakości danych jest najbardziej kosztownym i czasochłonnym zadaniem w projekcie migracji danych. Badania Bloor Research (Howard, 2011) wskazują, że niska jakość danych jest główną przyczyną (53,4%) przekroczenia zakładanego czasu realizacji projektów.

W obliczu wyzwań związanych z zarządzaniem dużymi zbiorami danych produktowych, niezbędne jest podkreślenie znaczenia jakości tych danych dla sukcesu przedsiębiorstw. Nie tylko wpływają one bezpośrednio na operacje biznesowe, jak pokazali Liu i inni (2019) oraz Putri i Pujani (2019), ale również na percepcję firmy przez klientów i jej zdolność do efektywnego wykorzystania informacji strategicznych. W dłuższej perspektywie, jak zauważają Niemir i Mrugalska (2022a), dane produktowe mają potencjał do szerokiego zastosowania w różnorodnych aspektach działalności przedsiębiorstwa, od marketingu po logistykę. Dlatego też, wysoka jakość danych produktowych nie jest już tylko kwestią wewnętrznej efektywności, ale staje się kluczowym elementem zapewniającym konkurencyjność i innowacyjność na dynamicznie zmieniającym się rynku. Brak adekwatnej jakości i standaryzacji danych może prowadzić do znaczących strat finansowych i utraty zaufania klientów, a także ograniczać możliwości wykorzystania nowoczesnych technologii, co podkreśla Żuchowski (2022). Zatem, w kontekście zarządzania danymi produktowymi, priorytetem dla organizacji powinno być dążenie do utrzymania wysokiej jakości danych, co jest fundamentem do budowania długotrwałego sukcesu i rozwoju przedsiębiorstwa.

## **2.5. Przegląd metod pozyskiwania danych o produktach**

Istnieje wiele metod pozyskiwania danych produktowych, które mogą być wykorzystywane w zarządzaniu katalogiem elektronicznym w zależności od skali i potrzeb budowanego rozwiązania, a także od wymagań dotyczących czasu i jakości zebranych informacji. Najczęściej spotyka się następujące metody:

- A. **Ręczne wprowadzanie danych** - dane pojedynczo może wprowadzać do systemu przeszkolony personel, dane można skanować z etykiet produktowych przy wykorzystaniu odpowiednich narzędzi.
- B. **Dane dostawców (producentów, dystrybutorów, hurtowników)** - dostawcy często udostępniają specyfikacje i opisy produktów, które można bezpośrednio importować do katalogu elektronicznego. Dane wprowadza przeszkolony personel.
- C. **Źródła zintegrowane tzw. „product feed”** - czyli wszelkie interfejsy umożliwiające integrację katalogu danych z zewnętrznym źródłem świadczącym usługi udostępnienia informacji o produktach, platformami handlowymi, dużymi serwisami e-commerce.

- D. **Automatyczne zbieranie danych, tzw. „web scraping”** – czyli techniki, które pozwalają na automatyczne pozyskiwanie danych o produktach ze stron internetowych dostawców lub konkurencji, z reguły bez ich wiedzy. Jest to szczególnie przydatne w przypadku monitorowania cen i dostępności produktów, ale może służyć także do pobierania danych o produktach.
- E. **Crowdsourcing** - zbieranie danych od użytkowników lub klientów, którzy mogą dodawać informacje o produktach, recenzje czy oceny. Jest to szczególnie wartościowe w kontekście tworzenia bogatej, wielowymiarowej bazy danych.

Każda z tych metod ma swoją specyfikę, a efektywne zarządzanie katalogiem elektronicznym często wymaga kombinacji różnych technik. Interesującym podejściem w tym kontekście jest crowdsourcing, który wykorzystuje potencjał zbiorowej inteligencji i angażuje szerszą społeczność w proces zbierania danych, oferując tym samym nowe możliwości wzbogacenia i weryfikacji informacji produktowych w sposób, w który tradycyjne metody mogą nie być w stanie zapewnić. Ta technika ma też dodatkową zaletę – w przeciwieństwie do technik Web Scrapingu, można udzielić informacji zwrotnej (ang. feedback) wprowadzającemu dane, jeżeli te nie będą spełniały odpowiedniego kryterium. Jest ona też istotna z punktu widzenia niniejszej dysertacji. Katalog będący przedmiotem badań posiada elementy katalogu crowdsourcingowego.

Przybliżając klasyczną definicję crowdsourcingu - to proces, w którym zadania, projekty lub problemy, zwykle wykonywane przez indywidualnych pracowników lub specjalistyczne firmy, są rozdzielane i przekazywane do szerokiej, często anonimowej grupy osób (Brabham, 2008; Prpic i in., 2017) poprzez publiczne ogłoszenie, zwykle za pośrednictwem Internetu. Termin ten, to połączenie słów „crowd” – tłum i „outsourcing” – czyli praktyki delegowania zadań na zewnątrz (Hirth i in., 2011). Został on wprowadzony przez Jeffa Howe’a, dziennikarza magazynu „Wired”, w artykule "Rise of Crowdsourcing" z 2006 roku, który opisywał, w jaki sposób firmy wykorzystywały Internet do "zlecenia pracy tłumowi".

Crowdsourcing, wg definicji umożliwia zatem zlecenie społeczności szerokiego spektrum zadań, od prostych mikro-zadań po bardziej skomplikowane projekty. Listę typowych zadań realizowanych przez crowdsourcing przedstawiono w Tabeli 4.

Tabela 4. Zadania wykonywane poprzez crowdsourcing

Kategoria	Opis
Zbieranie Danych	Gromadzenie informacji, wprowadzanie danych, ankiety, badania rynku, gdzie wymagane jest zebranie dużej ilości danych od różnych osób.
Rozwiązywanie Problemów i Innowacje	Wykorzystanie „tłumu” do generowania pomysłów, rozwiązywania problemów technicznych, naukowych lub biznesowych, na przykład przez konkursy innowacyjne lub platformy typu InnoCentive.
Testowanie i Ocena Jakości	Crowdsourcing może być używany do testowania oprogramowania, stron internetowych, aplikacji mobilnych, a także do oceny jakości usług lub produktów przez zewnętrzne grupy użytkowników.
Projekty Kreatywne	Zadania takie jak projektowanie graficzne, tworzenie treści, pisanie, tłumaczenia, a nawet rozwój koncepcji marketingowych i kampanii reklamowych.
Mikro-zadania	Proste, powtarzalne zadania, które można łatwo podzielić na mniejsze części, jak tagowanie zdjęć, klasyfikacja obrazów, proste zadania edycyjne, które są realizowane przez platformy takie jak Amazon Mechanical Turk.
Zbieranie Funduszy (Crowdfunding)	Pozyskiwanie kapitału na projekty, przedsięwzięcia biznesowe, cele charytatywne lub twórcze przez zgromadzenie wkładów finansowych od licznej grupy osób, często za pośrednictwem platform takich jak Kickstarter czy Indiegogo.
Opinie i Badania Rynkowe	Zbieranie opinii klientów, testowanie produktów, przeprowadzanie badań rynkowych, aby uzyskać wgląd w preferencje konsumentów i tendencje rynkowe.
Opracowywanie Treści i Zarządzanie Mediami Społecznościowymi	Tworzenie treści dla stron internetowych, blogów, mediów społecznościowych oraz zarządzanie interakcjami z użytkownikami na platformach społecznościowych.
Zadania Edukacyjne i Szkoleniowe	Rozwój materiałów edukacyjnych, kursów online, szkoleń i warsztatów z wykorzystaniem wiedzy i doświadczenia „tłumu”.

Źródło: opracowanie własne

Crowdsourcing jako metoda szybkiego pozyskiwania i przetwarzania danych zyskuje na popularności. W swojej analizie Dziuba (2012) uwydatnia, że crowdsourcing stał się wręcz kluczowym elementem przekształcającym działanie firm w erze cyfrowej, wpływając na kształt elektronicznych rynków pracy i ma to charakter globalny. Zauważył, że jako model biznesowy i strategia e-biznesowa, crowdsourcing przynosi znaczące korzyści społeczno-ekonomiczne, między innymi poprzez umożliwienie firmom redukcji kosztów, zwłaszcza w obszarze badań i rozwoju, oraz przyspieszenia realizacji projektów dzięki zaangażowaniu szerokiego „tłumu” osób. Crowdsourcing ewoluował jako nowatorska forma współpracy na rynku globalnym, pozwalając na kooperację firm z rozległymi zewnętrznymi grupami i jednostkami. Jednakże, jak podkreśla, przedsiębiorstwa narażone są na znaczne ryzyko, przenosząc odpowiedzialność

na „tłum” z sieci. Potwierdza to Zheng (2014), który podkreśla znaczenie „aktywnej komunikacji w procesie crowdsourcingu”, wskazując, że interakcja między sponsorem a osobami rozwiązującymi zadania ma kluczowy wpływ na jakość opracowanych rozwiązań, co okazuje się być bardziej efektywne niż systemy nagród.

Temat crowdsourcingu w katalogach elektronicznych nie jest zagadnieniem często poruszonym, co ilustruje sieć powiązań tematyki publikacyjnej (Rysunek 1). Związane jest to z faktem, że o ile dane nie są wprowadzane przez wykwalifikowaną kadrę pracowników danej firmy, są najczęściej pobierane z Internetu za pośrednictwem wspomnianej techniki „web scrapingu”. Jest to metoda bardzo szybka i prosta w implementacji, zważywszy, że podstawowe dane dotyczące produktów są specjalnie oznaczone na stronach internetowych w postaci odpowiednio zakodowanych metadanych, w celu łatwiejszego indeksowania przez wyszukiwarki internetowe takie jak Google czy Bing (Schema.org Community Group, 2015). Zatem, pomijając aspekty etyczne i prawne, o ile dane są dostępne publicznie w Internecie, z łatwością można je selektywnie pobrać i przetwarzać na własne potrzeby. Jest to technika powszechnie stosowana w różnych dziedzinach, takich jak analiza konkurencji, badania rynkowe, monitorowanie cen, agregacja treści oraz w projektach badawczych związanych z przetwarzaniem danych. Niestety, dane uzyskane za pomocą techniki web scrapingu często zawierają błędy, są niespójne i niedokładne, ponieważ pochodzą z różnych źródeł i nie przechodzą żadnego procesu weryfikacji. W przypadku katalogów elektronicznych, gdzie precyzja i aktualność informacji są kluczowe, opieranie się wyłącznie na danych zebranych w ten sposób może być więc ryzykowne. Web scraping, podobnie jak procesy pobierania i agregacji danych z rozmaitych, nieautoryzowanych źródeł, charakteryzują się także kolejną istotną wadą: nie istnieje możliwość zapewnienia informacji zwrotnej ani korekty danych przez ich pierwotnych autorów. W związku z tym, wszelkie mechanizmy automatyzacji muszą polegać wyłącznie na informacjach już dostarczonych, bez potencjału do ich modyfikacji czy walidacji.

Crowdsourcing jako alternatywna metoda pozyskiwania danych, pozwala na włączenie do procesu osób posiadających aktualne i praktyczne informacje o produktach. Pozwala to na uzupełnienie i weryfikację zebranych danych, co jest szczególnie ważne w przypadkach, gdy dane produktowe szybko się zmieniają lub gdy wymagana jest wysoka dokładność informacji. Pomimo wyzwań związanych z kontrolą jakości danych wprowadzanych przez użytkowników, crowdsourcing może znacząco przyczynić się do poprawy jakości i wartości informacji w katalogach elektronicznych.



W rozważaniu na temat źródeł danych o produktach i metod ich pozyskania warto zwrócić szczególną uwagę na podobieństwa i różnice pomiędzy definicją typowego modelu crowdsourcingowego, a zbieraniem danych od szerokiego grona dostawców. Crowdsourcing zwykle odnosi się do procesu, w którym informacje o produktach są zgłaszane i uzupełniane przez grupę użytkowników, którzy nie muszą być bezpośrednio związani z firmą lub produktem. Użytkownicy ci mogą dostarczać dane na podstawie osobistego doświadczenia, obserwacji rynku lub innych nieformalnych źródeł. Z kolei zbieranie danych od dostawców polega na gromadzeniu informacji bezpośrednio od producentów, dystrybutorów, hurtowników czy innych oficjalnych źródeł. Dane te są zwykle bardziej formalne, obejmując specyfikacje produktu, ceny, informacje o dostępności i inne dane techniczne. Te informacje są zazwyczaj bardziej wiarygodne, dokładne i są aktualizowane w sposób bardziej systematyczny niż informacje uzyskiwane przez crowdsourcing. Jednak, jak pokazują badania (Karpishek i in., 2011; Niemir i Mrugalska, 2022a), dane, nawet jeśli pochodzą faktycznie bezpośrednio od producentów, nadal zawierają liczne błędy. Niemir i Mrugalska (2023a) wymieniają kilka powodów, dla których ma to miejsce:

- a) błąd, literówki, zaniedbania
- b) brak przeszkolenia personelu
- c) usterki w oprogramowaniu
- d) brak standaryzacji danych.

Można zatem założyć, że o ile dane nie pochodzą od wąskiego grona dostawców, ale dostarczane są przez szeroką społeczność dostawców, to problemy z jakością i sposoby ich rozwiązania będą zbliżone do omawianego modelu crowdsourcingowego. W niniejszej pracy, przedmiotem badań jest katalog produktów napelniany przez kilkadziesiąt tysięcy producentów oraz dystrybutorów, czyli w tym ujęciu od szerokiego grona dostawców.

## **2.6. Podstawowe dane produktowe opisujące i identyfikujące produkt**

Pojęcie danych produktowych jest powszechnie rozumiane jako obejmujące wszystkie informacje związane z produktem (Kropsu-Vehkaperi i Haapasalo, 2011), które można odczytać, zmierzyć i ustrukturyzować w odpowiednim formacie. Odnosi się to do fizycznych i funkcjonalnych atrybutów produktu, w tym jego szczegółowych informacji technicznych, ale także informacji abstrakcyjnych i koncepcyjnych (Saaksvuori i Immonen, 2008). Atrybuty produktu mogą obejmować jego wymiary, wagę, materiał, z którego jest wykonany,

specyfikacje techniczne, takie jak moc czy pojemność, oraz niematerialne aspekty, takie jak marka, kategoria, zgodność z normami i różne parametry użyteczności, które wpływają na percepcję produktu przez konsumentów. Dane dotyczące produktu można również podzielić według następujących obszarów: identyfikacja, klasyfikacja i opis (Schemm i Legner, 2008). Często określa się je jako tzw. "Master Data", czyli dane służące do uspoźnienia informacji (Popa i Duica, 2011). Mogą być one podzielone na dane neutralne i dane zależne od relacji. Dane neutralne są to informacje, które są zazwyczaj współdzielone pomiędzy wieloma stronami i są niezależne od relacji. Te dane można podzielić na trzy kategorie:

- A. Dane podstawowe produktu (ang. Core Product Data) - to podstawowe atrybuty danych, które dotyczą wszystkich instancji jakiegokolwiek produktu. Przykłady to opis, nazwa marki, opakowanie, wymiary itd.
- B. Dane specyficzne dla kategorii (ang. Category Specific Data) - to atrybuty danych, które dotyczą tylko określonych kategorii produktów. Na przykład: kolor, rodzaj winogron i moc dla butelki wina.
- C. Dane rynku docelowego (ang. Target Market Data) - to atrybuty danych, które są specyficzne dla produktu na konkretnym rynku. Na przykład: wskaźniki opakowania w konkretnym kraju.
- D. Dane zależne od relacji (ang. Relationship Dependent Data) - to atrybuty danych dotyczące wszystkich warunków uzgodnionych dwustronnie i komunikowanych pomiędzy partnerami handlowymi, takich jak warunki marketingowe, informacje o cenach i rabatach, umowy logistyczne i inne.

Oprócz wspomnianej definicji „Core Product Data” (Popa i Duica, 2011), w literaturze można znaleźć wiele innych definicji jak i nazw oznaczających podstawowe dane produktu. Na przykład (Schäffer i Stelzer, 2017) definiują je jako zestaw danych reprezentujących produkt, (Brechan, 2006) jako podstawowe atrybuty produktu, a (Madlberger, 2011) jako główne dane przedmiotu. Taki zestaw danych może opisywać różne rodzaje cech czy właściwości produktu - zarówno fizyczne, strukturalne, jak i składników (Schäffer i Stelzer, 2017). Należy jednak zaznaczyć, że „core product data” nie jest tożsamy z „basic data” i może być jego podzbiorem. Dane „basic” mogą być w wielu systemach rozumiane znacznie szerzej i skategoryzowane jedynie po to, aby oddzielić je od danych logistycznych i sprzedażowych. Przykładowo, Krieger i Schorr (2019) opisuje, że atrybuty produktów są często klasyfikowane w oprogramowaniu ERP według tematycznych kategorii do której należą między innymi dane podstawowe („basic data”) które obejmują opis produktu, jednostki opakowaniowe, wymiary,

objętości, wagi brutto i netto, kody substancji niebezpiecznych i tym podobne, przy czym w dalszej części przy czym w dalszej części artykułu podkreślono, że tabela danych podstawowych pierwszej firmy zawiera 290 atrybutów, z czego wykorzystywanych jest 80, a 42% z nich to atrybuty dostosowane do konkretnych potrzeb klienta. Takie dane nie mogą być nazwane danymi podstawowymi, ponieważ informacje dostosowane specyficznie dla klienta zwykle należą do kategorii danych niestandardowych lub konfigurowalnych, które są dopasowywane do indywidualnych wymagań klientów i często są dodatkowe w stosunku do uniwersalnych danych podstawowych. Dane podstawowe zazwyczaj więc obejmują ogólne informacje o produkcie, które są niezbędne do jego identyfikacji i klasyfikacji w ramach standardowych procesów biznesowych i nie zmieniają się w zależności od klienta. Niemir i Mrugalska (2021) opisują dwa, kluczowe atrybuty danych produktowych:

- A. Nazwa produktu, która jednoznacznie i w pełni identyfikuje konkretny produkt, biorąc pod uwagę jego wariant i markę, bez potrzeby znajomości innych atrybutów produktu, bez konieczności przeglądania zdjęć czy fizycznego oglądania produktu.
- B. Unikalny identyfikator - jednoznacznie reprezentujący produkt w łańcuchu dostaw, opracowany według jednego standardu i interpretowany w ten sam sposób przez wszystkie systemy informacyjne.

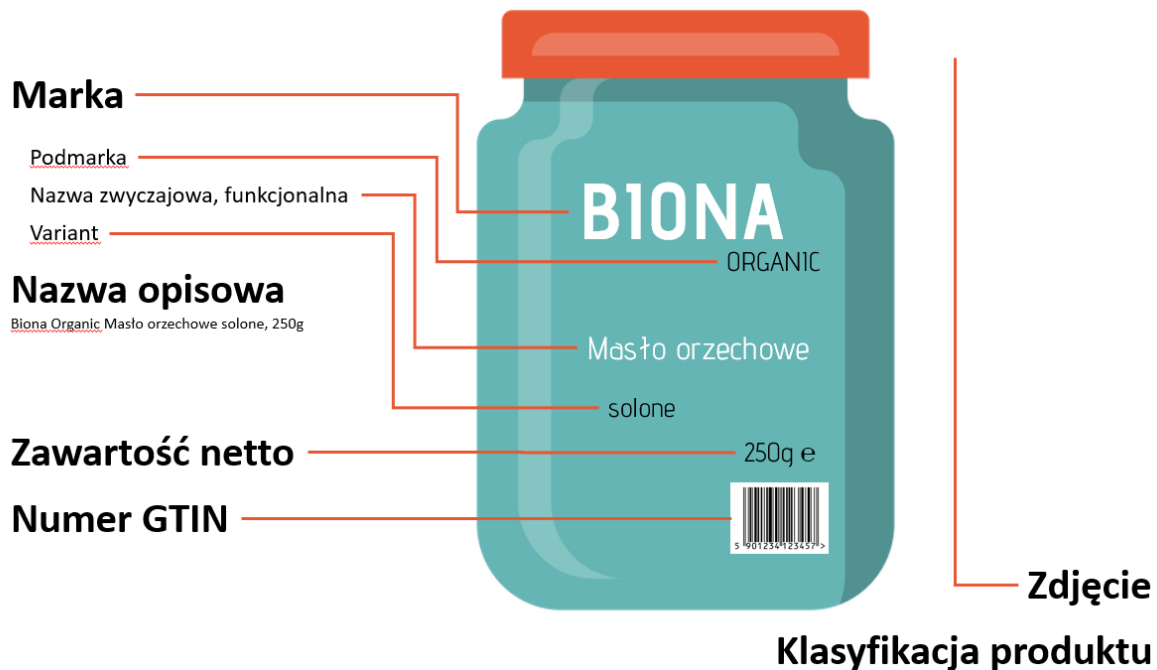
Rozpowszechnionym numerem, który spełnia wymagania unikalnego identyfikatora produktu jest GTIN (Global Trade Item Number) wydawany przez organizację GS1 (GS1, 2021), która zrzesza ponad 2 miliony firm na całym świecie. Numery GTIN obecnie znajdują się na 100 milionach różnych produktów i są zakodowane za pomocą kodów kreskowych (np. popularny w Europie EAN-13, czy w Stanach Zjednoczonych UPC), dzięki czemu możliwe jest efektywne śledzenie, identyfikacja i zarządzanie ogromną różnorodnością towarów na całym świecie. Ta wszechobecność numerów GTIN umożliwia organizacjom oraz konsumentom szybką i dokładną weryfikację informacji o produkcie, co jest kluczowe w zarządzaniu łańcuchami dostaw, logistyce i w handlu detalicznym. Z racji prostoty i powszechności w użyciu, korzystanie z numerów GTIN ułatwia także automatyzację procesów biznesowych, poprawiając dokładność danych i efektywność operacji, jak również umożliwia lepszą komunikację między różnymi podmiotami. Numery GTIN są coraz popularniejsze również w e-commerce, pozwalając na łatwe i jednoznaczne identyfikowanie produktów w systemach online, co przekłada się na lepsze doświadczenia zakupowe dla konsumentów i efektywne zarządzanie zapasami przez sprzedawców. Jak opisuje Niemir i Mrugalska (2021), organizacja GS1 intensywnie dąży do usystematyzowania danych

produktowych w skali globalnej nie tylko w sferze numeru identyfikacji. Jednym z kluczowych przedsięwzięć w tym zakresie jest implementacja Globalnego Modelu Danych (Global Data Model, GDM), który definiuje zestawy atrybutów dla różnych branż (GS1, 2023c). Jak zauważają Osmólski i Muszyński (2020), Global Data Model służy jako fundament dla wprowadzania platform technologicznych i innych standardów mających na celu podniesienie jakości informacji o produktach. Dzięki standaryzacji atrybutów produktowych, GDM staje się narzędziem harmonizującym dane na szeroką skalę, co jest niezbędne dla efektywnej i przejrzystej koordynacji łańcuchów dostaw.

Od roku 2019, wszyscy uczestnicy Systemu GS1, korzystający z Globalnych Numerów Handlowych (GTIN), mają obowiązek wprowadzania podstawowych informacji o swoich produktach do globalnego rejestru produktów - Global Registry Platform (GRP). Jest to niezbędne do uzyskania nowego numeru GTIN (Niemir i Mrugalska, 2021). W odpowiedzi na zapotrzebowanie wyrażone przez Consumer Group Forum (CGF), w ramach GRP, zainicjowano usługę "Verified by GS1", która umożliwia weryfikację podstawowych danych produktowych za pomocą interfejsu programistycznego (API). Celem tej inicjatywy, która rozwinęła się z projektów pilotażowych, jest walidacja unikalnych numerów GTIN i kluczowych atrybutów produktów, co ma przyczynić się do osiągnięcia wyższej jakości danych, większej kompletności i łatwiejszych sposobów dzielenia się informacjami (GS1.pl, 2023). W ramach rejestru GRP, producenci są zobowiązani do dostarczenia następujących podstawowych atrybutów produktów:

- A. GTIN (Globalny Numer Handlowy z ang. Global Trade Item Number).
- B. Nazwa opisowa (możliwa do powtórzenia z uwzględnieniem kodu języka).
- C. Nazwa marki (możliwa do powtórzenia z uwzględnieniem kodu języka).
- D. Zdjęcie produktu (możliwy do powtórzenia z uwzględnieniem kodu języka).
- E. Zawartość netto i jednostka miary (możliwe do powtórzenia z uwzględnieniem jednostki).
- F. GPC (Globalna Klasyfikacja Produktów, z ang. Global Product Classification).
- G. Kraje sprzedaży / rynek docelowy (w formie listy).

Rysunek 7 wizualizuje koncepcję dotyczącą źródła pobierania atrybutów podstawowych dla danych produktowych.

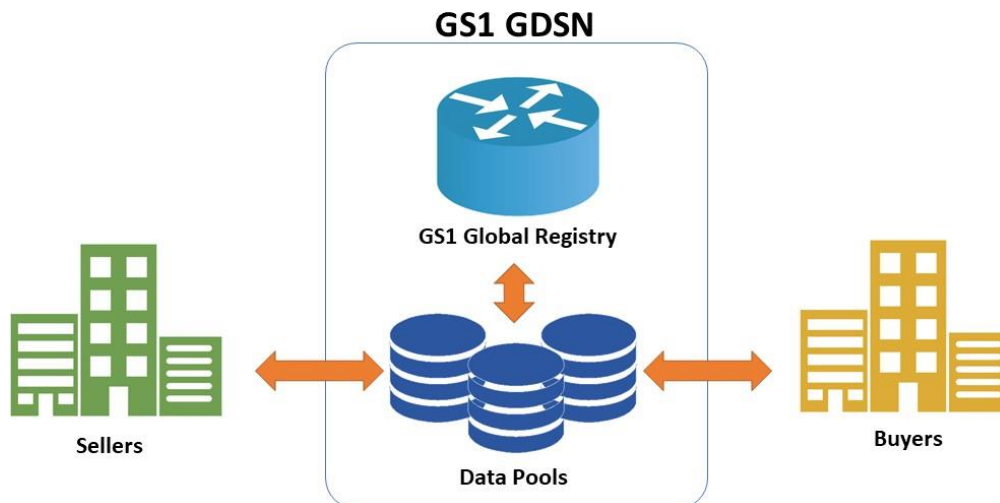


Rysunek 7. Podstawowe atrybuty produktu na etykiecie produktu  
Źródło: opracowanie własne

Organizacja GS1, odpowiedzialna za rozwój GRP, wykorzystwała zestaw sprawdzonych, istniejących standardów, bazując na rozwiązaniach zaimplementowanych wcześniej w sieci GDSN - Global Data Synchronization Network. Ten ruch skutkował znacznym podobieństwem terminologii i znaczenia atrybutów w obu systemach, choć model danych GRP uległ znaczącemu uproszczeniu w porównaniu do GDSN (Niemir i Mrugalska, 2021).

GDSN, funkcjonująca od 2004 roku, jest internetową siecią interoperacyjnych pul danych, które są zgodne ze standardami GS1 (GS1, 2023a). Sieć ta umożliwia globalną wymianę standardowych danych podstawowych o produktach między partnerami biznesowymi. GDSN charakteryzuje się szerokim zakresem danych produktowych, które są jednak rozproszone i przechowywane w tzw. "Data Pools", co przedstawiono na Rysunku 8. Na świecie funkcjonuje 44 takich pul danych, zarządzanych przez zewnętrzne firmy. W ramach tych pul danych, setki atrybutów produktów są przypisane do określonych grup produktowych, co umożliwia tworzenie modeli danych specyficznych dla poszczególnych branż i efektywną komunikację między partnerami w łańcuchu dostaw. W przeciwieństwie do GDSN, GDSN Global Registry przechowuje jedynie odniesienia do danych centralnie. Choć terminologia

i znaczenie atrybutów w GRP są podobne do tych w GDSN, GRP nie jest częścią infrastruktury GDSN, a role GDSN Global Registry i Global Registry Platform różnią się od siebie.



Rysunek 8. Struktura sieci GDSN  
Źródło: opracowanie własne

Z punktu widzenia logistyki i handlu, rozwiązania organizacji GS1, w tym GRP, mają istotne znaczenie. Skala przedsięwzięcia związana z agregacją danych wokół Globalnego Rejestru Danych może sprawić, że GRP stanie się największym rejestrem produktów na świecie, a atrybuty wymagane przez ten system, takie jak globalny unikalny identyfikator produktu, unikalna nazwa opisująca produkt, nazwa marki, zdjęcie produktu, zawartość netto, klasyfikacja produktu i kraje sprzedaży, mogą stać się referencyjne dla innych katalogów i systemów zarządzania produktami.

Podsumowując, precyzyjne określenie podstawowych danych produktowych jest niezbędne dla efektywności i dokładności w wielu obszarach biznesowych, w tym w logistyce, handlu detalicznym i zarządzaniu łańcuchami dostaw. Choć nie ma powszechnie akceptowanej definicji dotyczącej atrybutów, które powinny być uznawane za podstawowe, to jednak istotne jest uznawanie zarówno nazwy produktu, jak i numeru GTIN za kluczowe elementy tych danych. Dodatkowo, należy zwrócić uwagę na atrybuty używane w Globalnym Rejestrze Danych (GRP) organizacji GS1, które mogą służyć jako model dla przyszłych podstawowych danych produktowych. GRP może stać się kluczowym źródłem wiarygodnych i jednolitych danych produktowych, co byłoby ogromnym krokiem naprzód w zapewnieniu spójności i efektywności w zarządzaniu danymi produktowymi.

## 2.7. Główne źródła błędów w danych produktowych

W ramach rosnącego zainteresowania zarządzaniem jakością danych, szczególnie istotne staje się zrozumienie podstawowych przyczyn błędów w danych produktowych.

Jak wskazują Dasu i Johnson (1999) oraz Rahm i Do (2000), przyczyny niepoprawności danych często powiązane są ze źródłami danych. Wśród typowych przyczyn błędów autorzy wyróżniają kilka kluczowych kategorii, które, choć nie odnoszą się bezpośrednio do danych produktowych, mogą być z łatwością zaadaptowane i zastosowane do analizy jakości danych produktowych w różnych systemach zarządzania informacją. Uzyskujemy zatem następujące kategorie:

- A. **Błędy transkrypcji dla danych wprowadzanych ręcznie.** Te błędy pojawiają się podczas manualnego wprowadzania informacji o produkcie, na przykład podczas przenoszenia specyfikacji z dokumentów papierowych do systemów elektronicznych. Często są to błędy ludzkie, takie jak literówki w nazwach produktów czy pomyłki w wartościach, które mogą prowadzić do nieścisłości w katalogach produktowych.
- B. **Błędy instrumentacji.** Dotyczą one problemów związanych z pomiarami fizycznych właściwości produktów oraz wprowadzaniem tych danych do systemów informacyjnych. Błędy te mogą wynikać z nieprawidłowego kalibrowania urządzeń pomiarowych (np. wag, miarek, urządzeń do pomiaru objętości), błędów w oprogramowaniu wykorzystywanym do automatycznego zbierania danych (np. skanery kodów kreskowych, systemy do automatycznego rozpoznawania obrazu) lub nieprawidłowej interpretacji danych surowych (np. niepoprawne przeliczanie jednostek miary).
- C. **Błędy w kompilacji danych.** Odnoszą się do procesu agregacji danych o produkcie z różnych źródeł, takich jak bazy danych dostawców czy arkusze kalkulacyjne z danymi sprzedaży. Błędy mogą wystąpić na etapie integracji tych danych, generując niespójności i błędy w specyfikacjach produktów, opisach, cenach czy dostępności.
- D. **Niejednoznaczność w specyfikacji danych.** W tej kategorii błędy wynikają z niejasnych lub wieloznacznych definicji i specyfikacji produktów, co prowadzi do różnorodnych interpretacji i potencjalnych nieścisłości. Może to obejmować niejednoznaczność w nazewnictwie produktów, ich klasyfikacji czy opisach cech, co utrudnia jednoznaczne zidentyfikowanie i porównanie produktów.

E. **Błędy transformacji.** Występują podczas przenoszenia informacji o produkcie z jednego systemu lub formatu do innego, na przykład podczas eksportu danych z systemu zarządzania magazynem do sklepu internetowego. Problemy te mogą wynikać z niezgodności formatów danych, błędów w mapowaniu atrybutów produktów czy utraty informacji podczas konwersji, co może prowadzić do błędów w prezentacji produktów online.

Mao i inni (2019) zauważają, że problemy jakości danych mogą wywodzić się z wielu źródeł, w tym z systemów norm jakości danych, metod zbierania danych, stopnia współdzielenia i ponownego wykorzystywania danych, mechanizmów zarządzania jakością danych, poziomu operacyjnego operatorów, a także stopnia wykorzystania informacji przez użytkowników. Ponadto, autorzy publikacji podkreślają, że problem jakości danych może wystąpić w dowolnym momencie cyklu życia danych, od ich utworzenia, poprzez przechowywanie, aż po użytkowanie końcowe. W kontekście danych produktowych, szczególnie tych dostępnych w Internecie, często można zauważyć konsekwencje ponownego wykorzystywania danych, co prowadzi do propagacji błędów i nieścisłości.

Taleb i inni (2021) powołując się na badania Fürber i Hepp (2011) twierdzą, że problemy z jakością danych pojawiają się ogólnie wtedy, gdy nie są spełnione określone wymagania jakościowe w stosunku do wartości danych. Problemy te są wynikiem różnych czynników lub procesów, które mają miejsce na wielu poziomach organizacji i procesów przetwarzania informacji. Autorzy identyfikują szereg poziomów, na których mogą występować błędy danych:

- A. **Poziom źródła danych:** W tym kontekście błędy mogą wynikać z niepewności co do dokładności danych, niskiego poziomu zaufania do źródeł danych, problemów związanych z kopiowaniem danych, niespójnością danych między wieloma źródłami, a także z ograniczeniami dziedziny danych.
- B. **Poziom tworzenia danych:** Błędy mogą pojawiać się podczas ręcznego wprowadzania danych, z odczytów sensorów, z mediów społecznościowych, a także w wyniku obecności danych nieustrukturyzowanych i brakujących wartości.
- C. **Poziom procesu:** W tym obszarze, błędy mogą wynikać z procesów nabycia danych, w tym zbierania i przesyłania danych, co może prowadzić do problemów z jakością danych w dalszych etapach przetwarzania.



Autorzy dokonują także syntezy przyczyn słabej jakości danych powołując się na inne publikacje (Chen i in., 2012; Corrales i in., 2018; Laranjeiro i in., 2015). Syntezę przedstawiono w Tabeli 5.

Tabela 5. Źródła problemów z jakością danych.

Źródło danych	Poziom	Problem	Dokładność	Pełność	Spójność		
<b>Pojedyncze źródła danych</b>	Poziom indywidualnej komórki	Brakujące dane	•	•			
		Błędne dane i referencje, błędy, literówki	•				
		Nieistotne dane			•		
		Przestarzałe dane	•				
		Błędnie wprowadzone i sprzeczne wartości	•	•	•		
	Poziom schematu zbioru danych	Ograniczenia domeny i unikalności, naruszenie zależności	•				
		Nieprawidłowy typ danych, słaby projekt schematu			•		
		Naruszenie integralności referencyjnej	•	•	•		
		<b>Wielorakie źródła danych</b>	Poziom indywidualnej komórki	Różne jednostki, reprezentacje, konflikty			•
				Różne poziomy agregacji, niespójna agregacja	•		•
Niespójność czasowa	•						
Poziom schematu zbioru danych	Heterogeniczne modele danych i projekt schematu	•	•	•			
	Różne formaty kodowania			•			

Źródło: opracowanie własne na podstawie Taleb i in. (2021)

Podane źródła doskonale odnoszą się do rzeczywistych problemów z danymi produktowymi i stanowią kluczowy materiał do analizy.

Pantziarka (2005) zauważa, że w obliczu ogromnego wzrostu dostępnych danych w różnych dziedzinach oraz różnorodności ich źródeł, można ogólnie sklasyfikować błędy na dwie kategorie: błędy syntaktyczne i błędy semantyczne. Błędy syntaktyczne dotyczą poszczególnych elementów danych. Wśród nich wyróżniamy:

- A. Błędy typu danych, takie jak wartości całkowite zamiast liczby zmiennoprzecinkowej, dane tekstowe zamiast liczbowych.
- B. Błędy zakresu, na przykład procenty przekraczające 100% czy liczba godzin w tygodniu przewyższająca 168.

- C. Błędy formatu, w tym nieprawidłowy format daty, niestandardowe kodowanie numerów telefonów.
- D. niespójne reprezentacje, jak różne formy zapisu „Street” (St, Str, Street).
- E. Problemy z granularnością danych, na przykład wprowadzenie adresu jako pojedynczego elementu tekstowego zamiast sekwencji odrębnych elementów (numer domu, ulica, miasto, stan).

Analiza tych błędów zazwyczaj jest prostym procesem sprawdzania, czy każdy element odpowiada pożądanemu typowi i zakresowi wartości (de Almeida i in., 2013; Singh i Vashishtha, 2015). Problemy z granularnością danych, znane również jako problemy z „wbudowanymi danymi”, są bardziej złożone (Rahm i Do, 2000).

Druga klasa to błędy semantyczne. Odnoszą się do błędów w znaczeniu zmiennych. Każda zmienna w zbiorze danych jest ograniczona przez regułę pomiaru, określającą związek między treścią zmiennej a rzeczywistą cechą lub obiektem, który ta zmienna ma reprezentować. Etykieta zmiennej często jawnie koduje regułę pomiaru, na przykład w przypadku zmiennej „Indeks Dow Jones” z wartościami reprezentującymi dzienną cenę zamknięcia. Analiza semantyczna bada, czy zmienna rzeczywiście reprezentuje to, co deklaruje, sprawdzając, czy treść zmiennej została zmierzona zgodnie z regułą pomiaru. Przykłady błędów semantycznych obejmują:

- A. Nieprawidłowe jednostki pomiaru, jak temperatura rejestrowana w Fahrenheitach zamiast w Celsjuszach.
- B. Nieprawidłowy okres, na przykład roczny koszt zamiast miesięcznego.
- C. Błędy zakresu, zawierające wartości, które powinny zostać wykluczone.
- D. Podwójne liczenie, gdy zmienne zawierają dane już uwzględnione w innych zmiennych.
- E. Nieprawidłowe skalowanie, na przykład koszty zebrane w tysiącach zamiast w jednostkach.

W przeciwieństwie do sprawdzania składni, które można uznać za proces „atomowy” – stosujący reguły składni do pojedynczych, izolowanych zmiennych – sprawdzanie semantyczne rzadko jest tak proste. Jedna wartość w zmiennej zazwyczaj nie wystarcza, aby stwierdzić czy przestrzegano reguły pomiaru; wymaga to odpowiedzi na pytanie, czy zmienna „ma sens” w swoim kontekście. Jeśli błąd semantyczny można rozpoznać, patrząc na pojedynczą wartość, można go zakodować jako regułę składniową. Chociaż rozróżnienie

między błędami syntaktycznymi a semantycznymi jest kluczowe w kontekście walidacji danych, literatura na temat jakości danych rzadko formułuje to rozróżnienie w sposób jawny. Ten podział doskonale wpisuje się w zagadnienia badane w niniejszej dysertacji, gdzie zauważono, że poleganie jedynie na walidacji składni to zdecydowanie za mało, natomiast analiza semantyczna może kończyć się niepowodzeniem ze względu na trudności w doborze walidatorów.

Badania przeprowadzone przez Ge i Helferta (2007) szeregują problemy jakościowe z perspektywy danych i z perspektywy odbiorcy w macierzowym połączeniu z zależnościami kontekstu. Jest to podobne ujęcie semantyki i syntaktyki, ale przedstawione w sposób bardziej złożony i wielowymiarowy, który pozwala na głębsze zrozumienie interakcji między danymi a ich użytkownikami oraz kontekstem, w jakim dane są wykorzystywane. Wyniki tych badań zaprezentowano w Tabeli 6.

Tabela 6. Problemy z jakością informacji w danych produktowych

	Perspektywa danych	Perspektywa odbiorcy
<b>Niezależne kontekstowo</b>	Błąd pisowni	Informacja jest niedostępna
	Brak danych	Informacja jest niebezpieczna
	Zduplikowane dane	Informacja jest trudno dostępna
	Nieprawidłowa wartość	Informacja jest trudna do agregacji
	Niespójny format danych	Błędy w transformacji informacji
	Przestarzałe dane	
	Niekompletny format danych	
	Naruszenie składni	
	Naruszenie unikalności wartości	
	Naruszenie ograniczeń integralności	
	Formatowanie tekstu	
<b>Zależne kontekstowo</b>	Naruszenie ograniczeń domeny	Informacja nie opiera się na faktach
	Naruszenie zasad biznesowych organizacji	Informacja jest wątpliwej wiarygodności
	Naruszenie przepisów firmy i rządu	Informacja prezentuje bezstronny punkt widzenia
	Naruszenie ograniczeń narzuconych przez administratora bazy danych	Informacja jest nieistotna dla pracy
		Informacja składa się z niespójnych znaczeń
		Informacja jest niekompletna
		Informacja nie jest zwarto przedstawiona
		Informacja jest trudna do manipulacji
	Informacja jest trudna do zrozumienia	

Źródło: opracowanie własne na podstawie Haug i in. (2011)

Przedstawiona „perspektywa danych” obejmuje techniczne aspekty informacji. Problemy takie jak błędy pisowni, brak danych, czy naruszenie składni, są zazwyczaj łatwiejsze do

zidentyfikowania i skorygowania poprzez techniczne środki, takie jak walidacja danych i automatyczne narzędzia korekcyjne. Perspektywa odbiorcy, z drugiej strony, odnosi się do sposobu, w jaki informacje są postrzegane i wykorzystywane przez odbiorców. Problemy takie jak informacje, które są "trudne do agregacji" lub "nie opierają się na faktach", wymagają bardziej złożonego podejścia, które może obejmować edukację użytkowników, zarządzanie oczekiwaniami i rozwijanie kompetencji krytycznego myślenia. Podział na kwestie niezależne od kontekstu i zależne od kontekstu podkreśla, że niektóre problemy jakości informacji można rozwiązać bez względu na kontekst ich użycia, podczas gdy inne wymagają rozumienia i analizy specyficznego środowiska biznesowego lub operacyjnego.

Przegląd literatury dotyczącej źródeł błędów w danych produktowych ujawnia złożoność problemów, podkreślając, że pochodzą one z wielowymiarowego spektrum przyczyn. Źródła nieprawidłowości obejmują nieścisłości na etapie pozyskiwania danych, błędy generowane w trakcie ich tworzenia, a także wyzwania proceduralne i komplikacje wynikające z korzystania z różnych baz danych. Skutkują one szerokim zakresem problemów, w tym niedokładnościami, brakami w danych, pomyłkami podczas wprowadzania informacji, niespójnościami oraz ograniczeniami związanymi ze specyfiką danej dziedziny. Kluczowym wnioskiem jest jednak to, że skuteczna walidacja danych wymaga podejścia wykraczającego poza tradycyjne ramy systemów informatycznych i baz danych, co stanowi wyzwanie, zwłaszcza przy ograniczonej kontroli nad źródłami danych i ich jakością.

## **2.8. Systemy zarządzania danymi produktowymi i wyzwania w kontekście problematyki jakości**

Istotnym rozwiązaniem problemu jakości danych produktowych jest implementacja odpowiednich procedur kontroli (Haug i Stentoft Arlbjörn, 2011). W celu usprawnienia procesu kontroli i korekty danych oraz w celu zminimalizowania jego kosztów ważne jest, aby rolę tę przejął w jak największym stopniu scentralizowany system informacyjny. Źródła literatury naukowej identyfikują kilka kluczowych grup narzędzi i systemów stosowanych w tym celu, jednakże, autorzy publikacji wskazują również na wyzwania i ograniczenia tych systemów.

Najnowsze prace naukowe opisują stosunkowo nowe systemy dedykowane zarządzaniu informacjami o produkcie, tzw. PIM (Product Information Management, PIMS, Product Information Management System). Są to systemy informatyczne wykorzystywane do zarządzania informacjami o produktach skierowanych na klienta poprzez unifikację

i synchronizację rozproszonych danych o produkcie (Abraham, 2014; Boyd, 2006; Informatica, 2019). Jak opisują Battistello i współautorzy (2021), główna idea tych systemów polega na tym, że informacje o produktach są scentralizowane i przechowywane w PIMS, skąd następnie mogą być dystrybuowane bez konieczności ręcznego wprowadzania informacji do różnych systemów. Korzyści z wdrożenia PIMS obejmują krótszy czas wprowadzania produktu na rynek, poszerzoną gamę produktów, jednolite doświadczenie klienta w różnych kanałach, lepsze zarządzanie złożonością, kontrolowaną dystrybucję treści oraz zgodność z przepisami prawnymi. Ponadto, przynoszą one obniżone koszty, szybkość odzyskiwania informacji, czyszczenie danych i błędy logistyczne, a także mniej zwrotów i zapytań informacyjnych (Abraham, 2014; Informatica, 2019).

Systemy PIMS (Product Information Management Systems), choć blisko związane z systemami PDM (Product Data Management) (Do, 2018) i PLM (Product Lifecycle Management) (David i Rowe, 2016), różnią się od nich głównie koncentracją na zarządzaniu informacjami o produkcie z perspektywy sprzedaży i marketingu, podczas gdy PDM i PLM obejmują dane związane z produkcją i rozwojem produktu (Abraham, 2014). Systemy PDM zajmują się zarządzaniem danymi technicznymi produktu, takimi jak dane CAD, specyfikacje i dokumentacja techniczna, które są kluczowe w procesie projektowania i produkcji (Cheung i Schaefer, 2010). Z kolei systemy PLM obejmują szerszy zakres, zarządzając całym cyklem życia produktu od koncepcji, przez produkcję, aż do wycofania z rynku, co obejmuje integrację danych, procesów, systemów biznesowych oraz ludzi w całym przedsiębiorstwie (David i Rowe, 2016). W przeciwieństwie do PIMS, które zawierają informacje o sprzedaży i marketingu produktu, PDM i PLM informatyzują zagadnienia związane z wewnętrznym rozwojem produktu i jego produkcji, co stanowi zasadniczą różnicę między tymi systemami i nie jest bezpośrednio w obszarze badawczym niniejszej pracy.

Nie znaleziono źródeł naukowych opisujących problemy bezpośrednio z utrzymaniem jakości danych w systemach PIMS, natomiast literatura opisuje częste przypadki zaniechania wdrożeń tego typu systemów (Abraham, 2014), a jednym z głównych przyczyn podawana jest jakość danych (Battistello i in., 2021). Schäffer i Stelzer (2017) wskazują także, że przy wdrażaniu narzędzi informatycznych programiści i analitycy zazwyczaj koncentrują się na tworzeniu zaawansowanych struktur danych i na sposobie ich przechowywania, czasem zwracając uwagę na procesy pracy. Często pomija się natomiast konsultacje w zakresie technicznych rozwiązań, które mogłyby poprawić jakość danych w kontekście międzyorganizacyjnym. Battistello i współautorzy (2021) opisują, że dane produktowe,

z którymi firmy się spotykają, są zwykle nieustrukturyzowane i błędne. Dodatkowo, uzgadnianie zakresu danych pomiędzy różnymi interesariuszami jest samo w sobie wyzwaniem i często wymaga znalezienia kompromisu. W efekcie, kontrola jakości danych często jest zaniechywana, a rozwijane rozwiązania PIMS zazwyczaj posiadają jedynie podstawowe narzędzia do kontroli danych.

Oprócz PIMS, dane produktowe często znajdują się w systemach ERP (Enterprise Resource Planning). Początkowo sądzono, że nowoczesne systemy ERP będą w stanie rozwiązać problem jakości danych. Systemy ERP to zintegrowane oprogramowanie służące do zarządzania podstawowymi procesami biznesowymi w organizacji, często w czasie rzeczywistym i przy użyciu wspólnej bazy danych. Jednakże, zaobserwowano, że skomplikowane powiązania między modułami ERP mogą powodować, że wprowadzenie danych o niskiej jakości w jednym module negatywnie wpływa na funkcjonowanie pozostałych (Drd'gan i Metz, 2017; Haug i in., 2011).

Salem i Boufares (2019) zwracają uwagę, że ilość danych nieustannie rośnie, podobnie jak ryzyko anomalii. Większość narzędzi służących do migracji danych nie wspiera użytkownika podczas procesu czyszczenia danych, stąd interesujące jest tworzenie nowych narzędzi do skuteczniejszej ich integracji. Autorzy wspominają tu o ogólnej grupie rozwiązań ETL (Extract-Transform-Load), służących do wyodrębniania danych z różnych źródeł, ich transformacji do odpowiedniego formatu oraz ładowania do wspólnego systemu, oraz MDM (Master Data Management) jako narzędzi zarządzania danymi głównymi organizacji, które zapewniają spójność, dokładność i integralność danych poprzez centralne zarządzanie i synchronizację. Automatyczna korekta anomalii staje się coraz ważniejszym tematem zarówno w biznesie, jak i świecie akademickim (Peralta, 2006).

Podsumowując, rozwiązania informatyczne takie jak PIMS, ERP, ETL oraz MDM odgrywają kluczową rolę w zarządzaniu jakością danych produktowych w organizacjach. Każdy z tych systemów spełnia unikalne funkcje, wspierając unifikację, zarządzanie i utrzymanie danych produktowych, jednocześnie stwarzając wyzwania związane z jakością danych i ich integracją. PIMS przechowują informacje o produkcie z perspektywy sprzedaży i marketingu, ERP wspomaga zarządzanie procesami biznesowymi, ETL i MDM obejmują procesy przetwarzania i zarządzania danymi. Wspólnym wyzwaniem dla tych systemów jest zapewnienie wysokiej jakości danych, co jest kluczowe dla skutecznego zarządzania informacjami o produkcie. Często napotykane problemy, takie jak nieustrukturyzowane

i błędne dane, wskazują na potrzebę ciągłego rozwoju technologii i metodologii, aby poprawić efektywność zarządzania danymi i minimalizować ryzyko błędów.

## **2.9. Taksonomia sposobów oczyszczania danych produktowych**

W literaturze naukowej podkreśla się, że integralność i jakość danych są fundamentem efektywnego procesu ekstrakcji istotnych informacji ze zgromadzonych zbiorów danych, gdzie spójność, dokładność i aktualność informacji stanowią kluczowe parametry (Dasu i Johnson, 1999; Hipp i in., 2001; Raman i Hellerstein, 2001). W związku z tym, proces oczyszczania danych, polegający na identyfikacji i eliminacji danych niezgodnych z ustalonymi kryteriami jakości, staje się kluczowym elementem zarządzania danymi. Ten proces nabiera szczególnej wagi, gdy dane pochodzą z różnych źródeł, takich jak bazy danych różnych systemów, czy arkusze kalkulacyjne (Rahm i Do, 2000; Raman i Hellerstein, 2001). Dane te mogą być pozbawione kontekstu, łączone z innymi danymi i analizowane przez osoby nieposiadające wiedzy specjalistycznej związanej z pierwotnymi źródłami danych co staje się źródłem błędów i wielokrotnie propaguje się na kolejne zbiory danych.

Chu i współautorzy (2016) w swojej publikacji kompleksowo omawiają proces oczyszczania danych w kontekście baz danych zawierających błędy, tzw. „brudnych baz”. W ich pracy przedstawiono taksonomię jakościowego wykrywania błędów, która wyraźnie definiuje trzy główne pytania, na które należy odpowiedzieć w procesie oczyszczania danych: „Co Wykryć”, „Jak Wykryć” oraz „Gdzie Wykryć”. Sposób przedstawienia zagadnień poprzez formułowanie pytań jest niezwykle trafny i efektywnie oddaje temat w kontekście dysertacji. Autorzy doskonale ujęli problematykę, prezentując kompleksowy przegląd metod wykrywania i naprawy błędów w danych.

Analizując „Co Wykryć?”, autorzy wskazują na różnorodność typów błędów, które mogą być wykryte przez techniki jakościowego wykrywania błędów. Te techniki są klasyfikowane nie tylko przez typ wykrywanych błędów, ale także przez języki używane do opisu wzorców lub ograniczeń legalnych instancji danych. Przykłady takich integralnych ograniczeń (IC) obejmują zależności funkcjonalne (FD) (Bohannon i in., 2005) oraz ograniczenia odmowy (DC) (Chu i in., 2013b), które stanowią część logiki pierwszego rzędu i służą do uchwycenia reguł jakości danych. Wyzwaniem jest projektowanie tych IC, które wymaga dużej wiedzy specjalistycznej i jest czasochłonne, podkreślając potrzebę technik automatycznego odkrywania różnych IC (Chu i in., 2013a).

W aspekcie „Jak Wykryć?”, Chu i współautorzy kategoryzują proponowane metody zależnie od stopnia zaangażowania człowieka w proces wykrywania błędów. Wiele z tych metod jest w pełni automatycznych, umożliwiając wykrywanie naruszeń zależności funkcjonalnych bez interwencji człowieka (Bohannon i in., 2005). Inne z kolei, jak identyfikacja duplikatów rekordów, wymagają ludzkiego osądu i interwencji (Wang i in., 2012).

Ostatni aspekt, „Gdzie Wykryć?”, dotyczy identyfikacji etapu w stosie inteligencji biznesowej (BI), na którym należy wykrywać błędy. Chociaż wiele technik koncentruje się na wykrywaniu błędów w oryginalnej bazie danych, niektóre błędy mogą ujawnić się znacznie później, w dalszych etapach przetwarzania danych (Chalamalla i in., 2014). Jest to szczególnie istotne, kiedy dostępne są bardziej zaawansowane semantyki i logika biznesowa, które pozwalają na egzekwowanie bardziej złożonych ograniczeń, na przykład ograniczeń dotyczących całkowitego budżetu, które mogą być w pełni zrozumiałe i egzekwowane tylko po agregacji kosztów i wydatków.

Podobnie jak w przypadku wykrywania błędów, Chu i współautorzy (2016) dokonali klasyfikacji technik związanych z możliwością naprawy bazy "brudnych danych". Rozwijając tę klasyfikację, autorzy wskazują, że każda technika musi odpowiedzieć na trzy główne pytania: „Co Naprawić?”, „Jak Naprawić?” oraz „Gdzie Naprawić?”.

W aspekcie „Co Naprawić?”, analizie podlegają różne założenia dotyczące danych i reguł jakości. Niektóre algorytmy naprawcze opierają się na założeniu, że zadeklarowane ograniczenia integralności są nienaruszalne, co implikuje, że jedynie dane wymagają aktualizacji w celu usunięcia błędów (Chu i in., 2013a). Inne podejścia zakładają pełne zaufanie do danych i dopuszczają możliwość złagodzenia ograniczeń, aby dostosować się do zmieniających się schematów bazy danych i przestarzałych reguł biznesowych (Golab i in., 2008). Z kolei niektóre metody rozważają możliwość modyfikacji zarówno danych, jak i ograniczeń (Beskales i in., 2013), co otwiera drogę do bardziej elastycznego podejścia do zarządzania jakością danych.

Analizując „Jak Naprawić?”, metody te można podzielić ze względu na stopień automatyzacji. Wiele z nich wykorzystuje algorytmy, które minimalizują różnice między oryginalną a zmodyfikowaną instancją bazy danych, stosując różne funkcje kosztu (Yakout i in., 2011). Jednakże istnieje również spora grupa technik, które wymagają lub przynajmniej pozwalają na interakcje z użytkownikami, którzy mogą weryfikować proponowane naprawy,



sugerować własne rozwiązania lub uczestniczyć w trenowaniu modeli uczenia maszynowego służących do automatycznego podejmowania decyzji naprawczych.

W kontekście „Gdzie Naprawić?”, istnieje podział na podejścia naprawiające bazę danych w miejscu, co może prowadzić do destrukcji oryginalnych danych oraz te, które tworzą modele opisujące różne możliwe naprawy. Te drugie umożliwiają wykonywanie zapytań na bazie modeli danych, wykorzystując techniki takie jak próbkowanie z puli możliwych napraw oraz probabilistyczne mechanizmy odpowiadania na zapytania (Beskales i in., 2010) . Wykorzystanie takich modeli może znacznie zwiększyć efektywność i elastyczność procesu naprawczego, pozwalając na równoczesne eksplorowanie wielu potencjalnych rozwiązań problemu.

Proces korygowania informacji w zbiorach danych, które są niekompletne, nieprawidłowe, nieprecyzyjne, nieaktualne, bądź powielone, może obejmować szereg różnorodnych działań. Ilyas i Chu (2019) w swojej książce wymieniają następujące:

- A. **Identyfikacja błędów lub niespójności** - wykrywanie i rozpoznawanie anomalii, błędów typograficznych, nieścisłości oraz brakujących wartości w danych.
- B. **Usunięcie lub poprawa błędnych danych** - korekta błędów, takich jak literówki, błędne formatowanie, nieprawidłowe daty czy adresy.
- C. **Imputacja** - uzupełnianie brakujących danych poprzez estymację, interpolację lub wykorzystanie wartości domyślnych.
- D. **Deduplikacja** - wykrywanie i eliminowanie zduplikowanych rekordów, które mogą wprowadzać błędy w analizie.
- E. **Normalizacja** - przekształcanie różnych formatów i jednostek do standardowej formy, co ułatwia analizę i porównywanie danych.
- F. **Standaryzacja** - ujednolicanie danych z różnych źródeł w celu zapewnienia ich spójności i kompatybilności.
- G. **Walidacja** - sprawdzanie zgodności danych z określonymi regułami, standardami lub wymaganiami.

Należy zwrócić uwagę, że każde z tych działań ma swoje miejsce, kolejność wykonywania i znaczenie w procesie zarządzania danymi, ale zastosowanie musi być starannie rozważone w kontekście specyficznych wymagań i ograniczeń danego projektu. Nie wszystkie działania mogą być zaimplementowane bezpośrednio lub w pełni automatycznie, zwłaszcza w środowiskach, gdzie kluczowa jest odpowiedzialność za dane. Niezbędne jest w takich

sytuacjach opracowanie efektywnych i etycznych metod informowania użytkowników o zmianach w danych, jak również umożliwienie im kontroli nad tymi zmianami. To podejście wymaga nie tylko technicznych rozwiązań, ale również ustanowienia odpowiednich procedur i polityk zarządzania danymi, które uwzględniają zarówno potrzeby analizy danych, jak i odpowiedzialność za te dane.

## **2.10. Rola walidacji danych produktowych**

Walidacja jest kluczowym aspektem w zarządzaniu jakością danych, jednym z głównych elementów procesu oczyszczania danych. Zrozumienie jej roli, metod i wyzwań jest istotne w kontekście zapewnienia wiarygodności i użyteczności danych.

Według IV edycji Project Management Body of Knowledge (2021), walidacja to proces zapewnienia, że produkt, usługa lub system spełnia potrzeby klienta i innych zidentyfikowanych interesariuszy. Standard ISO 9000:2015 (2015) opisuje, że walidacja to potwierdzenie, za pomocą obiektywnych dowodów, że wymagania dla konkretnego zamierzonego użycia lub zastosowania zostały spełnione. Słowo "zwalidowany" oznacza, że produkt lub usługa przeszła pomyślnie proces walidacji.

W języku polskim często mylone są pojęcia walidacji i weryfikacji, lub stosowane są zamiennie. Podobnie, często dochodzi do pomyłek podczas automatycznego translowania tekstu angielskiego na polski. W odróżnieniu od walidacji - weryfikacja, zgodnie z Project Management Institute (2021), to ocena, czy produkt, usługa lub system jest zgodny z przepisami, wymaganiami, specyfikacjami lub narzuconymi warunkami. Standard ISO 9000:2015 (2015) definiuje weryfikację jako potwierdzenie, że określone wymagania zostały spełnione. Słowo "zweryfikowany" używane jest do oznaczenia produktu lub usługi, który przeszedł proces weryfikacji. Różnica zatem pomiędzy walidacją a weryfikacją polega głównie na celu i kontekście ich stosowania. Walidacja jest procesem zewnętrznym, a weryfikacja wewnętrznym. Walidacja koncentruje się na "prawidłowości" danych w kontekście ich zastosowania. Odpowiada na pytanie, czy dane są odpowiednie do celów, dla których mają być wykorzystane. Weryfikacja ma na celu zapewnienie, że proces przetwarzania danych jest poprawny i zgodny ze standardami. Koncentruje się na pytaniu, czy dane zostały poprawnie przetworzone lub zebrane.

Pantziarka (2005), w kontekście danych produktowych, rozszerza definicję walidacji jako procesu sprawdzania wewnętrznej spójności, poprawności oraz jakości zbioru danych.

Według autora, walidacja obejmuje weryfikację, czy każda zmienna ma odpowiedni typ, właściwe jednostki miary, została zebrana w odpowiednim przedziale czasu, mieści się w dopuszczalnych granicach oraz czy nie są naruszane logiczne, ilościowe i funkcjonalne relacje między zmiennymi. Jest to iteracyjny, dwuetapowy proces polegający na identyfikacji problematycznych elementów danych oraz ich korekcie lub usunięciu. Co istotne, walidacja danych niekoniecznie oznacza identyfikację i eliminację wartości odstających. W wielu rzeczywistych scenariuszach takie skrajne wartości mogą być ważne i mogą dostarczać więcej informacji niż wartości w zakresie normalnym (Liu i in., 2002). Identyfikacja wartości odstających jest bardziej właściwie postrzegana jako część procesu eksploracji danych, który zachodzi po procesie walidacji. Jednakże powinno być także jasne, że w niektórych przypadkach wartości odstające rzeczywiście reprezentują nieprawidłowe elementy danych, ale ocena, do której klasy należy dany element, nie jest możliwa bez jasnego zrozumienia zbioru danych.

Należy jednak podkreślić, że walidacja nie zajmuje się problemami jakości danych wynikającymi z operacji łączenia baz. Jest to problem znany jako Merge/Purge, czyli problem identyfikacji zduplikowanych rekordów i jest aktywnym obszarem badań od wielu lat (Ashraf i in., 2020; Hernández i Stolfo, 1998; Lee i in., 1999; Monge, 2000; Singh i Dwivedi, 2020), natomiast prawidłowo zwalidowane dane mogą ułatwić pewne aspekty problemu Merge/Purge, zwłaszcza w identyfikacji duplikatów. Problem łączenia danych w ostatnich latach poszerzył się o zagadnienia tzw. łączenia encji (Entity Linking, zwane czasem Entity Resolution lub Record Linkage). Jest to proces identyfikacji i łączenia różnych odniesień do tej samej jednostki (np. osoby, miejsca, organizacji) pochodzących z różnych źródeł danych. Tak jak Merge/Purge koncentruje się bardziej na czyszczeniu i konsolidacji danych w ramach jednego źródła danych lub źródła o podobnej charakterystyce i strukturze, tak Entity Linking zajmuje się powiązaniem i rozpoznawaniem tych samych encji pochodzących z różnych źródeł danych. Techniki związane z łączeniem encji obecnie często wykorzystują zaawansowane modele uczenia maszynowego w tym uczenie głębokie i są szeroko poruszane w literaturze (Ao i Chirkova, 2019; Brunner i Stockinger, 2020; Kruse i in., 2020; Li i in., 2020).

Warto wspomnieć także o innych klasach problemów z jakością danych, takich jak konflikty schematów (Hernández i Stolfo, 1998; Rahm i Do, 2000) oraz naruszenie unikalności (Rahm i Do, 2000), które są również istotne, ale wykraczają poza zakres terminologii walidacji.

W kontekście katalogów elektronicznych, walidacja danych stanowi kluczowy element zapewnienia jakości i wiarygodności informacji. Proces ten wymaga nie tylko odpowiednich

narzędzi i technik, ale także głębokiego zrozumienia specyfiki danych i kontekstu ich użycia. Implementacja skutecznych metod walidacji jest zatem niezbędna dla osiągnięcia celów związanych z doskonaleniem jakości danych produktowych.

## **2.11. Przegląd metod walidacji danych**

Walidacja danych produktowych stanowi kompleksowy proces, który wymaga precyzyjnego planowania, zastosowania zróżnicowanych technik oraz ciągłego monitorowania. Istotne jest połączenie różnych metod walidacyjnych, takich jak walidacja ręczna i automatyczna, aby wykorzystać zalety każdej z nich. Manualna walidacja okazuje się skuteczna w przypadku złożonych lub subiektywnych problemów, gdzie konieczne jest ludzkie rozeznanie, natomiast automatyzacja jest kluczowa przy obsłudze dużych zbiorów danych, umożliwiając efektywne i szybkie przetwarzanie. Kluczowym aspektem procesu walidacyjnego jest adaptacja do specyfiki danych produktowych. Różnorodność danych wymaga zastosowania odmiennych metod walidacji. Przykładowo, dane techniczne często lepiej nadają się do walidacji za pomocą reguł biznesowych, podczas gdy opisy produktów mogą wymagać głębszej analizy językowej, aby upewnić się, że są zrozumiałe i adekwatne. Ponadto, procesy walidacyjne powinny być regularnie przeglądane i aktualizowane, aby dostosować je do ewoluujących standardów, trendów rynkowych oraz nowych technologii. Takie podejście zapewnia, że dane są stale aktualne i wiarygodne. Równie istotna jest kwestia zapewnienia spójności i dokładności danych. Należy dążyć do tego, aby dane były nie tylko precyzyjne, ale również konsekwentne w całym katalogu, co wymaga stosowania jednolitych formatów, terminologii i standardów. W procesie walidacji niezwykle ważny jest także feedback od użytkowników i interesariuszy. Umożliwia on ciągłe doskonalenie jakości danych, dostosowując je do aktualnych potrzeb i oczekiwań. Taki zintegrowany i dynamiczny proces walidacji danych produktowych jest kluczowy dla utrzymania ich wysokiej jakości, co jest niezbędne w efektywnym zarządzaniu informacjami produktowymi w środowisku e-commerce.

Najprostszą klasyfikacją walidacji jest podział walidacji na ręczną i automatyczną. Oba typy mają swoje specyficzne zastosowania, zalety i wady, które sprawiają, że są one komplementarne w procesie zapewniania jakości danych.

### **Walidacja Ręczna**

Ta forma walidacji polega na manualnym przeglądaniu danych przez specjalistów. Jest to szczególnie użyteczne w przypadkach, które wymagają głębokiej wiedzy fachowej lub oceny

subiektywnej. Na przykład, może to obejmować ocenę zgodności opisu produktu z jego rzeczywistymi cechami, co często wymaga zrozumienia niuansów specyficznych dla danego produktu lub branży. Zaletą walidacji ręcznej jest jej dokładność i zdolność do radzenia sobie ze złożonymi i subiektywnymi problemami, które trudno jest zautomatyzować. Wada to znaczne nakłady czasu i kosztów, szczególnie w przypadku dużych zbiorów danych.

### **Walidacja Automatyczna**

Polega na użyciu algorytmów i narzędzi programistycznych do identyfikacji i korygowania błędów w danych. Metoda ta jest skuteczna w szybkim przetwarzaniu dużych zbiorów danych, gdzie ręczna walidacja byłaby niepraktyczna lub zbyt kosztowna.

Główną zaletą walidacji automatycznej jest jej wysoka efektywność oraz zdolność do szybkiego przetwarzania i korekcji danych, co ma szczególne znaczenie w przypadku przetwarzania dużych wolumenów informacji. Niemniej jednak, w sytuacjach wymagających interpretacji kontekstu lub charakteryzujących się niejednoznacznością, walidacja automatyczna może być mniej skuteczna. Ponadto, istnieje ryzyko pominięcia bardziej subtelnych lub złożonych błędów, które mogą nie zostać poprawnie zidentyfikowane przez algorytmy automatyczne.

Kombinacja obu metod jest często najlepszym rozwiązaniem, pozwalającym na wykorzystanie zalet każdej z nich. Ręczna walidacja może być stosowana do bardziej złożonych problemów i jako kontrola jakości dla wyników walidacji automatycznej, podczas gdy automatyzacja może szybko przetwarzać i wstępnie oceniać duże ilości danych. Takie połączenie pozwala na optymalizację procesu walidacji danych, maksymalizując efektywność przy jednoczesnym zachowaniu wysokiej jakości danych.

Według Prasad i współautorów (2011) dane przedsiębiorstwa są zawsze w ruchu, a nowe dane ciągle napływają do przedsiębiorstwa z różnymi cechami. Autorzy potwierdzają, że poprawa jakości danych jest ciągłym procesem, który musi być realizowany w celu obsługi zmienności towarzyszącej rosnącym danym. Ponadto, poprawa jakości danych jest zadaniem wymagającym dużej ilości ręcznej pracy przy tworzeniu nowych słowników i reguł w celu automatyzacji tego procesu. Słowniki i reguły są tworzone przez ekspertów domenowych i zmieniają się w zależności od klienta, wymagając znaczącej modyfikacji z klienta na klienta. Ilość wysiłku, jaki należy włożyć w dostosowanie zależy od rodzaju danych. Autorzy wspominają nasilający się problem zwłaszcza w dziedzinie walidacji danych produktowych, które są rejestrowane w różnych formach przez różnych klientów, gdzie prawie zawsze

konieczne jest opracowanie nowego zestawu reguł i słowników dla każdego klienta osobno. Autorzy argumentują, że czyszczenie danych zostało szczegółowo zbadane (Cong i in., 2007; Rahm i Do, 2000; Sung i in., 2002), jednak niewiele prac koncentruje się na wsparciu szybkiej personalizacji zestawów reguł dotyczących jakości danych oraz narzędzi umożliwiających szybki rozwój słowników. Autorzy twierdzą, że obecne podejście stosowane przez praktyków jakości danych polega na analizie danych i generowaniu raportów o częstotliwości występowania tokenów i wzorców, które kierują dodawanie wpisów do słowników i reguł. Proces ręcznego dodawania nowych reguł jest żmudny i podatny na błędy, a problemem w obsłudze tokenów jest fakt, że ekspert domenowy nie zawsze zna znaczenie badanych tokenów. Według Prasad i współautorów, problem żmudnego i podatnego na błędy procesu manualnego dodawania nowych reguł oraz trudności związanych z interpretacją znaczenia tokenów można rozwiązać poprzez wprowadzenie ram zwanych ripple down rules (RDR). Metodyka ta pozwala na lepszą organizację reguł jakości danych poprzez ich strukturalizację w formie drzewa decyzyjnego, gdzie na szczytach umieszcza się najbardziej ogólne reguły, a specyficzne przypadki i wyjątki tworzą gałęzie (Dani i in., 2010). Ta hierarchiczna organizacja reguł nie tylko ułatwia śledzenie i zrozumienie procesu decyzyjnego, ale również umożliwia autorom reguł szybką modyfikację istniejących reguł lub dodawanie nowych wyjątków w odpowiednich miejscach. Dodatkowo autorzy proponują narzędzia do interpretacji nieobsłużonych tokenów przez wydobycie kontekstu z otaczających tokenów, co pozwala na zrozumienie i właściwe przypisanie znaczenia nawet w przypadku, gdy ekspert domenowy nie jest w stanie zidentyfikować znaczenia tokenów, tak jak w przykładzie z indyjskimi adresami. To podejście skutecznie adresuje problem różnorodności językowej i skrótów stosowanych w różnych regionach.

Podobne zdanie dotyczące problemów z ręczną walidacją wyraża Pantziarka (2005). Autor zwraca uwagę na fakt, że choć proces walidacji danych jest kluczowym elementem wielu projektów analizy i eksploracji danych, to często jest on realizowany ręcznie, co wiąże się z szeregiem wyzwań. Zauważa, że te zadania mogą być monotonne i czasochłonne, odciągając analityków od bardziej twórczych aspektów pracy, jak eksploracja danych, odkrywanie wiedzy i modelowanie. Wskazuje również na kilka istotnych problemów związanych z ręcznym procesem walidacji danych. Zmęczenie lub znużenie analityka może wpływać na jakość walidacji, a brak spójnego i systematycznego procesu często prowadzi do podejścia opartego na próbach i błędach. Brak wiedzy domenowej może sprawić, że niektóre zasady walidacji, które powinny być odkryte w danych, pozostają niezauważone. Istniejące zasady walidacji

również mogą nie być dostatecznie często aktualizowane na podstawie doświadczeń z różnymi zestawami danych. Pantziarka podkreśla też, że brak technik radzenia sobie z niepewnością i skorygowania nieprawidłowych wartości może prowadzić do niskiej jakości walidacji danych. To z kolei może mieć bezpośrednie konsekwencje, takie jak niedokładne korekty danych, wykluczenie pełnych zestawów danych z analizy z powodu problemów walidacyjnych, czy konieczność powrotu do zadania walidacji w późniejszym etapie projektu. To ostatnie skraca czas dostępny na bardziej twórcze i interesujące zadania analizy. W związku z tym Pantziarka sugeruje, że zautomatyzowane narzędzia walidacyjne mogą znacząco poprawić proces walidacji. Takie narzędzia nie tylko usunęłyby część monotonii związanej z walidacją danych, ale również poprawiłyby ogólną jakość zestawów danych. To z kolei pozwoliłoby analitykom poświęcić więcej czasu na bardziej twórcze i interesujące aspekty analizy danych, zwiększając tym samym wartość i efektywność ich pracy. W swojej analizie Pantziarka zauważa, że analitycy danych i inni praktycy wykorzystują różnorodne techniki ad hoc, aby poradzić sobie z tym wyzwaniem związanym z poprawą jakości danych. Wymienia kilka powszechnych metod:

- A. **Składniowe sprawdzanie surowych danych.** Jest to elementarny etap, często integrowany ze wstępnym przetwarzaniem danych, polegający na identyfikacji i korekcie błędów takich jak literówki, błędy formatowania i inne niespójności.
- B. **Dopasowywanie wzorców.** Ta technika zakłada porównywanie nowych zestawów danych z wcześniej zweryfikowanymi danymi, co pomaga w identyfikacji anomalii, wymaga jednak dostępu do wiarygodnych danych referencyjnych.
- C. **Metody statystyczne i analiza wariancji.** Wykorzystanie narzędzi statystycznych do wykrywania wartości odstających i anomalii, choć niesie ryzyko błędnego zaklasyfikowania ważnych danych jako nieprawidłowych.
- D. **Analiza trendów.** Polega na identyfikacji wzorców i trendów, szczególnie w danych czasowych, co jest pomocne w wykrywaniu zmian takich jak sezonowe wahania czy długoterminowe tendencje.
- E. **Kontekstualizacja.** Wykorzystanie wiedzy eksperckiej i domenowej do interpretacji danych, zależne jednak od dostępności i jakości tej wiedzy, co może być ograniczające.

F. **Koncentracja na kluczowych zmiennych.** Użycie wiedzy domenowej do ograniczenia zakresu problemu walidacji, koncentrując się na zmiennych kluczowych dla modelu lub procesu biznesowego.

G. **Porównanie z danymi testowymi:** Porównywanie wyników modelu z oczekiwanymi wynikami uzyskanymi z wcześniej zdefiniowanych danych testowych, co pozwala na ocenę czy model działa zgodnie z oczekiwaniami w kontrolowanych warunkach.

Pantziarka (2005) podkreśla, że metody walidacji danych produktowych są zróżnicowane i zależne od specyfiki danych oraz dostępnych zasobów. Nie istnieje uniwersalna metoda odpowiednia dla wszystkich scenariuszy, dlatego ważne jest, aby praktycy rozważali kombinację tych metod, dostosowując je do konkretnych potrzeb i wyzwań.

W automatycznej walidacji danych produktowych warto wspomnieć o użyciu zaawansowanych technologii uczenia maszynowego, które znacznie poszerzają możliwości i skuteczność tego procesu. Uczenie maszynowe (ML) i sztuczna inteligencja (AI) mogą być wykorzystane do identyfikacji wzorców, anomalii i błędów w danych, które mogą nie być łatwo dostrzegalne w tradycyjnych metodach walidacji. Pantziarka (2005) w swojej pracy dokonuje przeglądu kilku technik, które mogą być stosowane do różnych aspektów problemu walidacji danych. Wymienia algorytmy i programowanie genetyczne, sieci neuronowe, reguły asocjacyjne, sieci bayesowskie oraz programowanie indukcyjno - logiczne. Szczegółowe informacje na temat poszczególnych rodzajów algorytmów i możliwości ich zastosowań przedstawiono w Tabeli 7.

Tabela 7. Rodzaje algorytmów możliwe do wykorzystania do walidacji danych

Rodzaj algorytmu	Opis zastosowania
<b>Algorytmy Genetyczne (GA)</b>	Algorytmy genetyczne, inspirowane procesami ewolucji biologicznej, polegają na symulacji naturalnej selekcji i adaptacji w celu rozwiązywania złożonych problemów. GA wykorzystują mechanizmy takie jak selekcja, krzyżowanie i mutacja genetyczna, aby iteracyjnie ulepszać i optymalizować potencjalne rozwiązania. Te algorytmy działają poprzez tworzenie populacji potencjalnych rozwiązań, ocenianie ich przydatności i generowanie nowych populacji z najlepszych kandydatów. W kontekście walidacji danych produktowych, GA mogą służyć do tworzenia i ulepszania reguł i procedur walidacyjnych. Przez iteracyjne doskonalenie zestawów reguł, algorytmy te mogą znacząco zwiększyć skuteczność wykrywania błędów i nieścisłości w danych produktowych. GA są szczególnie przydatne w sytuacjach, gdzie tradycyjne metody walidacji mogą być niewystarczające, na przykład w przypadku bardzo złożonych lub wielowymiarowych danych produktowych.



<b>Rodzaj algorytmu</b>	<b>Opis zastosowania</b>
<b>Programowanie Genetyczne (GP)</b>	Programowanie genetyczne to rozszerzenie idei algorytmów genetycznych, które polega na automatycznym tworzeniu i ewolucji programów komputerowych i algorytmów. GP symuluje ewolucję programów poprzez zastosowanie operacji genetycznych, takich jak mutacja, krzyżowanie i selekcja, w celu wytworzenia skutecznych algorytmów. Ta technika jest szczególnie użyteczna w sytuacjach, gdzie optymalna forma algorytmu nie jest znana z góry. GP może być wykorzystywane do tworzenia zaawansowanych procedur walidacji danych produktowych. Przez automatyzację procesu tworzenia algorytmów, GP umożliwia generowanie skutecznych i złożonych reguł walidacyjnych, które mogą adaptować się do różnorodnych typów błędów danych. To podejście jest szczególnie przydatne w sytuacjach, gdzie konieczne jest uwzględnienie dużej ilości zmiennych lub złożonych zależności między atrybutami produktu.
<b>Sieci Neuronowe (NN)</b>	Sieci neuronowe to zaawansowane modele uczenia maszynowego, które naśladują procesy zachodzące w ludzkim mózgu. NN uczą się rozpoznawać wzorce i zależności w danych poprzez trening z wykorzystaniem dużych zbiorów danych. Są zdolne do identyfikacji subtelnych wzorców i adaptacji do nowych danych, co czyni je potężnym narzędziem w analizie danych. NN mogą być wykorzystywane do skutecznego identyfikowania nieregularności i anomalii w danych produktowych. Dzięki swojej zdolności do uczenia się i adaptacji, sieci te są wyjątkowo skuteczne w wykrywaniu skomplikowanych błędów, które mogą być niewidoczne dla innych metod. NN mogą również pomóc w przewidywaniu potencjalnych problemów z jakością danych na podstawie wcześniej zaobserwowanych wzorców.
<b>Reguły Asocjacyjne (AR)</b>	Reguły asocjacyjne to techniki analizy danych, które pozwalają na odkrywanie interesujących związków między różnymi elementami w dużych zbiorach danych. Metoda ta polega na identyfikacji częstych wzorców, zależności lub asocjacji między zestawami danych, co może ujawnić ukryte relacje i tendencje. W kontekście danych produktowych, AR mogą być użyte do wykrywania i weryfikowania wzorców i zależności, co jest kluczowe w identyfikowaniu niespójności i błędów danych. Te reguły mogą pomóc w rozpoznawaniu nieoczekiwanych lub nietypowych kombinacji atrybutów produktu, które mogą wskazywać na potencjalne błędy lub nieprawidłowości.
<b>Sieci Bayesowskie (BN)</b>	Sieci Bayesowskie to graficzne modele probabilistyczne, które stosują wnioskowanie statystyczne do modelowania zależności i niepewności. BN są szczególnie skuteczne w reprezentowaniu złożonych zależności przyczynowo skutkowych i w zarządzaniu niepewnością w danych. BN mogą być stosowane do analizy prawdopodobieństwa poprawności różnych aspektów danych produktowych, umożliwiając modelowanie złożonych zależności i ocenę wpływu różnych czynników na jakość danych. Są szczególnie użyteczne w sytuacjach, gdzie dane są niepewne lub niekompletne, co pozwala na bardziej świadome podejmowanie decyzji dotyczących walidacji.

Rodzaj algorytmu	Opis zastosowania
<b>Programowanie Indukcyjno-Logiczne (ILP)</b>	ILP to metoda uczenia maszynowego, która wykorzystuje logikę i indukcję do wyciągania wniosków z danych. Polega na generowaniu reguł i wzorców na podstawie dostarczonych przykładów, wykorzystując wiedzę tła i logikę predykatów do formułowania hipotez. ILP może być używane do automatycznego generowania złożonych reguł walidacji, które mogą identyfikować subtelne i złożone błędy w danych produktowych. Dzięki zdolności do analizowania wzorców i relacji między atrybutami produktu, ILP może efektywnie uzupełniać tradycyjne metody walidacji, zapewniając głębsze zrozumienie i dokładniejszą analizę danych.

*Źródło: opracowanie własne*

Każda z tych technik ma swoje unikalne zalety i może być stosowana do różnych aspektów walidacji danych. Ważne jest, aby pamiętać, że żadna pojedyncza technika nie jest rozwiązaniem wszystkich problemów walidacji danych. W praktyce, kombinacja różnych technik może być stosowana w zależności od specyfiki danych i wymagań walidacyjnych. Przyszłe badania mogą dotyczyć eksploracji tych technik, ich integracji i optymalizacji w kontekście konkretnych zastosowań walidacji danych.

W przeglądzie rozwiązań sztucznej inteligencji do poprawy jakości danych produktowych wpływających na bezpieczeństwo konsumenta Niemir, Muszyński i Skwarek (2022) wskazują na konkretne algorytmy mogące usprawnić wybrane pola danych produktowych obszaru żywności, wchodzących w tzw. „master data”, czyli kluczowych informacji z tego zakresu:

- A. Klasyfikacja produktu
- B. Nazwa produktu
- C. Nazwa regulowana produktu
- D. Nazwa funkcjonalna
- E. Opis produktu (szczegóły produktu)
- F. Czy produkt zawiera substancje niebezpieczne dla zdrowia i życia
- G. Zawartość alergenów
- H. Maksymalna data spożycia
- I. Informacje żywieniowe / wartość kaloryczna
- J. Informacje żywieniowe / zawartość białka, cukrów, soli itp.
- K. Ostrzeżenie o alergenach na opakowaniu produktu

Algorytmy te wymagają jednak regularnej kalibracji oraz treningu z wykorzystaniem aktualnych i reprezentatywnych zbiorów danych, aby zachować wysoką dokładność i wiarygodność. Ponadto, ważne jest zapewnienie, że systemy te są zaprojektowane z uwzględnieniem etycznych i prawnych aspektów ochrony danych konsumentów oraz że są w stanie identyfikować i poprawnie przetwarzać informacje o potencjalnie niebezpiecznych substancjach czy alergenach. Wyzwanie stanowi również zapewnienie, że algorytmy są w stanie działać w realnych warunkach, gdzie dane produktowe mogą być niekompletne, niejednoznaczne lub podatne na szybkie zmiany, co wymaga od systemów sztucznej inteligencji zdolności do adaptacji i uczenia się w dynamicznym środowisku. Tak więc, chociaż wykorzystanie sztucznej inteligencji do poprawy jakości danych produktowych oferuje obiecujące perspektywy, to jednak wymaga ono ciągłego monitorowania, analizy i dostosowania, aby efektywnie wspierać bezpieczeństwo i dobrostan konsumentów.

W kontekście przetwarzania baz danych o wielkości rzędu milionów rekordów, gdzie kluczowe znaczenie mają zarówno aspekty semantyczne, jak i syntaktyczne, niezbędne jest zastosowanie metod automatycznych. Obejmuje to zarówno proste rozwiązania oparte na regułach, jak i zaawansowane technologie sztucznej inteligencji, które umożliwiają efektywne zarządzanie oraz analizę danych na dużą skalę. Wykorzystanie tych technologii umożliwi nie tylko efektywne uporządkowanie informacji, ale również na odkrywanie wartościowych wniosków i zależności, co jest niemożliwe do osiągnięcia przy użyciu manualnych metod pracy w tej skali.

## **2.12. Przegląd koncepcji i rozwiązań walidacji danych**

W ramach przeglądu literaturowego dokonano podziału na dwie grupy publikacji: teoretyczne, omawiające problematykę jakości danych, ramy, metryki i koncepcje, oraz praktyczne, gdzie autorzy wskazują narzędzia rozwiązujące wybrany problem jakości danych, lub sami tworzą modele do poprawy konkretnych problemów, a efekty testują na wycinku zgromadzonych informacji.

### **2.12.1. Przegląd rozwiązań teoretycznych**

W artykułach (Haug i in., 2011; Haug i Stentoft Arlbjörn, 2011; Hüner i in., 2011) autorzy koncentrują się na różnych aspektach związanych z jakością danych w łańcuchach dostaw i zarządzaniu danymi podstawowymi (Master Data). Hüner i in. (2011) analizują proces identyfikacji defektów krytycznych danych produktowych w firmie Beiersdorf AG. Ich

celem jest określenie i monitorowanie defektów danych produktowych kluczowych dla działalności biznesowej. Autorzy identyfikują typowe defekty danych oraz proponują zestaw siedmiu metryk jakości danych i 32 reguły walidacji, które mogą być wykorzystane przez inne firmy do monitorowania i poprawy jakości danych produktowych. Haug i Stentoft Arlbjörn (2011) badają bariery utrudniające osiągnięcie wysokiej jakości danych podstawowych (Master Data) w firmach. Przeprowadzone badanie kwestionariuszowe wśród 90 duńskich firm wskazuje, że brak odpowiedzialności za zarządzanie danymi podstawowymi ma największy wpływ na ich jakość. Autorzy identyfikują pięć głównych barier: brak delegacji odpowiedzialności, brak nagród za zapewnienie jakości danych, brak rutynowych kontroli, brak kompetencji pracowników oraz nieprzyjazność oprogramowania zarządzającego danymi. Haug i in. (2011) koncentrują się na analizie kosztów związanych ze złą jakością danych w firmach. Argumentują, że celem powinna być optymalna, a nie doskonała jakość danych. Przedstawiają model obliczania optymalnego poziomu wysiłku na utrzymanie danych, który uwzględnia różne kategorie kosztów związanych ze złą jakością danych. Model ten pomaga firmom określić, jak zrównoważyć nakłady na utrzymanie jakości danych i unikanie kosztów wynikających z ich niskiej jakości.

W pracach autorstwa Bradji i Boufaida (2011) oraz Madlberger (2011) autorzy koncentrują się na zwiększeniu jakości danych jako środka do osiągnięcia konkretnych celów: w przypadku Bradji i Boufaida, jest to poprawa procesu czyszczenia danych, natomiast w przypadku Madlberger, jest to przełamanie efektu pingwina poprzez poprawę jakości danych produktów. Celem badań Bradji i Boufaida (2011) było opracowanie systemu zarządzania regułami, który zwiększa jakość czyszczenia danych poprzez usprawnienie jakości reguł i procesu pozyskiwania wiedzy. System ten ma na celu poprawę jakości danych poprzez efektywne zarządzanie regułami i zapewnienie spójności w reprezentacji wiedzy domenowej. Z kolei Madlberger (2011) zbadała, czy poprawa jakości danych produktów może pomóc w przezwyciężeniu efektu pingwina. Efekt pingwina odnosi się do sytuacji, w której jednostki w grupie czekają, aż inni podejmą inicjatywę, zanim sami zaczną działać, co prowadzi do opóźnień lub stagnacji w przyjmowaniu nowych standardów czy technologii. W kontekście pul danych produktów, efekt pingwina ogranicza adopcję standardów, mimo ich potencjalnych korzyści. Madlberger sugeruje, że poprawa jakości danych może zwiększyć wykorzystanie pul danych i pomóc w przełamywaniu tego efektu.

Autorzy prac (Borek i in., 2015; Cai i Zhu, 2015; Chung i in., 2016; Liu i in., 2018; Yang i in., 2018) koncentrują się nie tylko na metodach oceny jakości danych, wyzwaniach

związanych z jakością danych w erze Big Data oraz metodykach estymacji liczby błędów w zbiorach danych, ale także na szerokim zakresie zagadnień związanych z poprawą jakości danych, zarządzaniem nimi oraz dostosowywaniem metod oceny do specyficznych wymagań użytkowników i aplikacji. Borek i in. (2015) przedstawiają przegląd i klasyfikację metod oceny jakości danych, tworząc taksonomię problemów związanych z jakością danych. Autorzy identyfikują istniejące metody, umiejscawiają je w kontekście określonych problemów oraz wskazują luki w aktualnych metodach, które wymagają nowych rozwiązań. Celem tej pracy jest stworzenie struktury, która pozwala na lepsze zrozumienie i wybór odpowiednich metod oceny jakości danych dla różnych zastosowań. Cai i Zhu (2015) skupiają się na wyzwaniach związanych z jakością danych w erze Big Data. Autorzy analizują specyficzne cechy i trudności związane z oceną jakości danych w środowisku Big Data, takie jak ogromna objętość danych, ich różnorodność oraz szybkość generowania. Proponują hierarchiczną ramę jakości danych oraz dynamiczny proces oceny jakości, który jest elastyczny i adaptacyjny, co pozwala na skuteczniejsze zarządzanie jakością danych w dużych zbiorach danych. Chung, Krishnan i Kraska (2016) opracowali metodykę estymacji liczby niewykrytych błędów w zbiorach danych po czyszczeniu z udziałem tłumu (crowdsourcing). Autorzy zauważają, że tradycyjne metody estymacji są wrażliwe na błędne odpowiedzi i wprowadzają nowy estymator oparty na "przełączeniach" w konsensusie tłumu, który jest bardziej odporny na szумы i błędy. Celem ich badań jest stworzenie dokładniejszej metody estymacji, która może być wykorzystana do oceny i poprawy jakości danych w różnych zastosowaniach. Liu, Chen i Cai (2018) przedstawiają metodykę oceny jakości danych, która koncentruje się na wymaganiach użytkowników aplikacji. Ich badania mają na celu dokładne dostosowanie procesu oceny do specyficznych potrzeb użytkowników systemów informatycznych. Proces oceny jakości danych składa się z czterech etapów: badanie potrzeb, przygotowanie programu oceny, implementacja procesu oceny oraz przygotowanie raportu końcowego. W pierwszym etapie autorzy przeprowadzają dokładną analizę potrzeb użytkowników, aby zidentyfikować kryteria oceny jakości danych. Następnie tworzą szczegółowy plan oceny, który jest realizowany w etapie implementacji. Kończącym produktem jest raport, który zawiera wyniki oceny oraz rekomendacje dotyczące poprawy jakości danych. Yang, Yuan i Li (2018) skupiają się na identyfikacji i opisie kluczowych czynników, które powinny być uwzględnione w procesie oceny jakości danych. Ich celem jest stworzenie kompleksowej metodologii, która integruje różne aspekty oceny jakości danych. Autorzy podkreślają znaczenie systematycznego podejścia do oceny jakości, które obejmuje takie elementy jak celowość, zasady ogólne, charakterystyki danych oraz funkcje pomiarowe. W swojej pracy opisują różnorodne metody oceny, wskazując

na ich zalety i wady w kontekście różnych zastosowań. Proponowana metodologia ma na celu zapewnienie, że ocena jakości danych jest dokładna, spójna i dostosowana do specyficznych potrzeb organizacji.

Autorzy publikacji (Drđgan i Metz, 2017; Taleb i in., 2018) koncentrują się na znaczeniu jakości danych w kontekście cyfryzacji i innowacji oraz na wyzwaniach związanych z oceną jakości danych w erze Big Data. W pracy Drđgan i Metz (2017) autorzy analizują znaczenie jakości danych w procesach cyfryzacji i innowacji, zwracając uwagę na niepewność jakości danych jako zagrożenie dla procesów biznesowych i generowanej wartości. Zidentyfikowanie i rozwiązanie problemów związanych z jakością danych jest kluczowe dla efektywnego wykorzystania danych w działalności biznesowej. Autorzy proponują podejście do poprawy jakości danych za pomocą systemów ERP, które integrują zarządzanie danymi i ich walidację. Taleb, Serhani i Dssouli (2018) w swojej pracy podkreślają, że jakość danych jest kluczowym czynnikiem w kontekście Big Data, gdzie ogromna ilość i różnorodność danych stwarzają unikalne wyzwania. Autorzy proponują ramy zarządzania jakością danych, które obejmują identyfikację problemów z jakością danych na każdym etapie cyklu życia danych oraz konieczność ciągłego doskonalenia procesów w celu zapewnienia wysokiej jakości danych.

Autorzy prac (Krieger i Schorr, 2019) oraz (Onyeabor i Ta'a, 2019) koncentrują się na poprawie jakości danych produktowych w systemach ERP oraz na problematyce jakości Big Data w całym łańcuchu wartości. Celem ich badań jest zmniejszenie nakładu pracy potrzebnego do uzyskania informacji o zawartości, strukturze i jakości danych produktowych oraz opracowanie modelu oceny jakości Big Data. Krieger i Schorr (2019) prezentują model referencyjny i najlepsze praktyki do systematycznego i efektywnego profilowania danych produktowych, natomiast Onyeabor i Ta'a (2019) rozwijają model transformacji jakości danych, który ocenia jakość Big Data od etapu zbierania po wizualizację.

W publikacji (Cichy i Rass, 2019) Autorzy dostarczają przegląd różnych ram jakości danych. Celem tej pracy jest porównanie istniejących metodologii oraz oferowanie przewodnika decyzyjnego do wyboru odpowiednich frameworków jakości danych dla różnych środowisk biznesowych. Cichy i Rass analizują różnorodne podejścia do zarządzania jakością danych, identyfikując ich mocne i słabe strony, co pomaga organizacjom w wyborze najefektywniejszych narzędzi i praktyk dostosowanych do specyficznych potrzeb i kontekstów operacyjnych. Jedno z ciekawych podejść w kontekście ram zarządzania jakością a także model pomiaru jakości zawarto w publikacji (Mao i in., 2019). Autorzy wprowadzają nowe ramy

zarządzania jakością danych (DQMPF) oraz model problemów i pomiaru jakości danych (DQPMM). Ramy te mają na celu zintegrowanie technologii informacyjnej z ekonomią społeczną, co pozwala na lepsze zrozumienie wpływu jakości danych na różne aspekty działalności biznesowej. DQMPF obejmuje strukturalne podejście do identyfikacji, oceny i zarządzania problemami jakości danych w organizacjach. Model DQPMM z kolei, koncentruje się na specyficznych problemach związanych z jakością danych, takich jak brakujące dane, nieścisłości i duplikaty, proponując metody ich pomiaru i monitorowania. Podejście to umożliwi organizacjom bardziej efektywne zarządzanie jakością danych, co w konsekwencji prowadzi do poprawy procesów decyzyjnych, optymalizacji operacji oraz zwiększenia konkurencyjności na rynku. Koncepcję innych ram zarządzania jakością (BDQ), które umożliwiają efektywne zarządzanie jakością danych na każdym etapie cyklu życia ze szczególnym uwzględnieniem Big Data zaprezentowano w publikacji (Taleb i in., 2021). Autorzy podkreślają konieczność ciągłego doskonalenia procesów w celu zapewnienia wysokiej jakości danych, co jest kluczowe w środowisku Big Data ze względu na ogromne ilości, różnorodność oraz szybkość generowania danych. Proponowane ramy mają na celu poprawę niezawodności i dokładności danych, co jest niezbędne dla podejmowania trafnych decyzji biznesowych i operacyjnych.

Zahari, Hussin, Yussof (2015) i współautorzy podkreślają znaczenie procesu czyszczenia danych jako kluczowego elementu w projektach migracji danych. Ich badania koncentrują się na wyzwaniach związanych z zapewnieniem wysokiej jakości danych podczas migracji, wskazując na liczne zagrożenia związane z otrzymywaniem, integrowaniem, utrzymywaniem, przetwarzaniem i ładowaniem danych. Autorzy podkreślają, że proces czyszczenia danych, który obejmuje wykrywanie i usuwanie błędów, niespójności oraz redundancji, jest niezbędny do zapewnienia, że migrowane dane są dokładne i spójne. To z kolei minimalizuje ryzyko wystąpienia problemów w nowych systemach, do których dane są migrowane.

W publikacji (Battistello i in., 2021) autorzy identyfikują wyzwania związane z fazą definiowania zakresu projektów PIMS (Product Information Management System). Badanie wskazuje na złożoność poznawczą, czasową i organizacyjną jako kluczowe czynniki wpływające na skuteczność wdrożeń PIMS. Autorzy podkreślają, że rozpoznanie i zrozumienie tych wyzwań jest niezbędne dla skutecznego planowania i realizacji projektów PIMS, co może prowadzić do lepszej jakości zarządzania informacją o produkcie oraz bardziej efektywnych procesów biznesowych.

Przedstawione publikacje obejmują szeroki zakres tematów związanych z jakością danych, od specyficznych studiów przypadków i rozwiązań technologicznych po bardziej ogólne ramy i modele zarządzania jakością danych. Wspólnym celem jest poprawa jakości danych i efektywne zarządzanie nimi, co ma istotne znaczenie w różnych kontekstach biznesowych i technologicznych. W Tabeli 8 zaprezentowano syntetyczny przegląd zebranych publikacji, uwzględniając streszczenie, cel badawczy, problem badawczy oraz opis zaproponowanych rozwiązań.

Tabela 8. Przegląd rozwiązań teoretycznych

Publikacja	Streszczenie	Cel	Problem	Opis rozwiązania
Product data quality in supply chains: the case of Beiersdorf (Hüner, K M.; Schierning, A; Otto, B; Österle, H 2011)	Autorzy przedstawiają studium przypadku procesu identyfikacji defektów krytycznych danych produktowych w firmie Beiersdorf AG oraz proponują zestaw metryk do monitorowania tych defektów. Analiza wykorzystuje typowe defekty danych produktów i podaje propozycje metryk jakości danych do ich monitorowania.	Określenie i monitorowanie defektów danych produktowych, które są krytyczne dla działalności biznesowej.	Brak odpowiednich metryk do identyfikacji i monitorowania defektów krytycznych danych produktowych w łańcuchach dostaw.	Autorzy opisują zestaw siedmiu metryk jakości danych, pozwalający na monitorowanie defektów oraz określają 32 reguły walidacji wykorzystujące powszechne atrybuty danych produktów, które mogą być adaptowane przez inne firmy.
Barriers to master data quality (Haug, Anders; Stentoft Arlbjörn, Jan 2011)	Zidentyfikowano bariery utrudniające osiągnięcie wysokiej jakości danych podstawowych (master data) w firmach. Wskazują, że brak delegowania odpowiedzialności za zarządzanie danymi podstawowymi ma największy wpływ na ich jakość. Przeprowadzili badanie kwestionariuszowe wśród 90 duńskich firm, aby ocenić znaczenie tych barier.	Zidentyfikowanie najważniejszych barier dla wysokiej jakości danych podstawowych w firmach, aby ułatwić ich przezwyciężanie i osiągnięcie wyższego poziomu jakości danych.	Firmy nie przykładają odpowiedniej uwagi do jakości danych podstawowych, co wpływa negatywnie na ich efektywność.	Autorzy opisują pięć głównych barier jakości danych: brak delegacji odpowiedzialności, brak nagród za zapewnienie ważności danych podstawowych, brak rutyn kontroli danych, brak kompetencji pracowników, brak przyjazności dla użytkownika oprogramowania zarządzającego danymi.
The costs of poor data quality (Haug, A.; Zachariassen, F.; Van Liempd, D. 2011)	Omówiono koszty złej jakości danych w firmach, argumentując, że nie doskonała, ale optymalna jakość danych powinna być celem. Koncentrują się na identyfikacji optymalnego poziomu jakości danych.	Określenie optymalnego poziomu jakości danych, który powinien być osiągnięty w firmach w kontekście kosztów.	Zaniedbywanie prac związanych z jakością danych prowadzi do znacznych kosztów w wielu firmach.	Rozwiązaniem jest model obliczania optymalnego poziomu wysiłku na utrzymanie danych, gdzie suma kosztów utrzymania danych i kosztów wynikających ze złej jakości danych jest najniższa. Model uwzględnia różne kategorie kosztów złej jakości danych.



Publikacja	Streszczenie	Cel	Problem	Opis rozwiązania
A rule management system for knowledge-based data cleaning (Bradji, L; Boufaida, M 2011)	Przedstawiono system zarządzania regułami dla czyszczenia danych oparty na wiedzy, który łączy cechy systemów opartych na regułach i frameworkach czyszczenia danych opartych na regułach. Zaproponowany system ma na celu zwiększenie jakości reguł, co bezpośrednio przekłada się na jakość czyszczenia danych.	Opracowanie systemu zarządzania regułami, który zwiększa jakość czyszczenia danych poprzez usprawnienie jakości reguł i procesu pozyskiwania wiedzy.	Brak praktycznych metodologii dla systemów regułowych czyszczenia danych (RBDC), które zapewniałyby jakość reguł oraz brak formalnych relacji reprezentacji reguł do logiki.	System pozwala na jednolitą i zjednoczoną reprezentację wiedzy domenowej i zarządzanie regułami, co przekłada się na poprawę jakości danych. System jest autonomiczny, rozszerzalny i niezależny od platformy, co ułatwia jego wdrożenie.
Can data quality help overcome the penguin effect? The case of item master data pools (Madlberger, Maria 2011)	Przeanalizowano wpływ jakości danych na przełamywanie efektu pingwina w kontekście pul danych dotyczących produktów. Wnioski opierają się na badaniach jakościowych wśród wiodących detalistów w sektorze spożywczym i farmaceutycznym w Austrii. Sugeruje, że poprawa jakości danych może przyczynić się do zwiększenia używania pul danych i pokonania efektu pingwina.	Zbadanie czy poprawa jakości danych produktów może pomóc w przewyżczeniu efektu pingwina w kontekście pul danych produktów.	Efekt pingwina ogranicza adopcję standardów, pomimo ich potencjalnych korzyści, co obserwuje się w małym stopniu adopcji pul danych produktów w wielu branżach.	Autorka proponuje, że puli danych mogą poprawić jakość danych produktów przez wspólne działania na rzecz zapewnienia jakości danych. Badanie eksploracyjne wśród liderów rynku detalicznego w Austrii potwierdza tę teorię.
A Classification of Data Quality Assessment Methods (Borek, A; Woodall, P; Oberhofer, M; Parlikad, AK 2015)	Zaprezentowano przegląd obecnych metod oceny jakości danych (DQ) i klasyfikują te metody w istniejącą taksonomię problemów związanych z jakością danych. Zawarte są przykłady umiejscowienia każdej z metod w taksonomii i dlaczego jest ona odpowiednia dla danego zagadnienia. Ujawnione luki w taksonomii wskazują, gdzie wymagane są nowe metody.	Zidentyfikowanie i sklasyfikowanie istniejących metod oceny jakości danych oraz wskazanie luk, gdzie nowe metody są potrzebne.	Brak kompleksowych metod oceny jakości danych dla określonych problemów, takich jak brakujące dane czy istnienie synonimów i homonimów.	Klasyfikacja wskazuje, gdzie obecne metody są skuteczne i gdzie pojawiają się luki, co może kierować przyszłym badaniem i rozwojem narzędzi DQ.
Data quality issues in data migration (Zahari, NM; Hussin, WYW; Yussof, MYM; ... 2015)	Autorzy przedstawiają, że głównym kryterium sukcesu projektów migracji danych jest jakość danych. Jakość może być zagrożona przez sposób otrzymywania, integrowania, utrzymywania, przetwarzania i ładowania	Pokazanie znaczenia czyszczenia danych jako kluczowego elementu zapewniającego wysoką jakość danych w projektach migracji danych.	Jakość danych może być kompromitowana na wielu etapach procesu migracji, od ekstrakcji po ładowanie do nowego systemu.	Czyszczenie danych opisano jako proces wykrywania i usuwania błędów, niespójności i redundancji, co ma kluczowe znaczenie dla poprawy jakości danych.

Publikacja	Streszczenie	Cel	Problem	Opis rozwiązania
	danych. Czyszczenie danych to kluczowa czynność w projektach migracji, polegająca na wykrywaniu i usuwaniu błędów, niespójności i redundancji.			
The challenges of data quality and data quality assessment in the big data era (Cai, L; Zhu, Y 2015)	Zaprezentowano przegląd badań jakości danych oraz analizują charakterystyki i wyzwania jakości danych w środowisku Big Data. Proponują hierarchiczną ramę jakości danych i konstruują dynamiczny proces oceny jakości danych, który jest elastyczny i adaptacyjny.	Zbadanie standardów jakości i metod oceny jakości danych w kontekście Big Data.	Brak kompleksowej analizy i badań standardów jakości oraz metod oceny jakości dla Big Data.	Ramy jakości składają się z wymiarów, charakterystyk i indeksów jakości Big Data. Proces oceny jest dynamiczny, posiada mechanizm informacji zwrotnej, jest rozszerzalny i adaptacyjny.
A data quality metric (DQM): how to estimate the number of undetected errors in data sets (Chung, Y; Krishnan, S; Kraska, T 2016)	Zaprezentowano metodykę estymacji liczby błędów pozostałych w zbiorze danych po oczyszczeniu z udziałem tłumy (crowdsourcing). Stawiają hipotezę o związku pomiędzy metryką jakości danych a estymacją gatunków. Odkrywają, że tradycyjne metody estymacji gatunków są wrażliwe na błędy, i proponują alternatywny estymator oparty na "przełączeniach" w konsensusie tłumy.	Estymacja liczby wszystkich wykrywalnych błędów w zbiorze danych po czyszczeniu przez tłum.	Problem wysokiej wrażliwości tradycyjnych estymatorów jakości danych na błędne odpowiedzi podczas czyszczenia danych przez tłum.	Autorzy formalizują związek między metryką jakości danych a estymatorami gatunków. Demonstrują wrażliwość istniejących metryk na błędy i wprowadzają nowy estymator oparty na "przełączeniach", który jest bardziej dokładny w obecności szumu.
Analysis and improvement of data quality. Improvement of data quality for business purposes (Drd'gan, MA; Metz, D 2017)	Przedstawiono znaczenie jakości danych w kontekście cyfryzacji i innowacji, zwracając uwagę na niepewność jakości danych jako zagrożenie dla procesów biznesowych oraz generowanej wartości. Artykuł ma na celu nakreślenie wyzwań związanych z zapewnieniem jakości danych i zarządzaniem nimi, a także proponuje podejście do poprawy jakości danych za pomocą systemów ERP.	Zidentyfikowanie wyzwań związanych z zapewnieniem jakości danych i zarządzaniem nimi oraz przedstawienie podejścia do poprawy i optymalizacji jakości danych w celu efektywnego wykorzystania danych biznesowych.	Niepewność jakości danych w procesach digitalizacji stanowi zagrożenie dla wartości generowanej przez dane oraz może prowadzić do znacznych kosztów związanych z automatyzacją procesów biznesowych.	Rozwiązanie polega na wykorzystaniu systemów ERP do poprawy jakości danych, co ma zapewnić efektywną pracę oprogramowania i powodzenie procesów biznesowych. Autorzy sugerują zintegrowane podejście do zarządzania danymi i ich walidacji.
Big Data Quality: A Survey (Taleb, Ikkal; Serhani,	W dobie Big Data jakość danych jest kluczowa dla ich przydatności i przytoczonych wniosków.	Przedstawić wyzwania związane z oceną jakości Big Data	Istniejące podejścia do zarządzania jakością danych są nie do końca dojrzałe i nie	Model zakłada identyfikację problemów z jakością danych na każdym

Publikacja	Streszczenie	Cel	Problem	Opis rozwiązania
Mohamed; Dssouli, Rachida 2018)	Mimo postępu w technologiach komunikacyjnych i wzroście ilości danych, zarządzanie ich jakością pozostaje wyzwaniem.	i zaproponować ramy zarządzania jakością.	dostosowane do specyfiki Big Data.	etapie cyklu życia i podkreśla konieczność ciągłego doskonalenia procesów w celu zapewnienia wysokiej jakości łańcucha wartości danych.
Application of requirement-oriented data quality evaluation method (Liu, Z; Chen, Q; Cai, L 2018)	Przedstawiono metodykę oceny jakości danych, podkreślając jej znaczenie dla analiz i aplikacji systemów informatycznych. W pracy określone zostały kryteria oceny jakości danych, oparte na potrzebach użytkowników oprogramowania, co zostało zweryfikowane w przypadku praktycznym, osiągając pozytywne wyniki.	Zaprezentowanie metody oceny jakości danych opartej na wymaganiach aplikacji.	Wyzwaniem jest ocena jakości danych, które bezpośrednio determinują dalsze analizy i zastosowania. Istniejące metody i techniki oceny jakości danych wymagają doprecyzowania.	Proces oceny podzielony jest na cztery części: badanie potrzeb, przygotowanie programu, implementacja oceny i przygotowanie raportu. Definiowanie kryteriów oceny opiera się na zdefiniowanych wymaganiach użytkowników. Wyniki oceny oraz przygotowanie raportu uwzględniają analizę i potwierdzenie wyników.
Data Quality Evaluation: Methodology and Key Factors (Yang, Y; Yuan, Y; Li, B 2018)	Przedstawiono kluczowe czynniki oceny jakości danych, które stają się zinstytucjonalizowanym etapem w cyklu życia danych, zwłaszcza w zaawansowanych aplikacjach informacyjnych. Podkreślają znaczenie metodyki i standardów w zarządzaniu jakością danych.	Przedstawienie kluczowych czynników do oceny i pomiaru jakości danych oraz zalecenia metodyczne na podstawie analizy głównych metodologii i standardów.	Brak zinstytucjonalizowanego podejścia do oceny jakości danych, które integruje cele, zasady ogólne, charakterystykę i funkcje pomiarowe.	Autorzy opisują metodykę oceny jakości danych, uwzględniając celowość, zasady ogólne, charakterystyki, funkcje pomiaru itd., co stanowi podstawę systematycznego podnoszenia efektywności.
A Reference Model for Product Data Profiling in Retail ERP Systems (Krieger, Rolf; Schorr, Christian 2019)	Zaprezentowano podejście do poprawy jakości danych produktowych w systemach ERP za pomocą metod uczenia maszynowego, które wspierają zarządzanie danymi głównymi. Opracowanie takich procedur wymaga bezbłędnych danych szkoleniowych, co oznacza konieczność profilowania danych. W tym celu prezentują model referencyjny i najlepsze praktyki do systematycznego i efektywnego profilowania i eksploracji danych produktowych.	Zmniejszenie nakładu pracy potrzebnego do uzyskania informacji o zawartości, strukturze i jakości danych produktowych, uwzględniając ich specyficzne właściwości i wiedzę dziedzinową.	Wyzwania związane z profilowaniem i czyszczeniem danych produktowych w systemach ERP, w tym duża liczba pozycji i kategorii produktowych, różnorodność typów produktów, wiek danych, optymalizacja procesów biznesowych, złożony model danych i zmiany w modelu danych.	Model referencyjny i najlepsze praktyki mają na celu systematyczną i efektywną eksplorację danych produktowych, z uwzględnieniem ich złożoności oraz potrzebnych czasu i zasobów.

Publikacja	Streszczenie	Cel	Problem	Opis rozwiązania
A model for addressing quality issues in big data (Onyeabor, GA; Ta'a, A 2019)	<p>Prezentowano problematykę jakości Big Data (BD) w całym łańcuchu wartości, od zbierania danych po ich wizualizację. Zaproponowany model transformacji jakości danych (DQ) ocenia jakość BD, począwszy od jakości samego procesu, aż do jakości danych. Badanie jest w toku, a model ma zostać zweryfikowany przy użyciu wymiarów jakości danych (DQDs) i jakości procesu.</p>	<p>Opracowanie modelu oceny jakości Big Data obejmującego cały łańcuch wartości od zbierania danych po wizualizację.</p>	<p>Istniejące wyzwania związane z jakością BD w różnych etapach jej przetwarzania, w tym zebranie danych, integracja i zapewnienie jakości.</p>	<p>Model składa się z sześciu faz: zbierania BD, przed i po ocenie jakości BD, przetwarzania i analizy BD oraz ocenie jakości wizualizacji BD, z wykorzystaniem różnych metryk DQD i wymiarów jakości procesu.</p>
An overview of data quality frameworks (Cichy, Corinna; Rass, Stefan 2019)	<p>Podkreślono znaczenie wysokiej jakości danych w biznesie, traktując je jako cenny atut. Różnorodność wymagań jakościowych w różnych organizacjach prowadzi do wyzwań w wyborze odpowiednich metod oceny i poprawy jakości danych. Praca dostarcza przegląd i porównanie ram jakości danych z naciskiem na metodyki uniwersalne dla różnych środowisk biznesowych oraz oferuje przewodnik decyzyjny do wyboru frameworków jakości danych.</p>	<p>Przegląd i porównanie ram jakości danych, z naciskiem na metody oceny i poprawy jakości, oraz pomoc w procesie decyzyjnym dotyczącym wyboru odpowiednich metod.</p>	<p>Wyzwanie wyboru odpowiednich metod oceny i poprawy jakości danych z uwagi na różnorodność wymagań organizacji.</p>	<p>Przewodnik ten pomaga zawęzić wybór metodologii jakości danych, bazując na serii określonych kryteriów, które działają podobnie do drzewa decyzyjnego, pozwalając na selekcję najbardziej odpowiednich ram dla danej sytuacji.</p>
Data quality management and measurement (Mao, X; Gong, B; Su, F; Xu, K; Xian, K; Liu, D; ... 2019)	<p>Głęboka integracja technologii informacyjnej z ekonomią społeczną, co prowadzi do ekspansji danych. Jako że dane pochodzą z wielu źródeł, są one heterogeniczne i niskiej jakości, co blokuje wymianę informacji i współpracę międzyoperacyjną. Problem jakości danych poważnie ogranicza zdolność do analizy danych i wsparcia decyzji. Proponowane są ramy zarządzania jakością danych (DQMPF) oraz model problemów i pomiaru jakości danych (DQPMM).</p>	<p>Zrozumienie i rozwiązanie problemu jakości danych poprzez zarządzanie i pomiar jakości danych.</p>	<p>Dane są heterogeniczne i niskiej jakości, co uniemożliwia wymianę informacji i współpracę międzyoperacyjną.</p>	<p>Autorzy konstruują nowe ramy DQMPF, wskazując, że testowanie powinno być włączone w proces dystrybucji danych dla ciągłej kontroli jakości. Proponują model DQPMM, składający się z domeny problemu oraz domeny wymiaru jakości danych. Model przekształca problemy jakości danych w analizę tablicową.</p>

Publikacja	Streszczenie	Cel	Problem	Opis rozwiązania
Implementation of product information management systems: Identifying the challenges of the scoping phase (Battistello, Loris; Haug, Anders; Suzic, Nikola; Hvam, Lars 2021)	Omówiono wyzwania występujące w fazie definiowania zakresu projektów wdrożeniowych systemów zarządzania informacją o produkcie (PIMS), poprzez studium przypadku międzynarodowej firmy. Zidentyfikowano osiemnaście głównych wyzwań i ich przyczyny, związane z jakością danych, złożonością poznawczą, czasem i złożonością organizacyjną.	Prezentowanie wyzwań związanych z fazą definiowania zakresu projektów PIMS oraz przyczyn tych wyzwań.	Słabo rozpoznany obszar implementacji PIMS w literaturze akademickiej i wyzwania z tym związane, które są słabo adresowane pomimo rosnącego zainteresowania firm w zarządzaniu informacją o produkcie.	Autorzy ilustrują fazę definiowania zakresu projektu PIMS w firmie produkującej wełnę mineralną, identyfikując 18 wyzwań i odpowiadając im przyczyny. Przyczyny te zostały zredukowane do czterech głównych czynników wpływających na fazę definiowania zakresu projektu PIMS.
Big data quality framework: a holistic approach to continuous quality management (Taleb, Ikkal; Serhani, Mohamed Adel; Bouhaddioui, Chafik; Dssouli, Rachida 2021)	Zdefiniowano obszar Big Data jako kluczowy dla instytucji rządowych, prywatnych i innych organizacji, podkreślając, że niska jakość danych może prowadzić do nieprzewidywalnych konsekwencji. Zaprezentowano Ramy Zarządzania Jakością Big Data (BDQ), wykorzystujące profilowanie jakości i weryfikację danych.	Opracowanie Ram Zarządzania Jakością BDQ umożliwiających efektywne zarządzanie jakością w cyklu życia Big Data.	Ryzyko obniżenia jakości danych ze względu na ich objętość, różnorodność źródeł i szybką generację.	Rozwiązanie to obejmuje profilowanie jakości danych i próbkowanie w celu szybszego i efektywniejszego oszacowania jakości, stosując zdefiniowane metryki i generując reguły jakości.

Źródło: opracowanie własne

Podsumowując przegląd literatury w zakresie rozwiązań teoretycznych, można wskazać na kilka kluczowych wniosków. Autorzy badań wielokrotnie podkreślają, że jakość danych ma bezpośredni wpływ na skuteczność podejmowania decyzji biznesowych. Dane o wysokiej jakości są kluczowe dla precyzyjnej analizy, prognozowania i strategicznego planowania. Nowoczesne technologie, w tym sztuczna inteligencja i systemy zarządzania regułami, mogą odegrać kluczową rolę w identyfikacji, monitorowaniu i poprawie jakości danych. Autorzy proponują różne koncepcje ram zarządzania jakością danych na różnym poziomie szczegółowości, które uwzględniają zarówno aspekty techniczne, jak i organizacyjne. W zależności od warunków, potrzeb, ograniczeń adaptacyjnych i środowiska mogą skutecznie wspomóc zarządzanie jakością danych a także jej oceną. Autorzy zidentyfikowali także szereg wyzwań implementacyjnych i organizacyjnych, szczególnie tych związanych ze

definiowaniem zakresu projektów systemów zarządzania informacją o produkcie (PIMS). W kontekście tematyki pracy - zarządzania dużym katalogiem produktowym, gdzie dane są wprowadzane przez społeczność, nie znaleziono jednoznacznych odpowiedzi dotyczących skutecznych metod nadzoru i walidacji danych, a także ogólnego modelu usprawniającego zarządzanie.

### **2.12.2. Przegląd rozwiązań praktycznych**

W publikacjach naukowych (Arasu i in., 2011; Prasad i in., 2011; Salem i Boufares, 2019) autorzy prezentują konkretne rozwiązania informatyczne do poprawy jakości danych. Prasad (2011) i jego zespół prezentują narzędzie przeznaczone do poprawy jakości danych w dużych przedsiębiorstwach. Narzędzie to identyfikuje warianty i synonimy jednostek w danych, co ułatwia pisanie reguł jakościowych. Wykorzystuje ono ramy reguł "ripple down" (RDR) do utrzymania i łatwego dodawania nowych reguł, prezentując typowy przepływ pracy procesu poprawy jakości danych. Narzędzie analizuje charakterystykę jednostek danych, co wspomaga pisanie reguł jakościowych i redukuje wysiłek usługowy przy dodawaniu nowych reguł. Prace te koncentrują się na minimalizacji wysiłku programistycznego i usługowego w procesie czyszczenia danych oraz na zapewnieniu elastyczności i łatwości implementacji w różnych kontekstach. Arasu (2011) i współautorzy przedstawiają platformę opracowaną przez Microsoft Research, która jest niezależna od dziedziny i zawiera zestaw podstawowych reguł oraz narzędzi projektowych. Platforma ta umożliwia programistom tworzenie zaawansowanych rozwiązań do czyszczenia danych przy minimalnym wysiłku programistycznym. Narzędzia te obejmują moduły oczyszczania danych oparte na regułach do wyszukiwania tekstów podobnych oraz moduły segmentacji danych. W publikacji (Salem i Boufares, 2019) autorzy przedstawiają narzędzie o nazwie iDQMS (Intelligent Data Quality Management System Tool) do zarządzania jakością danych, które wykorzystuje metody bazujące na zrozumieniu semantycznym danych w celu wykrywania i naprawiania błędów oraz poprawy ogólnej jakości danych w systemach informatycznych. Autorzy podkreślają, że skuteczne zarządzanie jakością danych wymaga nie tylko technik wykrywania i korekty błędów, ale także głębokiego zrozumienia struktury i semantyki danych. Narzędzie pozwala na bardziej efektywne zarządzanie jakością danych poprzez zautomatyzowane procesy analizy i korekty, co znacząco redukuje czas i wysiłek potrzebny do utrzymania wysokiej jakości danych.

Schelter (2018) przedstawia koncepcję systemu typu „low code”, co oznacza, że system ten umożliwi użytkownikom programowanie przy minimalnym udziale tradycyjnego kodowania. Kluczowym jego elementem jest tzw. API deklaratywne, czyli interfejs, w którym użytkownicy mogą definiować zasady i reguły walidacji danych w sposób deklaratywny, bez konieczności pisania złożonego kodu programistycznego. API deklaratywne umożliwia użytkownikom określenie, co chcą osiągnąć (np. sprawdzenie spójności danych), a nie jak to zrobić. Te deklaratywne zadania walidacji są następnie przekształcane na zapytania agregujące, które są wykonywane w środowisku Apache Spark. Apache Spark to platforma open-source zaprojektowana do przetwarzania danych na dużą skalę w rozproszonym środowisku. Umożliwia ona szybkie i równoległe przetwarzanie dużych zbiorów danych, co jest kluczowe dla efektywnej walidacji jakości danych w przypadku bardzo dużych i dynamicznie zmieniających się zbiorów danych. Dzięki przekształceniu zadań walidacji na zapytania agregujące, system może efektywnie skalować proces walidacji, zapewniając jednocześnie wsparcie dla walidacji przyrostowej. Oznacza to, że system może analizować i weryfikować tylko te części danych, które uległy zmianie, zamiast przetwarzać cały zbiór danych od początku. To podejście znacząco zwiększa wydajność i redukuje czas potrzebny na walidację danych, szczególnie w środowiskach, gdzie dane są stale aktualizowane i zmieniane.

Semenov (2019) koncentruje się na zarządzaniu spójnością danych produktowych poprzez metodę inkrementalnej walidacji semantycznej. Metoda inkrementalnej walidacji semantycznej polega na sprawdzaniu jedynie tych części danych, które zostały zmienione, zamiast ponownej walidacji całego zbioru danych. Dzięki temu proces walidacji jest bardziej efektywny i mniej czasochłonny, co jest szczególnie istotne w dynamicznych środowiskach, gdzie dane są często aktualizowane. Wykorzystuje graf zależności oparty na specyfikacji modelu danych w języku EXPRESS, specjalistycznym języku modelowania danych używanym do definiowania struktury i reguł dla złożonych danych. Graf zależności w EXPRESS reprezentuje relacje między różnymi elementami danych, co pozwala na dokładne śledzenie, które części danych są dotknięte przez daną transakcję i jakie reguły walidacyjne powinny być zastosowane. W praktyce, gdy dochodzi do zmiany danych, system identyfikuje, które "lokalne" reguły walidacyjne (spot rules) powinny zostać zastosowane na podstawie grafu zależności. To podejście pozwala na szybkie i precyzyjne sprawdzenie spójności danych, zachowując pełną zgodność z zasadami ACID (atomowość, spójność, izolacja, trwałość), które są kluczowe dla niezawodnego przetwarzania danych. Dzięki temu możliwe jest znaczące

przyspieszenie procesu walidacji, zwłaszcza w kontekście krótkich transakcji typowych dla aplikacji działających w czasie rzeczywistym.

Krima, Feeney i Foufou (2014) poruszają kwestie dynamicznych modeli danych produktów co jest szczególnie istotne, w przypadku zarządzania dużą liczbą różnych atrybutów zależnych od kategorii produktów. Autorzy koncentrują się na wykorzystaniu zaawansowanych narzędzi, takich jak ontologie i SPARQL Inference Notation (SPIN), które pomagają w sprawdzaniu poprawności danych i ich integracji w systemach zarządzania cyklem życia produktu. Ontologie organizują i interpretują informacje, ułatwiając zrozumienie i współdzielenie danych, natomiast SPIN automatycznie sprawdza poprawność danych i pomaga w ekstrakcji nowych informacji z istniejących danych. Autorzy pokazują, że dzięki tym narzędziom można dynamicznie dostosowywać modele danych produktów do zmieniających się wymagań. Używają również narzędzia OntoSTEP, które przekształca standardowe modele danych do bardziej elastycznego formatu OWL (Web Ontology Language). Dzięki temu powstaje spójny i zintegrowany system zarządzania danymi, co ułatwia ich utrzymanie i aktualizację w całym cyklu życia produktu.

Redyuk, Kaoudi, Markl i Schelter (2021) koncentrują się na automatyzacji walidacji jakości danych w dynamicznie zmieniających się zbiorach danych, takich jak jeziora danych. Ich rozwiązanie polega na obliczaniu statystyk opisowych dla nowych partii danych oraz użyciu metod wykrywania nowości opartych na uczeniu maszynowym do monitorowania jakości danych i identyfikowania odchyleń od normalnych charakterystyk danych. To podejście jest kluczowe w kontekście dużych, nieustrukturyzowanych zbiorów danych, które są regularnie aktualizowane i w których brak jest eksperckiej wiedzy domenowej.

Badania Karpishek, Michahelles i Fleisch'a (2012, 2014) skoncentrowane są na problemie niepoprawnych nazw produktów w publicznych źródłach danych. Autorzy tych badań używają modeli klasyfikacyjnych oraz metod uczenia maszynowego do identyfikacji i oceny błędnych nazw produktów, ze szczególnym uwzględnieniem GTIN (Global Trade Item Number). Celem jest zwiększenie jakości danych produktowych i zaufania konsumentów, a także ocena skali problemu. W badaniu z 2012 roku, autorzy skupiają się na jakości danych głównych produktów dostępnych publicznie w kontekście GTIN. Autorzy badają, jak dużym problemem jest niepoprawność nazw produktów w publicznie dostępnych źródłach w niemieckim i szwajcarskim sektorze dóbr konsumpcyjnych. Używają modelu klasyfikacyjnego, który łączy dopasowanie ciągów i uczenie nadzorowane do identyfikowania niepoprawnych nazw produktów, osiągając wysoką wartość recall (0,99 do 1,0) przy



akceptowalnej precyzji, co pokazuje przydatność metody. W kontynuacji badań z 2014 roku, autorzy analizują poprawność nazw produktów w internetowych źródłach danych głównych produktów, ponownie korzystając z GTIN. W tej pracy badane są dane z ośmiu publicznie dostępnych źródeł dla zestawu numerów GTIN skanowanych przez użytkowników aplikacji do skanowania kodów kreskowych na urządzeniach mobilnych. Metoda używa dopasowania ciągów i uczenia maszynowego do estymacji liczby błędnych nazw produktów. Wyniki wskazują, że około 2% nazw produktów jest nieprawidłowych, co podkreśla skalę problemu.

Wang (2020) i współautorzy koncentrują się na walidacji wartości atrybutów tekstowych w katalogach e-commerce przy wykorzystaniu ograniczonej ilości danych oznakowanych. Ich podejście polega na formalizacji problemu walidacji jako zadania wnioskowania w języku naturalnym, integracji meta-uczenia i zmiennych ukrytych w jednolitym modelu oraz wprowadzeniu nowej funkcji celu wspierającej konsystencję dystrybucji między danymi oznakowanymi i nieoznakowanymi. Celem jest zwiększenie dokładności walidacji atrybutów tekstowych.

Schmidts (2020) i jego zespół w swojej pracy zajęli się integracją danych produktowych o niskiej jakości z różnych źródeł do jednolitego katalogu. Proponują metodologię Attribute Label Ranking (ALR), która wykorzystuje modele uczone na rzeczywistych danych do przewidywania wielu etykiet dla każdego atrybutu i klasyfikowania ich w celu znalezienia najlepszego dopasowania do schematu katalogu. Jest to szczególnie przydatne dla małych i średnich przedsiębiorstw (MŚP), które muszą radzić sobie z różnicami w formatach i niską jakością danych produktowych. W kolejnym artykule Schmidts (2021) i współautorzy rozszerzają swoje badania o integrację heterogenicznych i często zmieniających się danych produktowych. Skupiają się na wyzwaniach związanych z integracją danych pochodzących z wielu różnych, dynamicznie zmieniających się źródeł. Proponują zastosowanie zaawansowanej architektury sieci neuronowej oraz funkcji językowych do poprawy procesu mapowania i klasyfikacji atrybutów. Metodologia ta umożliwi lepsze radzenie sobie ze zmiennością danych oraz ich heterogenicznością, co jest kluczowe w dynamicznym środowisku e-commerce. Dzięki temu, systemy mogą lepiej adaptować się do zmian w danych, co zwiększa dokładność i spójność zintegrowanych katalogów produktowych.

Kim, Lee i Shim (2021) koncentrują się na rozwijaniu modeli uczenia głębokiego, które są „świadome” danych wejściowych. Ich badania obejmują selekcję, transformację i filtrowanie danych wejściowych w celu poprawy dokładności klasyfikacji produktów. Modele te wykorzystują techniki uczenia głębokiego zintegrowane z procesami selekcji odpowiednich

danych treningowych, transformacji strukturalnej danych oraz usuwania niepotrzebnych informacji. Celem jest stworzenie systemu, który nie tylko klasyfikuje produkty z wysoką precyzją, ale także adaptuje się do różnorodnych i dynamicznie zmieniających się zestawów danych.

W Tabeli 9 zaprezentowano syntetyczny przegląd zebranych publikacji, zawierający streszczenie, cel badawczy, problem badawczy oraz opis zaproponowanych rozwiązań.

Tabela 9. Przegląd rozwiązań praktycznych

Publikacja	Streszczenie	Cel	Problem	Opis rozwiązania
Towards a domain independent platform for data cleaning (Arasu, A; Chaudhuri, S; Chen, Z; Ganjam, K; ... 2011)	Autorzy prezentują platformę do czyszczenia danych niezależną od dziedziny, opracowaną przez Microsoft Research. Platforma zawiera zestaw podstawowych reguł i narzędzi projektowych, które pozwalają programiście tworzyć zaawansowane rozwiązania do czyszczenia danych z minimalnym wysiłkiem programistycznym.	Prezentacja możliwości projektu	Trudność w opracowaniu efektywnych rozwiązań do czyszczenia danych, które byłyby zarówno elastyczne, jak i łatwe do zaimplementowania w różnych kontekstach.	Narzędzia składają się głównie z modułów oczyszczania danych opartych na regułach do wyszukiwania teksów podobnych oraz modułu segmentacji danych.
Data cleansing techniques for large enterprise datasets (Prasad, KH; Faruque, TA; Joshi, S; ... 2011)	Autorzy przedstawiają narzędzie do poprawy jakości danych, które poprzez identyfikację wariantów i synonimów jednostek w danych, ułatwia pisanie reguł jakościowych. Narzędzie to wykorzystuje ramy reguł "ripple down" (RDR) do utrzymania i łatwego dodawania nowych reguł, prezentując typowy przepływ pracy procesu poprawy jakości danych oraz jego użyteczność na różnych etapach.	Prezentacja narzędzia do poprawy jakości danych w dużych przedsiębiorstwach, które pomaga w procesie standaryzacji danych i eliminacji duplikatów.	Trudność w poprawie jakości danych w różnych domenach i geografiach, konieczność dużej indywidualizacji narzędzi czyszczących dane przy przejściu między klientami i domenami.	Narzędzie analizuje charakterystykę jednostek danych, pomagając w pisaniu reguł jakościowych. Stosuje ramy reguł RDR do utrzymania reguł i redukuje wysiłek usługowy przy dodawaniu nowych reguł. Prezentuje przepływ pracy i użyteczność narzędzia na każdym etapie, oraz eksperymentalne wyniki potwierdzające redukcję wysiłku.
The not so unique global trade identification number: product master data quality in publicly available sources (Karpischek, S; Michahelles, F; Fleisch, E 2012)	Badanie dotyczy jakości danych głównych produktów w kontekście GTIN, globalnego systemu identyfikacji artykułów handlowych. GTINy, wcześniej stosowane głównie w łańcuchach dostaw, obecnie służą także konsumentom do uzyskiwania informacji o produktach. Autorzy analizują problem niepoprawnych nazw produktów w publicznych źródłach w niemieckim i szwajcarskim sektorze dóbr konsumpcyjnych.	Zidentyfikowanie niepoprawnych nazw produktów i ocena skali problemu związanej z jakością danych głównych produktów w publicznie dostępnych źródłach.	Problem niepoprawności nazw produktów w publicznie dostępnych źródłach, który wpływa zarówno na relacje B2B, jak i B2C, wpływając na wizerunek marki i zaufanie konsumentów.	Użyto modelu klasyfikacyjnego, który łączy dopasowanie ciągów i uczenie nadzorowane do identyfikowania niepoprawnych nazw produktów. Wysoka wartość recall (0,99 do 1,0) przy akceptowalnej precyzji pokazuje przydatność metody.
Detecting incorrect product names in online sources for product master data (Karpischek, S; Michahelles, F; Fleisch, E 2014)	Autorzy analizują poprawność nazw produktów w internetowych źródłach danych głównych produktów, korzystając z GTIN (Global Trade Item Number). W pracy badane są dane z ośmiu publicznie dostępnych źródeł dla zestawu GTINów skanowanych przez użytkowników aplikacji do	Ocena poprawności nazw produktów w internetowych źródłach oraz kwantyfikacja problemu jakości danych produktów.	Wykrycie błędnych nazw produktów w online'owych źródłach danych i ocena jakości tych danych w kontekście wykorzystania GTIN.	Metoda używa dopasowania ciągów i uczenia maszynowego do estymacji liczby błędnych nazw produktów. Wyniki wskazują, że około 2% nazw produktów jest nieprawidłowych.

Publikacja	Streszczenie	Cel	Problem	Opis rozwiązania
	skanowania kodów kreskowych na urządzeniach mobilnych.			
Dynamic customization, validation and integration of product data models using semantic web tools (Krima, S; Feeney, AB; Foufou, S 2014)	Autorzy przedstawiają potrzebę dynamicznych modeli danych produktów wspierających różnorodne wymagania domenowe przez cały cykl życia produktu. Prezentują ulepszenia metodologii ISO wykorzystując ontologie i notację SPARQL Inference Notation (SPIN) do walidacji danych produktów oraz sposób użycia tych ontologii do integracji danych PLM poprzez Linked Data.	Wprowadzenie dynamicznych modeli informacji produktu i ulepszenie integracji danych PLM.	Wykorzystanie standardowych modeli danych produktu, które muszą wspierać różnorodne wymagania specyficzne dla danej dziedziny oraz rozwijać się, aby wspierać nowe wymagania.	Użycie OntoSTEP do transformacji modeli informacji STEP i danych do OWL, co pozwala na jednolitą architekturę i pełne wykorzystanie OWL. SPIN zapewnia utrzymanie spójności pomimo OWA (Open World Assumption) w OWL.
Automating large-scale data quality verification (Schelter, Sebastian; Lange, Dustin; Schmidt, Philipp; Celikel, Meltem; Biessmann, Felix; Grafberger, Andreas 2018)	Autorzy przedstawiają system do automatyzacji weryfikacji jakości danych na dużą skalę, który łączy powszechne ograniczenia jakości z użytkownikiem zdefiniowanym kodem walidacji. System wykorzystuje API deklaratywne i tłumaczy zadania walidacji na zapytania agregujące w Apache Spark.	Opracowanie systemu do automatycznej weryfikacji jakości danych, który skaluje się do dużych zbiorów danych i spełnia wymagania przypadków użycia w produkcji.	Wyzwaniem jest ręczne weryfikowanie jakości danych, co jest czasochłonne i niesie ryzyko błędów.	Rozwiązanie obejmuje API deklaratywne umożliwiające definiowanie ograniczeń, wykonanie walidacji przez zapytania agregujące i wsparcie dla walidacji przystosowanej na rosnących zbiorach danych. Wykorzystuje również uczenie maszynowe do sugerowania ograniczeń i wykrywania anomalii.
IDQMS: An Intelligent Data Quality Management System Tool (Salem, AB; Boufares, F 2019)	Zaprezentowano koncepcję narzędzia iDQMS do zarządzania jakością danych, które umożliwia lepsze zrozumienie definicji danych oraz wykrywanie i korygowanie anomalii poprzez kategoryzację semantyczną.	Zaprezentowanie projektu narzędzia iDQMS, które pomoże użytkownikowi w podejściu do jakości danych poprzez zrozumienie ich semantyki przed dokonaniem korekty.	Rosnąca ilość rozproszonych i heterogenicznych danych oraz trudności z manipulacją danych bez wiedzy o ich strukturze i semantyce.	iDQMS ma na celu lepsze rozumienie definicji danych, w tym semantycznej nazwy kolumny, typu danych i pewnych ograniczeń, oraz wykrywanie i korygowanie anomalii.
Effective consistency management for large-scale product data (Semenov, Vitaly; Ilyin, Denis; Morozov, Sergey; Tarlapan, Oleg 2019)	Autorzy przedstawiają metodę inkrementalnej walidacji semantycznej dużych zbiorów danych produktowych. Zaproponowana metoda polega na stosowaniu "lokalnych" reguł (spot rules) dotkniętych przez transakcje, wykorzystując graf zależności oparty na specyfikacji modelu danych w języku EXPRESS.	Zwiększenie wydajności walidacji semantycznej i zarządzania skomplikowanymi danymi produktowymi w pełnej zgodności z zasadami ACID.	Trudność w zapewnieniu spójności skomplikowanych danych zgodnie ze standardami, przy czym pełna walidacja semantyczna jest kosztowna czasowo.	Rozwiązanie umożliwia zwiększenie wydajności walidacji semantycznej poprzez analizę reguł dotkniętych przez transakcję, z wykorzystaniem pomocniczej struktury danych w postaci grafu zależności. Zastosowana metodyka pozwala na znaczne przyspieszenie walidacji, zwłaszcza dla krótkich transakcji typowych dla aplikacji w czasie rzeczywistym.
Automatic Validation of Textual Attribute Values in E-commerce Catalog by Learning with Limited Labeled Data (Wang, Yaqing; Xu, Yifan; Ethan; Li, Xian; Dong, Xin Luna; Gao, Jing 2020)	Autorzy przedstawiają podejście do automatycznej walidacji wartości atrybutów tekstowych w katalogach e-commerce, wykorzystując ograniczoną ilość danych oznakowanych. Istotne jest zweryfikowanie poprawności tych wartości, by poprawić doświadczenie kupujących i umożliwić skuteczniejsze rekomendacje produktów. Ze względu na ogromną liczbę produktów potrzebne jest	Celem jest opracowanie metody automatycznej walidacji poprawności tekstowych wartości atrybutów produktów.	Głównym problemem jest zawartość katalogów informacyjnych zawierających błędne fakty ze względu na samo zgłaszanie przez detalicznych sprzedawców oraz brak wystarczającej ilości danych oznakowanych dla wszystkich kategorii produktów.	Opis rozwiązania obejmuje: (1) formalizację problemu walidacji atrybutów tekstowych jako zadania wnioskowania w języku naturalnym w ustawieniu uczenia z małą ilością danych, (2) integrację meta-uczenia i zmiennych ukrytych w jednolitym modelu, (3) nową funkcję celu wspierającą konsystencję dystrybucji

Publikacja	Streszczenie	Cel	Problem	Opis rozwiązania
	skuteczne podejście automatycznej walidacji.			między danymi oznakowanymi i nieoznakowanymi.
Catalog Integration of Low-quality Product Data by Attribute Label Ranking: (Schmidts, Oliver; Kraft, Bodo; Winkens, Marvin; Zündorf, Albert 2020)	Autorzy przedstawiają wyzwanie integracji danych produktowych z różnych źródeł do jednolitego katalogu, co jest trudne dla MŚP z powodu różnic w formatach i niskiej jakości danych. Zaproponowano metodykę Attribute Label Ranking (ALR) do mapowania schematu danych.	Ułatwienie integracji danych produktowych dla MŚP.	Trudność w integracji zmiennych, heterogenicznych danych produktowych o niskiej jakości do jednolitego katalogu.	ALR to model uczony na rzeczywistych danych, przewidujący wiele etykiet dla każdego atrybutu i klasyfikujący je w celu znalezienia najlepszego dopasowania do schematu katalogu.
Automating Data Quality Validation for Dynamic Data Ingestion. (Redyuk, S; Kaoudi, Z; Markl, V; Schelter, S 2021)	Autorzy opisują sposób na automatyczne sprawdzanie jakości danych, co jest ważne w systemach opierających się na danych. Błędy w danych mogą powodować niespodziewane problemy w działaniu systemów i zależnych od nich usług. Głównym wyzwaniem jest ręczne i czasochłonne naprawianie błędów, szczególnie kiedy duże i stale rosnące ilości danych są regularnie dodawane do dużych, nierelacyjnych baz danych, takich jak jeziora danych.	Opracowanie metody automatycznej walidacji jakości danych dla dynamicznie zmieniających się zbiorów danych i dla których brak jest eksperckiej wiedzy domenowej.	Ręczne procesy walidacji danych są żmudne i trudne, zwłaszcza w kontekście dużych, nieustrukturyzowanych zbiorów danych, które ulegają zmianom.	Rozwiązanie polega na obliczaniu statystyk opisowych dla nowych partii danych i użyciu metody wykrywania nowości opartej na uczeniu maszynowym do monitorowania jakości danych i identyfikowania odchyleń od normalnych charakterystyk danych.
Catalog integration of heterogeneous and volatile product data (Schmidts, O; Kraft, B; Winkens, M; Zündorf, A 2021)	Autorzy przedstawiają wyzwanie integracji często zmieniających się danych produktowych od różnych producentów do jednolitego katalogu, co jest problematyczne dla małych i średnich firm e-commerce.	Uprościć proces integracji katalogów dla praktyków przez zastosowanie silnika rekomendacji Attribute Label Ranking (ALR).	Trudność w integracji heterogenicznych i zmiennych danych produktowych z różnych źródeł do jednolitego katalogu.	ALR używa różnych architektur sieci neuronowych i funkcji językowych do poprawy procesu integracji danych, oceniany na podstawie wpływu na produkcję i ograniczeń.
Developing data-conscious deep learning models for product classification (Kim, Y; Lee, HJ; Shim, J 2021)	Autorzy przedstawiają modele uczenia głębokiego świadome danych wejściowych zawierających tekstową informację o produkcie, które wykorzystują znane modele uczenia głębokiego zintegrowane z procesami selekcji, transformacji i filtrowania danych. Demonstrują praktyczność tych modeli poprzez eksperymenty z rzeczywistymi danymi informacji produktowych.	Opracowanie modeli uczenia głębokiego, które są świadome danych wejściowych i poprawa klasyfikacji produktów w e-commerce.	Niedokładna klasyfikacja produktów w systemach handlu elektronicznego spowodowana nieodpowiednim doborem i przetwarzaniem danych wejściowych dla modeli uczenia głębokiego.	Architektura modelu zawiera etapy selekcji odpowiednich danych jako danych treningowych, transformacji strukturalnej (a nie semantycznej) wybranych danych dla ich wygodnego zastosowania w treningu, oraz filtrowania, aby usunąć słowa uznane za niepotrzebne.

Źródło: opracowanie własne

Na podstawie przedstawionych publikacji można sformułować kilka kluczowych wniosków dotyczących jakości danych, czyszczenia danych, walidacji i integracji danych produktowych. Istnieje wyraźny trend w kierunku automatyzacji procesów walidacji danych. Automatyzacja, jak wskazują Schelter i współautorzy (2018) oraz Redyuk i współautorzy

(2021), pomaga w skutecznym skalowaniu procesów weryfikacji jakości danych, zwłaszcza w kontekście dużych zbiorów danych. Dynamiczne modele danych produktów, jak wskazują Krima, Feeney, Foufou (2014) oraz Semenov i współautorzy (2019), są kluczowe dla obsługi różnorodnych wymagań domenowych. Wykorzystanie zaawansowanych narzędzi, takich jak ontologie i SPIN, podkreśla potrzebę elastycznych i adaptacyjnych podejść do walidacji danych. Badania Karpischeka, Michahellesa i Fleischa (2012, 2014) ukazują, jak ważna jest poprawność nazw produktów w źródłach publicznych i online, ze względu na wpływ na relacje B2B oraz B2C i wizerunek marki. Prace Schmidtsa i współautorów (2020, 2021) oraz Kim, Lee i Shima (2021) ukazują wyzwania w integracji zmiennych, heterogenicznych danych produktowych. Zastosowanie nowoczesnych technik, takich jak modele uczenia głębokiego, jest kluczowe dla usprawnienia tego procesu.

Podsumowując, przedstawione badania podkreślają rosnącą rolę zaawansowanych technologii i automatyzacji w zarządzaniu jakością danych, co jest kluczowe dla efektywności operacyjnej w różnych sektorach przemysłu. Istotne jest również ciągle dostosowywanie metod i narzędzi do dynamicznie zmieniających się wymagań i kontekstów danych. Publikacje poruszały problematykę z różnych perspektyw, jednakże wciąż istnieje luka w rozwiązaniach dostosowanych do dużych katalogów produktowych i globalnych problemów związanych z danymi podstawowymi. Przykładowo, nie znaleziono metod poprawy semantycznej jakości nazw produktów (zaproponowane metody walidacji nazw produktów opierały się głównie na porównaniu nazw do wzorca, oraz do wykorzystaniu metod klasycznych bez rozumienia kontekstu danych) ani predykcji poprawności zawartości netto na podstawie nazwy produktu. Brakuje sposobów na walidację tego, co znajduje się na zdjęciach produktów oraz weryfikacji poprawności klasyfikacji GPC (Global Product Classification). Ponadto, nie rozpoznano, jak efektywnie wyuczyć modele z wykorzystaniem danych zastanych albo gotowych modeli, które od kilku lat całkowicie zmieniły obraz sztucznej inteligencji na świecie. Istnieje na przykład potencjał w najnowszych rozwiązaniach z wykorzystaniem zaawansowanych modeli, takich jak modele GPT (Generative Pre-trained Transformer). Modele te mogą znacznie poprawić jakość danych poprzez lepsze zrozumienie kontekstu i semantyki nazw produktów, co otwiera nowe możliwości w zakresie automatyzacji i inteligentnej analizy danych. Jest to więc obszar niewyeksplorowany, który wymaga dalszego zgłębienia.

### **3. Eksploracja standardów i wytycznych w kontekście podstawowych danych produktowych – badania własne**

W dobie cyfryzacji i rosnącej popularności handlu elektronicznego, precyzyjne i wiarygodne dane produktowe nabierają szczególnego znaczenia. Kwestia ta została szczegółowo omówiona w Rozdziale 2.4, gdzie podkreślono, że dokładność informacji o produkcie jest nieodzowna nie tylko dla skuteczności procesu sprzedaży, ale również dla budowania zaufania konsumentów oraz optymalizacji łańcucha dostaw. Utrzymanie wysokiej jakości danych produktowych staje się zatem priorytetem, co wymaga zrozumienia czy rynek stosuje i egzekwuje wspólne zasady dotyczące zapewnienia odpowiednich standardów. Takie ujednolicone wytyczne mogłyby stanowić podstawę do opracowania uniwersalnych narzędzi walidacyjnych oraz mechanizmów normalizacji danych produktowych, które byłyby użyteczne nie tylko w obrębie jednego katalogu, ale także dla szerokiego spektrum rozwiązań e-commerce, zapewniając wysoką jakość danych w różnych środowiskach. Niniejszy rozdział pracy skupia się na analizie sektora e-commerce pod kątem rozumienia, czym są i jak powinny być wypełniane poszczególne atrybuty danych produktowych. Celem było zbadanie, w jaki sposób różne platformy e-commerce oraz powszechnie stosowane rozwiązania technologiczne adaptują i wdrażają uznane zasady oraz standardy dotyczące danych produktowych, oraz czy mogą one służyć jako wzór dla procesów walidacji tych danych.

Wyniki badań zaprezentowano i opublikowano w ramach 13th International Conference on Applied Human Factors and Ergonomics (Niemir i Mrugalska, 2022b), a także opublikowano w European Research Studies Journal (Niemir i Mrugalska, 2021).

#### **3.1. Materiały i metoda badania**

W ramach niniejszego badania przyjęto do analizy następujące atrybuty, które uznano za kluczowe w kontekście zarządzania danymi produktowymi:

##### **A. Globalny unikalny identyfikator produktu (Global unique product ID).**

Rozważono znaczenie i sposób implementacji globalnych identyfikatorów produktów, które umożliwiają jednoznaczną identyfikację produktów na różnych platformach e-commerce.

- B. **Unikalna nazwa opisująca produkt (Unique name that describes the product).** Analizowano, jak nazwy produktów są używane do ich identyfikacji i jak różnią się one w zależności od platformy, biorąc pod uwagę ich unikalność i zrozumiałość.
- C. **Nazwa marki (Brand name).** Zbadano, jak różne platformy e-commerce podchodzą do atrybutu nazwy marki, jego standaryzacji i wpływu na identyfikację produktu.
- D. **Zdjęcie produktu (Product image).** Skoncentrowano się na wymaganiach dotyczących zdjęć produktów, w tym na standardach jakości, formatach i wytycznych dotyczących prezentacji wizualnej produktów.
- E. **Zawartość netto (Net content).** Analizowano, w jaki sposób informacje o zawartości netto produktu są prezentowane i weryfikowane na różnych platformach e-commerce.
- F. **Klasyfikacja produktu (Product Classification).** Badano, jak klasyfikacja produktów jest realizowana na różnych platformach, z uwzględnieniem istniejących systemów klasyfikacji i ich wpływu na organizację danych produktowych.

Dobór atrybutów był wynikiem analizy dotyczącej podstawowych danych produktowych, zawartej w Rozdziale 2.6. Oprócz wyboru grupy referencyjnej podstawowych atrybutów produktu, konieczne było wybranie odpowiednich przedstawicieli rynku, jak również platform i narzędzi powszechnie używanych w handlu elektronicznym (e-commerce) w celu dokonania analizy porównawczej interpretowalności atrybutów danych produktowych w stosunku do rozwiązania Global Registry Platform zaproponowanego przez Organizację GS1 (GS1.pl, 2023).

Dokonano tego na podstawie:

- A. Analizy statystyk ruchu internetowego na świecie i w Polsce w kategorii e-commerce, opartej na rankingu ruchu internetowego Similarweb.com (2021) (statystyki na dzień 01 marca 2021),
- B. Analizy popularności używania platform e-commerce, opartej na danych z BuiltWith.com (2021) (wyniki opublikowane 16 kwietnia 2021),
- C. Własnych doświadczeń dotyczących powszechnie używanych narzędzi w e-commerce (Google Merchant, Schema.org).

W Tabeli 10 zaprezentowano zbiór wytypowanych przedstawicieli rynku, ich charakterystykę, misję, a także kluczowe informacje o ich działalności i popularności.

Tabela 10. Przedstawiciele rynku, platformy i powszechnie używana narzędzia w e-commerce

Nazwa	Opis
<b>Schema.org</b>	Schema.org to wspólna aktywność społecznościowa, której misją jest tworzenie, utrzymywanie i promowanie schematów dla strukturalnych danych w Internecie, na stronach internetowych, w wiadomościach e-mail i poza nimi. Założona przez Google, Microsoft, Yahoo i Yandex, słowniki Schema.org są rozwijane w otwartym procesie społecznościowym, za pomocą listy mailingowej public-schemaorg@w3.org i przez GitHub. Użycie strukturalnych danych na stronach internetowych wpływa na wyniki wyszukiwania oraz sposób wyświetlania stron produktów e-sklepów w wyszukiwarkach. Jest to szeroko używany standard w e-commerce.
<b>Google Merchant Center</b>	Google Merchant Center to cyfrowa platforma umożliwiająca sprzedawcom internetowym przesyłanie danych produktowych, stanowiących kluczowy element systemu reklamowego Google Shopping (dawniej Google Product Search). Platforma ta nie tylko pozwala na przekazywanie informacji o produktach, ale również umożliwia prezentowanie szczegółowych danych dotyczących sklepów e-commerce. Główną funkcją Google Merchant Center jest zapewnienie firmom możliwości efektywnego zarządzania informacjami o produktach, w tym ich zdjęciami i cenami, co zwiększa ich widoczność w wynikach wyszukiwania Google Shopping. W rezultacie Google Merchant Center działa jako platforma agregująca dane, oferując narzędzia niezbędne do optymalizacji i skutecznej prezentacji produktów w ekosystemie Google.
<b>Najważniejsze platformy marketplace: Amazon, eBay, Allegro.</b>	Amazon to największy na świecie sprzedawca internetowy, który sprzedaje bezpośrednio lub jako marketplace. Numer 1 w rankingu ruchu internetowego w kategorii e-commerce i zakupów na świecie (SimilarWeb, 2021). eBay to platforma zakupowa, najbardziej znana z aukcji oraz transakcji między konsumentami, jednak pełni również funkcję popularnego marketplace'u dla sprzedawców internetowych, stanowiąc istotny kanał sprzedaży. Pod względem ruchu internetowego w kategorii e-commerce i zakupów na świecie, zajmuje drugie miejsce (SimilarWeb, 2021). Allegro to najpopularniejsza platforma zakupowa w Polsce i jedna z najpopularniejszych stron e-commerce w Europie (SimilarWeb, 2021).
<b>Porównywarka cen Ceneo</b>	Ceneo to najpopularniejsza w Polsce porównywarka cen, prezentująca ofertę ponad 18 000 sklepów internetowych. Zajmuje pierwsze miejsce w rankingu ruchu internetowego w kategorii porównywarek cen w Polsce, a na świecie drugie (SimilarWeb, 2021).
<b>Najpopularniejsze platformy e-commerce: WooCommerce, Shopify, Magento.</b>	Badanie przeprowadzone przez BuiltWith.com (2021) dostarcza istotnych danych dotyczących popularności platform e-commerce. Z analizy wynika, że najczęściej stosowaną platformą, z udziałem rynkowym wynoszącym 30% i obejmującym 35 887 stron internetowych w 2021 roku była WooCommerce. Jest to konfigurowalna, oparta na WordPress, otwarta platforma e-commerce, która wyróżnia się dostępnością i możliwością personalizacji. Shopify zajął drugie miejsce z 18% udziałem rynkowym i 22 285 stronami internetowymi. To łatwa w obsłudze platformę do budowy sklepów internetowych, która posłużyła do wytworzenia ponad miliona sklepów. Magento, z 9% udziałem rynkowym i 10 778 stronami internetowymi, uplasował się na trzeciej pozycji. Jest to platforma e-commerce oparta na otwartym kodzie. Przedstawione platformy różnią się pod wieloma względami. Różnice obejmują model biznesowy, sposób udostępniania oprogramowania (Software as a Service - SaaS vs. rozwiązania hostowane samodzielnie), stopień otwartości kodu źródłowego oraz możliwości rozszerzania funkcji poprzez dodatkowe moduły i wtyczki. Różnorodność tych podejść oferuje szeroki wybór dla różnych typów przedsiębiorstw, od małych firm po duże korporacje, umożliwiając dostosowanie platformy e-commerce do specyficznych potrzeb biznesowych i technicznych.

Źródło: opracowanie własne



### 3.2. Rezultaty badań własnych

Wyniki badań usystematyzowano względem atrybutów produktów i zaprezentowano w formie zestawień tabelarycznych o charakterze referencyjnym. Opisy, wyjaśniające znaczenie i ograniczenia poszczególnych atrybutów zostały maksymalnie ograniczone, aby wyeksponować kluczowe różnice w implementacjach, zachowując jednocześnie przejrzystość prezentacji danych. Nie zawierają więc szczegółowych informacji np. w odniesieniu do ograniczeń ustawodawczych danego kraju czy specyfiki prezentacji i opisu konkretnych wyrobów. Do analizy wykorzystano materiały dostępne na oficjalnych stronach internetowych twórców rozwiązań. W przypadku rozwiązań oferowanych przez GS1 była to strona GS1 Global Office (GS1, 2023b) w której prezentowano rozwiązania dotyczące GRP. Analiza ponadto opierała się o standardy GS1 Web Vocabulary Standard (GS1, 2015), zawierające kompleksowy wykaz atrybutów wykorzystywanych przez GS1 w internetowych rozwiązaniach, oraz na GS1 Global Data Dictionary (GDD) - repozytorium elementów danych zdefiniowanych we wszystkich standardach GS1 (GS1, 2022a). Posługiwano się również wytycznymi dotyczącymi umieszczania zdjęć w katalogach GS1 GDSN (GS1, 2022b). W analizie Schema.org użyto wytycznych ze strony Schema.org z przedziału danych strukturalnych zdefiniowanych dla elementu „Product” (Schema.org Community Group, 2015). Podkreślenia wymaga fakt, iż analiza nie objęła rozszerzenia schematu "GS1 SmartSearch" ze względu na jego ograniczone rozpowszechnienie oraz fakt, iż rozszerzenie to dodaje, a nie zastępuje podstawowe atrybuty produktu w strukturze danych. Szczegółowe informacje o "GS1 SmartSearch" dostępne są na stronach Schema.org oraz GS1. Do celów analizy dokumentacji Google Merchant, wykorzystano obszerny przewodnik online (Google.com, 2021), prezentujący z dużą precyzją ograniczenia w użyciu pól, wraz z przykładami ich zastosowania. Podobnie dokumentacja platformy Amazon (Amazon.com, 2021) wykorzystana w badaniu, zawierała szereg przykładów i informacji dotyczących jakości danych. W celu analizy dokumentacji platform eBay i Allegro użyto podobnego podejścia związanego z analizą przewodników (Allegro.pl, 2021; eBay Inc., 2018) , natomiast brakujące elementy uzupełniono bezpośrednio poprzez przeglądanie funkcjonalności platform. W przypadku porównywarki cen Ceneo, analiza opierała się na dokumentacji API (Ceneo.pl, 2021). Dokumentacje platform e-commerce takich jak platform jak Magento, WooCommerce i Shopify (Magento.com, 2020; Shopify.com, 2021; WooCommerce.com, 2021) dostarczyły ograniczonej ilości szczegółów implementacyjnych w kontekście podstawowych danych produktowych. Warto również zauważyć, że w procesie zbierania informacji o danych

produktowych dla tych platform nie brano pod uwagę możliwości modyfikacji kodu programu, instalacji dodatkowych wtyczek czy rozszerzeń.

W kolejnych podrozdziałach zaprezentowano wyniki badań przeprowadzone dla każdego z wymienionych atrybutów danych produktowych.

### 3.2.1. Globalny unikalny identyfikator produktu

Analiza platform wykazała, że jedynym powszechnie uznawanym globalnym identyfikatorem produktów jest GTIN, będący międzynarodowym standardem numeracji wykorzystywanym do jednoznacznej identyfikacji produktów i usług w globalnym łańcuchu dostaw (GS1, 2021). Szczegółowe informacje na temat powszechności użycia tego identyfikatora zostały omówione w Rozdziale 2.6. Każdy numer GTIN jest unikatowy na skalę światową, co zapewnia, że żaden inny produkt nie może posiadać tego samego numeru, z wyjątkiem niewielkiej przestrzeni celowo wyłączonej z numeracji do innych celów. GTIN może być reprezentowany w różnych formatach liczbowych, w zależności od rodzaju produktu i lokalizacji. Dopuszczalne formaty to 8, 12, 13 lub 14 cyfr. Długość ciągu definiuje typ numeru GTIN: GTIN-14, GTIN-13, GTIN-12 (numery amerykańskie UPC), GTIN-8. Każdy numer niższego typu można zaprezentować w typie wyższym dodając zero nieznaczące z lewej strony ciągu. Struktura numeru GTIN obejmuje prefiks firmy (przydzielony przez GS1), numer referencyjny przedmiotu oraz cyfrę kontrolną. Na Rysunku 9 zaprezentowano strukturę najpowszechniejszego w handlu detalicznym numeru GTIN-13 ze wskazaniem występowania prefiksu.



Rysunek 9. Struktura numeru GTIN-13

Źródło: opracowanie własne

Cyfra kontrolna jest wyliczana matematycznie na podstawie pozostałych cyfr w numerze GTIN niezależnie od długości samego numeru, co pomaga weryfikować poprawność numeru podczas skanowania kodu kreskowego.

$$S = \sum_{i=1}^{n-1} ((2 - (-1)^i) * d_i)$$

gdzie  $d_i$  to kolejna cyfra kodu, a  $n$  to długość kodu.

Cyfrą kontrolną będzie wartość o jaką należy dopełnić otrzymaną sumę, by ta była podzielna przez 10.

$$c = -S \text{ mod } 10$$

Prefiks kraju jest słownikiem, dostępnym w postaci tabeli na stronach organizacji GS1, która zarządza wydawaniem kodów i dba o jego unikalność w skali całego świata. Słownik zmienia się – są np. dodawane nowe prefiksy, natomiast dzieje się to stosunkowo rzadko. Tabela zawiera informacje zarówno o używanych prefiksach w danym kraju, jak również o rezerwacjach prefiksów do konkretnych zastosowań, do których nie mogą być przypisywane produkty w obrocie detalicznym, lub mogą być używane tylko wewnątrz sieci handlowej, a numery mogą dublować się w skali świata.

Prefiksy firm są przydzielane przez lokalne biura GS1, które działają w różnych krajach i regionach. Proces przydziału prefiksu firmy jest związany z rejestracją firmy w systemie GS1 i zależy od wielkości firmy oraz zakresu jej działalności. Po otrzymaniu prefiksu, firma może generować własne numery GTIN dla swoich produktów, co pozwala na ich jednoznaczną identyfikację w całym łańcuchu dostaw, od produkcji po punkt sprzedaży detalicznej. Zazwyczaj wydawane są prefiksy firm o długości od 4 do 8 cyfr co odpowiada możliwości rejestracji od 10 do 100000 produktów. Dłuższe prefiksy są często przydzielane mniejszym firmom, które potrzebują mniejszej liczby unikalnych numerów GTIN, podczas gdy większe firmy, produkujące większą liczbę różnorodnych produktów, mogą otrzymać krótsze prefiksy umożliwiające generowanie większej liczby unikalnych numerów GTIN.

Warto wspomnieć, że standard GTIN włączył Międzynarodowy Znormalizowany Numer Książki (ISBN), Międzynarodowy Numer Wydawnictw Ciągłych (ISSN), Międzynarodowy Numer Muzyczny (ISMN), Międzynarodowy Numer Artykułu (który obejmuje Europejski Numer Artykułu i Japoński Numer Artykułu) oraz Universal Product Codes (UPC) do uniwersalnej przestrzeni numerów, dlatego, o ile nie zaznaczono inaczej, pole

GTIN zawiera wymienione standardy. Numery GTIN mogą mieć 8, 12, 13 lub 14 cyfr. Krótkie numery mogą być kodowane jako GTIN-14 przez dodanie zer wiodących.

W Tabeli 11 podsumowano możliwości platform/narzędzi w zakresie zarządzania danymi identyfikacyjnymi produktu. Oddzielnie - dla identyfikatorów globalnych i identyfikatorów o ograniczonym zastosowaniu w łańcuchu dostaw. Tabela nie obejmuje własnych identyfikatorów platform, np. wewnętrzny ID, ASIN (Amazon Standard Identification Number), ePID (eBay) itp., natomiast uwzględnia numery producenta w kolumnie „Inne ID”. Są to: SKU i MPN. SKU - Stock Keeping Unit number, to unikatowy numer przypisany do produktu między innymi w celu śledzenia stanów magazynowych. MPN - Manufacturer Part Number to seria unikatowych numerów i/lub liter przypisanych do przedmiotu lub części przez producenta.

Tabela 11. Porównanie identyfikatorów produktów

Platforma / narzędzie	Globalny unikalny ID produktu	Komentarze / wymagania	Inny zdefiniowany identyfikator
GS1 GRP	GTIN*	GTIN w 14-cyfrowym formacie, nie ma wariantów produktu o tym samym GTIN, numer jest weryfikowany pod kątem cyfry kontrolnej, prefiksu kraju, prefiksu firmy - pewność, że produkt należy do firmy.	
Produkt - Schema.org Type	GTIN, ISBN, GTIN8, GTIN12, GTIN13, GTIN14	Można jednocześnie użyć jednego lub więcej ID i wielokrotnie. Wszystkie pola to pola tekstowe o dowolnej długości. W przypadku pola GTIN, może on przyjąć nie tylko wartość numeryczną, ale także być zgodny ze specyfikacją GS1 Digital Link i wyrażony w formie URL z dodatkowymi wartościami.	SKU, MPN
Google Merchant Center	GTIN+	Wymagane (dla wszystkich nowych produktów z przypisanym przez producenta GTIN). Opcjonalne, ale zalecane dla wszystkich innych produktów. Maksymalnie 50 znaków numerycznych (maksymalnie 14 na wartość). Numer jest weryfikowany pod kątem cyfry kontrolnej i prefiksu kraju. Powtarzalne pole do 10 razy.	SKU* (id) max 50 znaków, MPN+ Max 70 znaków, Wymagane, jeśli brak GTIN
Amazon	GTIN+	Wymagane w większości przypadków jako identyfikator produktu przy tworzeniu nowych stron produktów lub dopasowywaniu do istniejących stron w katalogu. Weryfikuje GTIN, ale nie określa w jakim zakresie.	SKU* Max 40 znaków
eBay	GTIN+	Musi być wprowadzone, ale można wpisać "Nie dotyczy". Weryfikuje GTIN, ale nie określa w jakim zakresie.	MPN+ Max 65 znaków. Musi być wprowadzone, ale można wpisać "Nie dotyczy".

Platforma / narzędzie	Globalny unikalny ID produktu	Komentarze / wymagania	Inny zdefiniowany identyfikator
Allegro	GTIN+	Obowiązkowy w wybranych kategoriach.	
Ceneo	GTIN* ISBN	Obowiązkowy, ale dla ISBN dane powinny być wprowadzone w oddzielnym polu.	MPN
WooCommerce	[konfigurowalny]	Można zainstalować odpowiednią wtyczkę, która pozwala na wprowadzanie kodów GTIN.	SKU max 255 znaków
Shopify	kod kreskowy	Pole tekstowe bez kontroli jakości. Istnieją dodatki, które pozwalają na wypełnienie pola GTIN.	SKU max 16 znaków
Magento	[konfigurowalny]	Istnieją wskazówki, jak ustawić dodatkowe pole do wprowadzania GTIN, ale będzie to zwykłe pole bez kontroli jakości.	SKU* max 64 znaków

\* Pola wymagane; + Pola wymagane w części przypadków  
*Źródło: opracowanie własne*

Wyniki analizy przedstawione w Tabeli 11 wskazują, że pomimo rozpoznawalności standardu GTIN, nie zawsze jest on wymagany ani poprawnie walidowany. W konsekwencji pojawiają się wątpliwości co do wiarygodności tak wprowadzanych danych. Prawdopodobnie wynika to z faktu, iż pole GTIN nie jest kluczowe dla realizacji podstawowego procesu sprzedaży w e-commerce. Istnieje przypuszczenie, że jest to celowe działanie - jego obligatoryjne wprowadzenie mogłoby ograniczyć możliwości użytkownika narzędzia lub platformy, co nie leży w interesie ich twórców. Z drugiej strony brak lub niepewna informacja co do globalnego identyfikatora produktu właściwie przekreśla możliwości wiarygodnej synchronizacji i agregacji danych.

W przypadku struktury danych produktu w Schema.org, podano tylko jej znaczenie dla pola GTIN, bez odniesienia do zewnętrznej dokumentacji lub opisu struktury tego standardu GS1. Jako że Schema.org definiuje tylko struktury danych bez ich walidacji, daje to pełną swobodę w popełnianiu błędów. Co więcej, istnieje całkowicie zbędna różnorodność nazw pól w strukturze, biorąc pod uwagę typ numeru (GTIN8, GTIN12, GTIN13, GTIN14) i ogólne pole GTIN (które może zawierać wszystkie pozostałe). Jest to jeszcze bardziej szkodliwe dla spójności.

Inną kwestią, na którą warto zwrócić uwagę analizując wyniki badania, jest możliwość wprowadzenia wielu numerów GTIN dla pojedynczego opisanego produktu. Z punktu widzenia standardu GS1 nie jest to odpowiednie - każdemu produktowi musi być przypisany indywidualny numer. W przeciwnym razie nie będzie wiadomo, jaki produkt faktycznie zamówił klient i co należy pobrać z magazynu. Takie architektoniczne rozwiązanie

w systemach IT dowodzi, że GTIN jest tylko drugorzędnym atrybutem wzbogacającym opis produktu, a nie identyfikatorem podstawowym (kluczem głównym w bazie danych) określającym produkt.

Na koniec analizy istotne jest zwrócenie uwagi na brak standaryzacji pod względem długości pól dla SKU (Stock Keeping Unit) i MPN (Manufacturer Part Number). Ten brak jednolitości wynika z faktu, że różne systemy e-commerce rozwijały się niezależnie, adaptując własne standardy i praktyki zarządzania tymi atrybutami. Każda platforma e-commerce, działając w ramach swojego indywidualnego środowiska technologicznego i biznesowego, ustaliła własne wytyczne dotyczące formatowania i długości tych identyfikatorów.

### 3.2.2. Unikalna nazwa opisująca produkt

„Nazwa produktu” jest niewątpliwie, obok identyfikatora, jednym z głównych atrybutów każdej bazy danych produktów. Umożliwia szybką identyfikację rekordu w bazie danych przez operatora (człowieka) w procesie zarządzania, ponieważ w kilku słowach opisuje, czym faktycznie jest produkt. Istnieją różne definicje nazwy produktu, które wpływają na sposób wypełniania tego pola w bazie danych. W modelach danych GDD występuje wiele atrybutów definiujących nazwę produktu, które powinny być wypełniane w zależności od potrzeb i wymagań partnerów biznesowych (GS1, 2022a). Atrybuty i opis przedstawiono w Tabeli 12.

Tabela 12. Porównanie różnych atrybutów "nazw produktów" w GDD

Nazwa Atrybutu	Długość	Opis
labelDescription	500	Dosłowne odtworzenie tekstu widniejącego na etykiecie produktu, w takiej samej kolejności słów, jak na froncie opakowania produktu. Może to nie być zgodne z opisem GTIN załadowanym przez dostawcę do pola opisu GTIN w GDSN.
functionalName	35	Opisuje sposób użycia produktu lub usługi przez konsumenta. Powinien pomóc wyjaśnić klasyfikację produktu związaną z GTIN.
regulatedProductName	500	Przepisana, regulowana lub ogólna nazwa produktu, która opisuje rzeczywistą naturę produktu i jest na tyle precyzyjna, aby odróżnić go od innych produktów, zgodnie z obowiązującymi przepisami krajowymi.
tradeItemDescription	200	Opis handlowy produktu powinien być jasny i precyzyjny, zawierać nazwę marki oraz inne istotne deskryptory. Atrybut ten należy wypełniać z minimalnym użyciem skrótów, zachowując odpowiednią długość. Opis powinien być znaczący i pełny, aby ułatwić przetwarzanie informacji. Detałści mogą wykorzystywać ten opis jako podstawę do pełnego zrozumienia takich cech jak marka, smak, zapach itp. dla konkretnego GTIN, co umożliwi im opracowanie opisu produktu odpowiedniego do ich wewnętrznych systemów.
descriptionShort	35	Krótki, swobodny opis handlowego przedmiotu, który może być używany do identyfikacji przedmiotu handlowego w punkcie sprzedaży.

Źródło: opracowanie własne

W handlu elektronicznym nazwa produktu ma znacznie większe znaczenie niż w tradycyjnym handlu. W rzeczywistym świecie klient może zostać zachęcony do zakupu produktu ze względu na jego atrakcyjny wygląd lub wygląd opakowania. W świecie wirtualnym nie ma fizycznego produktu, ale jego odbicie w postaci danych cyfrowych: nazwy, zdjęcia i opisu. Dlatego to właśnie dobra nazwa produktu może zadecydować o jego wyszukiwaniu i zakupie przez klienta. Nazwa powinna być jednoznaczna, opisowa i łatwa do zrozumienia, aby efektywnie przyciągnąć uwagę i wzbudzić zainteresowanie. W konkurencyjnym świecie online, gdzie wybór produktów jest ogromny, odpowiednio skonstruowana nazwa może zapewnić istotną przewagę rynkową. W Tabeli 13 przedstawiono porównanie atrybutów związanych z nazwą produktów w testowanych rozwiązaniach e-commerce w kontekście sposobu definiowania, ograniczeń i zaleceń przy wprowadzaniu danych.

Tabela 13. Porównanie nazw produktów

Platforma / narzędzie	Maksymalna długość pola	Nazwa pola i jej definicja
GS1 GRP	70 (zalecana)	Nazwa opisowa produktu. Zrozumiały, unikalny i użyteczny opis produktu, wykorzystujący kombinację kluczowych elementów, takich jak nazwa marki, podmarka (jeśli dotyczy), nazwa funkcjonalna, wariant i zawartość netto. Opis można wprowadzać w wielu językach.
Product - Schema.org Type	Nieokreślona	„Nazwa produktu”. Nazwa przedmiotu. Dokumentacja nie określa tej wartości.
Google Merchant Center	150	„Tytuł”. Dokładny opis produktu, zgodny z tytułem na stronie docelowej. Tytuły nie powinny zawierać tekstów promocyjnych, takich jak „darmowa wysyłka”, dużych liter ani nietypowych znaków. Tytuły powinny zawierać wyróżniającą cechę, taką jak kolor lub rozmiar dla różnych wariantów.
Amazon	200 (80 zalecane)	„Tytuł”. Każde słowo w tytule powinno zaczynać się dużą literą, z wyjątkiem spójników (and, or, for), przedimków (the, a, an) lub przyimków krótszych niż pięć liter (in, on, over, with). Dla zestawów produktów tytuły powinny zawierać liczbę przedmiotów. Tytuły nie powinny zawierać specjalnych znaków lub symboli, takich jak ©, dużych liter, informacji „o sobie” lub „o swojej firmie”, ceny i ilości, wiadomości promocyjnych, takich jak „sprzedaż” lub „darmowa wysyłka”, subiektywnych komentarzy, takich jak „Top Produkt” czy „Bestseller”.
eBay	80	„Tytuł”. Tytuły powinny zawierać tylko taką nazwę marki, której producent (lub marka) używa do identyfikacji konkretnego produktu. Nie powinny zawierać specjalnych, obcych ani nietypowych znaków, chyba że symbol jest częścią faktycznych danych o produkcie. Tytuły nie powinny zawierać przymiotników w stopniu najwyższym, tekstu promocyjnego, warunków użytkowania, cen, opcji wysyłki ani nieistotnych słów kluczowych, specjalnego formatowania, numerów telefonów, skrótów łączonych myślnikiem, nadmiernych białych spacji, skrótów stosowanych do kompresji tekstu, nie powinny zawierać słów obcych, nie powinny zawierać liczby mnogiej. Tytuły nie mogą zawierać powtarzających się tokenów, wielu synonimów ani informacji o dopasowaniu, nie mogą zawierać tagów HTML.

Platforma / narzędzie	Maksymalna długość pola	Nazwa pola i jej definicja
Allegro	50	„Nazwa przedmiotu”. Niedozwolone praktyki: powtarzanie słów kluczowych, używanie modnych fraz, nieadekwatnych do prezentowanego produktu, słów takich jak „polecam”, „musisz zobaczyć!”, „nowy”, „promocja”, „hit” itp., zawieranie dodatkowych informacji w tytule, na przykład swojego loginu, miasta, w którym można odebrać towar osobiście, informacji o fakturze lub szybkiej wysyłce, umieszczanie numerów magazynowych lub własnych oznaczeń w tytule, umieszczanie specjalnych znaków w tytule (np.: @,!, []).
Ceneo	150	„Nazwa”. Brak dokumentacji dotyczących wytycznych, najlepszych praktyk lub informacji o walidacji pola.
WooCommerce	Nieokreślona	„Nazwa”. Brak dokumentacji dotyczących wytycznych, najlepszych praktyk lub informacji o walidacji pola.
Shopify	150	„Tytuł”. Brak dokumentacji dotyczących wytycznych, najlepszych praktyk lub informacji o walidacji pola.
Magento	Nieokreślona	„Nazwa produktu”. Brak dokumentacji dotyczących wytycznych, najlepszych praktyk lub informacji o walidacji pola.

Źródło: opracowanie własne

W badaniu uwzględniono pola bezpośrednio odnoszące się do nazwy produktu, a także te, które pełnią taką funkcję. W niektórych wierszach tabeli pojawia się termin „tytuł”. Jest to atrybut wykorzystywany do opisu oferty aukcji jak i produktu. Ewolucja serwisów aukcyjnych w kierunku rynków typu marketplace oraz proces agregacji ofert wokół precyzyjnych danych produktowych, spowodowała zmianę pierwotnego znaczenia „tytułu oferty” na „tytuł produktu”. Przykładowo platforma Amazon implementuje rozwiązanie, w którym tytuł oferty może zostać zastąpiony przez nazwę produktu zapisaną w centralnym katalogu produktów.

W ramach analizy porównawczej atrybutów produktowych dotyczących nazwy produktu na różnych platformach e-commerce, istotnym aspektem, na który należy zwrócić uwagę, jest maksymalna długość pola. Zaobserwowano znaczącą różnicę, wynoszącą 400%, pomiędzy platformą Allegro, gdzie maksymalna długość wynosi 50 znaków, a platformą Amazon, gdzie limit ten sięga 200 znaków. Dla niektórych platform wartość ta w ogóle nie została określona. Wiele platform kładzie duży nacisk na formatowanie tekstu. Wymagania takie jak odpowiednie stosowanie wielkich liter, unikanie podwójnych spacji i znaków specjalnych, a także zakaz stosowania technik „spamowania słowami kluczowymi” (czyli używania popularnych, lecz niezwiązanych bezpośrednio z produktem fraz) są powszechne. Wyjątkiem jest wymóg platformy Amazon, gdzie „każde słowo w tytule powinno zaczynać się wielką literą”, co nie występuje na innych platformach. Mało uwagi poświęca się strukturze nazwy produktu, czyli doborowi i kolejności słów. Warto podkreślić najlepsze praktyki Google



Merchant Center (nie wszystkie uwzględnione w tabeli), które dostarczają przykładów poprawnie opisanych nazw produktów oraz dodatkowych wytycznych dotyczących zgodności z wymogami prawnymi w różnych krajach. Znaczenie tego pola i jego składowych jest także dobrze opisane w usłudze GS1 GRP.

### 3.2.3. Nazwa marki

Atrybut „nazwa marki” to kolejne, ważne pole z punktu widzenia między innymi identyfikacji, grupowania i wyróżnienia produktu na tle innych. Chociaż jego wypełnienie zazwyczaj nie jest obowiązkowe, jest powszechnie stosowane w e-commerce, co potwierdza zestawienie umieszczone w Tabeli 14. Warto również zwrócić uwagę na fakt, że połączenie nazwy marki z polem MPN (Manufacturer Part Number – numer części producenta) redukuje ryzyko błędów przy ustalaniu unikalności produktu, co zostało wykorzystane na platformie eBay. Niemniej jednak, nie gwarantuje to takiej samej pewności jak użycie numeru GTIN, ponieważ żadna instytucja nie nadzoruje ani nie standaryzuje numeru MPN. Dodatkowo, nie istnieje obowiązek rejestracji nazwy marki w urzędach patentowych na całym świecie, co oznacza, że nazwy mogą się powtarzać, a także występuje ryzyko błędów literowych podczas wprowadzania danych.

Tabela 14. Porównanie interpretacji nazw marek

Platforma i narzędzia	Typ pola	Maks. długość	Definicja
GS1 GRP	Pole tekstowe	70	Nazwa nadawana przez właściciela marki, rozpoznawalna dla konsumenta. Markę można wprowadzać w wielu językach.
Produkt - Schema.org Type	Wiele obiektów		Marka (lub marki) powiązana z produktem lub usługą, utrzymywana przez organizację lub osobę prowadzącą działalność. Marka może posiadać nie tylko nazwę, ale także logo, URL, identyfikator oraz motto.
Google Merchant Center	Pole tekstowe+	70	Wymagane dla wszystkich nowych produktów, z wyjątkiem marek filmów, książek i nagrań muzycznych. Pole powinno zawierać marki rozpoznawalne przez klientów, stworzone przez producentów. W przypadku braku marki, powinna to być nazwa producenta lub dostawcy. Nie akceptuje wartości typu „brak”, „N/A”, „Generic”.
Amazon	Pole tekstowe*	50	Charakterystyczny i rozpoznawalny symbol, skojarzenie, nazwa lub znak towarowy, służący do odróżnienia konkurencyjnych produktów lub usług. Może odnosić się do pojedynczego produktu, całej linii produktów lub firmy. Amazon musi zaakceptować nowo wprowadzoną markę, zanim będzie mogła być używana do listowania produktów.

Platforma i narzędzia	Typ pola	Maks. długość	Definicja
eBay	Pole tekstowe+	65	Wyłączna i wyróżniająca się nazwa lub symbol używany do identyfikacji towarów lub usług sprzedawcy. Nazwy marek mogą być znakami towarowymi i odnosić się do pojedynczego produktu, linii produktów lub całej firmy. Nazwa marki powinna dokładnie odpowiadać pisowni używanej przez producenta w odpowiednim języku, z uwzględnieniem wielkości liter w znakach towarowych. Nie należy używać symboli (®, ©, ™) ani skrótów („GmbH”, „Ltd.”), które nie są częścią nazwy marki. Nie należy używać nazwy producenta, lecz nazwę marki, pod którą produkt został określony. Pole wymagane razem z MPN jeśli nie ma podanego GTIN.
Allegro	Pole słownikowe+		Pole wymagane w zależności od kategorii (obowiązkowe w większości kategorii). Pole słownikowe, z możliwością wyboru opcji "inne" i wprowadzenia własnej wartości.
Ceneo	Pole słownikowe		Nie wymagane. Pole słownikowe, wartości zależą od wybranej kategorii.
WooCommerce			Pole może być swobodnie tworzone, ale domyślnie nie jest zawarte.
Shopify			Domyślnie nie zawarte. Pole może być dodane jako dodatkowe metadane. Często używane jest istniejące pole „sprzedawca”.
Magento			Pole może być swobodnie tworzone, ale domyślnie nie jest zawarte.

*Źródło: opracowanie własne*

Analiza danych przedstawionych w Tabeli 14 ukazuje brak spójności w maksymalnej dozwolonej długości pola „nazwa marki”, choć różnice te są niewielkie (50–70 znaków). Twórcy poszczególnych rozwiązań zgadzają się co do ogólnej zawartości tego pola. Z wyjątkiem specyfiki rozbudowanego podejścia Schema.org w schemacie danych strukturalnych „Product”, autorzy pozostałych rozwiązań sugerują, że nazwa marki powinna pochodzić od producenta, być rozpoznawalna na rynku i odnosić się do produktu, linii produktów lub całej firmy. Zauważono jednak, że w przypadku braku marki, Google Merchant Center rekomenduje wprowadzenie nazwy firmy lub dostawcy, co umożliwi uniknięcie pominięcia danych w tym polu, ale może prowadzić do niespójności. Dla porównania eBay wyraźnie zabrania użycia nazwy producenta zamiast marki. Wiele platform, takich jak Ceneo, Amazon i Allegro, zdecydowało się na zablokowanie możliwości wprowadzania dowolnych wartości tekstowych do pola „nazwa marki” i zastąpienie ich słownikiem marek, z dodatkową kontrolą na poziomie słownika (Amazon). Choć takie podejście sprzyja spójności danych w obrębie platform, ogranicza możliwość wymiany danych między różnymi systemami bazodanowymi.

Z analizy wynika, że każda z platform e-commerce przyjmuje własne, w dodatku odmienne podejście do kwestii definiowania i zarządzania markami produktów.

### 3.2.4. Zdjęcie produktu

Zdjęcie produktu w kontekście e-commerce stanowi kluczowy element w procesie prezentacji i identyfikacji produktów. Badaniem objęto analizę wymagań dotyczących zdjęć produktów na różnych platformach e-commerce, rozpatrując parametry techniczne, takie jak rozmiar zdjęcia, format, rozmiar pliku oraz parametry związane z treścią, w tym kolory tła, zawartość zabroniona, a także procentowe pokrycie kadru.

W Tabeli 15 przedstawiono wyniki analiz, uwzględniając różnice w wymaganiach dotyczących wymiarów zdjęcia produktu, dopuszczalnej liczby obrazów, formatu zapisu i wielkości pliku.

Tabela 15. Porównanie technicznych aspektów atrybutu "Zdjęcie produktu"

Platforma i narzędzia	Min ilość	Maks ilość	Min rozdzielczość	Maks rozdzielczość i rozmiar pliku	Format pliku
GS1 GRP	0	Wiele	900px × 900px	4800px × 4800px	JPG, PNG, GIF, TIFF
Produkt - Schema.org Type	0	Wiele			
Google Merchant Center	1	1+10 (dodatkowe)	100px × 100px, dla odzieży 250px x 250px	64 megapiksele, 16MB	GIF, JPG, PNG, BMP, TIFF
Amazon	1	9, wyświetlane tylko 7	500px, min zalecane 600px na dłuższym boku	10,000px na dłuższym boku	GIF, JPG, PNG, TIFF
eBay	1	12, dla branży motoryzacyjnej 24	1000px × 1000 px tylko ostrzeżenie, jeśli mniejsze	15000px × 15000px, 12MB	GIF, JPG, PNG, BMP, TIFF
Allegro	1	15	500px na dłuższym boku	2560px × 2560px, 2MB	JPG, PNG, BMP
Ceneo	1	Wiele			
WooCommerce	0	Wiele	min zalecane 800px × 800px		JPG, PNG
Shopify	0	250	1px × 1px	2048px × 2048px	JPG, PNG
Magento	0	Wiele		1200px × 1200px	JPG, PNG

Źródło: opracowanie własne

Zróznicowane wymagania dotyczące zdjęć produktów na różnych platformach e-commerce odzwierciedlają odmienne podejścia do prezentacji wizualnej produktów.

Minimalna liczba wymaganych zdjęć waha się od 0 do 1, natomiast maksymalna ilość, dla podanych wielkości: od 1 (Google Merchant Center) do 250 (Shopify) lub ograniczeń systemowych („wiele”). Minimalna rozdzielczość wynosi od 1px × 1px (Shopify) do 1000px × 1000px (eBay), a maksymalna rozdzielczość od 2048px × 2048px (Shopify) do 15000px × 15000px (eBay). Dozwolone formaty plików obejmują najczęściej JPG i PNG, ale niektóre platformy akceptują także GIF i co bardzo zaskakuje - BMP, oraz TIFF, które nie są powszechne w Internecie. Te różnice, szczególnie w odniesieniu do rozmiarów i formatów zdjęć, mogą stanowić wyzwanie dla sprzedawców, którzy starają się zoptymalizować swoje oferty pod kątem wielu rynków.

Dalsza analiza miała na celu ustalenie, czy istnieje jednolitość w wymogach dotyczących formy i zawartości obrazu produktu na różnych platformach e-commerce. Skoncentrowano się na platformach, które precyzyjnie określają takie wymagania. Wyniki w Tabeli 16 wskazują na różnorodność tych wymagań, obejmujących takie aspekty jak kolor tła, zakazana zawartość, stopień wypełnienia kadru oraz możliwość prezentacji produktu z różnych perspektyw na jednym obrazie. Analiza podkreśla różnice w podejściach platform do wizualnej prezentacji produktów, co ma kluczowe znaczenie dla ich spójnej i adekwatnej reprezentacji.

Tabela 16. Porównanie jakościowych aspektów atrybutu "Zdjęcie Produktu"

Platforma i narzędzia	Tło	Zabronione	Pokrycie	Wielokrotne widoki jednego produktu
GS1 GRP	Białe	Podpisy, znaki wodne	Powinno być 95%	Niedozwolone
	Należy użyć wysokiej rozdzielczości zdjęcia produktu, które wyraźnie przedstawia główną powierzchnię sprzedaży produktu. Zdjęcie powinno umożliwiać autentyczną identyfikację przedmiotu. Przyjmuje się, że główna powierzchnia sprzedaży jest równoznaczna z funkcjonalnym przodem produktu lub domyślnym przodem przedmiotu, jak określono w specyfikacjach GS1 (GS1, 2022b, 2022c).			
Google Merchant Center	Białe lub przezroczyste tło	Tekst promocyjny, znaki wodne lub ramki	75% - 90%	Niedozwolone
	Zdjęcie powinno przedstawiać pojedynczą jednostkę produktu. W przypadku użycia tzw. multipack, główne zdjęcie powinno przedstawiać pojedynczą jednostkę.			
Amazon	Białe	Tekst, logo, ramki, bloki kolorów, znaki wodne lub inne grafiki na wierzchu produktu lub w tle	Minimum 85%	Niedozwolone
	Główne zdjęcia powinny pokazywać produkty poza ich opakowaniem. Pudełka, torby lub etui nie powinny pojawiać się na zdjęciu, chyba że są ważną cechą produktu. Istnieją specjalne wymagania dotyczące wyświetlania odzieży i butów.			

Platforma i narzędzia	Tło	Zabronione	Pokrycie	Wielokrotne widoki jednego produktu
eBay	Między białym a jasnoszarym (akceptowalny jest również lekki cień, ale odbicia lustrzane są niedozwolone)	Znaki copyrightu, znaki wodne, odbicia lub gorące punkty, jakikolwiek tekst, który nie jest częścią oryginalnego produktu lub opakowania	80%-90%	Dozwolone, ale dwa obiekty na zdjęciu muszą pokazywać różne kąty tego samego produktu.
Główne zdjęcie powinno być widokiem z przodu produktu, prosto lub pod lekkim kątem. W ofertach dla multipacków główne zdjęcie musi wyraźnie pokazywać główny produkt. Zdjęcia z częściami ciała ludzkiego są akceptowalne, ale tylko na zdjęciach, które pokazują produkty do noszenia na ciele.				
Allegro	Białe	Logo		Dozwolone
Można pokazać zestaw produktów, jeśli są sprzedawane razem. Przedmiot może być wyświetlany obok oryginalnego opakowania producenta. Można pokazać produkt w kontekście użytkowania lub w aranżacji w wybranych kategoriach.				
Ceneo	Białe		Powinno wypełnić całą przestrzeń zdjęcia i być na jego środku	

Źródło: opracowanie własne

Większość platform e-commerce wymaga, aby tło zdjęcia produktu było białe. Jednakże eBay dopuszcza również jasnoszare tło, a na platformie GS1 GRP akceptowane jest tło przezroczyste. Zgodność między platformami widoczna jest w kwestii zakazanych elementów – powszechnie zabrania się dodatkowych logotypów, tekstów promocyjnych oraz znaków wodnych, które mogłyby zakłócić autentyczny wizerunek produktu.

Różnice między platformami e-commerce stają się wyraźne w odniesieniu do tego, co stanowi główny przedmiot zdjęcia, szczególnie w przypadku zdjęć reprezentatywnych. Na platformach takich jak eBay czy Allegro dopuszcza się prezentację produktu z różnych perspektyw na jednym obrazie, podczas gdy Amazon tego zabrania. Allegro umożliwia także prezentację produktu wraz z opakowaniem, co jest niedopuszczalne na GS1 GRP i akceptowane warunkowo na Amazonie. Google Merchant Center wymaga, aby zdjęcie produktów typu multipack przedstawiało pojedynczy produkt, podczas gdy eBay pozwala na obecność dodatkowych elementów w tle. Allegro umożliwia prezentację produktu w kontekście jego użytkowania lub aranżacji w wybranych kategoriach, Amazon wprowadza szczególne wymagania dotyczące odzieży i obuwia, a eBay stosuje restrykcyjne zasady

w zakresie wyświetlania ciała ludzkiego w wielu kategoriach. Dodatkowym wyzwaniem jest określenie odpowiedniego pokrycia kadru produktem, co jest kluczowe dla zgodności z wymogami platform. Optymalne pokrycie, zależnie od platformy, wynosi od 75% do 95% kadru, co podkreśla brak jednolitego standardu i konieczność dostosowywania zdjęć produktów do specyficznych wymagań.

Podsumowując, analiza wskazuje na znaczną różnorodność wymagań dotyczących zdjęć produktów w e-commerce, co stanowi wyzwanie dla sprzedawców starających się zoptymalizować prezentację swoich produktów na różnych platformach.

### **3.2.5. Zawartość netto**

W ramach badania zarządzania danymi produktowymi w różnych systemach e-commerce szczególną uwagę poświęcono analizie pola "zawartość netto". Zgodnie z definicją GRP GS1, odnosi się ono do ilości produktu zawartego w opakowaniu, wyrażonej w jednostkach miary (UoM), zazwyczaj wskazanych na etykiecie produktu dla danego rynku sprzedaży. Pole to jest opcjonalne i powtarzalne dla różnych jednostek miary, a lista kodów UoM opiera się na rekomendacji Europejskiej Komisji Gospodarczej (ang. United Nations Economic Commission for Europe – UNECE).

Analiza wykazała, że żadna z badanych platform e-commerce nie wdrożyła bezpośrednio pola "zawartość netto" jako standardowego elementu danych produktowych. Zamiast tego, platformy takie jak Amazon, Allegro, Ceneo, Shopify oraz schema.org w zakresie atrybutów produktu opcjonalnie uwzględniają wagę brutto. Wskazuje to, że głównym celem wdrożenia tego atrybutu jest optymalizacja procesu wysyłki, a nie dostarczanie szczegółowych informacji o produkcie. Ponadto, w niektórych systemach e-commerce wymagane jest podanie dodatkowych informacji dotyczących opakowania. Różnice w podejściu do implementacji pola wagi brutto wyrażają się również w zakresie stosowanych jednostek miary, co świadczy o braku jednolitego standardu w tym obszarze. W Tabeli 17 zaprezentowano zestawienie różnych podejść.

Tabela 17. Porównanie atrybutu zawartości netto produktu

Platforma i narzędzia	Zawartość	Jednostka miary (UoM)
GS1 GRP	Netto	UN/CEFACT Common Code
Produkt - Schema.org Type	Brutto	UN/CEFACT Common Code
Google Merchant Center	–	–
Amazon	Brutto	[gr], [kg], [lb], [mg], [oz]
eBay	–	–
Allegro	Brutto	[kg]
Ceneo	Brutto	[kg]
WooCommerce	–	–
Shopify	Brutto	[g], [kg], [oz], [lb]
Magento	–	–

Źródło: opracowanie własne

### 3.2.6. Klasyfikacja produktu

Analiza dotycząca klasyfikacji produktów w różnych systemach e-commerce ujawnia różnorodność podejść do tego aspektu zarządzania danymi. Zwraca uwagę zastosowanie różnych taksonomii oraz możliwość definiowania własnych klasyfikacji przez użytkowników. Różnice w podejściach są znaczące – od szczegółowych wymagań, takich jak Globalna Klasyfikacja Produktów (GPC) stosowana w GS1 GRP, przez bardziej elastyczne podejścia, jak w Google Merchant Center, po platformy takie jak WooCommerce, Shopify i Magento, które umożliwiają użytkownikom pełną swobodę w definiowaniu kategorii produktów.

Te różnorodne podejścia wskazują, że systemy e-commerce często nie opierają się na ujednoliconych globalnych standardach. Kategorie produktów są zazwyczaj tworzone samodzielnie przez użytkowników lub dostosowywane do specyficznych wymagań poszczególnych platform. Google Merchant Center wyróżnia się unikalnym podejściem, gdzie kategorie produktów są automatycznie definiowane przez sztuczną inteligencję na podstawie kategorii stworzonych przez użytkowników.

Po przypisaniu kodu klasyfikacji, niektóre platformy mogą wymagać dodatkowych informacji w celu sprecyzowania oferty lub spełnienia wymogów prawnych, takich jak ograniczenia dotyczące sprzedaży alkoholu, żywności czy odzieży.

Tabela 18 prezentuje porównanie klasyfikatorów produktów, uwzględniając taksonomie i możliwość definiowania własnych kategorii przez użytkowników na różnych platformach.

Tabela 18. Porównanie klasyfikatorów produktu

Platforma i narzędzia	Taksonomia	Opis klasyfikacji zdefiniowanej przez użytkownika
GS1 GRP	Globalna Klasyfikacja Produktów (GPC)*	
Produkt- Schema.org Type		Dowolna kategoria dla przedmiotu jako łańcuch tekstowy, URL lub inny obiekt. W przypadku wyboru tekstu, znaki większości lub ukośniki mogą być używane nieformalnie do wskazania hierarchii kategorii
Google Merchant Center	Kategoria produktu Google	Typ produktu z pełną kategorią zdefiniowaną przez użytkownika. Na przykład, zawierać „Dom/Kobiety/Sukienki/Maksy” zamiast tylko „Sukienki”
Amazon	Kategoria produktu Amazon*	
eBay	Kategoria eBay*,	
Kategoria eBay (2)	Kategoria Sklepu, Kategoria Sklepu 2 – zdefiniowana własna	
Allegro	Klasyfikacja Allegro*	
Ceneo	Kategoria Ceneo*	
WooCommerce		Własna zdefiniowana kategoria
Shopify		Własny zdefiniowany typ produktu
Magento		Własna zdefiniowana kategoria

\* Pole wymagane

Źródło: opracowanie własne

### 3.2.7. Pozostałe atrybuty

Opisane i porównane atrybuty odnoszą się bezpośrednio do produktu, a nie do oferty czy danych logistycznych. Z tego powodu nie porównywano pól takich jak wymiary opakowania, szczegółowy opis oferty czy cena oferty (rozumiana jako cena sugerowana przez producenta). Należy jednak zauważyć, że te pola, jak również specyficzne pola w zależności od typu produktu, są wymagane lub przynajmniej pożądane do wypełnienia przez wybrane platformy. Warto zwrócić także uwagę, że żadne z analizowanych rozwiązań e-commerce nie posiadało wyraźnie zdefiniowanych pól dotyczących "rynku docelowego" wspomnianego w Rozdziale 2.6 na poziomie zarządzania produktem. Wynika to z faktu, że sklepy internetowe zwykle definiują te aspekty na poziomie logistyki wysyłkowej, a język platformy e-commerce jest zazwyczaj dostosowany do języka użytkownika.



### 3.3. Wnioski

Przeprowadzone badania w ramach niniejszego rozdziału ujawniły znaczące zróżnicowanie podejść do kluczowych atrybutów produktowych w różnych systemach e-commerce. Standaryzacja tych atrybutów na poziomie globalnym jest nie tylko pożądana, ale wręcz niezbędna dla zapewnienia spójności i poprawy efektywności procesów handlu elektronicznego w przyszłości. Obecnie dane produktowe nie są ani poprawnie walidowane, ani spójne na szeroką skalę. Dodatkowo, brak powszechnego stosowania identyfikatora GTIN pogłębia ten problem, uniemożliwiając skuteczną synchronizację danych między platformami.

Dostosowując się do zestawionych wymagań przedstawionych w badaniach, można wywnioskować, że walidacja danych powinna obejmować nie tylko kontrolę poprawności technicznej (np. długość identyfikatorów czy format zdjęć), ale także poprawność semantyczną, umożliwiającą weryfikację tego, co powinno lub nie powinno zawierać dane pole tekstowe czy zdjęcie. W tym kontekście zastosowanie sztucznej inteligencji, zwłaszcza technik przetwarzania i rozumienia języka naturalnego (NLP, NLU) oraz analizy obrazów, może znacząco usprawnić proces walidacji. Algorytmy uczenia maszynowego mogą automatycznie wykrywać niespójności i sugerować poprawki, co przyczyni się do zwiększenia jakości danych.

Celowe wydaje się także opracowanie mechanizmu walidacji, który nie tylko będzie weryfikował i odrzucał dane, ale również informował o potencjalnych problemach związanych z wymogami poszczególnych platform. Podejście hybrydowe, łączące celowane reguły walidacyjne z zaawansowanymi algorytmami sztucznej inteligencji, pozwoli na skuteczniejsze wykrywanie i korektę błędów w danych produktowych. Takie połączenie zwiększy jakość i spójność informacji w systemach e-commerce, przyczyniając się do poprawy efektywności handlu elektronicznego na globalną skalę.

## **4. Analiza problemów jakościowych w danych produktowych dostępnych online – badania własne**

### **4.1. Cel badań**

W niniejszym rozdziale przeprowadzono szczegółową analizę danych produktowych, korzystając ze źródeł pochodzących z różnych baz danych, sieci, społeczności oraz platform cyfrowych dostępnych on-line. Cały proces analizy podzielono na kilka etapów, których głównym celem było dokładne zidentyfikowanie i głębokie zrozumienie wyzwań związanych z walidacją danych produktowych w świecie on-line. Ideą było wyjście poza ograniczenia pojedynczej bazy danych - pierwotnego obiektu badań, aby uzyskać bardziej rozległy ogólny obraz problematyki jakości i standaryzacji danych, a także aby zweryfikować, czy są to problemy o ograniczonym zasięgu, czy też na skalę globalną. Skoncentrowano się przy tym głównie na jakości i powtarzalności informacji zawartej w nazwie produktu w różnych bazach danych, a także poprawności stosowania numeru GTIN, ponieważ podmiot badań posiadał już zaawansowane walidatory w tym zakresie.

W pierwszym etapie przygotowano zbiór danych połączony ze źródeł, które nie dokonywały walidacji numerów GTIN lub realizowały ją w ograniczonym zakresie. Celem tego działania było zdiagnozowanie rodzajów błędów pojawiających się w tym kontekście. Analiza ta miała kluczowe znaczenie dla zrozumienia zakresu i natury problemów związanych z nieprawidłowym stosowaniem numerów GTIN, co jest istotne w kontekście tworzenia efektywnych metod walidacji danych produktowych zagregowanych wokół tego numeru. Przeanalizowano różnorodne przypadki, w tym niekompletne numery, błędy wynikające z niewłaściwego formatowania, a także przypadki całkowicie błędnych ciągów znaków nieodpowiadających żadnym zarejestrowanym produktom.

W drugiej części badania skupiono się na analizie danych produktowych pochodzących ze sklepów internetowych. Z uwagi na fakt, że właściciele e-sklepów często samodzielnie pozyskują dane o produktach, przeanalizowano, w jaki sposób interpretacja nazw produktów odzwierciedla rzeczywiste dane. Po zebraniu danych produktowych z kilku sklepów internetowych przy użyciu techniki web scraping, dokonano ich porównania. Następnie wybrano jeden z najczęściej oferowanych produktów, który pojawiał się w różnych sklepach, jako przedmiot szczegółowej analizy.

Trzeci etap obejmował badanie danych pochodzących z platform typu marketplace, będących miejscem współpracy producentów, dystrybutorów i sprzedawców. Tego rodzaju platformy stanowią często miejsce, w którym problematyka nazewnictwa produktów i ich agregacji jest szczególnie widoczna. Do badań wybrano największą polską platformę, a produkt do analizy wybrano na podstawie dostępnego zestawienia danych zagregowanych według numeru GTIN.

W czwartej części przeprowadzono badanie, polegające na porównaniu danych produktowych z trzech różnych katalogów produktów. Kryterium porównania stanowiły nazwy produktów odpowiadające tym samym numerom GTIN w każdej z baz danych. Wyselekcjonowane katalogi produktowe, tworzone we współpracy z producentami, posłużyły jako podstawa analizy, której celem było wykazanie, że problem jakości danych jest zjawiskiem powszechnym i występuje już na etapie samego źródła danych, czyli u producenta.

Ostatni scenariusz dotyczył głębszego problemu związanego z interpretacją tekstu. Zbadano, w jaki sposób nazwa produktu może wpływać na błędne wyobrażenia o nim. Wybór produktu do analizy został dokonany w sposób eksperymentalny, na podstawie osobistych obserwacji autora oraz korzystając z wyszukiwarki Google.

Wnioski płynące z przeprowadzonych badań mają fundamentalne znaczenie dla rozumienia problemu jakości danych produktowych i mogą stanowić bazę do opracowania skutecznych metod ich walidacji. Fragmenty wyników badań opublikowano w publikacji naukowej pt.: „Identifying the cognitive gap in the causes of product name ambiguity in e-commerce” (Niemir i Mrugalska, 2022a).

## **4.2. Poprawność numerów GTIN**

Podczas analizy zbiorów danych produktowych, składających się z kilkudziesięciu plików pochodzących z zewnętrznych źródeł, przede wszystkim arkuszy kalkulacyjnych klientów oraz baz danych dostępnych przez Internet, zidentyfikowano szereg przypadków niezgodności i błędów w zapisie numerów GTIN. Zidentyfikowano przypadki niezgodności numerów GTIN z ich standardową formą zapisu (od 8 do 14 cyfr) - w miejscu oczekiwanych poprawnych danych napotymano tekstowe wpisy, ciągi kilku numerów GTIN oddzielonych przecinkami lub numery o niestandardowej długości (np. 5 cyfr), co sugeruje, że mogły to być numery SKU (Stock Keeping Unit) używane do identyfikacji produktów. Zjawisko to występowało jednak stosunkowo rzadko. Podobnie, o ile był wprowadzony do pola poprawny

ciąg cyfr 8-14, powstały numer miał nieprawidłową cyfrę kontrolną. Wykryto także problemy wynikające z niezrozumienia standardu:

- A. Dodawanie cyfry kontrolnej w celu dopełnienia numeru 12-sto cyfrowego UPC (z cyfrą kontrolną) do postaci GTIN-13.
- B. Błąd importy numeru GTIN w formacie 14 cyfrowym z zerem na początku przez system przyjmujący dane 13-sto cyfrowe, gdzie bez ostrzeżenia końcowa cyfra była ucinana, a przedostatnia nadpisywana obliczoną cyfrą kontrolną.

W omawianych bazach częściej napotymano numery, które wydawały się zgodne ze standardem, a ich cyfry kontrolne, po weryfikacji, nie wzbudzały zastrzeżeń, lecz mimo to były niepoprawne z powodu niezgodności w ich wewnętrznej strukturze – zawierały nieprawidłowy prefiks kraju i/lub nieistniejący prefiks firmy. Najczęściej takie numery zawierały prefiks kraju znajdujący się w przestrzeni numerów służących do zastosowań wewnętrznych (np. sieci marketów), które niezgodnie z przeznaczeniem były używane poza nimi. Takimi numerami są np. numer GTIN-13 rozpoczynający się od cyfr 20. Zdarzały się także przypadki użycia numerów z przestrzeni zarezerwowanej do przyszłego użycia, np. z początkiem „591”. Przyczyny takiej sytuacji nie zostały jednoznacznie ustalone. Mogły to być błędy automatycznego generowania numerów przez systemy informatyczne lub celowe wprowadzanie do obiegu numerów niewpisanych do oficjalnych rejestrów w celu uniknięcia opłat. O ile prefiks kraju łatwo można zweryfikować na stronie Wikipedii lub stronie internetowej GS1 (<https://gs1pl.org/prefiks-gs1-firmy-a-prefiksy-gs1/>), o tyle numer przynależący do firmy nie jest publicznie dostępny w formie listy. Istnieje możliwość jedynie odpytania o konkretny numer GTIN w celu uzyskania informacji o jego rezerwacji.

W 2017 roku, podczas budowy agregatora danych produktowych, przeprowadzono szczegółowe badanie weryfikujące poprawność prefiksów firm na zweryfikowanych numerach GTIN dla 50 zewnętrznych źródeł danych pochodzących od brokerów informacji, sklepów stacjonarnych i internetowych. Pozyskane lub pobrane techniką web scrapingu ze stron internetowych dane, zostały umieszczone we wspólnej bazie agregatora, który łączył produkty po numerach GTIN. Badanie wykazało, że w przypadku ponad 3,5 miliona zebranych danych dotyczących produktów, 3% numerów było nielegalnych (nie istniały podane prefiksy firmowe), a 2% nie powinno być dostępnych w sprzedaży.

Podsumowując, badanie wykazało następujące problemy z numerami GTIN:

- a) niepoprawne znaki zamiast cyfr
- b) kilka numerów po przecinku w jednym polu
- c) nieprawidłowa długość numeru
- d) nieprawidłowa cyfra kontrolna
- e) nieprawidłowy prefiks kraju
- f) nieprawidłowy prefiks firmy

### 4.3. Nazwa produktu w wybranych sklepach elektronicznych

W ramach badania jakości i spójności danych dotyczących nazw produktów, przeanalizowano oferty kilkudziesięciu polskich sklepów internetowych specjalizujących się w sprzedaży produktów szybko zbywalnych (FMCG – Fast Moving Consumer Goods). Zebrane dane zostały zagregowane za pomocą numeru GTIN. Wyselekcjonowano 97 produktów o tych samych numerach GTIN, zarejestrowanych w różnych krajach (w tym w Niemczech, Wielkiej Brytanii, Grecji, Portugalii, Polsce, Węgrzech, Szwecji, Szwajcarii, Włoszech, Holandii i Austrii), obejmujących artykuły spożywcze i drogeryjne. Następnie przeprowadzono analizę zgodności nazw produktów w różnych ofertach, co wykazało, że każdy z badanych produktów posiadał odmienną nazwę w różnych sklepach. Wyniki umieszczono w Tabeli 19.

Tabela 19. Różnice w nazwach produktów

GTIN	Nazwa A	Nazwa B
0000090492341	Sok pomarańczowy 100%	Cappy Sok pomarańczowy 100% 330 ml
04000290908754	Pump Spray Repellent przeciw komarom	OFF! Atomizer - przeciwko komarom i kleszczom 100ml
04005808369867	POMADKA NIVEA 85083 FRUITY SHINE TRUSKAWKA	Pomadka Fruity Shine Strawberry 4,8g
04005808807017	NIVEA Creme Soft Kremowe mydło w płynie zapas 500 ml	Mydło w płynie Creme Soft - uzupełnienie
04007965560101	PASTA DO ZĘBÓW DLA DZIECI ELMEX 50ML	Pasta do zębów dla dzieci
04007965560200	PASTA DO ZĘBÓW SENSITIVE ELMEX 75ML	Pasta do zębów
04007965560606	PASTA DO ZĘBÓW JUNIOR 75ML ELMEX	Pasta do zębów dla dzieci od siódmego roku życia
04007965560804	PASTA DO ZĘBÓW 75ML MERIDOL	Pasta do zębów
04009900452090	Guma Orbit Wild Strawberry 25 drażetek	Guma do żucia w torebce 25 drażetek
04015400006756	Ultra Super Plus Duo Pack Podpaski 2x8szt	PODPASKI UTRA PLUS DUO CIENKIE16SZT
05010232964600	Prezerwatywy 12 sztuk	DUREX PREZERWATYWY EXTRA SAFE 12SZT
05010232964624	Prezerwatywy 12 sztuk	DUREX PREZERWATYWY CLASSIC 12SZT
05010232964822	Prezerwatywy 12 sztuk	DUREX PREZERWATYWY AROUSER 12SZT
05010232967823	Prezerwatywy 3 sztuki	Durex elite -prezerwatywy 3 szt
05010232967847	DUREX PREZERWATYWY EXTRA SAFE 3SZT	Durex Extra Safe Prezerwatywy 3 sztuki
05010232967861	Prezerwatywy 3 sztuki	DUREX PREZERWATYWY CLASSIC 3SZT
05010394984584	Pedigree DentaStix Medium Karma uzupełniająca 180 g (7 sztuk)	MEDIUM Przymak dla psów (10-25kg) 6szt + 1szt GRATIS
05038483167113	Żel intymny nawilżający	DUREX PLAY ŻEL INTYMNY NAWILŻAJĄCY 50ML
05038483382561	Żel do masażu i nawilżający żel intymny z witaminą E	DUREX PLAY ŻEL 2W1 DO MASAŻU 200ML
05038483432198	Prezerwatywy 3 sztuki	PREZERWATYWA DUREX TINGLE 3SZTUKI

GTIN	Nazwa A	Nazwa B
05201089011955	Szampon	Johnson & Johnson Baby Szampon dla dzieci zwykły 200 ml
05601028009218	Papier toaletowy Czarny 6 rolek	Papier toaletowy 6 szt. Renova czarny
05601028010108	Papier toaletowy Czerwony 6 rolek	Papier toaletowy 6 szt. Renova czerwony
05601028010481	Papier toaletowy Pomarańczowy 6 rolek	Papier toaletowy 6 szt. Renova pomarańczowy
05601028010573	Papier toaletowy w tubie 3 szt. Renova czerwony	Papier toaletowy Czerwony 3 rolki (tuba)
05601028010634	Papier toaletowy Zielony 3 rolki (tuba)	Papier toaletowy w tubie 3 szt. Renova zielony
05601028010900	Papier toaletowy Zielony 6 rolek	Papier toaletowy 6 szt. Renova zielony
05601028011952	Papier toaletowy Różowy 6 rolek	Papier toaletowy 6 szt. Renova różowy
05900017361000	KREM OCHRONNY 150ML BAMBINO	Krem ochronny
05900017814100	MYDŁO DLA DZIECI 100G BAMBINO	Mydło
05900095001577	Chusteczki do higieny intymnej Rumianek 10 szt	CHUSTECZKI DO HIGIENY INTYMNEJ RUICMIANEK 10 SZTUK CLEAN
05900095009146	Chusteczki odświeżające Pure&Glamour 15 szt	CHUSTECZKI ODŚWIEŻAJĄCE ACTIVE 15 SZTUK CLEANIC
05900259111531	Lays Chipsy Zielona Cebulka 225g	Chipsy Zielona Cebulka
05900344016697	Lisner Filety śledziowe w oleju a'la Matjas 220 g	Filety śledziowe w oleju a la Matjas
05900344801613	Pasta kanapkowa z tuńczyka z pomidorami	Lisner Pastella Pasta z tuńczyka z pomidorami 80 g
05900437081106	Babeczki Czekoladowe Ze Skórką Pomarańczową 335g	Babeczki czekoladowe z skórką pomarańczy
05900437081304	Dr. Oetker Ciasto Babeczki Jogurtowe 335g	Babeczki jogurtowe
05900437085265	Dr.oetker Ciasto Babka Cytrynowa 375g	Babka o smaku cytrynowym
05900477000976	Zupa fasolowa z boczkiem mrożona	Zupa Fasolowa Hortex 450g
05900512990569	Mlekovita Oselka Masło Polskie Ekstra 375g	Masło extra oselka
05900516300418	Podpaski Nova 10 szt	PODPASKI BELLA NOVA ZE SKRZYDEŁKAMI 10SZT
05900516300487	Podpaski Nova Maxi Air 10 szt	PODPASKI NOVA MAXI 10 SZT
05900516301071	Podpaski Ultra Orange 12 szt	PODPASKI PERFECTA ORANGE 12SZT
05900516301125	Podpaski Ultra Blue 10 szt	Podpaski PERFECTA BLUE 10SZT BELLA
05900516400040	Patyczki higieniczne 200 szt	BELLA PATYCZKI HIGIENICZNE 200SZT
05900552021148	Gellwe Krem do tortów smak ajerkoniakowy 120 g	Gellwe Krem Ajerkoniak 120g
05900562295515	Szynka z piersi indyka - plastry	Sokołów Klasyczny Polski Szynka Z Piersi Indyka Plastry 120g
05900571001176	Orzeszki ziemne solone	Orzeszki Felix Solone 400g
05900627012644	Vanish Płyn do płukania białych firanek 500 ml	Płyn do płukania białych firanek
05900627012910	Vanish Clean & Fresh Pranie mechaniczne Szampon do dużych powierzchni dywanów 500 ml	Szampon do dywanów - czyszczenie mechaniczne
05900749560016	Masa makowa z bakaliami	Masa Bakalland Makowa 850g
05900852000409	MLEKO BEBIKO1 OD URODZENIA 350G	1 Mleko modyfikowane początkowe - od urodzenia
05900951019814	Sucha karma dla psów z drobiem i wołowiną	Pedigree Vital Protection z wołowiną i drobiem Psy dorosłe Karma pełnoporcjowa 500 g
05900956100654	Syrop z dzikiej róży	SYROP DZIKA RÓŻA 420ML BT HERBAPOL
05900983000194	Duży budyń śmietankowy	Delecta Duży budyń smak śmietankowy 64 g
05900983000200	Duży budyń waniliowy	Delecta Duży budyń smak waniliowy 64 g
05900983000217	Duży budyń czekoladowy	Delecta Duży budyń smak czekoladowy 64 g
05900983002891	Ciasto Delecta Duża Blacha Czekoladowe 670g	Ciasto czekoladowe
05901588057323	Ptasie mleczko® o smaku kokosowym - Rajska Wyspa	E. Wedel Ptasie Mleczko Rajska Wyspa o smaku kokosowym w mocno mlecznej czekoladzie 380 g
05901680000609	Ser kozi twardy w plastrach	Ser Edamski Danmis Z Mleka Koziego 100g
05902180610336	Sonko Pop Cool Chipsy Popcornowe - Zielona Cebula 60g	Chipsy popcornowe - zielona cebula
05902180640333	Sonko Pop Cool Chipsy Popcornowe Fromage 60g	Chipsy popcornowe - fromage
05902180650332	Sonko Pop Cool Chipsy Popcornowe Cebula Z Serem 60g	Chipsy popcornowe - cebula z serem
05902180660331	Wafle błonnikowe z magnezem	Sonko Błonnikowe Żelazo Wafle Ryżowe Z Błonnikiem 100g
05902180680339	Wafle błonnikowe z żelazem	Sonko Błonnikowe Magnez Wafle Ryżowe Z Błonnikiem 100g
05902305000646	MYDŁO TOALETOWE SIARKOWE 100G	Antybakteryjne mydło siarkowe w kostce
05902620000116	Maxi Bułki do hot-dogów (4 szt.)	Bułki Schulstad Hot Dog 240g

GTIN	Nazwa A	Nazwa B
05902620000406	Bułki Schulstad Hot Dog 190g	Bułki do hot-dogów francuskich (3 szt.)
05902658506109	Viscoplast Wodoodporny Plastry z mikroperforacją 12 sztuk	PLASTRY VISCOPLAST WODOODPORNE 12SZT
05904906810778	Kabanosy Krakus 180g	Kabanosy
05905187001213	Chrupki Ziemiaczane	Chrupki Monster Munch 100g
05907078675183	Len mielony 200g OLEFARM	Len mielony + 200g GRATIS!
05907814660206	Ciecierzycza BIO	Bio Planet Ciecierzycza Bio 400g
05907814660602	Fasolka adzuki BIO	Bio Planet Fasolka Adzuki Bio 400g
05907814660701	Fasola mung BIO	Bio Planet Fasola Mung Bio 400g
05907814661104	Migdały blanszowane BIO	Bio Planet Migdały Blanszowane Bio 100g
05907814661302	Orzechy brazylijskie BIO	Bio Planet Orzechy Brazylijskie Bio 150g
05907814663931	Mieszanka studencka z żurawiną BIO	Bio Planet Mieszanka Studencka Z Żurawiną Bio 150g
05907814665003	Orzechy włoskie BIO	Bio Planet Orzechy Włoskie Bio 100g
05908230510335	Tarczyński Exclusive Kabanos Wieprzowy 120g	Kabanosy wieprzowe
05908275688051	Twaróg Delikatny President Śmietankowy 200g	Delikatny twaróg śmietankowy
05908275688136	Président Twaróg Delikatny z rzodkiewką 200 g	Delikatny twaróg z rzodkiewką
05908311868454	PASTA DO ZĘBÓW WHITENING 75ML+2-GA ZA 50% CENY SENSODYNE	Corega Super Mocny Krem Mocujący Do Protez Zębowych
05999518571513	Lactacyd Femina Chusteczki do higieny intymnej 20 sztuk	CHUSTECZKI LACTACYD 20SZT
07300400116534	Pieczywo Wasa Sport 275g	Pieczywo chrupkie Sport
07300400122078	Pieczywo Wasa Z Sezamem 220g	Pieczywo chrupkie z sezamem
07300400122573	Pieczywo Wasa Wieloziarniste 215g	Pieczywo chrupkie wieloziarniste
07613031395484	Karma dla dorosłych psów z wołowiną 10 kg	Friskies Active bogata w mięso Pełnoporcjowa karma dla dorosłych psów 10 kg
07613031395491	Karma dla dorosłych psów z kurczakiem i dodatkiem warzyw 10 kg	Friskies Balance z kurczakiem i dodatkiem warzyw Pełnoporcjowa karma dla dorosłych psów 10 kg
08002226006837	Ser Grana Padano 200g	Ser Grana Padano
08710438033762	McCain Country Potatoes Original Cząstki ziemniaczane ze skórką 750 g	Ćwiartki ziemniaka ze skórką
08710449002016	Frytki karbowane mrożone	Aviko Zig Zag Frytki karbowane do piekarnika 750 g
08710449002351	Frytki proste mrożone	Aviko Original Frytki proste do piekarnika 750 g
08710449986095	Frytki proste mrożone	Aviko Super Long Bardzo długie frytki do piekarnika 600 g
08710449986200	Ziemniaczki ze skórką (frytki) mrożone	Aviko Pati Parts Cząstki ziemniaków ze skórką do piekarnika 600 g
08712400762021	MLEKO BEBIKO HA 1 350G OD URODZENIA	HA1 Mleko modyfikowane hypoalergiczne początkowe-od urodzenia
09000100801911	Zawieszka do WC - Chlorine (2x50g)	Bref WC Power Aktiv Chlorine Zawieszka do muszli WC 2 x 50 g

Źródło: Opracowanie własne

W celu przeprowadzenia dokładniejszej analizy, wybrano popularny polski produkt spożywczy, który wyróżniał się szczególnie dużą liczbą wariantów nazw. Jego producent utrzymywał w swojej domenie internetowej oficjalną stronę produktu, lecz bez informacji o numerze GTIN i bez pełnej nazwy opisującej produkt. W nagłówku strony (wyróżniający się nagłówek H2 strony) widniała informacja „Aromat cytrynowy”, było także zdjęcie produktu. Produkt zarejestrowany był w oficjalnym rejestrze GS1 Polska jako „Aromat do ciast cytrynowy 9mlx20szt EXP”. Przeprowadzono analizę występowania tego produktu w sieci Internet pod różnymi nazwami, dla jawnie podanego, identycznego numeru GTIN na wszystkich stronach. Dla celów badawczych zanonimizowano nazwę marki producenta

wstawiając wartość „[MARKA]” pozostawiając przy tym oryginalną wielkość liter. Wyniki tej analizy przedstawiono w Tabeli 20.

Tabela 20. Różne nazwy przykładowego produktu

[MARKA] AROMAT CYTRYNOWY 10ML	AROMAT [MARKA] CYTR 9ML
[MARKA] Aromat Cytrynowy 9 MI	Aromat [Marka] cytrynowy
[MARKA] Aromat do ciast cytrynowy	Aromat [Marka] Cytrynowy 9ml [Producent]
[Marka] Aromat do ciast cytrynowy 10 ml	AROMAT CYTRYNA 10ML [MARKA]
[Marka] aromat do ciast cytrynowy 10ml but	AROMAT CYTRYNOWY
[Marka] Aromat do ciast cytrynowy 9 ml	AROMAT CYTRYNOWY [MARKA]
[MARKA] Aromat Do Ciast Cytrynowy 9 MI	AROMAT CYTRYNOWY 10ML *[MARKA]* EU
[MARKA] Aromat do ciast cytrynowy szkło	AROMAT CYTRYNOWY 10ML [MARKA]
[Marka] aromat do ciast, kremów i ponczu cytrynowy, 10 ml	AROMAT CYTRYNOWY 9ML [MARKA]
[Marka] Aromat do Ciasta Cytrynowy 10g.	AROMAT DO CIAST [MARKA] 9ML CYTRYNA IO
[MARKA] Cytrynowy Aromat do ciast, kremów i ponczu 10ml	AROMAT DO CIAST CYTRYNOWY [MARKA] 9ml
[MARKA]: AROMAT CYTRYNOWY	AROMAT DO CIAST CYTRYNOWY 9ML [MARKA]
[MARKA]-AROMAT DO CIAST CYTRYNOWY 10G A18.	

Źródło: Opracowanie własne

Warto nadmienić, że pod tym numerem GTIN znajduje się pojedynczy produkt, nie 20 szt. (opakowanie zbiorcze posiada inny numer GTIN) a zawartość netto produktu to 9 ml, nie 10 ml, 9 MI („megalitr”?) albo 10 g. Warto również wspomnieć, że podczas dodatkowego przeszukiwania stron wg nazwy produktu natrafiono na identyczny produkt z błędnie oznaczonym numerem GTIN. Z punktu widzenia wyszukiwarek internetowych, szczególnie tych inteligentnych, bardzo niekorzystną informacją są występujące terminy, np.: „szkło”, a szczególnie skrót, który jest jednocześnie innym wyrazem: „but” – jako prawdopodobnie skrót od „butelka”. Wpływa to niekorzystnie na trafność wyników. Dla algorytmów sztucznej inteligencji znaczenie ma również odmiana wyrazu i użycie go w kontekście. Czym innym jest „aromat cytrynowy” a czym innym „aromat cytryna”.

W tym przypadku problemem był nie tylko brak standaryzacji nazw produktów, który negatywnie wpływał na funkcjonowanie wyszukiwarek i agregatorów danych. Brak jednoznacznych standardów powodował, że operatorzy baz danych e-sklepów wprowadzali dane z etykiet produktów według własnego uznania lub zgodnie z wewnętrznymi regulacjami sklepu. W efekcie tekst rozpoczynał się od marki, nazwy powszechnej, przeznaczenia, a kończył na zawartości netto, rodzaju opakowania lub wewnętrznych oznaczeniach



sklepowych. Często brakowało jednak istotnych informacji, takich jak nazwa marki, a wprowadzane dane bywały nieprawdziwe lub niepełne.

W celu zbadania czy problem różnic w nazewnictwie jest specyficzny dla lokalnego rynku i języka, przeprowadzono analogiczną analizę dla wybranych produktów na rynkach niemieckim (Tabela 22) i brytyjskim (Tabela 21). Wyniki były zbliżone, co wskazuje na uniwersalny charakter zidentyfikowanych problemów.

Tabela 21. Przykład różnych nazw produktów na rynku brytyjskim

<b>Proszek do prania Ariel na 140 prań 9,1 kg, GTIN: 8001090395122</b>	
Ariel P&G Professional Powder Original 9.1Kg 140 Washes	<a href="https://www.gs1.org/services/verified-by-gs1/results?gtin=8001090395122">https://www.gs1.org/services/verified-by-gs1/results?gtin=8001090395122</a>
Ariel Professional Powder Detergent Washing Powder 140 Washes	<a href="https://www.onbuy.com/gb/ariel-professional-powder-detergent-washing-powder-140-washes~c17982~p50603772/">https://www.onbuy.com/gb/ariel-professional-powder-detergent-washing-powder-140-washes~c17982~p50603772/</a>
Ariel Washing Powder Professional Laundry Detergent 9.1KG 140 Wash	<a href="https://www.grelly.uk/itm/ariel-washing-powder-professional-laundry-detergent-91kg-140-wash-203817349321">https://www.grelly.uk/itm/ariel-washing-powder-professional-laundry-detergent-91kg-140-wash-203817349321</a>
Ariel 9.1kg Professional Washing Powder	<a href="https://www.ebay.co.uk/p/21020249473?iid=203480414772">https://www.ebay.co.uk/p/21020249473?iid=203480414772</a>
Ariel Washing Powder Professional Laundry Detergent 9.1KG 140 Wash	<a href="https://picclick.co.uk/Ariel-Washing-Powder-Professional-Laundry-Detergent-91KG-140-203817349321.html?refresh=1">https://picclick.co.uk/Ariel-Washing-Powder-Professional-Laundry-Detergent-91KG-140-203817349321.html?refresh=1</a>

Źródło: Opracowanie własne

Tabela 22. Przykład różnych nazw produktów na rynku niemieckim

<b>Gumy rozpuszczalne Maomam 150szt, GTIN: 4001686520789</b>	
STRIPES 1.05 Kilogramm	<a href="https://onlineshop.haribo.com/maoam/110/stripes">https://onlineshop.haribo.com/maoam/110/stripes</a>
Haribo Maoam Stripes, Dose, 150 Stück, 1050g	<a href="https://www.verpflegungslebensmittel-online-bestellen.de/shop/suessigkeiten/haribo-maoam-stripes-dose-150-stueck-1050g/">https://www.verpflegungslebensmittel-online-bestellen.de/shop/suessigkeiten/haribo-maoam-stripes-dose-150-stueck-1050g/</a>
MAOAM Kaubonbon-Streifen STRIPES	<a href="https://www.buero-bedarf-thuringen.de/maoam-kaubonbon-streifen-stripes-4001686520789">https://www.buero-bedarf-thuringen.de/maoam-kaubonbon-streifen-stripes-4001686520789</a>
Food Haribo Maoam Stripes 1050g	<a href="https://www.osma-werm.com/de_DE/kleinpreisartikel/getraenke-food-suessigkeiten/16452/food-haribo-maoam-stripes-1050g">https://www.osma-werm.com/de_DE/kleinpreisartikel/getraenke-food-suessigkeiten/16452/food-haribo-maoam-stripes-1050g</a>
Maoam Stripes 150er	<a href="https://www.worldofsweets.de/Maoam-Stripes-150er.100300.html">https://www.worldofsweets.de/Maoam-Stripes-150er.100300.html</a>
HARIBO MAOAM STRIPES, KAUBONBON, 150 STÜCK	<a href="https://suesswaren-grosshandel.de/bonbons/kaubonbon/weich-bonbons/haribo-maoam-stripes-kaubonbon-150-stueck.html">https://suesswaren-grosshandel.de/bonbons/kaubonbon/weich-bonbons/haribo-maoam-stripes-kaubonbon-150-stueck.html</a>

Źródło: Opracowanie własne

Przeprowadzone badania potwierdzają, że produkty o tym samym numerze GTIN mogą mieć różne nazwy w zależności od sklepu, w którym są sprzedawane. Przyczyną tego zjawiska jest przypuszczalnie brak standaryzacji nazw produktów, co negatywnie wpływa na spójność danych i prowadzi do wprowadzania informacji według indywidualnych zasad operatorów lub wewnętrznych regulacji sklepów. Analiza przeprowadzona na rynkach niemieckim

i brytyjskim wykazała, że problem różnic w nazewnictwie produktów, mimo korzystania z tego samego numeru GTIN, ma charakter uniwersalny i nie ogranicza się jedynie do rynku krajowego.

#### **4.4. Nazwa produktu/oferty na platformie zakupowej**

Badanie polegało na obserwacji różnic w danych produktu o podanym numerze GTIN, sprzedawanego pod wieloma nazwami ofert i kategoriami na jednej z platform zakupowych w Polsce. Celem było znalezienie i przeanalizowanie przypadku, w którym kupujący, próbując wyszukać różne produkty, otrzymywał liczne oferty zawierające ten sam produkt. W ramach badania przeanalizowano dane produktu – zabawki, mikrofonu z głośnikiem zasilanym bateryjnie, z możliwością podłączenia dodatkowego źródła dźwięku przez Bluetooth i wyposażonym w efekty świetlne. Dla tego produktu zidentyfikowano łącznie 771 par zawierających nazwę oferty (nazwę produktu) oraz kategorię produktu. Produkt ten został zaklasyfikowany do 10-ciu różnych kategorii i opisany 283-ma różnymi nazwami. W zależności od handlującego towarem i konkretnej oferty, produkt zaklasyfikowano jako:

- a) gadżet świetlny
- b) gadżet elektroniczny
- c) głośnik przenośny
- d) sprzęt do karaoke
- e) mikrofon bezprzewodowy
- f) akcesorium GSM
- g) zabawkę

W nazwach produktu wyodrębniono między innymi następujące dodatkowe zwroty poza nazwą produktu lub wręcz zamiast niej:

- a) dla dziewczynki
- b) do telefonu [podana nazwa modelu telefonu]
- c) karaoke świecący LED
- d) na prezent dla DZIECI
- e) lekkie głośne ledowe
- f) na imprezę
- g) pomysł na prezent na dzień ojca

Przedstawiony przykład ukazuje charakterystykę oraz konsekwencje celowego manipulowania nazwami produktów przez sprzedawców w celu dotarcia do szerszego grona odbiorców i zwiększenia sprzedaży. W takich przypadkach produkty często nie posiadały

ugruntowanej pozycji na rynku, brakowało im rozpoznawalnej marki, a ich producenci byli nowymi uczestnikami rynku. Choć ta strategia mogła przynieść sprzedawcom krótkoterminowe korzyści, jednocześnie przyczyniała się do pogłębienia globalnego problemu jakości danych. Dodatkowo, praktyki te mogły negatywnie wpływać na platformy e-commerce (marketplace), ponieważ liczne oferty tego samego produktu pod różnymi nazwami mogły zniechęcać klientów do dalszych poszukiwań.

Nowoczesne platformy e-commerce, takie jak Amazon, eBay czy Allegro, starają się obecnie rozwiązać ten problem, tworząc własne katalogi produktów, agregując dane na podstawie numerów GTIN i prezentując produkt pod jedną, wspólną nazwą. Mimo że jest to bardzo dobra inicjatywa, to efekt tego rozwiązania jest czasochłonny, jest rozwiązaniem wewnętrznym danej platformy i zależy od dostępności danych referencyjnych oraz spójności numerów GTIN w różnych źródłach danych.

W kontekście wymiany danych w e-commerce, taka sytuacja podkreśla konieczność wdrożenia bardziej efektywnych metod walidacji i standaryzacji danych produktowych. Dotyczy to przede wszystkim nazw produktów dla zapewnienia spójności i przejrzystości informacji, ale także innych atrybutów produktów, w tym identyfikatorów, co jest kluczowe dla poprawnej agregacji danych. Wyzwania związane z zarządzaniem danymi produktowymi wymagają opracowania nowych strategii i narzędzi, które będą w stanie skutecznie rozwiązywać te problemy.

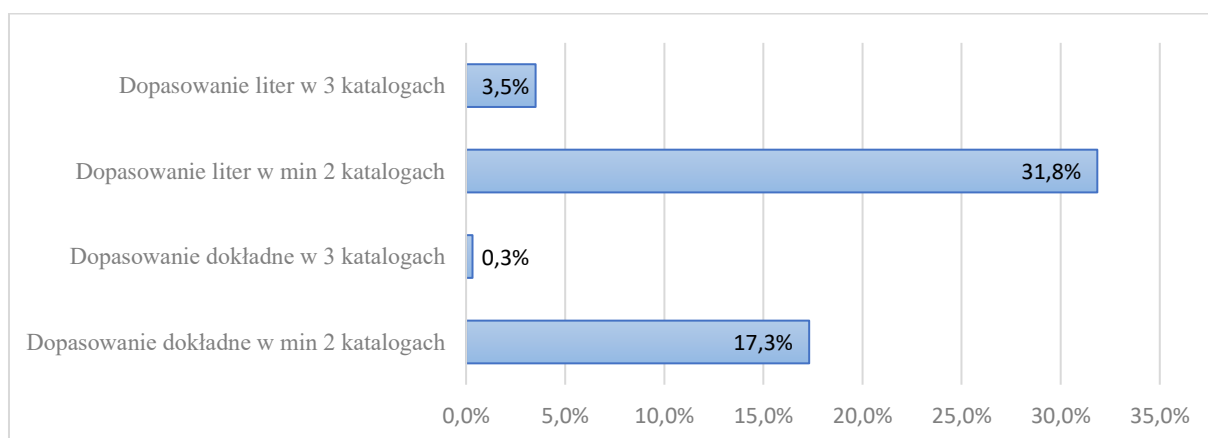
#### **4.5. Nazwy produktów w różnych katalogach produktowych**

W poprzednich badaniach dane do analizy pochodziły ze sklepów internetowych, systemów aukcyjnych oraz platform typu marketplace. W niniejszym badaniu skoncentrowano się na analizie, czy źródłem błędów w danych mogą być sami producenci. Wyselekcjonowano trzy katalogi danych, które świadczyły usługi dla rynku e-commerce, a ich informacje były weryfikowane we współpracy z producentami. Badanie obejmowało produkty, które występowały równocześnie we wszystkich trzech katalogach i posiadały identyczny numer GTIN. Dodatkowym kryterium doboru była dostępność produktów na polskim rynku, niezależnie od pochodzenia i kraju producenta.

W ramach badania, połączono dane dla 9266 produktów po ich numerze identyfikacyjnym GTIN, mając na celu ustalenie czy nazwy danych produktów są identyczne (dokładne dopasowanie ciągów tekstowych bez uwzględniania wielkości liter) lub podobne.

Do weryfikacji podobieństwa nazw zastosowano algorytm, który ignorował różnice w kolejności słów, liter i cyfr w nazwie, z zastrzeżeniem, że liczba liter i cyfr w obu przypadkach musiała się zgadzać.

Wyniki przeprowadzonego badania, przedstawione na Rysunku 10, dostarczają istotnych spostrzeżeń na temat spójności nazw produktów w katalogach danych weryfikowanych we współpracy z producentami. Uzyskane dane wskazują, że pełna zgodność nazw produktów we wszystkich trzech analizowanych katalogach była rzadkością, odnotowano ją dla zaledwie 30 produktów. Interesującym jest fakt, że gdy pominięto jeden z katalogów, stopień zgodności wzrósł do 17,3% całego zbioru, co nadal stanowi stosunkowo niski wynik. Testowanie podobieństwa nazw, oparte na porównywaniu tych samych liter i cyfr w obu nazwach, przyniosło lepsze rezultaty, choć należy uwzględnić ograniczenia zastosowanego algorytmu. Mimo to, ogólny poziom podobieństwa nazw wciąż pozostał stosunkowo niski.



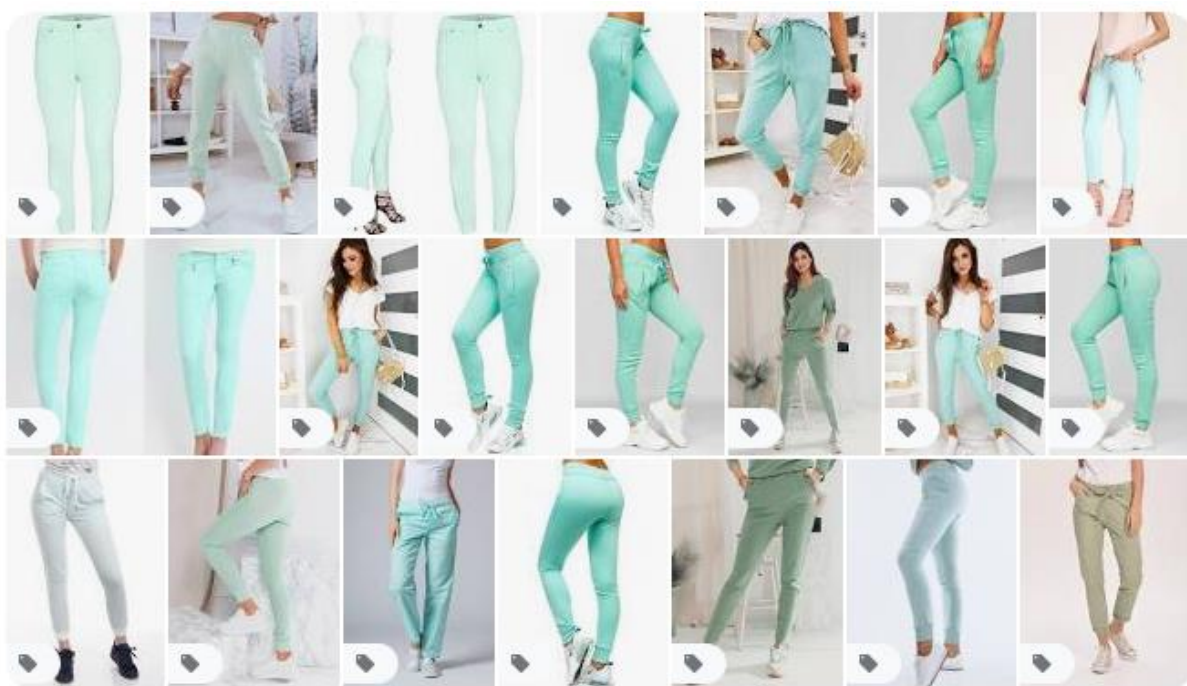
Rysunek 10. Zgodność nazwy produktu

Źródło: opracowanie własne

Analiza wyników wskazuje, że nawet w katalogach danych współtworzonych i weryfikowanych przez producentów, nie można zagwarantować pełnej spójności nazw produktów. To odkrycie podkreśla potencjalne wyzwania w zarządzaniu danymi produktowymi i skalę problematyki.

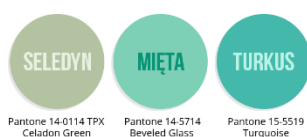
#### 4.6. Zgodność nazwy ze zdjęciem produktu

W kontekście analizy błędów w e-commerce, interesującym zagadnieniem była niezgodność między nazwą a zdjęciem produktu. Przykładem takiej sytuacji była sytuacja, w której nazwa produktu sugerowała określoną kolorystykę, która nie była odzwierciedlona na dołączonym zdjęciu. Aby zbadać to zjawisko, przeprowadzono eksperyment polegający na wpisaniu do wyszukiwarki Google frazy „miętowe spodnie damskie”. Dane pozyskano z różnych sklepów internetowych, gdzie każdy produkt był reprezentowany przez zdjęcie oraz nazwę zawierającą tę frazę. Wyniki wyszukiwania zaprezentowano na Rysunku 11, w kolejności ich wyświetlania na stronie.



Rysunek 11. Przykładowe zdjęcia produktów dla wyszukiwanej frazy  
Źródło: opracowanie własne na podstawie wyników wyszukiwarki Google

Pierwszy problem, który łatwo zaobserwować na zdjęciach, to duża różnica w kolorach spodni. Na Rysunku 12 zaprezentowano wzornik kolorów Pantone dla porównania definicji barw.



Rysunek 12. Przykłady nazw kolorów z przypisanymi numerami wzornika Pantone  
Źródło: opracowanie własne

Poruszone w tym badaniu zagadnienia dotyczą kilku kwestii. Po pierwsze - kompozycji zdjęcia produktu, obejmującego ujęcie, tło, liczbę produktów na jednej fotografii a także sposobu prezentacji – czy to na modelce, czy jako samodzielny element garderoby. Po drugie, odnoszą się do spójności pomiędzy nazwą produktu a jego zdjęciem. Jak pokazano na Rysunku 11, z samych zdjęć często nie wynika czym jest przedmiot zakupu – czy są to spodnie, komplet spodni z bluzką, torebka czy buty. Niejasna pozostaje również kwestia rzeczywistego koloru produktu, jaki klient otrzyma po dostawie, o czym wspomniano na początku badania. Wszystkie te problemy wynikają z braku standaryzacji danych, przede wszystkim jednoznacznych wytycznych dotyczących zawartości (Niemir i Mrugalska, 2022a) oraz braku jednolitego „paszportu produktu” gdzie nazwa produktu i zdjęcie byłoby elementem nierozdzielalnym, co prawdopodobnie choćby częściowo eliminowałoby problem spójności pomiędzy nazwą a zdjęciem.

#### **4.7. Wnioski**

Wyniki badań wskazują na złożoność problemów związanych z jakością danych produktowych, wynikającą z braku standaryzacji i ogólnodostępnych baz referencyjnych. W kontekście konkretnych rozwiązań informatycznych istnieje wyraźna potrzeba opracowania skutecznych metod walidacji, które obecnie są niewystarczające lub nieobecne.

Istotnym wyzwaniem jest brak spójności w nazewnictwie produktów. Potwierdzają to wyniki badań ofert sklepów internetowych, marketplace'ów oraz danych katalogowych współtworzonych z producentami. Sugerowanym rozwiązaniem jest wykorzystanie technik sztucznej inteligencji w obszarze rozumienia języka naturalnego (NLU), co ułatwi wykrywanie błędów i sugerowanie poprawek w danych. Stwierdzono również niezgodności między nazwami produktów a ich zdjęciami, co dodatkowo podkreśla konieczność wdrożenia zaawansowanych procedur walidacyjnych. Techniki uczenia maszynowego, takie jak sztuczna inteligencja w multimodalnej analizie obrazów i opisów tekstowych, mogą zostać zastosowane do wykrywania takich niezgodności. Umożliwi to identyfikację błędów, takich jak nieodpowiednie zdjęcia czy niezgodne opisy, zwiększając wiarygodność informacji prezentowanych klientom.

Jednak sztuczna inteligencja nie stanowi jedyne rozwiązanie dla wykrytych problemów. Przykładowo, analiza numerów GTIN ujawniła błędy, takie jak nieprawidłowe znaki, niewłaściwe długości numerów, błędne cyfry kontrolne oraz niezgodne prefiksy krajowe

i firmowe, dla których sztuczna inteligencja nie jest potrzebna. Błędy tego typu świadczą raczej o braku odpowiednich mechanizmów walidacyjnych na poziomie podstawowym, pomimo dostępności stosownych reguł od wielu lat. W związku z tym, oprócz zaawansowanych algorytmów sztucznej inteligencji, konieczne jest wdrożenie prostych reguł walidacyjnych, które dotychczas były często pomijane w systemach informatycznych.

Podsumowując, niezbędne wydaje się podejście hybrydowe, łączące celowane reguły walidacyjne z wykorzystaniem sztucznej inteligencji w sytuacjach wymagających bardziej zaawansowanej analizy. Takie połączenie pozwoli na skuteczniejsze wykrywanie i korektę błędów w danych produktowych, zwiększając ich jakość i spójność w systemach e-commerce.

W kontekście niniejszej pracy doktorskiej, identyfikacja problemów i uzyskane wyniki stanowią istotny wkład w opracowywanie narzędzi do poprawy jakości elektronicznych katalogów produktowych.

## 5. Analiza problemu jakości danych produktowych

W niniejszym rozdziale dokonano analizy problematyki jakości danych produktowych bazy produktów Rejestru GS1, stanowiącej podmiot badawczy. Jest to jednocześnie reprezentatywny katalog danych produktowych zasilany przez szerokie grono użytkowników zewnętrznych, gdzie występują problemy podobne do katalogów crowdsourcingowych, opisanych w Rozdziale 2.5. Celem analizy było zidentyfikowanie wyzwań związanych z procesem walidacji danych na wielomilionowym, rzeczywistym zbiorze danych.

### 5.1. Audyt danych i wywiad ekspercki

Proces analizowania przedmiotowych danych, znany jako tzw. „data profiling” (Naumann, 2014), rozpoczęto od dokładnego zbadania struktury i schematu bazy. Zidentyfikowano kluczowe tabele, relacje, typy danych i kolumn. W konsekwencji analizą objęto kilka najważniejszych atrybutów produktowych bazy takich jak:

- a) numer GTIN
- b) nazwa produktu (opis produktu na etykiecie)
- c) marka
- d) zdjęcie produktu
- e) zawartość netto wraz z jednostką zawartości netto
- f) klasyfikacja GPC produktu
- g) język opisu.

Zachowano w ten sposób zgodność z badaniem opisanym w Rozdziale 3, jednocześnie obejmując wszystkie wymagane atrybuty bazy uznawane za podstawowe (podstawowe dane produktowe omówiono w Rozdziale 2.6). Pozostałe atrybuty podmiotu badań, określone jako nieobowiązkowe w strukturze bazy, takie jak strona WWW produktu, opis produktu, kod wewnątrz – nie zostały uwzględnione w pracy. Szerszą perspektywę z uwzględnieniem wymienionych pól, a także atrybutów produktowych związanych z bezpieczeństwem konsumenta, omówiono w publikacjach (Niemir i in., 2022; Niemir i Mrugalska, 2023a).

Analizując bazę, zbadano występowanie pustych wartości, nieprawidłowości, błędów oraz niekompletności danych, wykrywając dane odstające oraz badając wzorce i trendy (Abedjan i in., 2015). Inspirowano się przy tym wcześniej zdefiniowanymi grupami potencjalnych błędów opisanych szczegółowo w Rozdziale 2.7 wykonując audyt danych pod kątem następujących typów potencjalnych naruszeń:

- a) błąd pisowni
- b) brak danych



- c) zduplikowane dane
- d) nieprawidłowa wartość
- e) niespójny format danych
- f) niekompletny format danych
- g) naruszenie składni
- h) naruszenie unikalności wartości
- i) naruszenie ograniczeń integralności
- j) formatowanie tekstu
- k) naruszenie ograniczeń domeny
- l) naruszenie zasad biznesowych organizacji
- m) naruszenie ograniczeń administracyjnych bazy danych.

Nie brano pod uwagę aktualności danych, a także kwestii bezpieczeństwa, wiarygodności i wydajności koncentrując się jedynie na tematyce jakości samych danych zastanych w zbiorze.

Analizowana baza danych posiadała już liczne zabezpieczenia dotyczące unikalności, integralności, naruszeń ograniczeń bazy danych oraz zabezpieczenia braku danych w stopniu zapewniającym jej działanie przez kilkanaście lat i obsługę z udziałem kilkudziesięciu tysięcy użytkowników, dlatego skoncentrowano się na szczegółowej analizie pozostałych błędów, które nie były dotychczas wykrywane. W celu pogłębienia wiedzy i uzyskania niezbędnych informacji przeprowadzono indywidualne wywiady półstrukturyzowane (Czakon, 2011) ze specjalistami pracującymi z analizowaną bazą, aby zidentyfikować problemy zauważone przez użytkowników i tym samym potwierdzić zbieżność z tymi, które zostały wykryte w badaniach w Rozdziale 4 lub ewentualnie rozszerzyć zakres audytu. Wśród respondentów wytypowano: administratora bazy i systemu informatycznego realizującego zgłoszenia błędów aplikacyjnych, administratora katalogu danych mającego bezpośredni kontakt z klientami, eksperta ds. e-commerce znającego specyficzne potrzeby rynku, oraz eksperta ds. jakości danych. Lista pytań obejmowała zakres atrybutów połączony z wymienionymi w rozdziale typami naruszeń. Odpowiedzi respondentów pozwoliły zidentyfikować szereg problemów, które mogą negatywnie wpływać na jakość i wiarygodność danych produktowych w bazie danych. Analiza odpowiedzi dotyczyła różnych aspektów danych produktowych, w tym problemów z formatowaniem tekstu, nieprawidłowym kategoryzowaniem produktów, niespójnością danych, błędami w nazwach i opisach produktów, a także problemach związanych z jakością i adekwatnością zdjęć produktów. Odpowiedzi pogrupowano tematycznie, a każda wypowiedź specjalisty odnosząca się do zauważonych problemów w danej kategorii została oznaczona symbolem . Wyniki odpowiedzi zaprezentowano w Tabeli 23.

Tabela 23. Problemy z jakością danych zgłoszone przez ankietowanych

Kategoria	Admin. Bazy i Systemu	Admin. Katalogu Danych	Ekspert ds. e-commerce	Ekspert ds. Jakości Danych
<b>Nieprawidłowości w zarządzaniu numerami GTIN</b>				
<b>Nieprawidłowości w danych słownikowych systemu IT</b>				
<b>Brak danych, niespójności w systemie IT</b>				
<b>Problem z funkcjonowaniem systemu przez nieprawidłowości w danych</b>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		
- ukryte znaki w tekstach atrybutów produktów				
- zbyt duże zdjęcia zawieszające system				
<b>Problem z poprawnym wyświetleniem nazwy i marki produktu</b>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
- wielokrotne spacje i pokreślenia w tekście, encje HTML, spacje przed/za tekstem, znaki tabulatora				
- tekst rozpoczynający się od cudzysłowu/apostrofu,				
- niepoprawnie zdekodowane znaki zamiast polskich liter,				
- formuła Excela zamiast treści,				
- język danych tekstowych niezgodny z zadeklarowanym				
<b>Niepoprawna nazwa produktu i marki</b>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>
- nazwa produktu jest nazwą powszechną wyrobu,				
- niezrozumiała nazwa produktu - kody producenta, skróty				
- duplikaty nazw produktów				
<b>Niestandardyzowane nazwy i marki produktu</b>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
- długie nazwy zawierające zbędne opisy i dane,				
- nazwy marek bardzo podobne do siebie, ale inaczej zapisywane				
<b>Niedostateczna jakość zdjęć produktu</b>			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
- zdjęcie nie przedstawia produktu,				
- produkt na niewłaściwym tle				
<b>Nieprawidłowe kategoryzowanie produktów</b>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
- wiele produktów w kategorii "tymczasowe",				
- niewłaściwe przypisanie do kategorii				
<b>Nieprawidłowa zawartość netto i jednostka</b>	<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>
- zawartość netto $\leq 0$ , ułamkowa zawartość netto dla towarów sprzedawanych na sztuki,				
- nieprawdopodobna zawartość netto i jednostka zawartości netto dla danej nazwy produktu				

Źródło: opracowanie własne

Respondenci wskazali na problemy związane z nieprawidłowościami w danych, które negatywnie wpływają na funkcjonowanie systemów zarządzania danymi, takich jak ukryte znaki kontrolne, czy zbyt duże zdjęcia. Istotne były także problemy z formatowaniem tekstu, które utrudniały poprawne wyświetlenie nazw i marek produktów, w tym wielokrotne spacje, znaki tabulatora oraz błędnie zdekodowane znaki. Niepoprawne nazwy produktów i marek, w tym użycie nazw powszechnych, jednowyrazowych nazw czy literówek, również zostały zauważone jako problematyczne. Nie poświęcono za to uwagi kwestiom związanym

z nieprawidłowościami z numerem GTIN, co sugeruje, że w tym zakresie system został odpowiednio zabezpieczony. Potwierdzono to wykonując szczegółową weryfikację systemu - zaimplementowano wszystkie walidatory opisane w Rozdziale 4.2, obejmujące: niepoprawne znaki zastępujące numer w polu, w tym występowanie kilku numerów rozdzielonych przecinkami w jednym polu, nieprawidłową długość numeru, nieprawidłową cyfrę kontrolną, nieprawidłowy prefiks kraju oraz nieprawidłowy prefiks firmy. Nie zidentyfikowano również żadnych problemów technicznych dotyczących pól, takich jak brak wymaganych danych, wprowadzenie danych niewystępujących w słowniku lub przekraczających ograniczenia bazy danych.

Odpowiedzi respondentów dotyczyły przede wszystkim zawartości informacyjnej danych, ich formatowania oraz logicznych zależności między nimi. Zwrócili oni uwagę na brak standaryzacji nazw i marek, co może prowadzić do dezorientacji i utrudniać identyfikację produktów. Istotną kwestią była również jakość zdjęć produktów, zwłaszcza w kontekście ich nieprawidłowego przedstawienia. Wskazane problemy podkreśliły nie tylko znaczenie właściwego formatowania danych, ale również problemu braku walidacji treści przechowywanych danych.

## **5.2. Analiza i ocena możliwości rozwiązania wykrytych błędów w danych**

Na podstawie przeprowadzonych wywiadów sporządzono listę zagadnień, obejmującą potencjalne problemy jakościowe, rozszerzoną o własne spostrzeżenia sporządzone na podstawie badań opisanych w Rozdziale 4 oraz na podstawie próbek danych zebranych podczas analizowania zbioru danych. Znalezione problemy odpowiednio sklasyfikowano, jednocześnie analizując przyczyny ich powstania i szacowaną wielkość w stosunku do całego zbioru danych, o ile na tym etapie było to możliwe do wyliczenia bez implementacji zaawansowanych algorytmów. Zestawienie przedstawiono w Tabeli 24, a następnie rozszerzono je o ocenę możliwości rozwiązania wykrytych problemów, co zaprezentowano w Tabeli 26.

Tabela 24. Zestawienie znalezionych problemów jakościowych w katalogu

Lp.	Atrybut danych	Wykryty problem	Uzasadnienie: Dlaczego jest to problem?	Rzeczywiste przykłady	Prawdopodobne źródło powstawania błędu	Szac. liczba danych
1	GPC ↔ Nazwa	Niepoprawna lub tymczasowa kategoria produktu	Prowadzi do nieporządku i nieefektywności w zarządzaniu danymi, utrudniając wyszukiwanie i filtrowanie produktów według ich faktycznych kategorii. Może także wpływać na analizę danych i algorytmy rekomendacyjne, zmniejszając ich skuteczność.	<ul style="list-style-type: none"> <li>▶ Zegar ścienny TERMOMETR + HIGROMETR</li> <li>▶ ŚRUBA MOCUJĄCA 80MM (K-16-80)</li> <li>▶ Wazon Black Bamboo</li> <li>▶ HP EliteBook 8540p 14,4V 5200mAh czarna</li> <li>▶ Ramka kompletna ZSP 04-108 2A31</li> </ul>	Niewystarczająca znajomość struktury kategorii lub próba obejścia wymagań systemu przez użytkownika.	ŚREDNIA
2	Marka	Marka produktu z nadmierną ilością słów	Utrudnia szybkie zrozumienie i identyfikację marki, może prowadzić do niejednoznaczności.	<ul style="list-style-type: none"> <li>▶ Zaawansowane Technologicznie Ekologiczne Rozwiązania Czyszczące i Pielęgnacyjne dla Domu i Przemysłu Super Clean &amp; Care</li> </ul>	Niezrozumienie zasad skutecznej komunikacji marketingowej; próba włączenia zbyt wielu słów kluczowych.	ŚREDNIA
3	Marka	Nazwa marki bez liter	Uniemożliwia identyfikację marki, co komplikuje wyszukiwanie i kategoryzację produktów.	<ul style="list-style-type: none"> <li>▶ E.65.10.00 N</li> <li>▶ 0,9</li> <li>▶ 0000.0541</li> <li>▶ 0-1671132-2</li> <li>▶ 42080043W</li> </ul>	Błędne wprowadzanie danych przez nieuwagę lub celowe ominięcie ograniczeń systemowych.	BARDZO MAŁA
4	Marka	Nadmiernie długa nazwa marki	Może utrudniać wyszukiwanie i sortowanie, a także wpływać negatywnie na prezentację w interfejsach użytkownika.	<ul style="list-style-type: none"> <li>▶ Butelka ze szkła barwnego z zakrętką z PE i kropłomierzem z PE lub Butelka z PET z kropłomierzem z PE z zakrętką z PE z pierścieniem gwarancyjnym</li> </ul>	Nieprawidłowe użycie pola nazwy marki do umieszczania np. opisów produktów.	MAŁA
5	Marka	Różne zapisy tej samej marki	Prowadzi do niekonsekwencji w danych, utrudniając agregację i analizę produktów danej marki.	<ul style="list-style-type: none"> <li>▶ Waweel / WAWEL / Wawel S.A. / Wawel SA / Wawelo</li> <li>▶ Scandia Cosmetics S.A / Scandia cosmetic S.A / Scandia Cosmetics S.A / Scandia Cosmetics SA / Scania Cosmetics S.A</li> <li>▶ KRYSALIA / KRYSTALIA / KYSTALIA</li> </ul>	Brak standaryzacji nazw marek w systemie danych, co może wynikać z niejednorodnych źródeł danych.	MAŁA
6	Marka	Niepoprawne nazwy marek (lista)	Wprowadza błędy i nieścisłości do bazy danych, utrudniając identyfikację i klasyfikację produktów.	<ul style="list-style-type: none"> <li>▶ GS1 Polska</li> <li>▶ POLSKA</li> <li>▶ własna</li> <li>▶ brak</li> <li>▶ inna</li> <li>▶ INNE</li> <li>▶ no name</li> </ul>	Niekompletne dane produktowe spowodowane nieuwagą lub niezrozumieniem struktury danych produktowych. Producent może nie używać marki produktu na co dzień, co utrudnia poprawne wprowadzenie danych.	MAŁA

Lp.	Atrybut danych	Wykryty problem	Uzasadnienie: Dlaczego jest to problem?	Rzeczywiste przykłady	Prawdopodobne źródło powstawania błędu	Szac. liczba danych
				<ul style="list-style-type: none"> <li>▶ Test</li> <li>▶ marka własna</li> <li>▶ NoName</li> </ul>		
7	Nazwa ↔ Marka	Identyczna nazwa produktu i marki	Może powodować błędy w klasyfikacji oraz wyszukiwaniu produktów.	—	Automatyczne generowanie nazw produktów na podstawie ograniczonych danych wejściowych lub nieuwaga przy wprowadzaniu danych.	ŚREDNIA
8	Nazwa	Początek tekstu to nieprawidłowy znak	Może utrudniać prawidłowe sortowanie i wyszukiwanie, a także wpływać negatywnie na prezentację danych.	<ul style="list-style-type: none"> <li>▶ "elki Zygzaki k□ka w kwa□nym cukrze 80g</li> <li>▶ .ŻARÓWKA ABILITE SMD-3528 E14 48LED</li> <li>▶ B.ZIMNY 150(Lm) ką t św. 120° 230V/2.5W (spotlight).</li> <li>▶ #_MAS CON MICRA 12002000</li> <li>▶ **PIÓRO WIECZNE COLLAGE Z NABOJAMI STK BE</li> </ul>	Błędy kodowania znaków przy imporcie danych z systemów o różnych ustawieniach regionalnych.	ŚREDNIA
9	Nazwa	Nadmiernie długa nazwa produktu	Utrudnia wyszukiwanie, sortowanie i może negatywnie wpływać na interfejsy użytkownika.	<ul style="list-style-type: none"> <li>▶ Boost-Serum z niacynamidem i GABA odbudowuje barierę ochronną skóry i stymuluje produkcję kolagenu. Natychmiastowo i długotrwanie redukuje zmarszczki oraz głębokie bruzdy. Zmniejsza TEWL, dzięki czemu zwiększa nawilżenie i elastyczność skóry oraz wygładza i nadaje skórze idealną miękkość. Rozjaśnia i wyrównuje koloryt skóry. Działa antyoksydacyjnie i zwalcza wolne rodniki. Formułą aktywna uzupełnio</li> </ul>	Próba umieszczenia nadmiernej ilości informacji w polu nazwy produktu zamiast odpowiednich pól opisowych.	ŚREDNIA
10	Nazwa	Nadmiernie długa nazwa produktu do niektórych zastosowań e-commerce	Ogranicza możliwość przekazywania informacji pomiędzy systemami	<ul style="list-style-type: none"> <li>▶ NaturalEco Organizer na Biurko - Ekologiczny, Biodegradowalny, 500g, Wielofunkcyjny, Z Recyklingu</li> </ul>	Próba umieszczenia nadmiernej ilości informacji w polu nazwy produktu zamiast odpowiednich pól opisowych.	ŚREDNIA
11	Nazwa	Nazwa produktu z nadmierną ilością słów	Utrudnia szybkie zrozumienie i identyfikację produktu, może prowadzić do niejednoznaczności.	<ul style="list-style-type: none"> <li>▶ Wkręty fosfatowane do płyt kartonowo-gipsowych (drewno) 3,5x25 opakowanie 1000szt. VisDisc Michał Archacki typ. wkręty azmet</li> </ul>	Niezrozumienie zasad skutecznej komunikacji marketingowej; próba włączenia zbyt wielu słów kluczowych.	ŚREDNIA

Lp.	Atrybut danych	Wykryty problem	Uzasadnienie: Dlaczego jest to problem?	Rzeczywiste przykłady	Prawdopodobne źródło powstawania błędu	Szac. liczba danych
12	Nazwa	Produkt z nazwą jednowyrazową	Może nie dostarczać wystarczającej ilości informacji o produkcie, utrudniając jego identyfikację.	<ul style="list-style-type: none"> <li>▶ AMMUS</li> <li>▶ -----</li> <li>▶ -----</li> <li>▶ 0A1AN5424PESX92A</li> </ul>	Nadmierne skracanie nazw produktów dla uproszczenia katalogowania, nieświadomość znaczenia zawartości informacyjnej pola, lub błędy przy wprowadzaniu danych.	DUŻA
13	Nazwa ↔ Zaw. netto + jednostka	Brak wskazania zawartości netto w nazwie produktu	Utrudnia porównywanie produktów i może wprowadzać konsumentów w błąd co do ich faktycznej zawartości.	▶ Napój gazowany Fanta	Pominięcie kluczowych informacji o produkcie ze względu na brak świadomości ich znaczenia lub błędy w procesie wprowadzania danych.	BARDZO DUŻA
14	Nazwa ↔ Marka	Brak marki w nazwie produktu	Utrudnia identyfikację i klasyfikację produktów, wpływając na wyszukiwanie i porównywanie.	▶ Chleb krojony 1kg	Niekompletne dane produktowe spowodowane nieuwagą lub niezrozumieniem struktury danych produktowych.	BARDZO DUŻA
15	Nazwa	Nazwa produktu pisana wyłącznie wielkimi literami	Może być trudniejsza do odczytania, a także może być postrzegana jako mniej profesjonalna prezentacja danych.	▶ AA PRESTIGE MORPHO CREATOR PRO-CALCIUM KREM REGENERUJĄCY NA DZIEŃ SPF6 50+ 50ML	Próba wyróżnienia produktu przez użycie wyłącznie wielkich liter, często ignorując zasady dobrego stylu lub ograniczenia systemu z którego dokonywano import danych.	DUŻA
16	Nazwa	Brak słów w nazwie produktu	Oznacza brak możliwości identyfikacji produktu, co jest krytycznym błędem w zarządzaniu danymi.	<ul style="list-style-type: none"> <li>▶ L-6</li> <li>▶ -----</li> <li>▶ do 13.07.05 dzb 0,2 tina</li> <li>▶ Ośc.PS Bezp-okl.synt</li> <li>▶</li> <li>skr.pn.wew.KN.gA.w3.P80.D/orz.t1.msr.bl.3zP.ra</li> <li>m.sMat.</li> <li>▶ W05.0X012-WW-PH-B METR.WALC.</li> </ul>	Niekompletne dane produktowe spowodowane nieuwagą lub niezrozumieniem struktury danych produktowych.	ŚREDNIA
17	GTIN ↔ Nazwa	Powtarzające się nazwy produktów dla różnych GTIN	Prowadzi do niejednoznaczności i błędów w identyfikacji produktów.	<ul style="list-style-type: none"> <li>▶ DRESS x 183372</li> <li>▶ System kominowy x 114064</li> <li>▶ Towar x 109999</li> </ul>	Używanie nazw powszechnych zamiast pełnej nazwy produktu	ŚREDNIA
18	Nazwa	Nazwa jest niezrozumiała	Utrudnia identyfikację produktu przez użytkowników i systemy, może prowadzić do niejasności co do funkcji	<ul style="list-style-type: none"> <li>▶ Towar x 109999</li> <li>▶ SZKL OF GL #6</li> <li>▶ Serduszka Haneczki</li> </ul>	Brak standardów nazewnictwa (organizacja nie posiada lub nie stosuje się do wytycznych)	—

Lp.	Atrybut danych	Wykryty problem	Uzasadnienie: Dlaczego jest to problem?	Rzeczywiste przykłady	Prawdopodobne źródło powstawania błędu	Szac. liczba danych
			lub przeznaczenia produktu. Niezrozumiałe nazwy mogą negatywnie wpływać na wyszukiwanie i klasyfikację, ograniczając efektywność zarządzania danymi.		dotyczących tworzenia nazw produktów, może dojść do tworzenia nazw niejasnych lub mylących), automatyczne tłumaczenie, nadmierne użycie żargonu branżowego lub technicznego.	
19	Język opisu ↔ Nazwa	Nazwa produktu w języku nieodpowiadającym wskazanemu językowi	Wprowadza niezgodność i może być myląca dla użytkowników oraz systemów przetwarzających dane, zwłaszcza w środowiskach wielojęzycznych. Może to utrudniać wyszukiwanie i klasyfikację produktów, a także obniżać jakość interakcji z użytkownikiem.	<ul style="list-style-type: none"> <li>▶ (20x85g)PL,LT,LV,EE - litewski</li> <li>▶ Kaczka Pekin z szyną i podrobami mrożona 1600g - niemiecki</li> <li>▶ KOSTIUM KĄPIELOWY FIONA KOLOR CZARNY ROZMIAR 42/85 F - rosyjski</li> <li>▶ Autositzbezüge Volkswagen LT 97-06 1+2 Vordersitze Schwarz Maß Sitzbezüge Auto - polski</li> </ul>	Błędna interpretacja pola rozumiana jako rynek docelowy, brak danych i ręczne uzupełnianie pola przez użytkownika, nieuwaga.	MAŁA
20	Nazwa, Marka	Nadmiarowe spacje w tekście	Utrudniają prawidłowe sortowanie, wyszukiwanie i mogą powodować problemy z formatowaniem.	<ul style="list-style-type: none"> <li>▶ SZKL OF GL #6</li> <li>▶ KURTKA NARCIARSKA DAMSKA COLD LIGHT GREY MELANGE S KUDN604 AW19</li> <li>▶ Plakat metalowy Mapa Navy Amsterdam M</li> <li>▶ R92616 __CZ38-DBOTEK</li> </ul>	Ręczne wprowadzanie danych z nieświadomym dodawaniem spacji lub automatyczna konwersja tabulatorów na spacje w procesie eksportu danych.	DUŻA
21	Nazwa, Marka	Encja HTML zamiast znaku specjalnego w tekście	Może być niepoprawnie wyświetlana w różnych interfejsach, wprowadzając nieścisłości i błędy wizualne.	<ul style="list-style-type: none"> <li>▶ "30 Stück Spanplattenschrauben 3 X 35/21 Kreuzschlitz POZI</li> <li>▶ ""Z"" (Teilgewinde), Senkkopf, Edelstahl A2"</li> </ul>	Automatyczne konwertowanie znaków specjalnych na encje HTML przez systemy CMS lub edytory tekstu wspierające formatowanie HTML.	ŚREDNIA
22	Nazwa, Marka	Podwójny znak cudzysłowu w tekście	Może powodować błędy w przetwarzaniu tekstu, np. w systemach baz danych lub interfejsach API.	<ul style="list-style-type: none"> <li>▶ Mandarynki "Primasol" 10 kg 55 + Turcja</li> <li>▶ "Nasadka do trymera FOX TANGO S do noża ""T"" 3mm"</li> <li>▶ Brelok ""beczka""</li> </ul>	Mechanizmy obrony przed atakami SQL lub błędne przetwarzanie ciągów znaków mogą powodować niezamierzone podwojenia cudzysłowów. Niektóre systemy eksportując dane tworzą podwójne cudzysłowy wewnątrz tekstu, aby odróżnić je od tych oznaczających początek i koniec ciągu.	BARDZO MAŁA

Lp.	Atrybut danych	Wykryty problem	Uzasadnienie: Dlaczego jest to problem?	Rzeczywiste przykłady	Prawdopodobne źródło powstawania błędu	Szac. liczba danych
23	Nazwa, Marka	Znak nowej linii w tekście jednolinijkowym	Powoduje problemy z formatowaniem i prezentacją danych w interfejsach użytkownika oraz raportach.	—	Nieintencjonalne wciśnięcie Enter podczas wprowadzania tekstu lub importowania plików CSV może być niewidoczne w niektórych programach, co sprawia, że użytkownicy nie zauważają tego błędu. Problem często występuje przy danych z Excela lub wprowadzanych na stronach internetowych.	MAŁA
24	Nazwa, Marka	Kod sterujący w tekście	Może być interpretowany przez systemy i aplikacje, prowadząc do nieoczekiwanych efektów lub błędów.	► □Keyring "Toruń" PL-11KR-007	Nieprawidłowe przetwarzanie lub czyszczenie danych wejściowych zewnętrznych systemów, pozwalające na wprowadzenie niepożądanych kodów sterujących.	BARDZO MAŁA
25	Nazwa, Marka	Znak tabulacji w tekście jednowierszowym	Podobnie jak znak nowej linii, wprowadza problemy z formatowaniem i prezentacją danych.	—	Niezamierzenie wprowadzone lub skopiowane przez użytkownika znaki tabulacji, które nie są widoczne w niektórych programach, np. Excel.	BARDZO MAŁA
26	Nazwa, Marka	Początek tekstu to cudzysłów lub apostrof	Może powodować błędy w przetwarzaniu tekstu i być źródłem niekonsekwencji w danych.	► 'InvisaFlex Łuki NiTi Full-Form okrągłe kosmetyczne góra .014"' (2 szt.)' ► "Serduszka Haneczki" Herbatniki w polewie kakaowej 1 kg ► "MALINKA " - fermentowany napój winopodobny,owocowy,aromatyzowany,	Sposób zapisu nazwy wyróżniający fragment nazwy własnej od opisu nazwy powszechnej.	ŚREDNIA
27	Nazwa, Marka	Formuła Excela zamiast treści	Oznacza brak właściwej wartości danych, co może prowadzić do błędów w przetwarzaniu i analizie.	► =VLOOKUP(A12,Arkusz1!A:B,2,0) ► =B918 VLOOKUP(M186,[1]Cennik!\$A\$2:\$B\$9249,2,FALSE)	Błąd przy eksporcie/importcie danych z Excela, gdzie formuły nie zostały przekonwertowane na wartości tekstowe.	BARDZO MAŁA
28	Nazwa, Marka	Nadmiarowe podkreślenia w tekście	Może negatywnie wpływać na czytelność tekstu i estetykę prezentacji danych.	► IS45201 __2241-MTRZEWIK PROFFLAM ANTISTATIC KURTKA TOOKU	Nieprawidłowe użycie znaków formatujących w próbie organizacji tekstu lub jako separatorów, nieprawidłowy eksport danych.	BARDZO MAŁA



Lp.	Atrybut danych	Wykryty problem	Uzasadnienie: Dlaczego jest to problem?	Rzeczywiste przykłady	Prawdopodobne źródło powstawania błędu	Szac. liczba danych
29	Nazwa, Marka	Niedozwolone znaki w tekście wynikające z błędnej konwersji znaków (lista)	Powodują problemy z czytelnością, mogą być źródłem błędów w przetwarzaniu danych.	<ul style="list-style-type: none"> <li>▶ OBRAZ PLAKAT W RAMIE 53x73 cm GÅŁy P958</li> <li>▶ DŁĄWIGNIE KLAMKI HAMULCA NEXELO ALUMINIUM NA 3 PALCE - SREBRNO-CZARNA</li> <li>▶ Odbojnica naroŁŁna z rÅŁŁnymi ramionami Łrednica 76,1 mm wysokoŁŁ 300 mm, dŁŁ, ugoŁŁ 3 ramion B=600 mm A=800 mm PRAWA</li> </ul>	Błędy eksportu danych z systemu użytkownika, głównie problem konwersji znaków diakrytycznych (polskich "ogonków")	BARDZO MAŁA
30	Nazwa, Marka, Jednostka	Spacja przed tekstem właściwym	Utrudnia właściwe sortowanie i wyszukiwanie, wprowadzając niekonsekwencje.	—	Niezamierzone dodawanie spacji przy ręcznym wprowadzaniu danych lub w wyniku kopiowania i wklejania tekstu z różnych źródeł.	MAŁA
31	Nazwa, Marka, Jednostka	Spacja po tekście właściwym	Podobnie jak spacja przed tekstem, wprowadza niekonsekwencje w danych.	—	Dodanie spacji na końcu tekstu podczas wprowadzania danych, często niezauważalne w interfejsie użytkownika.	BARDZO MAŁA
32	Zaw. netto	Przekroczona wielkość zawartości netto	Może wprowadzać konsumentów w błąd odnośnie do rzeczywistej zawartości produktu.	—	Błędna interpretacja standardów jednostek miar lub błędy przy wprowadzaniu danych.	BRAK
33	Zaw. netto	Zawartość netto $\leq 0$	Jest błędem krytycznym, uniemożliwiającym prawidłową identyfikację i klasyfikację produktów na podstawie ich zawartości.	<ul style="list-style-type: none"> <li>▶ Bluzka Dziewczęca Nike 116-146 0.000 szt</li> <li>▶ Karta Sta?ego Klienta 0.000 szt</li> </ul>	Nieuwaga przy wprowadzaniu danych	MAŁA
34	Zaw. netto	Niecałkowita liczba sztuk, par, egzemplarzy	Może prowadzić do nieporozumień i błędów w zamówieniach, szczególnie w przypadkach, gdy oczekuje się określonej liczby produktów. Niejasności w ilości mogą wpływać negatywnie na satysfakcję klienta i procesy logistyczne.	<ul style="list-style-type: none"> <li>▶ Kubek - 1940Harley Davidson Knucklehead 0.280 szt</li> <li>▶ 5523 Kalka kreślarska ROLKA 0,33 x20m 50 g/m2 15.130 szt</li> <li>▶ Urządzenie kryptograficzne CM-107E 205543.250 szt</li> </ul>	Nieprawidłowe zrozumienie wymagań dotyczących jednostek liczebnych w specyfikacjach produktów.	ŚREDNIA
35	Zaw. netto (jednostka) ↔ Nazwa lub GPC	Nieadekwatna jednostka zawartości netto w odniesieniu do kategorii/grupy produktów	Powoduje zamieszanie i może być myląca dla konsumentów, utrudniając porównywanie produktów. Nieprawidłowa jednostka może również wpływać na precyzję i skuteczność	<ul style="list-style-type: none"> <li>▶ Orzechy/nasiona - nieprzetworzone (łatwo psujące się) 17 m</li> <li>▶ Wyroby cukiernicze cukrowe/z substytutów cukru 3 m</li> </ul>	Nieuwaga w przyporządkowaniu jednostek miar, lub interpretacji atrybutu zawartości netto produktu jako danej logistycznej opakowania.	—

Lp.	Atrybut danych	Wykryty problem	Uzasadnienie: Dlaczego jest to problem?	Rzeczywiste przykłady	Prawdopodobne źródło powstawania błędu	Szac. liczba danych
			procesów magazynowych oraz logistycznych.	<ul style="list-style-type: none"> <li>▶ Sosy do gotowania (trwale) 1 m</li> <li>▶ Odzież sportowa - nakrycia głowy 422 kg</li> <li>▶ Dorsz atlantycki filet bez skóry 1g</li> </ul>		
36	Zaw. netto ↔ Nazwa lub GPC	Nieadekwatna zawartość netto w odniesieniu do kategorii/grupy produktów	Może wprowadzać w błąd co do ilości lub wielkości produktu, co jest szczególnie problematyczne w kategoriach, gdzie zawartość netto ma kluczowe znaczenie dla oceny wartości produktu (np. żywność, kosmetyki).		Nieuwaga przy wprowadzaniu danych	—
37	Zdjęcie	Przekroczona wielkość pliku	Może powodować problemy z wydajnością systemu i dostępnością danych, szczególnie w środowiskach o ograniczonej przepustowości.	—	Nieświadomość standardów branżowych dotyczących jakości zdjęć produktów w handlu elektronicznym.	MAŁA
38	Zdjęcie	To samo zdjęcie dla różnych produktów	Wprowadza w błąd konsumentów i utrudnia prawidłową identyfikację produktów.	—	Nieuwaga lub celowe działanie w wyniku braku właściwych danych	—
39	Zdjęcie	Zbyt niska rozdzielczość zdjęcia dla większości zastosowań e-commerce	Utrudnia właściwą prezentację produktu, wpływając negatywnie na jego atrakcyjność i rozpoznawalność. Zdjęcia o niskiej rozdzielczości mogą nie spełniać wymogów platform e-commerce, co ogranicza możliwości sprzedaży.	—	Nieświadomość standardów branżowych dotyczących jakości zdjęć produktów w handlu elektronicznym.	—
40	Zdjęcie	Zbyt niska rozdzielczość zdjęcia	Utrudnia właściwą prezentację produktu, może wpływać na jego atrakcyjność i rozpoznawalność w e-commerce.	—	Nieświadomość standardów branżowych dotyczących jakości zdjęć produktów w handlu elektronicznym.	—
41	Zdjęcie	Zbyt duża rozdzielczość zdjęcia	Może powodować problemy z wydajnością ładowania strony i dostępnością danych w środowiskach o ograniczonej przepustowości.	—	Nieświadomość standardów branżowych dotyczących jakości zdjęć produktów w handlu elektronicznym.	—
42	Zdjęcie	Zdjęcie produktu na niejednorodnym lub kolorowym tle	Utrudnia automatyczną obróbkę i jednolitą prezentację w katalogach, potencjalnie wpływając na estetykę i spójność prezentacji.	—	Ignorowanie lub nieświadomość wytycznych dotyczących prezentacji produktów, które zalecają jednolite tło dla zdjęć.	ŚREDNIA

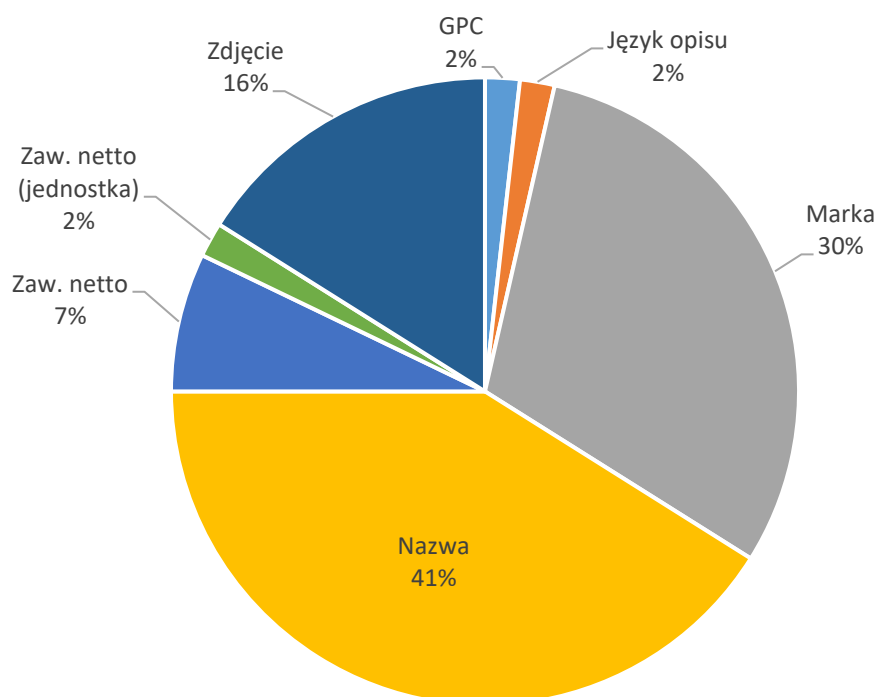
Lp.	Atrybut danych	Wykryty problem	Uzasadnienie: Dlaczego jest to problem?	Rzeczywiste przykłady	Prawdopodobne źródło powstawania błędu	Szac. liczba danych
43	Zdjęcie	Zdjęcie produktu na zbyt dużej powierzchni tła	Wpływa na estetykę i spójność prezentacji.	—	Ignorowanie lub nieświadomość wytycznych dotyczących prezentacji produktów, które zalecają jednolite tło dla zdjęć.	ŚREDNIA
44	Zdjęcie	Zbyt niska rozdzielczość zdjęcia dla niektórych zastosowań e-commerce	Może nie spełniać specyficznych wymagań niektórych platform lub kanałów dystrybucji, co ogranicza zasięg i efektywność promocji produktu. Niedostosowanie do standardów może również wpływać na postrzeganie marki jako mniej profesjonalnej.	—	Nieświadomość standardów branżowych dotyczących jakości zdjęć produktów w handlu elektronicznym.	—
45	Zdjęcie	Zdjęcie nieprzedstawiające produktu do sprzedaży	Wprowadza w błąd konsumentów, może być przyczyną niezadowolenia i reklamacji.	—	Nieświadomość standardów branżowych dotyczących jakości zdjęć produktów w handlu elektronicznym lub celowe działanie w wyniku braku właściwych danych	—

Źródło: opracowanie własne

W kolumnach Tabeli 24 zestawiono informacje o atrybutach, do których się odnoszą, opisie konkretnego problemu wraz z uzasadnieniem, przykładach, prawdopodobnego źródła powstawania błędu oraz szacowanej liczby danych, których może dotyczyć problem. W pierwszej kolumnie symbolem „↔” oznaczono zależności pomiędzy atrybutami. Znak „-” w przykładach wskazuje na brak konieczności lub możliwości wizualizacji problemu, które można umieścić w takim zestawieniu. Skala szacowania liczby danych nie odzwierciedla bezpośrednio wagi problemu, ale potencjalny wpływ na dane. Składa się z kategorii:

- A. **Bardzo mała** (< 30 tys.): Dotyczy zbiorów z mniej niż 30 tys. rekordów.
- B. **Mała** (>= 30 tys. - < 300 tys.): Zbiory od 30 tys. do poniżej 300 tys. rekordów.
- C. **Średnia** (>= 300 tys. - < 3 mln): Od 300 tys. do poniżej 3 mln rekordów.
- D. **Duża** (>= 3 mln - < 15 mln): Zbiory od 3 do poniżej 15 mln rekordów.
- E. **Bardzo duża** (>= 15 mln): Ponad 15 mln rekordów.
- F. — (Trudne do oceny na wstępnym etapie): Brak możliwości oszacowania.

Na podstawie przeprowadzonej analizy zidentyfikowano łącznie 45 specyficznych problemów jakościowych. Na wykresie kołowym (Rysunek 13) przedstawiono je w podziale na poszczególne atrybuty. Ponieważ niektóre z problemów dotyczyły kilku atrybutów, łącznie wykres zawiera 56 elementów. Znaczną część problemów, dokładnie 23, stanowiły kwestie związane z nazwą produktu. Dalej, 17 dotyczyło marek, następnie zdjęć produktów oraz zawartości netto.



Rysunek 13. Procentowy rozkład wykrytych problemów  
 Źródło: opracowanie własne

Analizując zestawienie zauważonych problemów jakościowych można zwrócić uwagę na częste występowanie zależności pomiędzy polami danych. Problemy, takie jak nieprawidłowa kategoryzacja produktów, nieprawidłowy język opisu, nieadekwatne zawartości netto, czy zdjęcia nieprzedstawiające produktu, podkreślają złożoność i potencjalne pułapki, które nie sposób wychwycić bez holistycznego podejścia do całego zbioru danych. W obszarze problemów wynikających jedynie z relacji między atrybutami, uporządkowano je i przedstawiono w Tabeli 25.

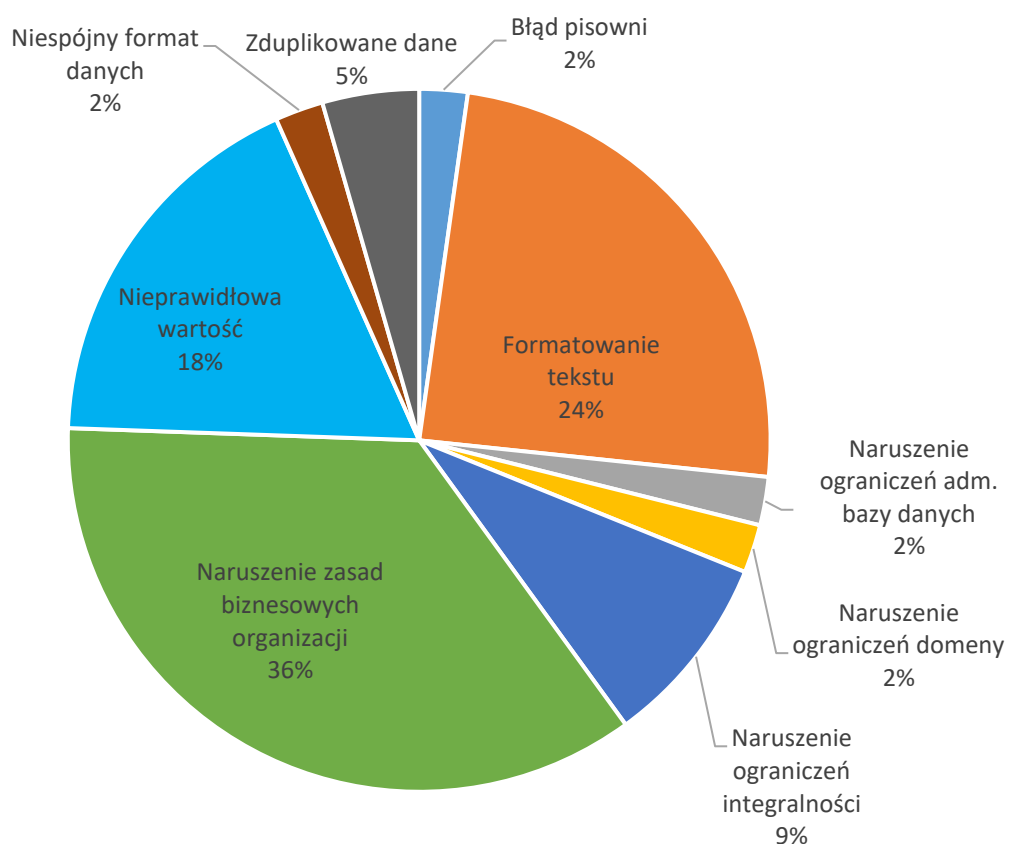
Tabela 25. Zidentyfikowane zależności między atrybutami danych.

Opis problemu	Liczba problemów
GPC ↔ Nazwa	1
Marka ↔ Nazwa	2
Język opisu ↔ Nazwa	1
Zaw. netto (jednostka) ↔ Nazwa	1
Zaw. netto ↔ Nazwa	1
GTIN ↔ Nazwa	1
Nazwa ↔ Zaw. netto + jednostka	1

Źródło: opracowanie własne

Liczba wykrytych problemów wynikających z relacji stanowi aż 14% wszystkich wykrytych problemów jakościowych. Interesujący jest również fakt, że wszystkie posiadają zależność od nazwy produktu.

Wykryte problemy zostały także sklasyfikowane zgodnie z systematyką przedstawioną w Rozdziale 2.7, a ich wyniki zaprezentowano na wykresie kołowym, widocznym na Rysunku 14. Dzięki analizie wykazano, że problemy dotyczyły różnych aspektów, takich jak błąd pisowni, formatowanie tekstu, naruszenie ograniczeń administracyjnych bazy danych, naruszenie ograniczeń domeny, naruszenie ograniczeń integralności, naruszenie zasad biznesowych organizacji, nieprawidłowa wartość, niespójny format danych oraz zduplikowane dane. Najwięcej wyzwań stwarzały naruszenia zasad biznesowych organizacji, z 16. zidentyfikowanymi przypadkami. Liczba problemów związanych z formatowaniem tekstu wyniosła 11, co stanowiło drugą co do wielkości, znaczną część wykrytych problemów, podczas gdy nieprawidłowa wartość została zauważona 8 razy. Naruszenia ograniczeń integralności zarejestrowano w czterech przypadkach. Ogółem problemy zagregowano w 9 różnych grup, co podkreśla różnorodność i złożoność problemów z danymi, z jakimi organizacje mogą się spotkać.



Rysunek 14. Klasyfikacja wykrytych problemów  
Źródło: opracowanie własne

Szczególną uwagę zwracają sklasyfikowane problemy, które pozornie mogą być uznane za techniczne – dotyczące liczby liter w tekście, liczby słów, liczby wyrazów itd. Te problemy nie odnoszą się ani do domeny, ani bezpośrednio do treści w rozumieniu znaczenia zawartości danego pola, natomiast, co ciekawe, w podanych przykładach skutecznie ukazują anomalie, przy jednoczesnej prostocie samego badania. Ich rozwiązanie w postaci implementacji precyzyjnie dopasowanych algorytmicznie walidatorów może okazać się efektywne w wychwytywaniu błędów, pomyłek oraz kontroli standardów i unifikacji.

W badaniu przedstawiono rzeczywiste przykłady problemów z danymi produktowymi, które zostały zidentyfikowane podczas analizy danych. Niektóre z tych problemów mogłyby wydawać się niemożliwe do zaistnienia w rzeczywistych warunkach, jednak skala i społecznościowy charakter bazy, w której autorami danych są osoby o różnym poziomie przeszkolenia i wiedzy na temat danych produktowych przyczyniają się do częstego powstawania takich anomalii. To różnorodne tło użytkowników skutkuje zmienną jakością wprowadzanych danych, co z kolei podkreśla potrzebę opracowania wysoko zaawansowanych narzędzi do walidacji i standaryzacji danych, aby zapewnić ich spójność i dokładność w całej bazie.

W Tabeli 26 przedstawiono kompleksową ocenę możliwości utworzenia algorytmów dla wykrytych problemów. Powstała ona w celu oceny potencjalnej skuteczności i wykonalności algorytmów służących do poprawy jakości danych, oferując jednocześnie wgląd w możliwe wyzwania i ograniczenia związane z ich implementacją.

Tabela 26. Ocena możliwości utworzenia algorytmów w celu eliminacji wykrytych problemów

Lp.	Wykryty problem	Prawdopodobna złożoność algorytmu	Czy algorytm może korygować błędy bez ryzyka utraty treści istotnej?	Czy istnieje ryzyko, że algorytm wygeneruje fałszywie pozytywny wynik?	Czy dane można warunkowo zaakceptować, jeśli błąd nie zostanie usunięty?	Czy algorytm może podpowiadać prawidłową / lepszą wartość?
1	Niepoprawna lub tymczasowa kategoria produktu	Wysoka	NIE	TAK	TAK	TAK
2	Marka produktu z nadmierną ilością słów	Niska	NIE	NIE	NIE	NIE
3	Nazwa marki bez liter	Niska	NIE	NIE	NIE	NIE
4	Nadmiernie długa nazwa marki	Niska	NIE	NIE	NIE	NIE
5	Różne zapisy tej samej marki	Wysoka	NIE	TAK	TAK	TAK
6	Niepoprawne nazwy marek (lista)	Niska	NIE	NIE	NIE	NIE
7	Identyczna nazwa produktu i marki	Niska	NIE	NIE	NIE	NIE
8	Początek tekstu to nieprawidłowy znak	Niska	NIE	NIE	NIE	NIE
9	Nadmiernie długa nazwa produktu	Niska	NIE	NIE	NIE	NIE
10	Nadmiernie długa nazwa produktu do niektórych zastosowań e-commerce	Niska	NIE	NIE	TAK	NIE
11	Nazwa produktu z nadmierną ilością słów	Niska	NIE	NIE	NIE	NIE
12	Produkt z nazwą jednowyrazową	Niska	NIE	NIE	NIE	NIE
13	Brak wskazania zawartości netto w nazwie produktu	Niska	NIE	NIE	NIE	NIE
14	Brak marki w nazwie produktu	Niska	NIE	NIE	NIE	NIE
15	Nazwa produktu pisana wyłącznie wielkimi literami	Niska	NIE	NIE	TAK	TAK
16	Brak słów w nazwie produktu	Niska	NIE	NIE	NIE	NIE
17	Powtarzające się nazwy produktów dla różnych GTIN	Niska	NIE	NIE	NIE	NIE
18	Nazwa jest niezrozumiała	Wysoka	NIE	TAK	TAK	TAK
19	Nazwa produktu w języku nieodpowiadającym wskazanemu językowi	Wysoka	NIE	TAK	TAK	TAK
20	Nadmiarowe spacje w tekście	Niska	TAK	NIE	-	-
21	Encja HTML zamiast znaku specjalnego w tekście	Niska	TAK	NIE	-	-
22	Podwójny znak cudzysłowu w tekście	Niska	TAK	NIE	-	-
23	Znak nowej linii w tekście jednolinijkowym	Niska	TAK	NIE	-	-
24	Kod sterujący w tekście	Niska	TAK	NIE	-	-



Lp.	Wykryty problem	Prawdopodobna złożoność algorytmu	Czy algorytm może korygować błędy bez ryzyka utraty treści istotnej?	Czy istnieje ryzyko, że algorytm wygeneruje fałszywie pozytywny wynik?	Czy dane można warunkowo zaakceptować, jeśli błąd nie zostanie usunięty?	Czy algorytm może podpowiadać prawidłową / lepszą wartość?
25	Znak tabulacji w tekście jednowierszowym	Niska	TAK	NIE	-	-
26	Początek tekstu to cudzysłów lub apostrof	Niska	TAK	NIE	-	-
27	Formuła Excela zamiast treści	Niska	NIE	NIE	NIE	NIE
28	Nadmiarowe podkreślenia w tekście	Niska	TAK	NIE	-	-
29	Niedozwolone znaki w tekście wynikające z błędnej konwersji znaków (lista)	Niska	NIE	NIE	NIE	NIE
30	Spacja przed tekstem właściwym	Niska	TAK	NIE	-	-
31	Spacja po tekście właściwym	Niska	TAK	NIE	-	-
32	Przekroczona wielkość zawartości netto	Niska	NIE	NIE	NIE	NIE
33	Zawartość netto $\leq 0$	Niska	NIE	NIE	NIE	NIE
34	Niecałkowita liczba sztuk, par, egzemplarzy	Niska	NIE	NIE	NIE	NIE
35	Nieadekwatna jednostka zaw. netto w odniesieniu do kategorii/grupy produktów	Wysoka	NIE	TAK	TAK	TAK
36	Nieadekwatna zawartość netto w odniesieniu do kategorii/grupy produktów	Wysoka	NIE	TAK	TAK	TAK
37	Przekroczona wielkość pliku	Niska	NIE	NIE	NIE	NIE
38	To samo zdjęcie dla różnych produktów	Wysoka	NIE	TAK	TAK	NIE
39	Zbyt niska rozdzielczość zdjęcia dla większości zastosowań e-commerce	Niska	NIE	NIE	TAK	NIE
40	Zbyt niska rozdzielczość zdjęcia	Niska	NIE	NIE	NIE	NIE
41	Zbyt duża rozdzielczość zdjęcia	Niska	NIE	NIE	NIE	NIE
42	Zdjęcie produktu na niejednorodnym lub kolorowym tle	Wysoka	NIE	TAK	TAK	TAK
43	Zdjęcie produktu na zbyt dużej powierzchni tła	Wysoka	NIE	TAK	TAK	TAK
44	Zbyt niska rozdzielczość zdjęcia dla niektórych zastosowań e-commerce	Niska	NIE	NIE	TAK	NIE
45	Zdjęcie nieprzedstawiające produktu do sprzedaży	Wysoka	NIE	TAK	TAK	NIE

Źródło: opracowanie własne

Struktura Tabeli 26 organizuje dane w kolumny, które zawierają następujące informacje:

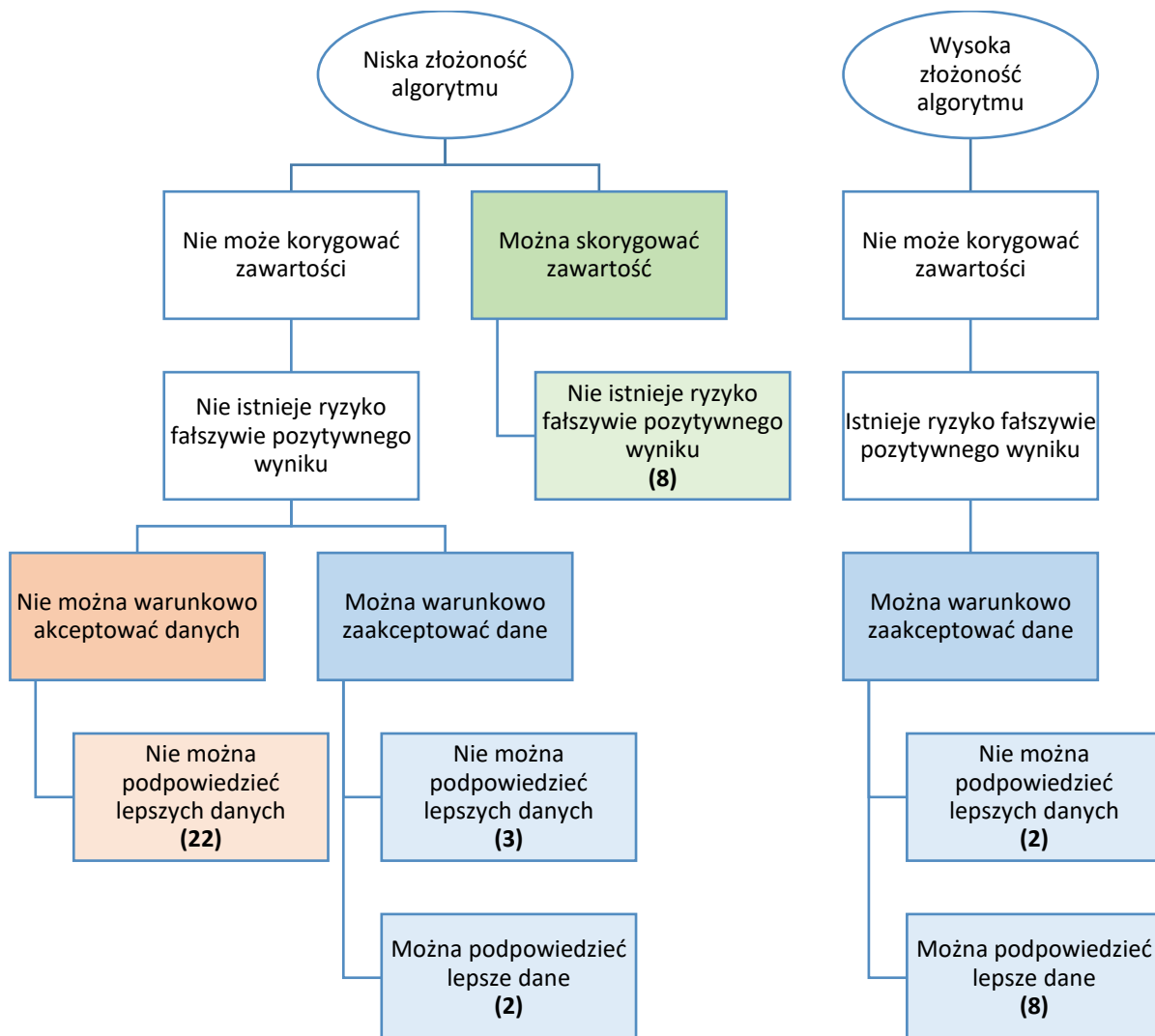
- A. **Wykryty problem jakościowy** – przedstawiono problemy zidentyfikowane podczas analizy danych, zgodnie z kolejnością z Tabeli 24.
- B. **Prawdopodobna złożoność algorytmu (Niska/Wysoka)** – oszacowano poziom trudności w opracowaniu algorytmów przeznaczonych do rozwiązania danego problemu.
- C. **Możliwość korygowania błędów bez ryzyka utraty istotnych informacji (Tak/Nie)** – określono, czy algorytm może naprawić błędy bez utraty kluczowych danych.
- D. **Ryzyko wygenerowania fałszywie pozytywnego wyniku przez algorytm (Tak/Nie)** – rozważono potencjalne ryzyko błędnej interpretacji danych, skutkującej nieprawidłowym oznaczeniem poprawnych danych jako błędnych.
- E. **Możliwość warunkowego zaakceptowania danych, jeśli błąd nie zostanie usunięty (Tak/Nie/—)** – oceniono możliwość tymczasowego zaakceptowania danych mimo wykrycia błędów, o ile nie wpływają one znacząco na ogólną jakość danych. Znak „—” oznacza, że pytanie nie dotyczy danej kategorii.
- F. **Możliwość podpowiadania prawidłowej/lepszej wartości przez algorytm (Tak/Nie/—)** – określono, czy algorytm mógłby sugerować bardziej odpowiednie wartości w przypadku wykrycia błędów, co mogłoby przyspieszyć korektę danych. Znak „—” oznacza, że pytanie nie ma zastosowania w danym wierszu.

Algorytmy o niskiej złożoności obejmują łącznie 34 problemy, które można efektywnie rozwiązać przy użyciu prostych i łatwych do wdrożenia reguł lub wzorców. Te problemy nie wymagają głębokiego zrozumienia kontekstu ani złożonych procesów decyzyjnych. Do typowych przykładów należą błędy pisowni, problemy z formatowaniem oraz naruszenia standardów danych. Algorytmy przeznaczone do ich rozwiązania charakteryzują się prostotą implementacji i minimalnym ryzykiem generowania fałszywie pozytywnych wyników, jeśli w ogóle można je rozpatrywać w taki sposób. Na przykład, kwestie takie jak nadmierna liczba słów w nazwie produktu, brak liter w nazwie marki, zbyt długie nazwy marek, czy niepoprawne nazwy marek mogą być efektywnie weryfikowane i korygowane za pomocą prostych funkcji kontrolujących długość ciągu znaków, występowanie określonych znaków lub porównania z predefiniowaną listą akceptowalnych wartości.

Algorytmy wysokiej złożoności, rozwiązujące 9 pozostałych problemów, wiążą się z potrzebą zaawansowanej analizy semantycznej, interpretacji kontekstu lub rozpoznawania wzorców w danych, które nie są jednoznaczne. **To problematyka, w której mogą się sprawdzić algorytmy sztucznej inteligencji.** Przykładem jest rozwiązanie problemu kategoryzacji produktów jako "tymczasowych" przy istniejącej odpowiedniej taksonomii, a który wymaga przetworzenia znaczenia i kontekstu nazw produktów, co z kolei wymusza zastosowanie zaawansowanych technik opartych o rozumienie języka. Różne zapisy tej samej marki lub nazwa produktu w języku nieodpowiadającym wskazanemu językowi również wymagają głębokiej analizy i interpretacji, znamiennej dla złożonych algorytmów sztucznej inteligencji. Podobnie nieadekwatność jednostki lub zawartości netto w odniesieniu do kategorii produktów, jak również problem z tym samym zdjęciem dla różnych produktów lub zdjęciem na niejednorodnym tle, są przykładami, przy których potrzebne jest zrozumienie kontekstu kategorii produktowej, analiza wizualna obrazów oraz potencjalnie złożone algorytmy dopasowania i klasyfikacji.

Na Rysunku 15 przedstawiono diagram ilustrujący przejścia między pytaniami, uwzględniający liczbę znalezionych problemów dla każdego z przejść. Dla problemów o niskiej złożoności algorytmicznej zidentyfikowano trzy typy przejść. Pierwszy dotyczy korekty danych bez ryzyka utraty istotnych informacji, reprezentując grupę normalizatorów danych. W tym przypadku, po korekcie, dane mogą zostać automatycznie zaakceptowane przez system bez potrzeby interwencji użytkownika. Drugie przejście odnosi się do walidacji danych, gdzie nie ma możliwości ich korekty - oznacza to, że dane są permanentnie błędne i nie mogą być ani warunkowo zaakceptowane, ani poprawione, podobnie nie można uzyskać żadnej propozycji podpowiedzi ze strony systemu. Trzecie przejście opisuje sytuacje, w których dane są błędne, ale akceptowalne; w jednym przypadku możliwe jest zasugerowanie poprawnych danych właściwej nazwy produktu bez użycia wielkich liter – w dwóch pozostałych, związanych z niedostatecznym rozmiarem zdjęć produktów nie jest dopuszczalne. Z kolei dla bardziej skomplikowanych algorytmów obserwujemy tylko jeden główny typ przejścia. W tej kategorii, wszystkie dane są akceptowane pomimo występowania błędów. Występuje ryzyko naruszenia znaczenia danych w przypadku autokorekty, a więc ryzyko interwencji w obszar, za który odpowiedzialność ponosi osoba wprowadzająca dane, dlatego danych nie można modyfikować. Istnieje też potencjalne ryzyko wygenerowania fałszywie pozytywnych wyników walidacji, dlatego do działania tych walidatorów należy podchodzić z zachowaniem szczególnej ostrożności. Mimo tych ograniczeń, algorytmy te, z wyjątkiem problemów

związanych ze zdjęciami produktów, mogą skutecznie zaproponować lepsze rozwiązania, a także pełnić rolę systemu ostrzeżenia przez ewentualną nieprawidłowością.



Rysunek 15. Diagram podziału zidentyfikowanych problemów z jakością danych produktowych  
 Źródło: opracowanie własne

### 5.3. Wnioski

Podsumowując badania, można zauważyć, że dla części problemów o niskiej złożoności algorytmicznej istnieje możliwość automatycznego przyjęcia danych po korekcie przez normalizatory danych, bez potrzeby interwencji użytkownika. Jest to optymalne rozwiązanie, gdy dane można skorygować bez ryzyka utraty ważnych informacji. Jednakże, w przypadkach,

gdy dane są permanentnie błędne i nie ma możliwości ich korekty, system z reguły nie oferuje żadnych propozycji podpowiedzi, co powoduje potrzebę manualnej interwencji lub odrzucenia danych. Dla bardziej złożonych problemów, z reguły tych, dotyczących zależności pomiędzy polami lub analizy semantycznej zwykle występuje możliwość autokorekty, ale istnieje ryzyko naruszenia integralności danych, co może prowadzić do nieprawidłowości. Co więcej, istnieje niepewność fałszywie pozytywnych wyników walidacji. Mimo to, algorytmy te posiadają zdolność do proponowania lepszych rozwiązań i mogą pełnić funkcję systemu ostrzegawczego przed ewentualnymi nieprawidłowościami. Takie zwrócenie uwagi na potencjalny błąd i sugerowanie zmiany na etapie wprowadzania danych przez użytkownika mogą skutecznie wpłynąć na poprawę jakości danych.

Przedstawione wnioski stały się podstawą do utworzenia modelu walidacji w niniejszej dysertacji.

## **6. Opracowanie modelu walidacji danych produktowych**

Wyniki dotychczasowych analiz stanowiły podstawę do opracowania modelu poprawy jakości danych produktowych, w którym kluczową rolę odgrywa zastosowanie sztucznej inteligencji (SI). Włączenie sztucznej inteligencji do modelu umożliwiło dodatkową automatyzację oraz zaawansowaną analizę semantyczną, co przełożyło się na bardziej efektywne działania. Głównym celem modelu było stworzenie rozwiązań umożliwiających skuteczną kontrolę danych produktowych wprowadzanych do bazy przy minimalnym udziale osób weryfikujących, a jednocześnie zapewnienie użytkownikowi szybkiej informacji zwrotnej dotyczącej nieprawidłowości. Jest to szczególnie istotne w przypadku dużych wolumenów danych, wprowadzanych przez osoby bez odpowiedniego przeszkolenia lub pochodzących z niepewnych źródeł, takich jak bazy danych tworzone przez społeczności, zrzeszone firmy czy platformy typu marketplace. W tym kontekście SI pełni funkcję narzędzia wspomagającego, dostarczając użytkownikom informacji dotyczących potencjalnych błędów, wyjaśnień, wskazówek co do ich poprawy oraz sugestii konkretnych zmian. Szczególnie zastosowanie generatywnej SI, która generuje nowe treści w oparciu o analizę danych i odpowiednio sformułowane zapytania, otwiera nowe możliwości w zakresie automatyzacji procesów komunikacyjnych oraz kontroli jakości danych. Jak wskazują badania Wach, Ejdyś i in. (2023), generatywna SI, taka jak ChatGPT, ma bardzo duży potencjał do wykorzystania w różnych sektorach i gałęziach gospodarki, ale niesie również pewne ryzyka. Istnieje ryzyko związane z niską jakością generowanych treści, dezinformacją oraz algorytmicznymi uprzedzeniami, które mogą negatywnie wpływać na wynik działania i proces decyzyjny. W związku z tym, w proponowanym modelu, generatywna SI została użyta w warstwie (poziomie walidacji), która nie musi modyfikować danych produktowych bezpośrednio, ale może jedynie sugerować poprawki, pozostawiając użytkownikowi lub administratorowi decyzję o ich wdrożeniu, co minimalizuje ryzyko niepożądanych zmian i zachowuje odpowiedzialność po stronie człowieka, nie maszyny.

### **6.1. Idea segmentacji względem podejmowanych działań**

W implementacji metod automatycznej walidacji kluczowe jest zachowanie odpowiedniej kolejności działań oraz zidentyfikowanie grup mechanizmów, które mogą zostać wykorzystane w zależności od skomplikowania algorytmów i ich wpływu na dane. Poszczególne mechanizmy, szczegółowo opisane w rozdziale 5, podzielono na trzy główne obszary na podstawie wyników walidacji: wpływu na dane lub na działanie użytkownika, który

je wprowadza. Ideą było rozszerzenie formy klasycznej walidacji prowadzącej do odrzucania danych, na formę, która w wyniku uruchomienia sugeruje działania i edukuje, a także automatycznie proponuje korektę danych o ile jest to możliwe i dopuszczalne w kontekście odpowiedzialności za dane.

Tradycyjne walidatory, które tylko odrzucają dane, mogą zniechęcać użytkowników do dalszych działań. Ponadto muszą być deterministyczne i pewne, co oznacza, że powinny dawać zawsze jednoznaczne i dobre wyniki. W przeciwnym razie wymagają zaangażowania administratora bazy danych w proces dodatkowej akceptacji specyficznych danych lub stosowania wyjątków, gdy dane są prawidłowe, ale system je blokuje. Tego typu walidatory muszą więc działać bezbłędnie, co ogranicza ich elastyczność, zwłaszcza przy wdrażaniu bardziej zaawansowanych technik, takich jak sztuczna inteligencja. Bez wykorzystania technik SI, systemy walidacji nie mogą efektywnie przewidywać potencjalnych błędów, co z kolei prowadzi do ograniczeń w wychwytywaniu błędów i ogólnie pojętym doskonaleniu danych. Z tego powodu do modelu dodano trzeci poziom walidacji – sugestię poprawy jakości. Podział działań normalizacyjno-walidacyjnych uwzględnia zatem następujące aspekty:

- A. **Normalizację danych**, gdzie dane są porządkowane i korygowane edycyjnie bez zmian w znaczeniu zawartości danego atrybutu czy pola. Normalizacja może obejmować:
  - a. **Czyszczenie danych** - usunięcie lub zamianę specyficznych znaków (np. zbędnych spacji, tabulacji, nowych linii)
  - b. **Konwersję i dekodowanie** - (np. dekodowanie encji HTML na postać UTF-8, konwersję jednostek)
  - c. **Standaryzację formatowania** (np. standaryzację numeru GTIN do postaci 14 cyfrowej, formatowanie dat do jednolitego formatu YYYY-MM-DD)
  - d. **Harmonizację terminologii** (np. ujednoczenie terminów poprzez zamianę synonimów i skrótów na standardowe określenia, np. "kilo" na "kg")
- B. **Walidację techniczno – logiczną**, gdzie dane są poddawane dogłębnej analizie pod kątem ich technicznej poprawności oraz logicznej spójności. Obejmuje to dwa główne podobszary:
  - a. **Walidację błędów podstawowych**, obejmującą weryfikację danych pod względem poprawności zgodnie z przyjętymi standardami i normami. Z reguły są to bezsporne błędy techniczne, takie jak niepoprawne formaty kodów numerycznych, daty wykraczające poza ustalone zakresy czasowe, niezgodności w systemach jednostek miar czy naruszenia schematów zdefiniowanych przez

formalne specyfikacje, takie jak schematy XML lub JSON. Weryfikacja ta może również obejmować kontrolę poprawności adresów URL, identyfikatorów GTIN produktów, a także sprawdzanie czy wartości numeryczne nie przekraczają zdefiniowanych ograniczeń lub czy ciągi znaków nie zawierają niedozwolonych symboli.

- b. Walidację błędów wynikających z określonych zależności**, wymagającą informacji z innych pól lub źródeł danych do oceny poprawności. Są to zazwyczaj błędy logiczne, na przykład czy dane w jednym polu są zgodne z danymi w innym polu, czy dla danej kategorii produktu może być przypisana dana jednostka netto, bezsprzeczne niezgodności między klasyfikacją produktu a jego opisem, niekonsekwencje w zestawach atrybutów dla produktów z tej samej kategorii, czy też nieprawidłowe powiązania między produktami a ich akcesoriami lub częściami zamiennymi. Inne błędy logiczne mogą obejmować sprzeczne informacje o dostępności produktu w magazynie w stosunku do danych o jego sprzedaży, czy też niespójności między krajem pochodzenia a przypisanymi taryfami celnymi. Walidacja w tym obszarze może również dotyczyć sprawdzania reguł biznesowych, takich jak zgodność cen produktów z określonymi schematami cenowymi lub zasadami promocji.
- C. Sugestię poprawy jakości**, rozumianą jako proces, który może nie dawać pewności co do powstałego błędu lub daje taką pewność, ale jest warunkowo dopuszczany, jeśli jej nie spełnia. System może wskazać prawdopodobieństwo błędu i dać sugestię osobie wprowadzającej dane lub przekazać informację kontrolerowi, a taka walidacja nie zablokuje wprowadzenia danych. W tej przestrzeni, oprócz klasycznych technik walidacji, znajduje się miejsce na walidację danych opartą na statystyce, heurystyce i sztucznej inteligencji. Przykładowo, algorytmy uczenia maszynowego mogą być wykorzystane do identyfikacji nietypowych wzorców danych, które mogą sugerować potencjalne błędy, takie jak anomalie w cenach lub w opisach produktów, które odbiegają od normy dla danej kategorii. Może to również obejmować systemy rekomendacji, które na podstawie historii poprawek danych sugerują najbardziej prawdopodobne wartości dla nowo wprowadzanych informacji, co pozwala na szybszą i bardziej precyzyjną korektę. W tej przestrzeni można wyróżnić podobszary:
  - a. Walidacji stawiającej ograniczenia** – gdy powyżej/poniżej danego progu dane nie będą mogły być użyte do pewnych rozwiązań (np. zbyt małe zdjęcie do zastosowań e-commerce)

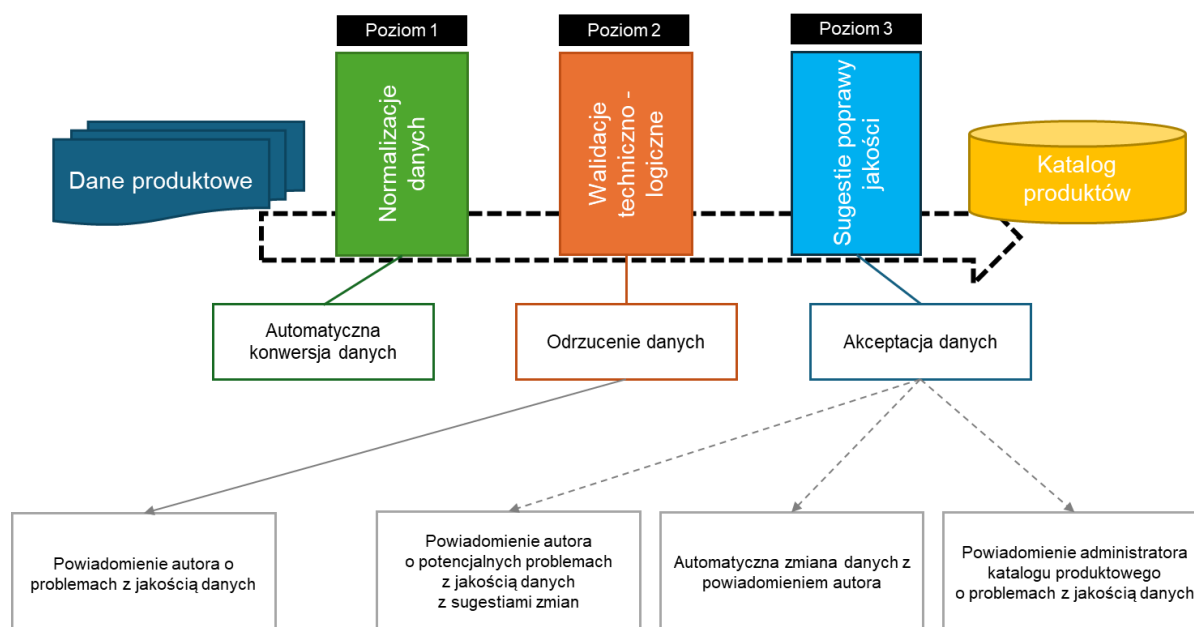


- b. Walidacji dokonującej predykcji** – gdy rozwiązanie dokonuje własnych obliczeń i przyrównuje je do danych wprowadzonych przez użytkownika (np. klasyfikacja produktu)
- c. Oceny zawartości** – gdy na podstawie danych algorytm dokonuje własnej oceny jakości (np. czy nazwa jest zrozumiała)

W kontekście wymienionych obszarów: normalizacji, walidacji techniczno - logicznej i sugestii poprawy jakości, ważne jest zachowanie kolejności działań, które w wielu wypadkach powinny być sekwencyjne, aby zapewnić maksymalną efektywność procesu walidacji danych, tworząc tzw. potoki (z ang. pipelines).

## 6.2. Omówienie proponowanego modelu

Wielopoziomowy model walidacji przedstawiono na Rysunku 16.



Rysunek 16. Struktura wielopoziomowego modelu walidacji danych produktowych  
 Źródło: opracowanie własne

Zaproponowany model skoncentrowany jest na problematyce automatycznej walidacji danych i nie zawiera pozostałych aspektów kontroli jakości. Wizualizuje proces akceptacji danych pochodzących ze źródła (lewa strona) do bazy danych (prawa strona). Rozpoczynając od normalizacji (Poziom 1), dane przygotowywane są do dalszych etapów walidacji,

zapewniając jednolitość i usuwając najbardziej oczywiste błędy formatowania, które mogłyby wpłynąć na rezultaty późniejszych kontroli. Sam proces musi być tak zaprogramowany, aby zachowywał odpowiednią kolejność wewnętrznych działań. Algorytmy nie mogą się wzajemnie wykluczać, ale dopełniać (na przykład zamiana znaku enter na spację w tekście może spowodować, że w tekście będą koło siebie dwie spacje, zatem normalizator usuwający spacje musi być uruchomiony po tym procesie, a nie przed nim). W ramach Poziomu 2, czyli walidacji techniczno – logicznej, następuje wprawdzie walidacja błędów podstawowych polegająca na identyfikacji i korygowaniu błędów, które są jednoznacznie klasyfikowane jako niezgodne ze standardami lub specyfikacjami, a następnie walidacja zależności, która wymaga skomplikowanego porównania danych z różnych pól i potencjalnie różnych źródeł. To podejście pozwala na wykrycie bardziej subtelnych błędów logicznych, które mogą być niewidoczne bez kontekstu. Ostatnim poziomem jest weryfikacja jakości i sugerowanie zmian, gdzie zaawansowane metody mogą sugerować możliwe ulepszenia danych, zwiększając tym samym szanse na podniesienie ich jakości przez wprowadzającego dane. Przyjęcie takiej sekwencyjnej metodyki pozwala na wydajne zarządzanie danymi i ich ciągłe ulepszanie. W kontekście wymienionych poziomów istotne są również czynności podejmowane przez system w celu zapewnienia integralności danych. System korzystający z takiego modelu musi być zaprojektowany tak, aby nie tylko pasywnie analizować dane, ale również interaktywnie uruchamiać procedury w przypadku wykrycia nieprawidłowości. Podczas normalizacji, system automatycznie standaryzuje i koryguje dane wejściowe, co nie wpływa na zakres merytoryczny danych, dlatego może być wykonywany bez udziału i wiedzy administratora danych jak i osoby wprowadzającej dane, w czasie rzeczywistym. W obszarze walidacji techniczno-logicznej, system wykonuje zestaw predefiniowanych walidacji, aby upewnić się, że wszystkie dane spełniają techniczne kryteria poprawności. Tutaj system nie może akceptować błędnych danych, dlatego wszelkie nieprawidłowości muszą zostać odrzucone lub przekazane do poprawy bez rejestracji danych w głównej bazie danych. Gdy w tym obszarze dochodzi do walidacji zależności, system analizuje powiązania między różnymi atrybutami i polami danych i wykrywa błędy, które nie są widoczne na poziomie pojedynczych wpisów. W tym miejscu system może również wykorzystać dane zewnętrzne do krzyżowej weryfikacji i zapewnienia koherencji danych. Ze względu na dopuszczenie zewnętrznych systemów do walidacji danych dane mogą być tymczasowo przechowywane do czasu wykonania testu, jednak, jeżeli wynik testu będzie wskazywał na błąd, takie dane muszą być oznaczone jako niepoprawne i niekwalifikowane do użycia. Na etapie sugerowania zmian, system wykorzystuje zaawansowane techniki mające na celu zwrócenie uwagi na potencjalny błąd. Takie działania

mogą obejmować informowanie autora danych o potencjalnych problemach, co umożliwia szybką reakcję i korektę. System może również sugerować konkretne zmiany w danych, które mogą zostać zastosowane przez autora danych lub automatycznie wprowadzone przez system, z jednoczesnym powiadomieniem autora o dokonanych modyfikacjach. Można wprowadzić również ocenę jakości danych, która może być przeprowadzona przez zaawansowane algorytmy analityczne, i w przypadku wykrycia potencjalnych nieprawidłowości, dane mogą być skierowane do manualnego przeglądu przez operatora systemu. Wszystkie te działania muszą być zaprojektowane tak, aby nawet w przypadku stwierdzenia niedoskonałości, system kontynuował proces akceptacji danych, jednocześnie zapewniając mechanizmy do ich poprawy. Jest to podejście, które zachęca do ciągłego doskonalenia jakości danych, jednocześnie zapewniając ich przepływ w systemie bez niepotrzebnych przestoju. Dzięki temu rozwiązaniu nawet jeśli dane zawierają pewne błędy, nie są one odrzucane, ale podlegają procesowi optymalizacji, co zapewnia płynność operacji biznesowych i pozwala na stopniową poprawę jakości informacji przechowywanych w bazie danych. Kluczowe jest, aby cały proces był dynamiczny i adaptacyjny, umożliwiając operatorowi i algorytmom naukę z poprzednich interakcji oraz dostosowywanie się do nowych danych i wzorców, co zwiększy skuteczność w miarę użytkowania. Ponadto system korzystający z modelu powinien być wyposażony w takie interfejsy, które umożliwiają operatorom danych łatwą interwencję i wprowadzanie poprawek, a także w mechanizmy raportowania i monitoringu, które informują o stanie procesu walidacji i jego wynikach. To również dobre źródło danych, które może zasilić mechanizmy tzw. grywalizacji (Richter i in., 2015).

Zaproponowany model oferuje szczegółową segmentację działań walidacyjnych niż te przedstawione w ramach przeglądu publikacyjnego (Mao i in., 2019; Taleb i in., 2021). Wprowadza także zaawansowane techniki sztucznej inteligencji do sugestii poprawy jakości danych co jest nowoczesnym podejściem zgodnym z ideą Jakości 4.0 (Wawak, 2022; Zonnenshain i Kenett, 2020) oraz wykracza poza tradycyjne metody walidacji. W porównaniu do istniejących podejść (Onyeabor i Ta'a, 2019) model bezpośrednio angażuje użytkowników wprowadzających dane w proces poprawy, co promuje ciągłe doskonalenie jakości danych.

### **6.3. Metodyka oceny wpływu walidatorów na dane w zaproponowanym modelu**

Ocena skuteczności modelu wielopoziomowej walidacji powinna być dobrana pod kątem specyficznych celów i wymagań, jakie model ma spełniać w danym kontekście. Warto zwrócić uwagę na fakt, że w modelu nie zakładano poprawy danych bezpośrednio na każdym poziomie walidacji, ale – dla poziomu trzeciego – możliwe było jedynie rekomendowanie zmiany, co samo w sobie wpływa jedynie na decyzje, a nie na poprawę jakości danych. W konsekwencji ocena podniesienia jakości danych zawierała niezbadaną zmienną - jak często wprowadzający dane zareaguje na rekomendacje algorytmów i od czego będzie zależeć jego decyzja. Takie badanie, nie będące przedmiotem niniejszej dysertacji, mogłoby dostarczyć wielu cennych wskazówek, również w kontekście wpływu rozwiązania informatycznego na wprowadzanie zmian, głównie w zakresie User Experience. W literaturze poruszono już kwestie tego typu, np. w pracy Rouse'a i Sage'a (2011). Badanie mogłoby obejmować ocenę użyteczności systemu, treści komunikatów oraz stopnia akceptacji modelu przez użytkowników. Pomiar musiałby opierać się również o analizę skuteczności sugestii dla różnych sposobów ich implementacji przez system informatyczny. Analiza mogłaby również uwzględnić różne sposoby wdrażania sugestii przez system informatyczny, jak np. jakość komunikatów, częstotliwość powiadomień, a także wpływ mechanizmów motywacyjnych, takich jak grywalizacja, nagrody czy kary (Richter i in., 2015).

Innym aspektem oceny może być pomiar redukcji kosztów związanych ze zmianą procesu walidacyjnego bazy, jednak dla badanego podmiotu, który nie prowadzi obecnie ręcznej walidacji, wyliczenie „as is” i „to be” (English, 1999) byłoby jedynie teoretyczne. Warto zauważyć, że ręczne zarządzanie danymi o tak dużej skali jest bardzo dużym wyzwaniem. Przy założeniach, że ręczna walidacja i poprawa jednego produktu mogą zająć 10 minut pracy wykwalifikowanej osoby, przy milionie produktów oznaczałoby to 100 lat pracy pracownika zatrudnionego na cały etat.

Do oceny skuteczności zaproponowanego modelu wybrano inną metodę – przeliczenia potencjału poprawy danych. Obliczona została suma wykrytych problemów jakościowych wszystkich zaproponowanych walidatorów dla każdego poziomu walidacji. Dzięki takiemu rozwiązaniu można było zaobserwować potencjalny zysk z zastosowania każdego z walidatorów osobno, ale także zysk na każdym poziomie walidacji. Z uwagi na to, że część algorytmów mogło nie dawać stuprocentowo poprawnych wyników, zdecydowano się na obliczanie odpowiednich metryk: precyzji i czułości (Witten i in., 2016) oraz w niektórych

przypadkach dokładności i miary  $F_1$  w celu lepszej ich oceny. Na podstawie precyzji odejmowano potencjalnie niepoprawne wyniki tych algorytmów od głównej sumy potencjału poprawy danych, czułość natomiast pozwalała określić jaka część rzeczywistych błędów została wykryta przez algorytmy. Dokładność przedstawiała ogólną poprawność działania algorytmu, pokazując stosunek prawidłowych predykcji do wszystkich przypadków. Miara  $F_1$  stanowiła zrównoważoną ocenę działania algorytmu, łącząc zarówno precyzję, jak i czułość, oferując pełniejszy obraz efektywności algorytmu, szczególnie w sytuacjach, gdy równowaga między tymi dwoma metrykami była kluczowa. Metryki dały zatem ogólny obraz działania algorytmów, pomagały w wyborze ich optymalnej wersji oraz wskazywały miejsca do poprawy. W przypadku, gdy użyte w algorytmach modele sztucznej inteligencji zwracały wyniki, w których istotne były negatywne wartości predykcji (np. w odpowiedzi na pytanie, „czy zdjęcie jest poprawne?”), interesujące były przypadki, gdzie zdjęcie zostało uznane za niepoprawne), skorzystano z dodatkowych metryk takich jak ujemna wartość predycyjna (NPV) oraz swoistość. NPV umożliwiła ocenę, jaka część negatywnych predykcji była rzeczywiście negatywna, co było istotne przy weryfikacji poprawności odrzuconych danych. Z kolei swoistość umożliwiła oszacowanie, jak skutecznie algorytm rozpoznawał przypadki rzeczywiście negatywne, co było kluczowe w walidacji zdjęć oraz innych atrybutów, gdzie istotne było wykrywanie błędów lub niepoprawnych danych. Dzięki temu, możliwe było dokładniejsze oszacowanie efektywności algorytmów w zakresie identyfikacji i odrzucania niepoprawnych danych, co stanowiło uzupełnienie klasycznych metryk takich jak precyzja, czułość i miara  $F_1$ . Metryki wyliczono korzystając ze wzorów:

$$[\text{czułość}] = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$[\text{swoistość}] = \frac{TN}{N} = \frac{TN}{FP + TN}$$

$$[\text{dokładność}] = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + FN + TN + FP}$$

$$[\text{precyzja}] = \frac{TP}{TP + FP}$$

$$[\text{NPV}] = \frac{TN}{TN + FN}$$

$$[F_1] = 2 * \frac{[Precyzja] * [Czułość]}{[Precyzja] + [Czułość]}$$

Gdzie:

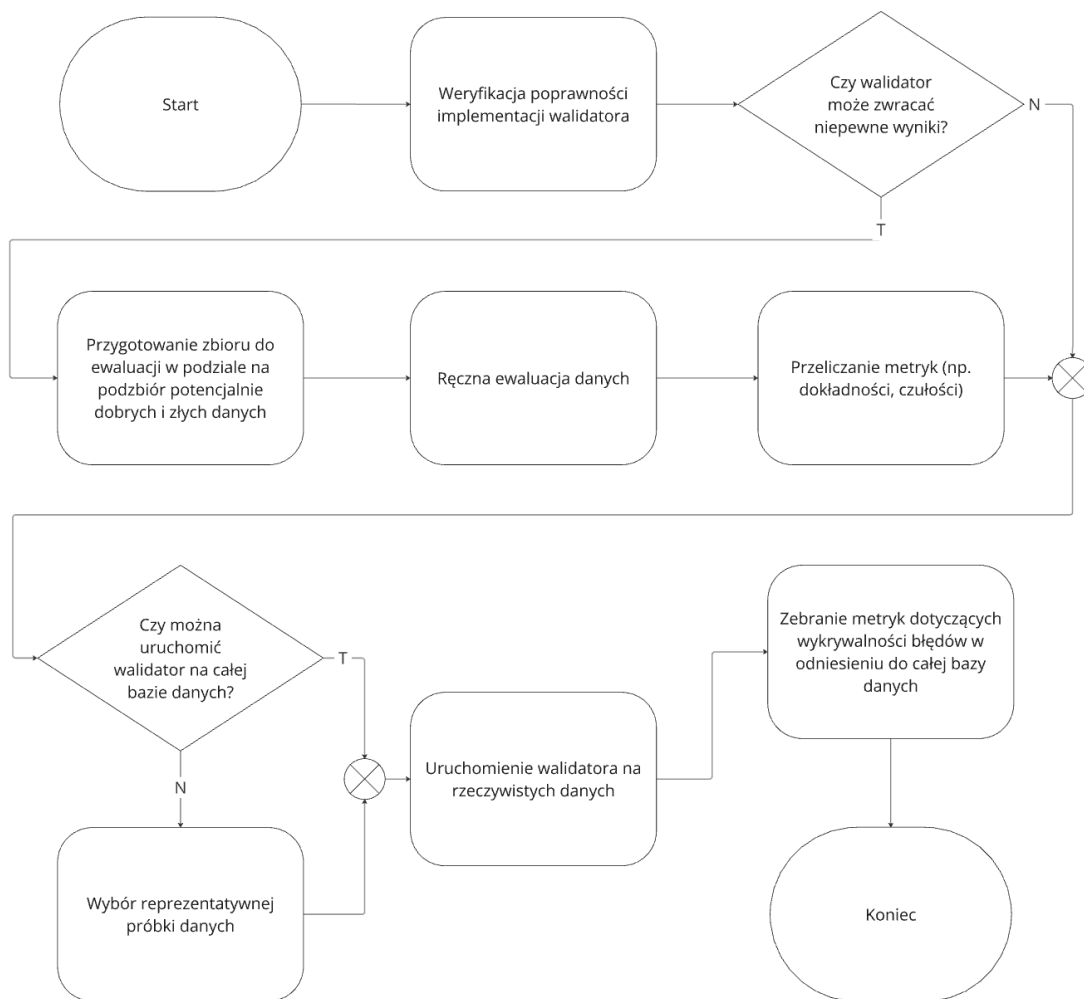
- TP (ang. *true positive*) – wartości prawdziwie pozytywne
- TN (ang. *true negative*) – wartości prawdziwie negatywne
- FP (ang. *false positive*) – wartości fałszywie pozytywne
- FN (ang. *false negative*) – wartości fałszywie negatywne
- $P = (TP + FN)$  – wszystkie wartości pozytywne
- $N = (TN + FP)$  – wszystkie wartości negatywne

W celu oceny skuteczności walidatorów w wykrywaniu błędów, przeprowadzono analizę na wybranych próbkach danych, które wcześniej zostały manualnie zweryfikowane przez ekspertów. Pozwoliło to na zminimalizowanie problemu tzw. „brudnych danych” w ocenie i umożliwiło zbadanie rzeczywistej efektywności działania algorytmów. Ze względu na skalę i nierównomierne rozłożenie danych, zdecydowano się na zastosowanie specyficznej, celowanej stratyfikacji zbioru testowego (Han i in., 2011). Stratyfikacja polegała na podziale danych na podzbiory reprezentujące wynik walidacji (np. pozytywny lub negatywny), a nie na proporcji rozkładu danych. Dzięki temu możliwa była bardziej szczegółowa analiza jakości danych w każdym podziorze, co pozwoliło na lepsze zrozumienie skuteczności walidatorów. Niestety, taka metoda analizy często daje gorsze wyniki precyzji i czułości niż stratyfikacja oparta na proporcjach lub losowy wybór danych, np. co dziesiątej próbki w sytuacji, gdy dane w części są zbliżone do siebie, co jest powszechne w katalogu produktowym. Z drugiej strony lepiej oddaje rzeczywistą jakość samej walidacji, co czyni ją bardziej wartościową w krytycznej ocenie skuteczności walidatorów.

W przypadku większości walidacji opartych na z góry określonych wytycznych (np. „*dlugość ciągu tekstowego musi być mniejsza niż ...*”), obliczanie metryk takich jak precyzja czy czułość nie było wymagane. Działanie takich algorytmów polegało na sztywnym egzekwowaniu określonych wytycznych, co oznaczało, że zgodność danych była sprawdzana w sposób deterministyczny. W związku z tym, po przeprowadzeniu weryfikacji wewnętrznej algorytmu walidatora przyjęto, że funkcjonuje on zgodnie z założeniami, a każde wykrycie problemu jakościowego przez taki walidator uznawano za rzeczywisty błąd danych w całej bazie. W sytuacji, gdyby walidator został użyty jednak niezgodnie z przyjętą wytyczną lub jego działanie opierało się jedynie na założeniu poprawności, a nie było potwierdzone, konieczne byłoby zastosowanie metryk takich jak precyzja i czułość. Pomiar tych metryk pozwoliłby na

ocenę jakości działania algorytmu, umożliwiając weryfikację skuteczności walidatora oraz minimalizację ryzyka błędów w procesie walidacji danych.

Ogólny schemat blokowy procesu ewaluacji poszczególnych walidatorów zaprezentowano na Rysunku 17.



Rysunek 17. Schemat blokowy procesu ewaluacji walidatorów  
Źródło: opracowanie własne

Proces rozpoczął się od weryfikacji poprawności implementacji walidatora. Na tym etapie przeprowadzano kontrolę zgodności reguł z założeniami projektowymi, a w przypadku modeli sztucznej inteligencji – ewaluację zgodnie z przyjętą metodyką na odpowiednim zbiorze danych testowych. W ramach tego procesu możliwe były wielokrotne korekty, jednak nie uwzględniano ich na schemacie, aby zachować jego przejrzystość. Po zakończeniu weryfikacji oceniano, czy walidator mógł zwracać wyniki niepewne. Jeśli istniała taka możliwość,

szczególnie w przypadku zastosowania modeli sztucznej inteligencji, walidator poddawano eksperckiej ocenie.

Następnie przygotowywano zbiór danych do ewaluacji, dzieląc dane na podzbiory zawierające potencjalnie poprawne oraz potencjalnie błędne informacje, co ułatwiało dokładniejszą analizę i identyfikację błędów. Przeprowadzano ręczną analizę tych danych przez ekspertów w celu ich oczyszczenia i klasyfikacji. Mimo że była to metoda czasochłonna, umożliwiała precyzyjną weryfikację, szczególnie że używane dane były tzw. "brudnymi" danymi, które nigdy wcześniej nie przeszły walidacji.

Po zakończeniu ręcznej oceny przeliczano metryki, co pozwalało na obiektywną ocenę skuteczności walidatora. Metryki takie jak dokładność i czułość służyły jako wskaźniki wykrywalności błędów i były kluczowe dla oceny efektywności algorytmu. Następnie walidator uruchamiano na całej bazie danych, o ile pozwalały na to zasoby obliczeniowe. W przypadku ograniczeń mocy obliczeniowej, analiza odbywała się na reprezentatywnej próbce 0,5 miliona rekordów. Zebrane metryki dotyczące wykrywalności błędów w odniesieniu do całej bazy były wykorzystywane do oceny skuteczności walidatora oraz jego użyteczności w rzeczywistych warunkach operacyjnych.



## 7. Weryfikacja koncepcji modelu w podziale na 3 warstwy

W niniejszym rozdziale omówiono proces weryfikacji koncepcji modelu wielopoziomowej walidacji danych produktowych na przykładzie wybranych walidatorów. Badaniem objęto wszystkie poziomy walidacji, jednak szczególną uwagę poświęcono trzeciemu poziomowi, który został dokładnie opisany, aby ukazać zastosowanie sztucznej inteligencji — kluczowego elementu niniejszej dysertacji.

### 7.1. Poziom 1: Normalizacja danych

Zgodnie z opracowanym modelem, poprawę jakościową bazy danych rozpoczęto od procesu normalizacji danych. Wykryte nieprawidłowości wyszczególnione w Tabeli 24 poddano szczegółowej analizie. Głównym wyzwaniem w realizacji tego zadania była identyfikacja jak największej liczby potencjalnych problemów i wytworzenie przekształceń nieuszkodzających danych. W opracowanej koncepcji nacisk położono przede wszystkim na zastosowaniu metod z wykorzystaniem wyrażeń regularnych (Stubblebine, 2008), które oprócz zwięzłego kodu, są interpretowane w identyczny lub zbliżony sposób w różnych językach programowania. W budowie reguł skorzystano także z ustandaryzowanych słowników grup znaków tabeli kodowej Unicode (Unicode, 2023), zamiast tradycyjnych rozwiązań polegających na identyfikacji takich elementów, jak spacje czy wielkie litery. Dzięki takiemu podejściu, wykryto znacznie więcej ukrytych znaków i liter w analizowanych ciągach tekstowych niż w klasycznym rozwiązaniu. Kolejnym wyzwaniem okazało się określenie kolejności wykonywania zadań normalizacyjnych, dzięki któremu możliwe było zapewnienie optymalnej sekwencji przetwarzania, mającej kluczowe znaczenie dla maksymalizacji efektywności procesów normalizacji. Na przykład, usuwanie znaków kontrolnych, takich jak niepożądane tabulatory czy znaki nowej linii, powinny być uruchamiane po dekodowaniu encji HTML. Taka kolejność zapobiega ryzyku, że po dekodowaniu pojawią się nowe, niepożądane znaki, które wymagałyby ponownego przetwarzania. Podobnie, redukcja nadmiernych spacji między słowami jest bardziej efektywna, gdy zostanie wykonana po korektach, które mogą wprowadzać dodatkowe spacje, jak na przykład łączenie fraz z dwóch linii, czyli zamiany znaku enter na spację.

Implementacja procesów normalizacyjnych przed dalszymi etapami walidacji danych ma zasadnicze znaczenie, a zaniechanie ich wdrożenia może skutkować znacznymi trudnościami w kontekście późniejszej prawidłowej analizy danych. Z tego powodu

zbudowano rozwiązania dla wszystkich problemów wymienionych w Tabeli 24, dla których przeanalizowano możliwość modyfikacji danych edycyjnych bez naruszenia ich zawartości semantycznej. W ramach przeprowadzonego badania przetworzono całą dostępną bazę produktów podmiotu badawczego (44,9 mln). W wyniku działania normalizatorów poprawiono ponad 2 miliony produktów. Szczegółową specyfikację zidentyfikowanych normalizatorów oraz ich bezpośredni wpływ na optymalizację jakości bazy danych przedstawiono w Tabeli 27.

Tabela 27. Zaimplementowane normalizatory danych

Problem jakości	Opis operacji	Wyrażenie regularne	Sekwencja wywołania	Wpływ na markę produktu	Wpływ na nazwę produktu
[W1.1] Encja HTML zamiast znaku specjalnego w tekście	Dekodowanie encji HTML w polach tekstowych	n.d.	1	11 182	29 116
[W1.2] Znak tabulacji w tekście jednowierszowym	Zastępowanie znaku tabulacji spacją we wszystkich atrybutach tekstowych	$\wedge\{t+m\}$	2	206	16 646
[W1.3] Kod sterujący w tekście	Usuwanie znaków kontrolnych z pól tekstowych, z wyjątkiem tabulatorów i znaków nowej linii ( $\wedge\{t\nr\}$ )	$\wedge\{p\{Cc\}(?![\wedge\{t\nr\}]+)\}$	3	3	181 956
[W1.4] Znak nowej linii w tekście jednolinijkowym	Zastąpienie znaku nowej linii ("enter") spacją w polach tekstowych jednolinijkowych	$\wedge\{p\{n\}r\}+$	4	1 060	67 520
[W1.5] Nadmiarowe spacje w tekście	Usuwanie wielokrotnych spacji i twardych spacji wewnątrz tekstu	$\wedge\{p\{Zs\}+\}$	6	35 560	1 934 469
[W1.6] Podwójny znak cudzysłowu w tekście	Redukcja zduplikowanych cudzysłowów	$\wedge\{p\{''\}+\}$	7	0	7407
[W1.7] Spacja przed tekstem właściwym	Usuwanie pustych miejsc na początku tekstu (z 17 znaków Unicode grupy "Separator spacji")	$\wedge\{p\{Zs\}+\}$	8	610	4413
[W1.8] Spacja po tekście właściwym	Usuwanie pustych miejsc na końcu tekstu (z 17 znaków Unicode grupy "Separator spacji")	$\wedge\{p\{Zs\}+\}\wedge\{s\}$	9	9 225	38 578

Źródło: opracowanie własne

Tabela przedstawia wyniki operacji normalizacyjnych przeprowadzonych na zbiorze danych produktowych, obejmujące szczegółową analizę wpływu poszczególnych normalizatorów na poprawę jakości danych. Kolumny tabeli przedstawiają się następująco:

- A. **Problem** – wskazuje na konkretny problem w danych, zidentyfikowany zgodnie z Tabelą 24, który został rozwiązany za pomocą danej operacji normalizacyjnej.
- B. **Opis operacji** - szczegółowy opis wykonanej operacji normalizacyjnej.
- C. **Wyrażenie regularne** - przedstawia wyrażenie regularne użyte do identyfikacji specyficznych wzorców w danych, zgodnie z opisem operacji.
- D. **Sekwencja wywołania** - określa, w jakiej kolejności operacje normalizacyjne zostały zastosowane do danych.
- E. **Wpływ na markę / nazwę produktu** - liczba korekt dokonanych przez daną operację normalizacyjną, wyrażona w ilości zmodyfikowanych atrybutów danych.

Dodatkową ewaluację przeprowadzono na wyselekcjonowanym zbiorze 1000 produktów, które poddano weryfikacji eksperckiej porównując przekształcenia normalizatorów do pierwowzoru. Ten etap umożliwił ocenę poprawności wprowadzonych korekt. Dzięki takiemu rozwiązaniu wychwycono zarówno błędy składniowe wyrażen regularnych, jak i doprecyzowano działania normalizatorów.

Przedstawione rozwiązania i ich realny wpływ na dane udowadniają kluczową rolę normalizacji w procesie poprawy jakości. Chociaż prezentowane walidatory dotyczyły jedynie danych tekstowych, to ich zastosowanie nie ogranicza się wyłącznie do tego typu operacji, ale obejmuje również dane cyfrowe, w tym formatowanie liczb z użyciem przecinków, spacji i kropek. Podobnie, może obejmować także konwersję numerów GTIN do formatu GTIN14, konwersję adresów URL, czy innych pól, standaryzując i ujednolicając zawartość. Rozszerzone badania w tym zakresie opublikowano w rozdziale książki *Human Factors in Engineering* (Niemir i Mrugalska, 2023b).

## **7.2. Poziom 2: Walidacja techniczno - logiczna**

Drugi poziom walidacji w zaproponowanym wielopoziomowym modelu to klasyczna grupa walidatorów, której celem była eliminacja nieprawidłowości. Walidatory opracowano na podstawie określonych reguł (opisy znajdują się w Tabeli 24) przy zachowaniu ostrożności, aby zbyt restrykcyjne zasady nie zablokowały wprowadzenia poprawnych danych. Odrzucanie danych na podstawie nieuzasadnionych, nieprzejrzystych lub niepewnych reguł wiąże się z dużym ryzykiem zaburzenia procesu wprowadzania danych, co było powodem, dla którego w tym przypadku całkowicie zrezygnowano z wykorzystania sztucznej inteligencji na tym poziomie walidacji, właśnie z uwagi na ryzyko niepewności decyzji podejmowanych przez modele.

Podmiot badań - walidowana baza danych, jak wspomniano we wcześniejszych rozdziałach, zawierała informacje o produktach, które zostały wprowadzone przez system informatyczny, który posiadał wbudowane wszystkie walidatory odnoszące się do numerów GTIN implementujące rozwiązania wskazane w badaniu dotyczącym tego numeru w Rozdziale 3.2. Podobnie, system posiadał również ograniczenia implementacyjne, wynikające z ograniczeń systemowych, na przykład ograniczenia maksymalnej długości ciągów znakowych oraz ograniczenia relacyjne bazy danych. Mimo to, wskazano aż 21 walidatorów do implementacji, zgodnie z propozycjami znajdującymi się w Tabeli 24 i późniejszym podziałem zwizualizowanym na Rysunku 15.

W Tabeli 28 zaprezentowano listę walidatorów wraz z ich sposobem działania oraz ogólnym opisem zastosowanych algorytmów. Najważniejszą zasadą, którą kierowano się przy doborze rozwiązań, była idea modelu wielopoziomowej walidacji, zgodnie z którą walidatory drugiego poziomu uruchamiane są wyłącznie w sytuacjach, w których istnieje pewność wystąpienia błędu lub akceptowalne, bardzo wysokie prawdopodobieństwo jego zaistnienia. W kolumnie „Problem jakości” przedstawiono opis konkretnego problemu jakościowego, który wymagał rozwiązania. Kolumna „Sposób realizacji” prezentuje podejście zastosowane do rozwiązania tego problemu. W kolumnie „Ogólny opis algorytmu” szczegółowo omówiono sposób, w jaki algorytm analizuje i przetwarza dane, a ostatnia kolumna „Wpływ na dane” prezentuje liczbę rekordów, które zostały odrzucone w wyniku działania danego walidatora w kontekście rzeczywistych danych produktowych podmiotu badawczego. Sposób wykonania walidatorów różnił się w zależności od typu danych i celu walidacji. Obejmował operacje na ciągach tekstowych, gdzie najczęściej stosowano wyrażenia regularne ze względu na ich precyzyjność i łatwość definiowania skomplikowanych reguł. Dodatkowo, przeprowadzono operacje na liczbach oraz z wykorzystaniem przeszukiwania bazy danych. W przypadku zdjęć wykorzystywano specjalistyczną bibliotekę do przetwarzania obrazów, Pillow, dostępnej dla języka Python.



<b>Problem jakości</b>	<b>Sposób realizacji - reguła lub istotny element działania walidatora</b>	<b>Ogólny opis algorytmu</b>	<b>Wpływ na dane (w tys.)</b>
[W2.17] Zbyt niska rozdzielczość zdjęcia	Operacja na bibliotece Pillow	Sprawdza, czy zdjęcie posiada minimalną rozdzielczość 320x320 pikseli, wymagane do zachowania odpowiedniej jakości.	5,5
[W2.18] Nadmiernie długa nazwa marki	Operacja na ciągach tekstowych	Oblicza długość nazwy marki i weryfikuje, czy nie przekracza maksymalnej dopuszczalnej liczby 70 znaków.	1
[W2.19] Nadmiernie długa nazwa produktu	Operacja na ciągach tekstowych	Sprawdza, czy długość nazwy produktu nie przekracza 500 znaków.	1
[W2.20] Zbyt duża rozdzielczość zdjęcia	Operacja na bibliotece Pillow	Sprawdza, czy zdjęcie zajmuje obszar powyżej 64 megapikseli, analizując jego wymiary.	0,16
[W2.21] Przekroczona wielkość pliku	Operacje na plikach	Weryfikuje, czy rozmiar pliku przekracza maksymalną wartość 16 MB.	0,1
[W2.22] Przekroczona wielkość zawartości netto	Operacja na liczbach	Sprawdza, czy zawartość netto produktu nie przekracza maksymalnej wartości 10 miliardów, z ograniczeniem do 5 miejsc po przecinku.	0,1

*Źródło: opracowanie własne*

W analizowanej bazie danych produktowych zidentyfikowano trzy główne problemy dotyczące jakości danych ([W2.1],[W2.2],[W2.3]). Wszystkie dotyczyły nazwy produktu i jego ogólnej formy zapisu. Był to problem braku marki w nazwie, braku zawartości netto w nazwie oraz problem duplikujących się nazw względem całej bazy / produktów oznaczonych różnymi numerami GTIN. W efekcie, dzięki prostym operacjom wyszukiwania ciągów tekstowych wykryto łącznie ponad 75 milionów wad w nazwach produktów, co oznacza, że kilka walidatorów zwróciło pozytywny wynik dla tej samej nazwy (walidatory nakładały się). Trzeci z walidatorów został utworzony w celu zapewnienia unikalności pary GTIN – Nazwa (szerzej problem opisano w rozdziale 2.6). Pomimo że celem było wykrycie duplikatów, zidentyfikowano w ten sposób przypadki, w których osoby wprowadzające dane, z powodu braku odpowiedniego przeszkolenia, ograniczyły się do wpisywania ogólnych nazw produktów, takich jak „chleb”, „bluzka” czy „klamka”. Należy jednak zwrócić uwagę, że wszystkie walidatory to efekt niezrozumienia oczekiwań co do kompletności i warunków unikalności nazwy produktu. Z tego powodu przy analizie związanej z oceną stanu bazy i jakości danych produktowych warto wyłączyć te walidacje, lub skategoryzować je jako efekt niezamierzonego błędu związanego z interpretacją pola „nazwa produktu”.

Walidacje dotyczące kwalifikacji i ograniczeń ilości wyrazów w nazwie ([W2.4], [W2.5], [W2.6], [W2.12]), stanowiły drugą co do wysokiej skuteczności grupę walidatorów (ponad 7,8 milionów wychwyconych błędów). Sam proces doboru ograniczeń wynikał z obserwacji danych i był wynikiem procesu eksploracji danych.

Walidacje liczbowe ([W2.7], [W2.13]), które wynikały z logiki (zachowanie liczby całkowitej dla sztuk, eliminacja wartości ujemnych dla zawartości netto) umożliwiły oznaczenie ponad 600 tysięcy błędnych przypadków.

Kolejne walidacje dotyczyły analizy występowania określonych znaków na początku oraz wewnątrz tekstu ([W2.8], [W2.10], [W2.11], [W2.15]). Dobór także wynikał z procesu eksploracji danych, tworzeniu statystyk i wizualizacji danych na potrzeby badań. Wytyczne sieci i e-sklepów nie specyfikują ograniczeń w tym zakresie, natomiast z obserwacji zbioru danych wynika, że używanie określonych wzorców i liter jest równoznaczne z nieprawidłowym zapisem nazwy i powinno podlegać odrzuceniu danych. Dla walidacji [W2.11] – „nieprawidłowa konwersja” – celowo dokonano symulacji wielokrotnej nieprawidłowej konwersji znaków i wyodrębniono te znaki, co do których upewniono się, że nie występują standardowo w nazwach. Walidacja [W2.10] – „nadmiarowe podkreślenia” – została użyta celowo w tej grupie walidacji, pomimo możliwości zastosowania normalizacji,

czyli eliminacji zduplikowanych podkreśleń. W wyniku analizy zbioru okazało się jednak, że tego rodzaju duplikaty wynikają najczęściej z problemu konwersji danych i systemowego zastępowania niezidentyfikowanych błędnych znaków znakiem podkreślenia.

Walidacje [W2.9] i [W2.16] to przykłady, gdzie zastosowano algorytmy do identyfikacji specyficznych wzorców. [W2.16] skutecznie identyfikuje formuły Excela, które z powodu błędów importu danych mogły być nieprawidłowo wprowadzone jako nazwy produktów. Z kolei [W2.9] wykorzystuje zasoby słownikowe do wykrywania określonych nazw lub wzorców nazw marek, które podlegają ograniczeniom w powszechnym użyciu. Wprowadzenie słownika było konsekwencją częstego wprowadzania nazw zastępczych, jak np. „brak”, „nie ma”, „własna” itd. Zostały tutaj zastosowane jednoznaczne reguły, bez implementacji rozwiązań opartych o algorytmy sztucznej inteligencji ze względu na ryzyko popełnienia błędu i braku kontroli nad działaniem algorytmu, szczególnie w kontekście nazw własnych, które niejednokrotnie bazują na skojarzeniach, kulturowych niuansach oraz specyfice językowej, które są trudne do jednoznacznego zdefiniowania i modelowania przez algorytmy uczenia maszynowego. W związku z tym wykorzystanie prostych technicznie, deterministycznych reguł słownikowych pozwala na utrzymanie wysokiego poziomu kontroli działania walidatora, minimalizując ryzyko fałszywych pozytywów i negatywów, które mogą wystąpić przy złożonych modelach uczenia maszynowego. Jest to kluczowe przy pracy z dużymi zestawami danych, gdzie stabilność procesu jest priorytetem, szczególnie w przypadku ryzyka odrzucania poprawnych danych. Jednakże można zasugerować rozwiązanie wykorzystujące potencjał sztucznej inteligencji w dwóch przypadkach: dzieląc walidator na dwie części – część, która odrzuca dane, i część, która sugeruje zmianę, czyli walidacji poziomu trzeciego, lub alternatywnie w przyszłości może być zasadne użycie rozwiązań sztucznej inteligencji, do procesu podpowiedzi kolejnych elementów słownikowych do omawianej walidacji poziomu drugiego. Odpowiednio skonfigurowany algorytm mógłby wówczas cyklicznie lub na żądanie oznaczać kolejne nazwy marek jako sugerowane administratorowi do wyeliminowania z powszechnego użycia.

Pozostałe walidatory zbudowano w wyniku analiz przeprowadzonych w badaniu 3.2, z których wynikały ograniczenia zarówno dotyczące długości ciągów znakowych, liczb jak i formatów i wielkości zdjęć.

Każdy z walidatorów testowano na docelowym zbiorze danych w celu całkowitej eliminacji niepoprawnych pozytywnych klasyfikacji błędów (false positive) przy minimalizacji błędnych negatywnych klasyfikacji (false negative). W celu ostatecznej weryfikacji



skuteczności każdego walidatora, końcowe wyniki działania były testowane na próbce składającej się z 200 losowo wybranych produktów. W tej próbce, połowa danych była wcześniej sklasyfikowana przez walidator jako niepoprawne (true positive), a druga połowa jako poprawne (true negative), co umożliwiło analizę i ocenę stosowanych metod. Wszystkie walidatory uzyskały 100% skuteczność w kontekście założonego działania. Należy jednak zaznaczyć, że nie oznacza to całkowitej eliminacji występujących problemów jakościowych, a jedynie oznacza skuteczność w zakresie wytypowanego rozwiązania dla zidentyfikowanych problemów w danej kategorii.

### **7.3. Poziom 3: Sugestie poprawy jakości**

Trzeci poziom walidacji danych produktowych stanowi klucz do zrozumienia innowacyjności w proponowanym podejściu do doskonalenia jakości danych. Na tym etapie wprowadzono metody, których celem było zwrócenie uwagi i zwiększenie zaangażowania użytkownika wprowadzającego produkty w proces poprawy danych, zamiast bezpośredniego korygowania czy odrzucania zawartości. Założono, że algorytmy stosowane na tym poziomie nie musiały jednoznacznie wykazywać błędu, co oznaczało, że mogły działać w oparciu jedynie o pewne prawdopodobieństwo w ocenie jakości danych. W konsekwencji otwarto więc możliwość skorzystania ze sztucznej inteligencji, dzięki której możliwe było wykrywanie potencjalnych błędów w danych i proponowanie sugestii, nawet jeśli nie było stuprocentowej pewności co do poprawności takich predykcji. W proponowanych rozwiązaniach priorytetem była wysoka precyzja algorytmów, nawet kosztem czułości, co oznaczało, że lepiej, aby algorytm nie zadziałał, niż miałby działać błędnie. Niemniej jednak, nawet błędne działanie algorytmu w niektórych przypadkach było akceptowalne. Wykorzystanie sztucznej inteligencji znacząco poprawiło proces identyfikacji potencjalnych błędów i proponowania poprawek, co zwiększyło efektywność walidacji danych.

W Tabeli 29 zaprezentowano wyniki analiz oraz proponowanych algorytmów wraz z ogólnym wpływem na bazę danych i oznaczeniem, czy do rozwiązania planowano lub użyto algorytmów z zakresu sztucznej inteligencji, oraz w jakim zakresie. W kolejnych podrozdziałach szczegółowo opisano poszczególne rozwiązania.

Tabela 29. Lista walidatorów trzeciego poziomu

Problem jakości	Główne algorytmy / modele / biblioteki	Ogólny opis wybranego algorytmu	Wpływ na dane	Czy użyto SI do walidacji?	Czy użyto SI do sugestii zmian?
[W3.1] Nazwa produktu pisana wyłącznie wielkimi literami	Wyrażenie regularne, następnie biblioteka SpaCy	Analizuje stosunek dużych liter do małych liter i proponuje poprawioną nazwę z odpowiednim formatowaniem liter.	11 401 826	Nie	Tak
[W3.2] Nadmiernie długa nazwa produktu do niektórych zastosowań e-commerce	Operacja na ciągu znaków, następnie LLaMA	Oblicza długości ciągu. Podpowiada krótszą nazwę.	5 271 882	Nie	Tak
[W3.3] Początek tekstu to cudzysłów lub apostrof	Wyrażenie regularne, następnie LLaMA	Sprawdza, czy początek tekstu jest cudzysłowem lub apostrofem. Podpowiada poprawioną nazwę.	29 759	Nie	Tak
[W3.4] Różne zapisy tej samej marki	Algorytm Jaro-Winklera, następnie LLaMA	Wyszukuje podpobieństw, następnie daje do przetworzenia LLaMA	10 421	Tak	Tak
[W3.5] Nazwa produktu w języku nieodpowiadającym wskazanemu językowi	FastText i Morfeusz2	Analizuje nazwę pod kątem języka, sprawdzając występowanie słów powszechnych i typowe sekwencje znaków dla języka polskiego.	4 787 554	Tak	Tak
[W3.6] Zbyt niska rozdzielczość zdjęcia dla niektórych zastosowań e-commerce	Biblioteka Pillow	Weryfikuje, czy zdjęcie ma odpowiednią rozdzielczość ( $\geq 0,3$ Mpix, szerokość i wysokość $>900$ px).	978 926	Nie	Brak ★
[W3.7] Zbyt niska rozdzielczość zdjęcia dla większości zastosowań e-commerce	Biblioteka Pillow	Sprawdza, czy zdjęcie zajmuje powierzchnię mniejszą niż 0,3 Mpix.	17 440	Nie	Brak ★
[W3.8] To samo zdjęcie dla różnych produktów	Algorytm skrótu percepcyjnego	Porównuje "odciski palca" zdjęć, aby sprawdzić, czy identyczne zdjęcia są używane dla różnych produktów.	1 465 797	Nie ★	Nie
[W3.9] Zdjęcie produktu na niejednorodnym lub kolorowym tle	Biblioteka Pillow i RemBG	Algorytm analizuje tło zdjęcia, badając jego kolorystykę i jednolitość. Wymaga białego lub przezroczystego tła.	559 740	Tak	Tak
[W3.10] Zdjęcie produktu na zbyt dużej powierzchni tła	Biblioteka Pillow i RemBG	Algorytm mierzy proporcję tła do produktu na zdjęciu.	7 252	Tak	Brak ★
[W3.11] Niepoprawna lub tymczasowa kategoria produktu	Algorytm BERT wyuczony do klasyfikacji GPC, następnie LLaMA	Generuje propozycje kategorii produktu na podstawie nazwy i weryfikuje zgodność przesłanych danych z listą kategorii.	1 663 071	Tak	Tak
[W3.12] Nieadekwatna jednostka zawartości netto w odniesieniu do kategorii/grupy produktów	Algorytm BERT wyuczony do klasyfikacji jednostek zawartości netto	Proponuje odpowiednią jednostkę netto na podstawie kategorii produktu.	3 903 393	Tak	Tak
[W3.13] Nieadekwatna zawartość netto w odniesieniu do kategorii/grupy produktów	Metody statystyczne	Ogranicza maksymalne wartości zawartości dla danej jednostki w danej kategorii GPC	2 641 815	Nie ★	Nie
[W3.14] Nazwa jest niezrozumiała	LLaMA	Model LLM ocenia czytelność nazwy i sugeruje poprawki zgodnie z wytycznymi, analizując składnię i sens.	9 509 789	Tak	Tak
[W3.15] Zdjęcie nieprzedstawiające produktu do sprzedaży	Moondream2	Analizuje zdjęcie produktu pod kątem jego użyteczności w e-commerce, oceniając, czy przedstawia faktyczny produkt do sprzedaży.	219 424	Tak	Tak

★ – rozważano użycie sztucznej inteligencji

Źródło: opracowanie własne

### 7.3.1. Optymalizacja formatowania nazwy produktu

Algorytmy rozwiązujące problemy dotyczące poprawek formatowania tekstu ([W3.1], [W3.2], [W3.3]) stanowią grupę algorytmów bazujących na ciągach tekstowych, podobnie jak dla walidacji poziomu drugiego. Różnicą jest fakt, że wynik działania walidatorów nie odrzuca danych, ale służy do informowania użytkownika o danym problemie w formie sugestii do poprawy / polepszenia danych.

Walidator [W3.1] „*Nazwa produktu pisana wyłącznie wielkimi literami*” opracowano w celu identyfikacji częstego problem dotyczącego sposobu zapisu nazwy produktu – wyłącznie dużymi literami. Algorytm zliczał procent dużych liter (minimum dwie występujące po sobie) w stosunku do małych i dużych, i jeśli wynosił on 80% lub więcej – sygnalizował nieprawidłowość. Dodatkowo algorytm proponował nazwę alternatywną. Cały tekst nazwy produktu był przekształcany na małe litery i przetwarzany przez bibliotekę SpaCy (Vasilev, 2020). SpaCy to zaawansowana biblioteka oparta na algorytmach sztucznej inteligencji do przetwarzania języka naturalnego (NLP), opracowana głównie w języku Python. Jest ona używana do szerokiego zakresu zadań związanych z lingwistyką komputerową, takich jak tokenizacja (czyli podział tekstu na mniejsze jednostki, np. słowa i zdania), rozpoznawanie części mowy, ekstrakcja nazw własnych, analiza składniowa oraz rozumienie zależności między elementami w zdaniach. W omawianym przypadku algorytm [W3.1] identyfikował i kapitalizował pierwsze słowo w tekście oraz te słowa, które nie należą do określonych części mowy, takich jak spójniki, przyimki czy partykuły, i jednocześnie nie są jednostkami miary (jak m, cm, kg). To umożliwiło zachowanie odpowiedniej formy nazw produktów, które mogą zawierać skomplikowane wyrażenia. Tekst wynikowy składał się z odpowiednio sformatowanych słów, przy czym każde słowo zachowało oryginalne odstęp. Dzięki takiemu podejściu możliwe było dynamiczne dostosowanie kapitalizacji w zależności od kontekstu przetwarzanego tekstu. Ponieważ jednak niektóre nazwy produktów, czy też precyzując - nazwy marek produktów będące częścią nazwy produktu, celowo posiadają zróżnicowany sposób zapisu liter, propozycja nowej nazwy może być jedynie ułatwieniem dla wprowadzającego dane użytkownika, ale nie może zastąpić jego ostatecznej weryfikacji i korekty przed zatwierdzeniem propozycji. Działanie walidatora [W3.1] nie weryfikowano pod względem poprawności sugestii nazw z uwagi na brak bazy referencyjnej. Zadanie ograniczono do weryfikacji poprawności implementacji algorytmu walidacji na bazie 200 próbek i subiektywnej oceny jego przydatności działania. We wszystkich badanych przypadkach algorytm zadziałał prawidłowo. Rysunek 18 przedstawia wykres procentowego udziału dużych

liter w nazwie produktu, zgodnie z implementacją algorytmu, gdzie występują minimum dwie duże litery po sobie. Charakterystyka wykresu wskazuje na dwa główne trendy w zapisywaniu nazw. Nazwy zapisywane są prawidłowo z użyciem małych i dużych liter, oraz nazwy zapisywane są w postaci jedynie dużych liter (np. w związku z ograniczeniami oprogramowania służącego jako źródło danych, np. baza leków BAZYL). Jako próg graniczny walidatora przyjęto wartość 80%, od którego system zwraca uwagę na problem dużych liter i sugeruje lepszą nazwę. Wartość ta została określona na podstawie oceny eksperckiej oraz analizy charakterystyki danych, uznając ją za optymalną w kontekście omawianego algorytmu. Łącznie w całej bazie danych wykryto 11 401 826 produktów do poprawy.



Rysunek 18. Procentowy udział dużych liter w nazwie produktu  
 Źródło: opracowanie własne

W Tabeli 30 zaprezentowano przykłady propozycji nazw zbudowanych za pomocą opisywanego algorytmu. Jak można zauważyć na przykładach, algorytm efektywnie poprawia formatowanie, dostosowując wielkość liter w zależności od pełnionej funkcji syntaktycznej i semantycznej poszczególnych wyrazów. Tym samym nazwy produktów stają się bardziej zgodne z przyjętymi standardami, co może pozytywnie wpływać na ich odbiór przez użytkowników końcowych.

Tabela 30. Przykłady proponowanych nazw produktów

Nazwa pierwotna	Udział dużych liter	Nazwa proponowana
Kazar OBUWIE MĘSKIE WZUWANE, PONIŻEJ KOSTKI Kolor: GRANATOWY 43	80	Kazar Obuwie Męskie Wzuwane, Poniżej Kostki Kolor: Granatowy 43
House SPODNIJE JEANSOWE DAMSKIE 1111A-55J-34 5903610767323	81	House Spodnie Jeansowe Damskie 1111a-55J-34 5903610767323
Artman ELF WALEC DUŻY NEUTRAL ŚWIECA PARAFINOWA	85	Artman Elf Walec Duży Neutral Świeca Parafinowa
D02329-7137-005 nr 38 - BOTEK ZAMSZ GRANAT TOP BLUE	92	D02329-7137-005 nr 38 - Botek Zamsz Granat Top Blue
BARLETTA 6G	100	Barletta 6g
ŁAPKI KURZE NATURALNE 500 G	100	Łapki Kurze Naturalne 500 g
OCIEPLACZ ASPEN 845 ROZ. 31 GRANAT HISERT	100	Ocieplacz Aspen 845 roz. 31 Granat Hisert
SPODNIJE K X_FUNK	100	Spodnie K X_FUNK
REKAWICZKI MISSION RED "S"	100	Rękawiczki Mission Red "S"
WYCIERACZKA MULTITYPE 500 MM	100	Wycieraczka Multitype 500 mm
NICEA W-2 90P KOLOR RF-07 DĄB SZARY	100	Nicea w-2 90P Kolor Rf-07 Dąb Szary
ZESTAW DO KANIULACJI DUŻYCH NACZYŃ TRZYKANALOWY 7F18, HYDROFILNY	100	Zestaw do Kaniulacji Dużych Naczyń Trzykanałowy 7F18, Hydrofilny
STRUCLA Z JAGODAMI 450 G	100	Strucla z Jagodami 450 g

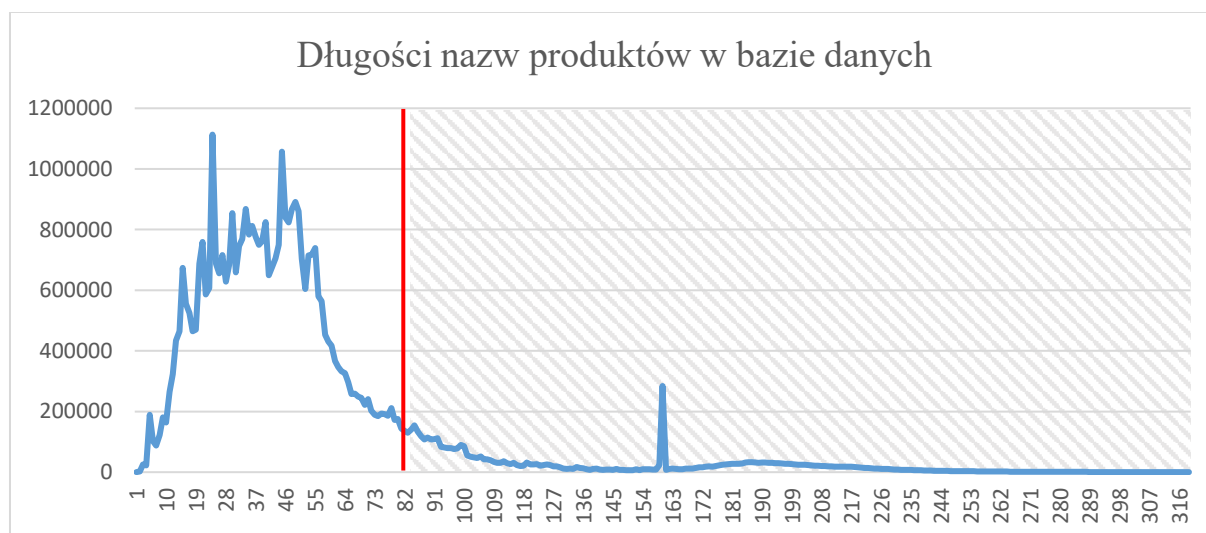
Źródło: opracowanie własne

Walidator oznaczony jako [W3.2] oblicza długość ciągu w nazwie produktu. Wartość maksymalną ustalono na poziomie 80 znaków. Została ona przyjęta na podstawie wytycznych rekomendowanych przez część badanych platform, jak opisano w Rozdziale 3.2.4. W trakcie implementacji algorytmu [W3.2] zidentyfikowano istotny problem związany z potencjalnym negatywnym wpływem skracania nazw produktów przekraczających 80 znaków. W założeniu nazwa produktu powinna obejmować markę, podmarkę, ogólną nazwę opisującą produkt, wariant oraz zawartość netto. W takiej sytuacji limit 80 znaków w niektórych przypadkach staje się zbyt rygorystyczny, szczególnie dla produktów typu mix zawierających kilka produktów w jednym opakowaniu lub takich, dla których występuje szczegółowa wariantowość i podmarki. Przykładowe, powyżej 80-cio znakowe nazwy produktów umieszczono w Tabeli 31, natomiast Rysunek 19 przedstawia liczbę produktów w badanej bazie danych w stosunku do ich długości nazw.

Tabela 31. Długość nazwy produktów

Nazwa opisowa produktu	Długość
Cleanic Nawilżany papier toaletowy z kwasem mlekowym i z ekstraktem z rumianku 40szt	84
Dr Irena Eris Body Art Regenerujący Krem Do Rąk Rozjaśniający Przebarwienia SPF 20 25ml	87
Kopernik Kostka piernikowa z nadzieniem o smaku śliwkowym w czekoladzie luz 2.5kg	81
Lirene Cera Naczynkowa krem witalizująco-przeciwzmarszczkowy na dzień i na noc 50ml	83
Mix Granulki do udrażniania rur Tytn 800g + żel do udrażniania rur Tytan 1 kg - display	87
NIVEA 24H NAWILŻENIA+REGENERACJA Odświeżający krem na noc cera normalna i mieszana 50ml	87
NIVEA 3w1 Pielęgnujące chusteczki micelarne do demakijażu wszystkie typy cery 25szt	83
SM Gostyń Mleko gostyńskie zagęszczone niesłodzone z błonnikiem light 4% 350 g/K-18 sztuk	89
Wawel Czekoladki Coctailowe Kwiaty - bombonierka z czekoladkami nadziewanymi 305g	81

Źródło: opracowanie własne



Rysunek 19. Liczba produktów w stosunku do długości nazw w bazie danych

Źródło: opracowanie własne

Na Rysunku 19 można zaobserwować, że większość nazw produktów w badanej bazie danych mieści się w przedziale od 10 do 100 znaków, z wyraźnymi szczytami w okolicach 30-50 znaków. Liczba nazw produktów znacząco maleje wraz z wydłużaniem się ich długości, co wskazuje na tendencję do tworzenia krótszych i bardziej zwięzłych nazw, natomiast nie kończy się na zalecanych 80-ciu znakach. Z tego powodu granica powinna być jedynie zaleceniem. Łącznie, algorytm [W3.2] wykrył 5 271 882 produktów o nazwach dłuższych niż 80 znaków.

W przedstawionym algorytmie zaimplementowano również funkcję podpowiadania krótszej nazwy za pomocą sztucznej inteligencji. Do skrócenia nazwy do wymaganego limitu użyto generatywnego modelu językowego LLaMA, którego szczegółowy opis znajduje się

w rozdziale poświęconym ocenie przydatności nazw produktów 7.3.9. Model ten może przetworzyć długie nazwy, zachowując kluczowe informacje i syntetyzując je w skróconą formę. Przeprowadzono badanie na próbie 100 losowo wybranych produktów o nazwach dłuższych niż 80 znaków. Po zastosowaniu odpowiedniego zapytania sterującego wykonywaniem zadania przez model: *„Skróć nazwy produktów do 80 znaków ze spacjami. Zachowaj czytelność dla konsumenta, markę produktu, zawartość netto”*, uzyskano pożądane rezultaty, z wyjątkiem jednego przypadku, dla którego algorytm wygenerował nazwę o długości 82 znaków. Wynikało to z faktu, że modele językowe oparte na sztucznej inteligencji generują tekst w sposób probabilistyczny, co oznacza, że nie zawsze mogą precyzyjnie kontrolować długość generowanej odpowiedzi. Chociaż algorytm stara się dostosować do zadanych ograniczeń, takich jak maksymalna liczba znaków, mogą występować drobne odchylenia. Tego typu różnice są wynikiem działania generatywnych modeli językowych, które dążą do zachowania sensu i czytelności tekstu, nawet jeśli prowadzi to do minimalnego przekroczenia limitów.

Walidator [W3.3] utworzono w celu identyfikacji cudzysłowów na początku tekstu. Rozpoczęcie nazwy produktu od cudzysłowu jest dopuszczalne, lecz nie jest to praktyka zalecana, szczególnie w handlu elektronicznym. Algorytm został poddany testom pod kątem poprawności implementacji i założeń na próbie kilkuset produktów. Dla algorytmu [W3.3] wykryto 29 759 produktów do ewentualnej poprawy. W tym rozwiązaniu ponownie przetestowano możliwość podpowiadania właściwej nazwy za pomocą modeli generatywnych typu LLM opisane w rozdziale 7.3.9. Oba testowane modele: GPT-3.5 jak i LLaMA 3.1:70b właściwie zinterpretowały polecenie *„Popraw nazwy produktów, aby nie zaczynały się od cudzysłowu. Inne cudzysłowy wewnątrz tekstu mogą pozostać.”*. Testy przeprowadzone na 200 losowych nazwach wykazały 100% skuteczność działania podpowiedzi, a nazwy produktów zachowały wewnętrzne cudzysłowy używane w celu wyodrębnienia tekstu lub jako osobny znacznik.

W Tabeli 32 przedstawiono przykłady nazw rozpoczynających się od cudzysłowu. Warto zwrócić uwagę na ostatni przykład, w którym nazwa marki celowo została ujęta w cudzysłów co może stanowić świadomą decyzję marketingową mającą na celu jej wyróżnienie.

Tabela 32. Przykłady nazw produktów rozpoczynających się od cudzośliwu

Przykłady nazw rozpoczynających się od cudzośliwu
"Bambusowy Kocyc Dziany z Kapturkiem Słoniki 80cm x 80cm z kaputrkciem"
"Filtr osadnikowy skośny z wbudowanym magneem 1" by RED"
"IZOHAN" folia w płynie PSB op.4 kg
"Magnetyczny separator zanieczyszczeń do instalacji c.o. Model: T 4 by RED"
"Masarnia u Józefa" Szyuka wieprzowa wędzona parzona, w siatce niejadalnej. 0.8kg
"MEBLE SKÓRZANE" WN21 - 5L
"Społem" PSS Kielce Chleb codzienny, pszenno-żytni krojony 500g
"Uniwersalny zestaw zaworów kątowych do kotłów gazowych ZB-KIT-08-U by RED"
"O" zmywak podwójny silver&gold marki "O"- 2 szt.- karton zbiorczy 60 szt

Źródło: opracowanie własne

### 7.3.2. Wykrywanie i sugerowanie podobnych marek

Algorytm wykrywania zbliżonych marek [W3.4] został opracowany zarówno w oparciu o techniki przetwarzania języka naturalnego (NLP) jak i zrozumienia języka naturalnego (NLU). Nazwy marek charakteryzują się dużą różnorodnością w zapisie: mogą być pojedynczymi słowami, zdaniami, mogą zawierać powszechnie znane wyrazy, być neologizmami, zawierać celowe błędy ortograficzne, a także być w obcym języku. Jak wspomniano w Rozdziale 3.2.3, nazwa marki może również odpowiadać nazwie firmy, co jest akceptowalne przy wprowadzaniu produktów na niektórych platformach e-commerce, zatem część użytkowników wprowadza dane rejestrowe firmy wraz z imieniem i nazwiskiem oraz formą prowadzenia działalności jako markę. Z uwagi na złożoność analizy tak różnorodnych przypadków, zdecydowano się na zaprojektowanie walidatora, który bada zbieżność nazw marki wprowadzanej przez użytkownika z istniejącymi w bazie danych bez analizy jej poprawności samej w sobie. Tym samym, algorytm oferuje sugestie nazw marek z listy najczęściej używanych i najbardziej zbliżonych w kontekście danej firmy, jeżeli zaistnieje podejrzenie co do poprawności wprowadzonej nowej marki.

Proces ten zaczyna się od normalizacji tekstu, czyli przekształcenia go do jednolitej formy, co obejmuje usunięcie znaków specjalnych i przekształcenie wszystkich liter na małe. Algorytm następnie analizuje podobieństwo między nazwami marek w obrębie danej firmy. Zastosowano algorytm Jaro-Winklera (Winkler, 1990), który jest efektywny w porównywaniu krótkich ciągów znaków. Algorytm Jaro-Winklera jest szczególnie użyteczny w kontekstach, gdzie drobne różnice na początku ciągu znaków mogą mieć znaczący wpływ na ogólną zbieżność, co zostało potwierdzone w literaturze przedmiotu (Cohen i in., 2003). Przyjęto doświadczalnie, że wartość współczynnika Jaro-Winklera wynosząca minimum 0,89 wyznacza



wysokie podobieństwo pomiędzy nazwami. Dodatkowo wprowadzono ograniczenia dotyczące zbioru danych do porównania. Porównywane nazwy marek muszą być co najmniej 1,5 krotnie częściej użyte w nazwach produktu niż analizowana nazwa. To ograniczenie ma za zadanie zintensyfikować działania walidatora na nowo wprowadzanych markach i może być regulowane lub zdjęte w zależności od analizowanego katalogu produktowego. Wynikiem działania algorytmu jest lista rekomendacyjna podobnych nazw w rozumieniu podobieństwa ciągu tekstu, bez jego rozumienia. Następnie algorytm przechodzi do kolejnego etapu, w którym wykorzystuje model językowy sztucznej inteligencji dużej skali (model LLaMA, omówiony szerzej w Rozdziale 7.3.9) do ostatecznej weryfikacji, czy wykryte podobieństwo może wskazywać na błędne wprowadzenie nazwy marki. W tym celu formułowane jest odpowiednie pytanie do modelu. Pytanie najlepiej wizualizuje rzeczywisty przykład: „Czy marka produktu 'opero CAP' to ta sama marka co ['opero®', 'opero', 'opero® DRAPE', 'opero DRAPE', 'opero PLUS'] tylko napisana z błędami?”. Model LLaMA analizuje kontekst i strukturę nazw, a następnie zwraca odpowiedź w formie [TAK/NIE]. W przypadku odpowiedzi twierdzącej, algorytm generuje listę alternatywnych nazw, które użytkownik może zaakceptować jako poprawne. Jeżeli natomiast odpowiedź brzmi „NIE”, algorytm przyjmuje, że dana nazwa jest poprawna i nie dokonuje zwrotu sugestii użytkownikowi.

Proponowane rozwiązanie charakteryzuje się wieloma zaletami. Po pierwsze, algorytm działa w sposób wydajny, stosując szybkie metody porównawcze w początkowej fazie, co pozwala na przetwarzanie dużych ilości danych w krótkim czasie. Po drugie, zaawansowane techniki zrozumienia języka naturalnego (NLU) są wykorzystywane wyłącznie do weryfikacji już wyselekcjonowanych danych, co znacząco redukuje wymagania obliczeniowe. Po trzecie, integracja modelu SI w procesie walidacji pozwala na wyeliminowanie podobnych nazw marek z różnicami semantycznymi, które mogą nie być wykryte jedynie za pomocą klasycznych algorytmów porównujących ciągi znaków.

Ewaluację algorytmu dokonano na podstawie analizy 100 przypadków różnych marek wskazanych jako sugerowane do poprawy przez algorytm. Z uwagi na brak referencyjnej bazy marek, oszacowano metodą ekspercką, czy algorytm w danej sytuacji powinien zwrócić listę sugestii ze względu na potencjalny błąd użytkownika. Wyniki przedstawiono kolejno w Tabeli 33 oraz w Tabeli 34.

Tabela 33. Weryfikacja ekspercka algorytmu proponowania lepszej marki

Wyniki na próbie (100)	Dane
Prawdziwie pozytywna (TP) – algorytm poprawnie wykrył problem jakościowy	95
Fałszywie pozytywna (FP) – algorytm błędnie wykrył problem jakościowy	5
Precyzja algorytmu dla stratyfikowanej próby	95,00%

Źródło: opracowanie własne

Tabela 33 przedstawia wyniki eksperckiej analizy algorytmu na próbie 100 przypadków. Podczas analizy odrzucono przypadki niewłaściwie wprowadzonych marek. W bazie występowały takie, które zostały wprowadzone jako numery wewnętrzne produktów (np. „\*2005CHO01”), a także połączenia marek z podmarkami („Nivea Men”), zawartością netto lub wariantami („dywan Dywilan kolekcja Atena”). W takich sytuacjach algorytm nie mógł wygenerować prawidłowych podpowiedzi. Algorytm poprawnie wykrył możliwości poprawy w 95% przypadkach. Wszystkie wykryte problemy jakościowe miały poprawnie zaproponowane alternatywy nazw.

Tabela 34. Wpływ walidatora proponowania lepszej marki na całą bazę danych

Wyniki na całej bazie	Nazwy marek	Wpływ na dane produktów
Liczba wykrytych problemów jakościowych w całej bazie	10 970	217 745
Liczba potwierdzeń braku problemów jakościowych	118 811	35 104 663
Zakładana liczba poprawnych wykryć problemów jakościowych	10 421	206 857

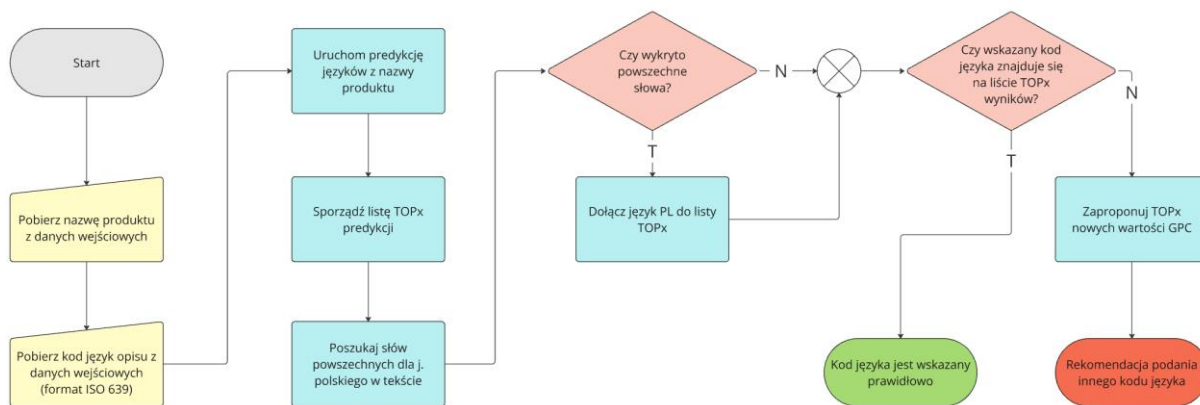
Źródło: opracowanie własne

W Tabeli 34 zaprezentowano wyniki działania algorytmu na całej bazie danych. Algorytm wykrył 10 970 podobnych nazw marek, które wpływały na 217 745 produktów. Brak podobieństw potwierdzono w 118 811 przypadkach, obejmujących 35 104 663 produktów. Oszacowano, że liczba poprawnych wykryć problemów jakościowych może wynosić około 10 421, natomiast błędnych wskazań może być około 548. Liczba wykrytych potencjalnych problemów jakościowych w całej bazie świadczy o użyteczności koncepcji i wykonania algorytmu. Choć nie ma pewności, czy w każdym przypadku algorytm wskazuje realne błędy, które bezwzględnie wymagają poprawy, to idea zwrócenia uwagi użytkownikowi może okazać się kluczowa w poprawie jakości, ponieważ w sposób znaczący wpływa na spójność bazy.

### 7.3.3. Szacowanie i sugerowanie języka opisu

Walidację języka opisu nazwy produktu (P5) oparto o dwie biblioteki, z których pierwsza – FastText, służy bezpośrednio do wykrywania języka, a druga, Morfeusz2, do wykrywania powszechnych słów w języku polskim. FastText (Joulin i in., 2016) jest narzędziem opartym na sztucznej inteligencji, które wykorzystuje algorytm uczenia maszynowego do automatycznego wykrywania języka tekstu. Modele FastText są trenowane na dużych zbiorach danych, gdzie każdy tekst przypisany jest do konkretnego języka, co umożliwia modelowi naukę zróżnicowanych wzorców językowych. Kluczową cechą FastText jest zastosowanie reprezentacji n-gramów, czyli sekwencji kolejnych znaków, co pozwala na lepsze uchwycenie morfologicznych i fonologicznych wzorców charakterystycznych dla każdego języka. Podczas klasyfikacji nowego tekstu, FastText generuje wektorowe reprezentacje zarówno słów, jak i n-gramów, które następnie są wykorzystywane przez model wytrenowany w oparciu o techniki sztucznej inteligencji do identyfikacji języka. Algorytm ten jest wysoce efektywny, co umożliwia szybkie rozpoznanie języka nawet na podstawie krótkich fragmentów tekstu, a dzięki zdolności przetwarzania danych w sposób zbliżony do podejścia probabilistycznego, FastText może dokładnie klasyfikować teksty w 176 językach. FastText wybrano ze względu na skuteczność (Gargova i in., 2022) jak i łatwość implementacyjną.

Niestety w przypadku nazw produktów trudno jest poprawnie oszacować język tekstu, szczególnie tam, gdzie występowało mieszanie języków, na przykład w nazwach produktów zawierających zarówno elementy polskie, jak i angielskie, oraz tam, gdzie przeważały nazwy własne. Kluczowym problemem była także długość nazwy – im krótsza i mniej opisowa, tym potencjalnie większe ryzyko niepewności wyników. Z tego względu zdecydowano się na zastosowanie dodatkowego wsparcia w wykrywaniu języka polskiego poprzez identyfikację słów powszechnie występujących w języku polskim. Należy wspomnieć, że język polski w przedmiotowej bazie danych był wiodący, dlatego takie rozwiązanie miało uzasadnienie w polepszeniu skuteczności wykrywania rodzimego języka. Do analizy nazw powszechnych wykorzystano bibliotekę Morfeusz2 – narzędzie do analizy morfologicznej tekstu dla języka polskiego (Kieraś i Woliński, 2017) . Na Rysunku 20 przedstawiono uproszczony schemat blokowy rozwiązania.



Rysunek 20. Poglądowy przebieg walidacji języka dla treści nazwy produktu  
 Źródło: opracowanie własne

Proces rozpoczyna się od pobrania danych wejściowych, w tym nazwy produktu oraz kodu języka opisu w formacie ISO 639. Następnie uruchamiana jest predykcja języków na podstawie nazwy produktu przy użyciu biblioteki FastText. Wynikiem tej analizy jest lista (TOPx) najbardziej prawdopodobnych języków, w których może być napisana nazwa produktu. W kolejnym kroku algorytm sprawdza, czy w nazwie produktu występują powszechne słowa w języku polskim. Jeżeli takie słowa są obecne, algorytm dołącza język polski (PL) do listy TOPx z wysokim prawdopodobieństwem. Następnie algorytm porównuje podany kod języka z listą TOPx wyników. Jeżeli kod języka znajduje się na tej liście, algorytm uznaje, że kod języka jest prawidłowy i kończy działanie. W przeciwnym razie, algorytm generuje rekomendację podania innego kodu języka z listy TOPx, sugerując najbardziej prawdopodobne języki dla nazwy produktu.

Weryfikacja danych przeprowadzona metodą ekspercką obejmowała analizę 1000 losowo wybranych produktów, pochodzących od różnych firm, co miało na celu wyeliminowanie powtarzalności produktów jednej firmy w próbie. Próbkę podzielono tak, aby połowa produktów wskazywała na błąd przy wstępnie skonfigurowanym walidatorze, natomiast druga połowa była wolna od błędów. Walidator wstępnie klasyfikował produkty na te, w których predykcja języka w TOP5 była zgodna z językiem wskazanym przez użytkownika, oraz na te, w których taka zgodność nie występowała. Zrezygnowano z prowadzenia osobnej analizy dokładności, precyzji i czułości algorytmu w odniesieniu do języka polskiego oraz innych języków, pozostawiając próbki w ich naturalnym, losowym rozkładzie.

Spośród analizowanych danych, 305 próbek zostało oznaczonych jako „niemożliwe do analizy” ze względu na brak dostatecznej czytelności w nazwach, przez co niemożliwe było poprawne wykrycie języka. Wśród nich, w 292 walidator uznał język podany przez wprowadzającego za niewłaściwy – czyli tym samym poprawnie wykazał błąd, natomiast w 13 przypadkach walidator nie wykazał błędu, ponieważ mimo niejasności w nazwie wskazał poprawny język.

W Tabeli 35 przedstawiono wyniki analizy poprawności działania algorytmu walidacji języka dla pozostałych przypadków:

Tabela 35. Wyniki działania algorytmu do wykrywania języka

Opis	Liczba nazw	Procent
<b>Niezgodność walidatora z danymi użytkownika</b>	<b>216</b>	<b>31,1%</b>
Użytkownik nie miał racji	<b>215</b>	<b>99,5%</b>
<b>Poprawna propozycja na pozycji 1</b>	206	95,8%
<b>Poprawna propozycja na pozycji 2</b>	5	2,3%
<b>Poprawna propozycja na pozycji 3</b>	2	0,9%
<b>Poprawna propozycja na pozycji 4</b>	1	0,5%
<b>Poprawna propozycja na pozycji 5</b>	1	0,5%
Użytkownik miał rację	<b>1</b>	<b>0,5%</b>
<b>Brak poprawnej propozycji</b>	1	100,0%
<b>Zgodność walidatora z danymi użytkownika</b>	<b>479</b>	<b>68,9%</b>
Użytkownik nie miał racji	<b>23</b>	<b>4,8%</b>
<b>Poprawna propozycja na pozycji 1</b>	12	52,2%
<b>Poprawna propozycja na pozycji 2</b>	11	47,8%
Użytkownik miał rację	<b>456</b>	<b>95,2%</b>
<b>Poprawna propozycja na pozycji 1</b>	439	96,3%
<b>Poprawna propozycja na pozycji 2</b>	15	3,3%
<b>Poprawna propozycja na pozycji 3</b>	2	0,4%
<b>Suma końcowa</b>	<b>695</b>	<b>100,0%</b>

Źródło: opracowanie własne

Analiza danych z Tabeli 35 wykazała, że w przypadkach niezgodności między predykcją algorytmu a danymi dostarczonymi przez użytkownika, algorytm konsekwentnie przedstawiał trafne propozycje języków w ramach TOP5, co stanowiło podstawę jego działania i pozwalało na zawężenie wyboru do kilku najbardziej prawdopodobnych języków. W 95,8% przypadków algorytm zaproponował właściwy język już na pierwszej pozycji, co wskazuje na jego wysoką dokładność. W trakcie realizacji prac podjęto decyzję o zmianie poszukiwania wskazanego przez użytkownika języka na liście TOP5 predykcji na wartość progową predykcji

na poziomie równiej lub wyższej niż 11,84%, poniżej której powinno zostać wyzwolone zdarzenie alertu. Ustalono, że określenie tej wartości progowej jest bardziej efektywne niż poleganie wyłącznie na top 5 wyników, ponieważ pozwala na precyzyjniejszą identyfikację potencjalnych błędów. Zaobserwowano również, że predykcje o wartości poniżej 9,7% nie będą wybierane przez użytkowników, co sugeruje, że nie muszą być wyświetlane w interfejsie. Usunięcie takich predykcji z widoku użytkownika może dodatkowo uprościć proces wyboru i poprawić ogólne doświadczenie użytkownika systemu, jednocześnie zachowując wysoką trafność i efektywność algorytmu.

Wykorzystanie biblioteki Morfeusz2 znacząco poprawiło jakość walidatora. W analizie 443 polskich nazw Morfeusz2 nie zidentyfikował nazw powszechnych w 7 przypadkach, podczas gdy FastText nie wskazał języka polskiego w 38 przypadkach, a w 12 przypadkach polski znajdował się na niższych pozycjach listy. Z tego powodu dostrojono algorytm w taki sposób, aby wykrycie powszechnego słowa w języku polskim przez bibliotekę Morfeusz2 skutkowało przypisaniem prawdopodobieństwa języka polskiego dla nazwy produktu na poziomie 50%, co mogło plasować język polski na drugim miejscu na liście najbardziej prawdopodobnych języków, w sytuacji dużego prawdopodobieństwa innego języka. Na testowym zbiorze taka sytuacja miała miejsce w 23 przypadkach.

W Tabeli 36 przedstawiono przykłady nazw produktów, które negatywnie wpływały na skuteczność działania algorytmu walidacyjnego. Każda z tych nazw stwarzała specyficzne wyzwania dla algorytmu, co uwidacznia jego ograniczenia w kontekście precyzyjnego rozpoznawania języka opisu produktu.

Tabela 36. Przykłady specyficznych nazw produktów

Nazwa opisowa produktu	Język wskaz. przez użytkownika	Komentarz
Medisept Medisept Velox Oxy ETA 5l - płyn (export wersja włoska) 5l	it	Wprowadzający dane pomylił język opisu z krajem rynku docelowego.
Aries Watches Smartwatch GWR101 RG DAMSKI	pl	Nazwa produktu zawiera zapożyczenia, a jedynym wskaźnikiem języka polskiego jest słowo „DAMSKI”, które algorytm traktuje jako nazwę własną, przez co preferuje język angielski.
Hantla 20 kg	pl	Błąd ortograficzny wpływa na skuteczność rozpoznania języka.
AGROMUSCA 20 SP 100g	pl	Niejasna, wieloznaczna nazwa, utrudniająca poprawną klasyfikację języka.
4F OBML260-25S-46	pl	Nazwa składa się z symboli i kodów, które nie dostarczają wystarczających informacji do jednoznacznego rozpoznania języka.

Nazwa opisowa produktu	Język wskaz. przez użytkownika	Komentarz
Armchair Glamour Turkusowy	pl	Połączenie dwóch języków w nazwie wprowadza algorytm w błąd.
Roastains Kawa Panama Beautiful Eyes FILTER 250g	pl	Połączenie dwóch języków (polskiego i angielskiego) komplikuje klasyfikację języka.
GARDEN-EQUIPMENT EIGHTY5 Construction MINI LAMPKA / ŻARÓWKA LED 5V / 1W Warm Light	pl	Silny wpływ nazwy marki na końcową predykcję języka, preferujący język angielski.
Star Wars karty 5 pack (Force Attax Movie)	pl	Wpływ nazw własnych na ocenę algorytmu, skutkujący preferowaniem języka angielskiego.
mozartt.com - 4x steel chainring bolts and nuts 4szt	en	Polska jednostka miary w angielskiej nazwie produktu wpływa na rozpoznanie języka.

Źródło: opracowanie własne

Jedyny w ewaluowanym zbiorze przypadek błędnej predykcji to znalezienie poprawnego języka dla produktu „*Aries Watches Smartwatch GWR101 RG DAMSKP*”. Algorytm błędnie wskazał język angielski z prawdopodobieństwem 83%, zamiast oczekiwanego rozpoznania języka polskiego. Jedynym słowem sugerującym język polski jest „*DAMSKP*”, jednak algorytm zidentyfikował je za nazwę własną ze względu na pisownię dużymi literami, co w połączeniu z dominującymi elementami w języku angielskim i brakiem rozpoznania jako nazwy powszechnej zapożyczonego słowa angielskiego pochodzenia „*Smartwatch*”, doprowadziło do nieprawidłowej klasyfikacji. Przykłady te ukazują ograniczenia algorytmu, szczególnie w sytuacjach, gdzie nazwy produktów zawierają błędy ortograficzne, literówki, brak znaków diakrytycznych, czy też są pisane dużymi literami. Dodatkowo, algorytm może preferować rozpoznawanie języka polskiego w przypadkach, gdy nazwy marek są jednocześnie nazwami powszechnymi w języku polskim, jak „*Motyl*” czy „*Koło*”, niezależnie od innych słów w nazwie produktu.

Pomimo wyzwań związanych z krótkimi i wieloznacznymi nazwami produktów, badanie wykazało, że możliwe było skuteczne wykrycie języka na podstawie opisowej nazwy, nawet przy ograniczonej liczbie słów. Poprawność odpowiedzi algorytmu zależy w dużej mierze od jakości samej nazwy, liczby słów oraz obecności specyficznych wyrazów, ponieważ algorytmy wykrywania języka osiągają lepsze wyniki na większych próbkach tekstu, podczas gdy krótkie nazwy produktów, zwłaszcza te, zawierające obcojęzyczne marki i niewielkie opisy, które brzmią podobnie w różnych językach, mogą prowadzić do mniej trafnych

rezultatów. Ostatecznie, decyzja rozstrzygająca tą kwestię zawsze jednak należy do osoby, która odbierze informację o potencjalnym problemie.

Szczegółowe wyniki ewaluacji przedstawiono w Tabeli 37, a wpływ walidatora na całą bazę produktów zaprezentowano w Tabeli 38.

Tabela 37. Weryfikacja ekspercka algorytmu walidacji języka

Wyniki na próbie (508/492)	Dane
Liczba poprawnie wykrytych problemów jakościowych	507
Liczba błędnie wykrytych problemów jakościowych	1
Liczba poprawnie potwierdzonych braków problemów jakościowych	456
Liczba błędnie potwierdzonych braków problemów jakościowych	36
Dokładność algorytmu	96,30%
Precyzja algorytmu	99,80%
Czułość algorytmu	93,37%
Poprawne rekomendacje algorytmu (TOP3)	42%, ale 99,6% dla poprawnych nazw

Źródło: opracowanie własne

Algorytm osiągnął dużą dokładność (96,30%) oraz wysoką precyzję (99,80%), co świadczy o jego zdolności do prawidłowego identyfikowania problemów jakościowych związanych z nazwami produktów. Czułość algorytmu, wynosząca 93,37%, potwierdza jego skuteczność w wykrywaniu rzeczywistych problemów, choć wykazuje pewne ograniczenia w pełnym ich wykryciu.

Warto zauważyć, że w próbie użytej do weryfikacji, liczba rekordów nie wyniosła 500/500, ponieważ analiza została dostosowana do procentowego rozkładu wyników (zamiast klasyfikacji wyłącznie w TOP5). Ponadto, poprawne rekomendacje algorytmu na poziomie TOP3 wyniosły 42%, co jest zadowalającym wynikiem, szczególnie biorąc pod uwagę, że w kontekście poprawnych nazw, jego skuteczność sięga aż 99,6%. Oznacza to, że algorytm z dużą pewnością dostarcza użytkownikom trafnych rekomendacji językowych o ile nazwy produktów są zrozumiałe.

Tabela 38. Wpływ walidatora języka na całą zawartość bazy danych

Wyniki na całej bazie	Wpływ na dane
Liczba wykrytych problemów jakościowych w całej bazie	4 787 554
Liczba potwierdzeń braku problemów jakościowych	40 171 746
Zakładana liczba poprawnych wykryć problemów jakościowych	4 778 129

Źródło: opracowanie własne



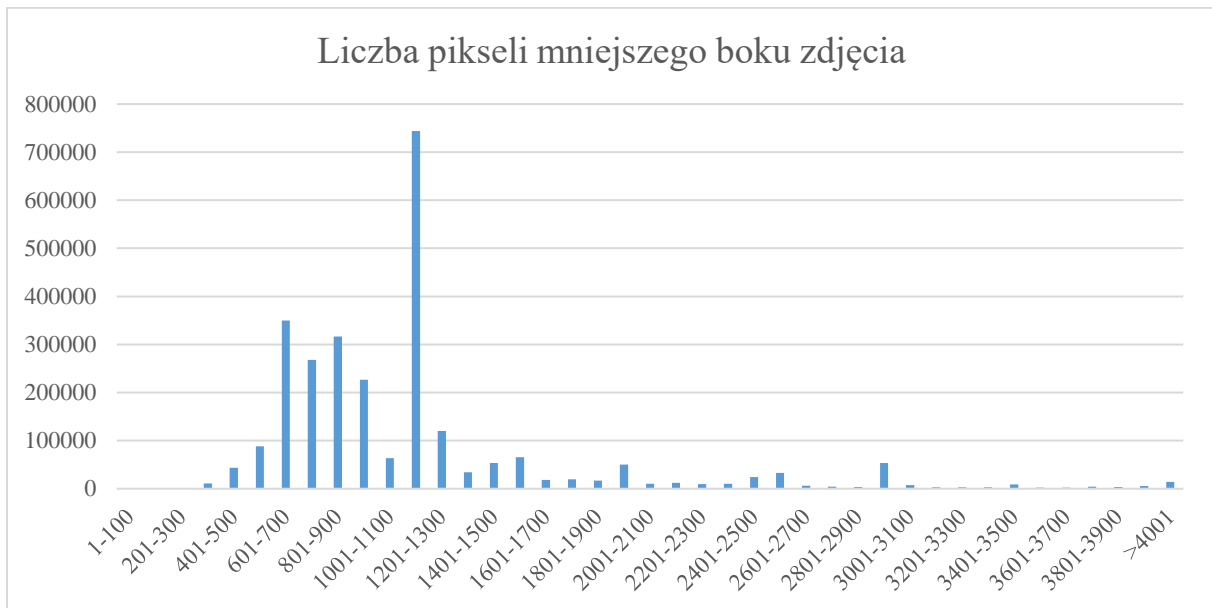
W odniesieniu do całej bazy danych, algorytm zidentyfikował 4 787 554 potencjalnych problemów jakościowych, co stanowi istotny wkład w poprawę jakości danych w bazie. Zakładana liczba poprawnych wykryć problemów jakościowych wyniosła 4 778 129. Szacuje się, że błędnych wskazań może być przy tym około 9 424, co wskazuje na bardzo niską liczbę fałszywych alarmów.

### **7.3.4. Kontrola wymiarów zdjęć produktów**

Operacje takie jak [W3.6] "Zbyt niska rozdzielczość zdjęcia dla niektórych zastosowań e-commerce" oraz [W3.7] "Zbyt niska rozdzielczość zdjęcia dla większości zastosowań e-commerce" mogą zostać zaimplementowane z wykorzystaniem większości współczesnych bibliotek do przetwarzania obrazów dostępnych w nowoczesnych językach programowania. Algorytm [W3.6] sprawdza, czy zdjęcie spełnia wymagania dotyczące rozmiaru, zajmując co najmniej 0,3 megapiksela, przy jednoczesnych wymiarach poniżej 900 pikseli zarówno dla szerokości, jak i wysokości. Parametry te zostały określone na podstawie specyficznych potrzeb i zastosowań, z uwzględnieniem wyników badania 3.2.4. Algorytm [W3.7] kontroluje, czy zdjęcie zajmuje obszar poniżej 0,3 megapiksela, co jest niewystarczające dla większości zastosowań e-commerce. Podobnie jak proponowane operacje na ciągach tekstowych opisane w Rozdziale 7.3.1, działania sprawdzające rozdzielczość zdjęć opierają się na jasno zdefiniowanych założeniach i regułach, a celem ich zastosowania jest zwiększenie świadomości autorów danych na temat wymagań jakościowych oraz dostarczenie administratorom precyzyjnych informacji o stanie jakości posiadanych danych. Komunikacja o niskiej jakości zdjęcia nie powinna więc blokować możliwości wprowadzenia takiego zdjęcia do bazy danych, ale autor powinien znać konsekwencje takich działań.

W ramach badania przeprowadzono analizę dotyczącą liczby zdjęć o zbyt niskiej rozdzielczości przechowywanych obecnie w bazie. Wyniki ujawniły, że 0,6% wszystkich zdjęć ma zbyt niską rozdzielczość do większości zastosowań e-commerce, a 36% zdjęć ma zbyt niską rozdzielczość do części z nich. Na Rysunku 21 przedstawiono rozkład liczby pikseli mniejszego boku zdjęcia. Wykres wyraźnie wskazuje, że największa liczba zdjęć ma mniejszy bok o długości około 1201-1300 pikseli, co potwierdza, że wiele zdjęć w bazie danych jest przygotowanych w standardowych rozdzielczościach, typowych dla średniej jakości obrazów stosowanych w e-commerce. Warto również zauważyć, że liczba zdjęć gwałtownie spada przy

mniejszych rozdzielczościach (poniżej 600 pikseli) oraz przy bardzo dużych rozdzielczościach (powyżej 2000 pikseli).



Rysunek 21. Liczba pikseli mniejszego z boków zdjęcia produktu  
Źródło: opracowanie własne

Na Rysunku 22 przedstawiono liczbę zdjęć w bazie danych w funkcji liczby megapiksela. Największa liczba zdjęć znajduje się w przedziale 2,4-2,7 megapiksela, co jest zgodne z obserwacją z Rysunku 21, gdzie dominująca liczba zdjęć miała mniejszy bok w zakresie 1201-1300 pikseli. Rozkład wskazuje również, że istnieje duża liczba zdjęć o bardzo niskiej rozdzielczości (poniżej 1 megapiksela), co może być problematyczne w kontekście wymagań e-commerce, szczególnie w przypadkach, gdy zdjęcia te muszą spełniać określone standardy jakości.



Rysunek 22. Liczba megapikseli zdjęć produktów  
 Źródło: opracowanie własne

Istnieją rozwiązania z dziedziny sztucznej inteligencji, które mogą powiększać rozdzielczości zdjęć, zachowując przy tym wysoki poziom szczegółowości i minimalizując artefakty (Wang i in., 2018). Algorytmy te potrafią nie tylko zwiększać liczbę pikseli w obrazie, ale także rekonstruować brakujące detale, co znacząco podnosi wizualną jakość zdjęć, zapewniając lepszą percepcję produktów przez klientów. Wykorzystanie takich technologii mogłoby stanowić istotne wsparcie jako dodatkowa propozycja automatycznej poprawy w ramach wdrożenia proponowanego walidatora. Jednakże zwiększanie rozdzielczości może wiązać się z ryzykiem wprowadzenia artefaktów, które mogą wpłynąć na wiarygodność prezentacji produktów, szczególnie przy dużym powiększeniu. Z tego względu, zdecydowano o rezygnacji z dalszych badań w tym kierunku.

### 7.3.5. Wykrywanie duplikatów zdjęć produktu

Proces wykrywania duplikatów zdjęć [W3.8] można zrealizować za pomocą różnych technik, w tym funkcji skrótu MD5 do porównywania binarnej formy zdjęcia, hashy percepcyjnych do analizy wizualnych cech obrazu, których długość może być dostosowana w zależności od specyficznych wymagań, lub hashowania neuronowego (deep hashing) (Liong i in., 2015), które wykorzystuje głębokie uczenie do tworzenia efektywnych i precyzyjnych reprezentacji binarnych, umożliwiających szybkie porównywanie obrazów. Wykorzystanie technik zależne jest od potrzeb i dokładności porównań.

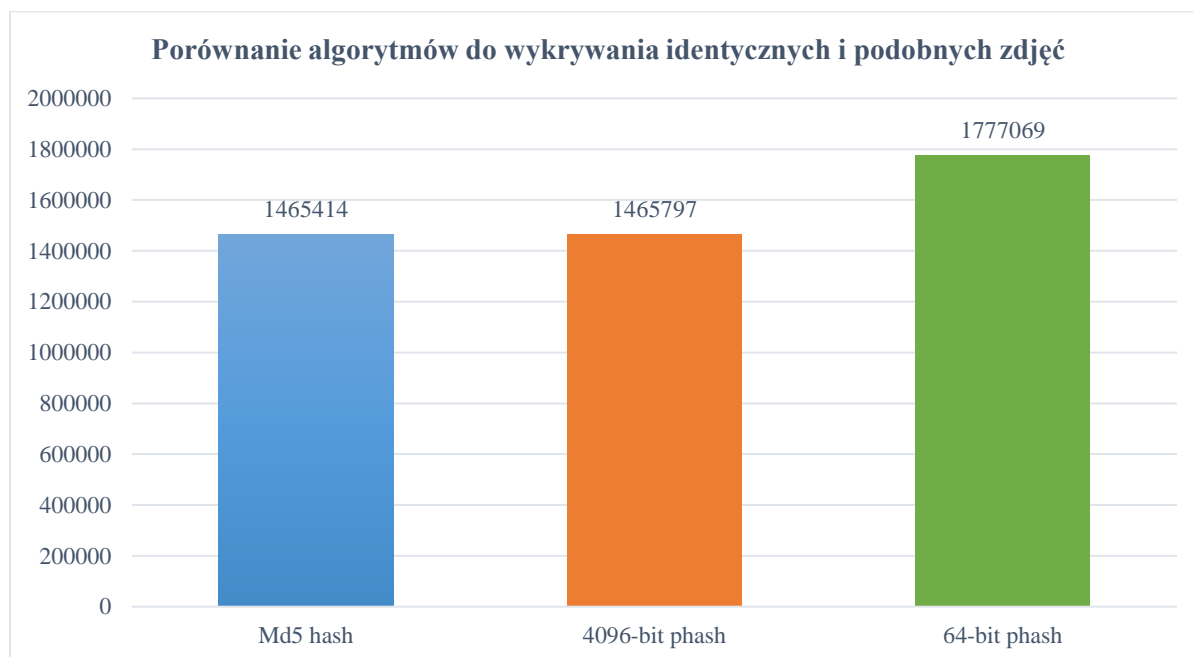
Funkcja MD5 (Message-Digest Algorithm 5) jest jedną z kryptograficznych funkcji skrótu, która przekształca dane wejściowe (w tym wypadku plik zdjęcia) w skrót o stałej długości wynoszącej 128 bitów (16 bajtów). Jest powszechnie stosowana do generowania sum kontrolnych i weryfikacji integralności danych. Proces generowania skrótu MD5 polega na przetwarzaniu danych w blokach o stałej wielkości i stosowaniu szeregu operacji bitowych, które prowadzą do uzyskania skrótu o określonej długości. Ten sam zestaw danych wejściowych zawsze wygeneruje identyczny skrót MD5. W przypadku użycia tej funkcji do hashowania zdjęć, MD5 będzie czuły na każdy piksel i/lub format zapisu zdjęcia generując zupełnie inny skrót. Z kolei hash percepcyjny to metoda generowania skrótu, która opiera się na analizie wizualnych cech obrazu, zamiast dokładnego przetwarzania danych wejściowych pliku, w którym to zdjęcie jest zakodowane, jak w przypadku działania MD5. Hashe percepcyjne, często nazywane „odciskami palca” obrazów, są tworzone przy użyciu technik takich jak transformacja kosinusowa. Ta metoda umożliwia ekstrakcję kluczowych informacji wizualnych z obrazu i ich zakodowanie w postaci ciągu bitów. Długość tego ciągu może być stała (np. 64 bity) lub dostosowana do specyficznych wymagań. Hashe percepcyjne mogą pomijać drobne różnice w zdjęciach, takie jak zmiany w jasności, kolorystyce czy rozdzielczości w zależności od długości ciągu generowanych bitów „odcisku palca”. Co więcej, obrazy o wysokim stopniu podobieństwa generują zbliżone skróty, różniące się jedynie kilkoma bitami. Pomiar podobieństwa między hashami można przeprowadzić za pomocą odległości Hamminga, która analizuje różnice bit po bicie, pozwalając na precyzyjne określenie stopnia podobieństwa między obrazami. Hashowanie neuronowe to zaawansowana technika, która łączy w sobie możliwości głębokiego uczenia i hashowania, tworząc wysoce efektywne i precyzyjne reprezentacje binarne obrazów. W odróżnieniu od tradycyjnych metod, takich jak MD5 czy hash percepcyjny, hashowanie neuronowe wykorzystuje sieci neuronowe do ekstrakcji cech z obrazu, a następnie przekształca te cechy w binarne wektory (hashe). Te wektory mogą być łatwo porównywane przy użyciu operacji bitowych, takich jak XOR, co umożliwia szybkie i dokładne porównywanie dużych zbiorów danych. Jedną z kluczowych zalet hashowania neuronowego jest jego zdolność do dostosowywania się do złożonych wzorców wizualnych w obrazach, co pozwala na lepsze rozróżnianie nawet subtelnych różnic między obrazami, które mogą nie być uchwycone przez tradycyjne metody. Dzięki temu, hashowanie neuronowe jest szczególnie użyteczne w systemach wymagających wysokiej precyzji w wykrywaniu podobieństw. Hashowanie neuronowe przewyższa hashowanie percepcyjne w wykrywaniu bardziej złożonych, abstrakcyjnych i semantycznych podobieństw między obrazami. Jest ono bardziej elastyczne i zdolne do radzenia sobie z transformacjami,

zmianami kontekstowymi, różnicami stylistycznymi oraz złożonymi zmianami w obrazach, które mogą być problematyczne dla hashy percepcyjnych.

W kontekście walidacji kluczowym zadaniem była identyfikacja duplikatów zdjęć oraz informowanie o zbyt częstym ich przypisywaniu do różnych produktów. Ustalono, że walidator miał wywołać zdarzenie błędu, gdy to samo zdjęcie zostało przypisane do więcej niż dziesięciu produktów o różnych numerach GTIN. Liczba ta została dobrana w konsultacji z ekspertami ds. zarządzania danymi, aby uniknąć nadmiernie restrykcyjnego podejścia. Taki próg umożliwiał akceptację zdjęć reprezentujących różne warianty tego samego produktu, np. odzieży w różnych rozmiarach, zapewniając jednocześnie, że powielanie zdjęć nie prowadziło do dezinformacji ani problemów związanych z unikalnością treści, co mogłoby negatywnie wpłynąć na doświadczenie kupujących.

Pomimo dostępnych nowoczesnych technologii, takich jak hasze neuronowe, nie miały one zastosowania w tym przypadku. Zastosowano tradycyjne metody porównywania skrótów, które w kontekście tego walidatora zapewniały wystarczającą dokładność przy znacznie niższym zapotrzebowaniu na zasoby obliczeniowe.

W celu wyboru optymalnego algorytmu przeprowadzono badanie efektywności trzech różnych wariantów skrótu: z wykorzystaniem MD5, 64-bitowego hash'a percepcyjnego (phash) oraz 4096-bitowego phash.



Rysunek 23. Porównanie algorytmów do wykrywania identycznych i podobnych zdjęć  
Źródło: opracowanie własne

Na Rysunku 23 przedstawiono wyniki porównania trzech algorytmów dla całej bazy zdjęć produktów. Algorytm kryptograficzny MD5, który generuje 128-bitowe skróty, uzyskał wynik 1 465 414 zduplikowanych zdjęć. Algorytm 64-bit phash - oparty na analizie percepcyjnej uzyskał wynik 1 777 069 podobnych zdjęć, co jest najwyższą wartością spośród badanych algorytmów. Jednakże, jak pokazano na Rysunku 23, 64-bitowy phash wykazywał tendencję do nadmiernego uogólniania zdjęć. Przykładowo, różne mieszanki przypraw oraz różne modele pokrowców na siedzenia samochodowe zostały sklasyfikowane jako zbyt podobne, co prowadziło do błędnych wniosków na temat ich podobieństwa. Nadmierna generalizacja spowodowała, że algorytm nie był w stanie prawidłowo rozróżnić wariantów produktów na zdjęciach. Aby rozwiązać problem nadmiernej generalizacji, zastosowano 4096-bitowy phash, który uzyskał wynik 1 465 797 wykrytych podobnych zdjęć. Dzięki większej długości skrótu, 4096-bitowy phash okazał się bardziej precyzyjny w identyfikacji różnic między obrazami. Algorytm ten zachował istotne detale wizualne, które umożliwiły prawidłową klasyfikację zdjęć, unikając błędów, które występowały przy użyciu 64-bitowego phash, oraz przy zachowaniu większej precyzji niż MD5, który bazował jedynie na danych binarnych zdjęć. W 228 przypadkach binarna forma zapisu zdjęcia jedno lub wielokrotnie różniła się dla wizualnie tego samego zdjęcia.



Rysunek 24. Przykładowe podobne zdjęcia z tym samym 64-bitowym odciskiem palca  
 Źródło: opracowanie własne


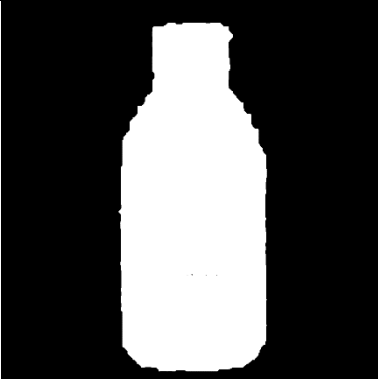



Analiza wyników potwierdziła skuteczność zastosowania porównywania duplikatów jako metody doskonalenia jakości danych wizualnych, co uwidoczniło się w dużej liczbie wykrytych problemów w bazie danych. Algorytmy percepcyjne, dzięki analizie rzeczywistej treści obrazu, okazały się bardziej precyzyjne w rozróżnianiu zdjęć niż algorytmy kryptograficzne, które koncentrowały się jedynie na analizie formy zapisu. Kluczowym czynnikiem wpływającym na skuteczność była właściwa długość skrótu, która znacząco wpływała na dokładność rozpoznawania podobieństw między obrazami. Dodatkową, lecz bardzo istotną funkcjonalnością algorytmu było automatyczne oznaczanie przypadków, w których zdjęciem produktu było logo firmy lub zdjęcie zastępcze, co jest częstym problemem w bazach danych tworzonych przez społeczności, gdzie takie obrazy są często powielane przy różnych produktach. W wyniku analizy zidentyfikowano ponad 18 tysięcy tego typu zdjęć.

### **7.3.6. Ocena jakości tła zdjęcia**





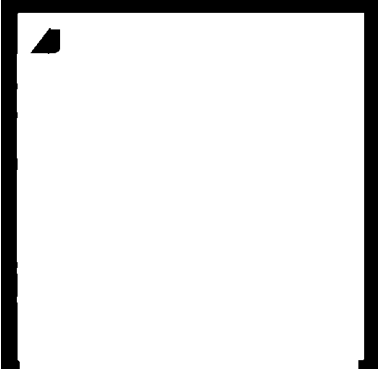

Zgodnie z przeprowadzonymi badaniami opisanymi w Rozdziale 3.2.4, zdjęcie produktu powinno być umieszczone na białym (w niektórych przypadkach jasnym szarym) lub przezroczystym tle, z zalecanym procentem pokrycia 75%, a nawet 95% - w zależności od wytycznych platform e-commerce. Opracowano dwa walidatory: [W3.9], który wylicza kolor i jednorodność tła, oraz [W3.10], który bada i sugeruje korekcję pokrycia. Dla oceny jednorodności tła w początkowej fazie badań opracowano algorytm, który dokonywał jedynie pomiaru skrajnych pikseli zdjęcia w każdym z narożników (5 pikseli per narożnik), aby w ten sposób ocenić kolor i jednolitość tła – zakładając, że produkt nie zajmuje 100% powierzchni zdjęcia. Taki walidator posiadał wiele niedoskonałości, natomiast był bardzo szybki w implementacji i uruchomieniu, dzięki czemu w krótkim czasie można było dokonać wstępnego pomiaru skali problemu. W wyniku pomiaru określono, że dotyczy on 40% wszystkich zdjęć. W dalszych pracach, przy konstruowaniu architektury dla obu walidatorów jako kluczowy element wskazano konieczność separacji tła od głównego obiektu. Badania odpowiednich narzędzi do realizacji tego celu rozpoczęto od użycia biblioteki OpenCV dostępnej w języku Python, gdzie separację można wykonać przy wykorzystaniu techniki segmentacji metodą „GrabCut” (używającej model mieszany Gaussa i algorytm grafowy). Algorytm okazał się, jednakże niewystarczająco skuteczny, zwłaszcza w przypadkach, gdy produkt był w ramce lub tło nie było jednolite. W ramach badania, skorzystano więc z bardziej zaawansowanej i skuteczniejszej metody – biblioteki RemBG. Jest to odpowiednio wytrenowany model sztucznej inteligencji oparty o sieć neuronową U2Net (Qin i in., 2020).

U2Net to z kolei zaawansowany model sieci neuronowej opracowany z myślą o precyzyjnym segmentowaniu obiektów na obrazach. Struktura tej sieci wykorzystuje innowacyjną architekturę "nested U-structures", czyli zagnieżdżone struktury w kształcie litery U, co jest rozwinięciem koncepcji znanej z sieci U-Net, szeroko stosowanej w przetwarzaniu obrazów medycznych i innych zadaniach segmentacji. Charakterystyczną cechą U<sup>2</sup>-Net jest jej zdolność do dokładnego wyodrębniania nawet bardzo skomplikowanych i drobnych szczegółów na obrazach. W efekcie zastosowania biblioteki uzyskano pożądane wyodrębnienie tła od obiektu/obiektów. Przykłady użycia i porównanie działania obu bibliotek przedstawiono w Tabeli 39.

Tabela 39. Przykłady wyodrębnienia tła z wykorzystaniem różnych metod

#	Źródło	GrabCut	RemBG
1			
2			

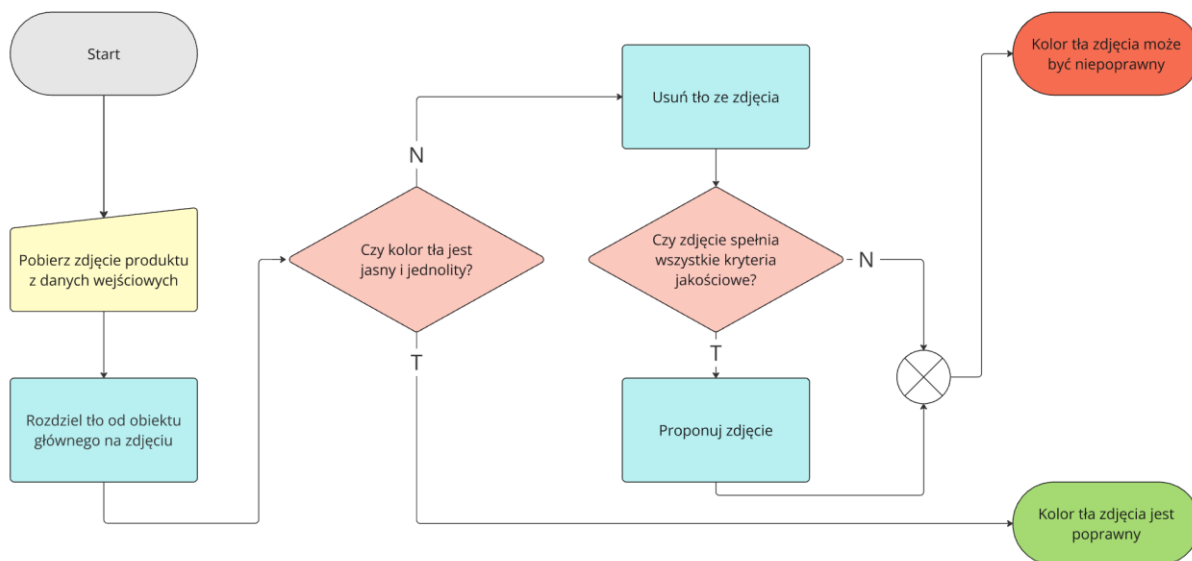


#	Źródło	GrabCut	RemBG
3			
4			

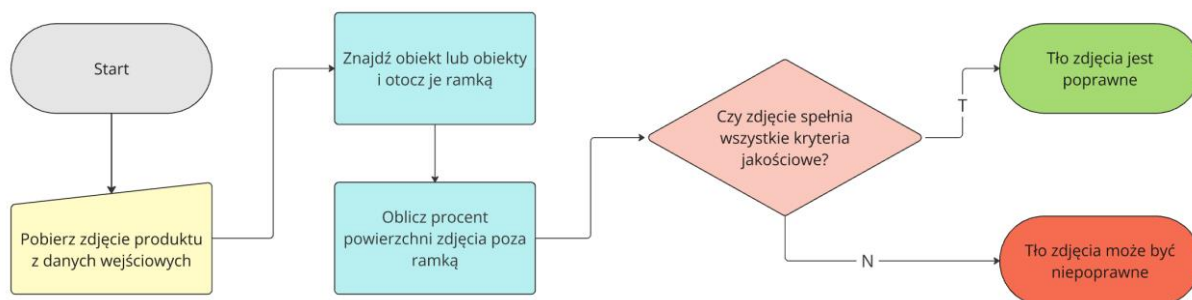
Źródło: opracowanie własne

Należy zauważyć, że dla przykładu czwartego, rezultat wyodrębnienia tła jest nieprawidłowy niezależnie od użytej biblioteki. Jest to celowa wizualizacja ograniczeń zastosowanego rozwiązania, mająca na celu podkreślenie wyzwań związanych z segmentacją w przypadku szczególnie trudnych obrazów.

Realizując dalsze prace nad wyodrębnionym tłem, przeprowadzono pomiary koloru, różnic kolorystycznych oraz zajętego obszaru tła w stosunku do produktu. W wyniku tych analiz utworzono dwa walidatory, których koncepcje zaprezentowano odpowiednio na Rysunku 25 i Rysunku 26.



Rysunek 25. Schemat blokowy procesu walidacji kolorystyki tła.  
Źródło: opracowanie własne



Rysunek 26. Schemat blokowy procesu walidacji powierzchni tła.  
Źródło: opracowanie własne

Walidator kolorystyki tła (Rysunek 25) ładuje obraz, wyodrębnia tło, a następnie analizuje jakość tła, obliczając średnią intensywność i jednolitość krawędzi obrazu. Obraz jest podzielony na segmenty obejmujące krawędzie (górna, dolna, lewa, prawa) oraz centralną część. Średnia intensywność kolorów pikseli w danym segmencie obrazu jest obliczana jako średnia wartość składowych koloru (RGB). Wyższa średnia intensywność może wskazywać na jaśniejsze tło, podczas gdy niższa intensywność sugeruje ciemniejsze tło. Jednolitość obrazu odnosi się do stopnia, w jakim piksele mają zbliżone wartości kolorów. Analiza jednolitości polega na obliczeniu odchylenia standardowego jasności pikseli w segmencie obrazu. Mniejsze odchylenie standardowe wskazuje na bardziej jednolite tło, co jest pożądanym wynikiem oceny.

jakości tła jest zwracany jako wartość liczbowa, która obliczana jest na podstawie różnicy między średnią intensywnością a jednolitością dla wszystkich segmentów obrazu. Tą wartość liczbową można interpretować jako miarę ogólnej jakości tła, gdzie wyższe wartości wskazują na lepszą jakość tła, charakteryzującą się odpowiednią jasnością i jednolitością bez nadmiernych szumów czy niejednorodności. Próg akceptowalnej jakości – 150 dla liczb od 0 do 255 ustalono na podstawie analizy wyników uzyskanych na bazie 500 zdjęć. Wybrano wartości, które najlepiej oddają granicę między akceptowalnym a nieakceptowalnym tłem, zapewniając, że obrazy z wartością powyżej progu są wystarczająco jasne i jednolite, a obrazy poniżej progu wymagają dalszej poprawy jakości. Była to miara z dużą przestrzenią tolerancji dla zdjęć tworzonych w warunkach naturalnych, z cieniami i gradientem, przy zachowaniu akceptowalnej jakości. Na Rysunku 27 przedstawiono wykres liczby zdjęć w funkcji poziomu jakości tła, gdzie można zaobserwować wysoką tolerancję dla przyjętej miary w stosunku do zauważalnych trendów. Łącznie, poza dopuszczalnym progiem zaklasyfikowano 559740 zdjęć do poprawy. W przypadku, gdy tło nie jest akceptowalne, algorytm usuwa tło ze zdjęcia, a wynik – nowe zdjęcie – poddawane jest walidacjom dotyczącym zdjęcia z poziomu 1, 2 i 3. Jeżeli walidacja nie wykaże błędów, nowe zdjęcie może być zaproponowane użytkownikowi wraz z informacją o problemach jakościowych istniejącego zdjęcia. W Tabeli 40 przedstawiono przykłady zdjęć niespełniających kryteriów jakości tła. Wśród nich są zarówno zdjęcia produktów w plenerze, jak i zdjęcia przedstawiające renderowane grafiki, lub też fragmenty produktów bez tła, co nie jest pożądane w przypadku zdjęcia produktowego. Poprawność rekomendacyjną algorytmu oceniono na podstawie stu próbek, dzieląc zdjęcia na błędnie wprowadzone przez użytkownika (nie przedstawiające produktu, np. wyżej wymienione zdjęcia grafik), zdjęcia akceptowalne do publikacji po wyczyszczeniu tła, oraz zdjęcia nieakceptowalne ze względu na nieprawidłowe wyczyszczenie tła. Skuteczność wyliczono na poziomie 68% na bazie 62 poprawnych zdjęć.



Rysunek 27. Liczba zdjęć produktów w funkcji poziomu jakości tła zdjęcia  
 Źródło: opracowanie własne

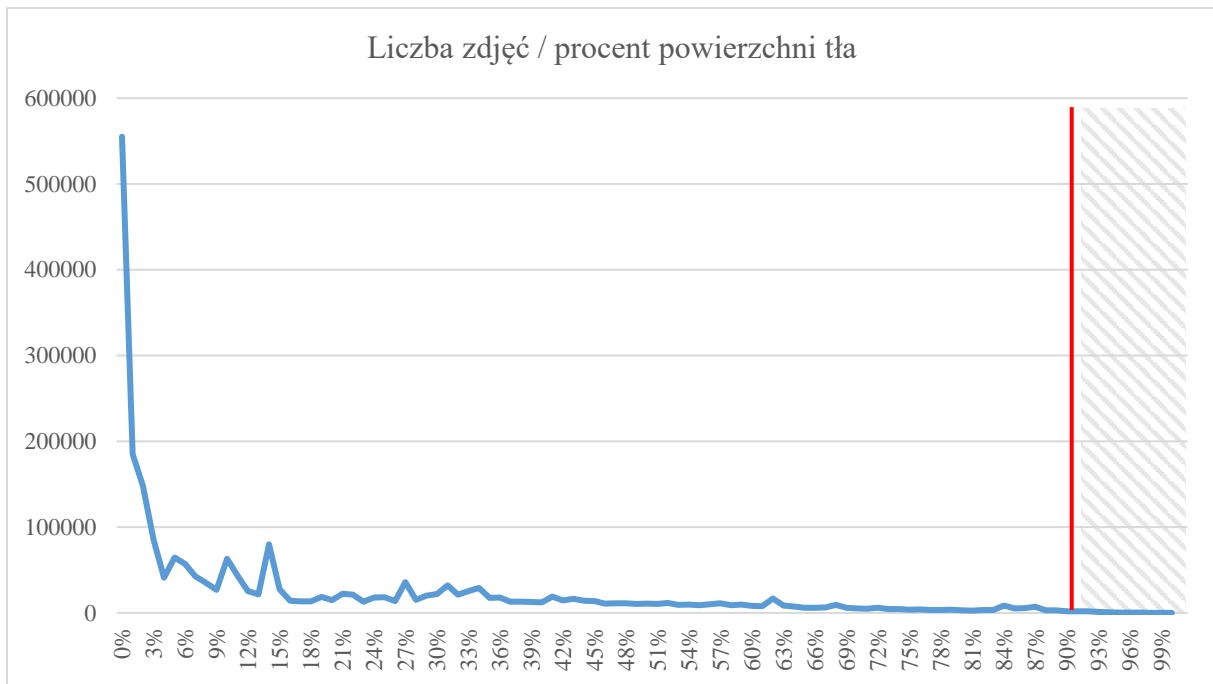
Tabela 40. Przykłady zdjęć niespełniających kryteriów jakościowych tła


Źródło: opracowanie własne

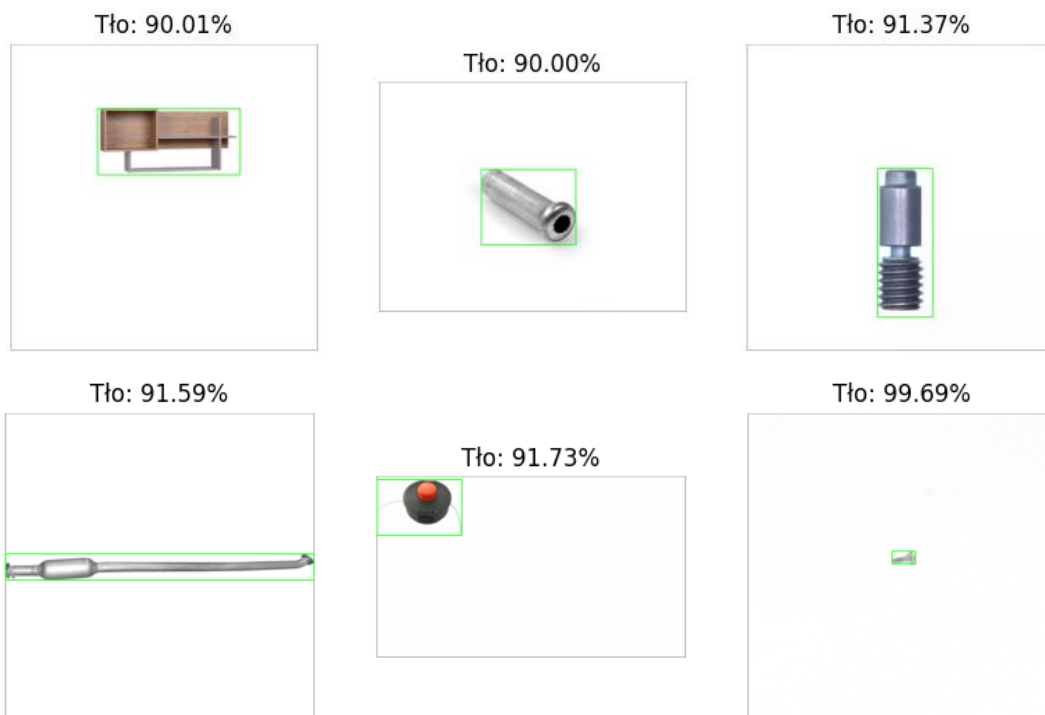
Walidator powierzchni tła, którego algorytm przedstawiono na Rysunku 26 działa podobnie do algorytmu badającego kolor i jednolitość tła, rozpoczynając działanie od załadowania zdjęcia produktu i podziału obrazu na tło i obiekt. Następnie obliczana jest powierzchnia tła w stosunku do całego zdjęcia. Sama idea algorytmu bazuje na wynikach badań przeprowadzonych w sekcji 3.2.4, które dotyczyły analizy standardów i wytycznych zamieszczania zdjęć na platformach e-commerce, gdzie jednym z parametrów był zalecany minimalny procent pokrycia zdjęcia produktem wahający się od 75% do 90% całej powierzchni. Poniżej tej wartości oceniano, że tło może dominować nad produktem, co

negatywnie wpływa na estetykę i czytelność zdjęć produktów, utrudniając ich atrakcyjną prezentację i identyfikację przez klientów.

Zgodnie z założeniami, początkowo ustalono próg dopuszczalnego udziału tła na poziomie 25%, co oznaczało, że co najmniej 75% zdjęcia powinno być zajęte przez produkt. Po analizie 2000 próbek zdjęć zarekomendowano podniesienie limitu, ustalając maksymalny dopuszczalny udział tła na poziomie aż 90%. Zmodyfikowano również sposób obliczania powierzchni tła, uwzględniając przestrzeń jedynie poza wyznaczoną ramką, w której zidentyfikowano produkt lub grupę produktów, zamiast obliczania całego tła. Powodem tej zmiany było dostosowanie rekomendacji do specyfiki fotografowania produktów o nieregularnych kształtach i długich wymiarach. Rzeczywiste warunki prezentacji produktów, takich jak igły, wędki, kije, listwy czy lampy, przedstawianych w orientacji pionowej lub poziomej, często odbiegały od założeń dotyczących udziału tła. Produkty te często były fotografowane na białym tle o proporcjach kwadratu, aby spełnić wymagania dotyczące minimalnej wielkości zdjęcia, co skutkowało większym udziałem tła niż przewidywały pierwotne zalecenia. Łącznie oznaczono 7257 zdjęć do poprawy. Na Rysunku 28 przedstawiono rozkład liczby zdjęć w zależności od procentu powierzchni tła. Najwięcej zdjęć (powyżej 600 tysięcy) ma niski udział tła (blisko 0%), a liczba ta gwałtownie spada przy wzroście procentu tła. Poziom tła w większości zdjęć nie przekracza 10%. Powyżej 90% tła (zaznaczone czerwoną linią) liczba zdjęć jest marginalna. Na Rysunku 29 przedstawiono przykłady zdjęć ze zbyt dużą powierzchnią tła.



Rysunek 28. Liczba zdjęć w funkcji procentu powierzchni tła  
 Źródło: opracowanie własne



Rysunek 29. Przykłady zdjęć ze zbyt dużą powierzchnią tła  
 Źródło: opracowanie własne

Pomimo stosunkowo niewielkiej liczby zdjęć, które zostały zaklasyfikowane do poprawy, zastosowanie tego algorytmu ma uzasadnienie w poprawie jakości danych. Istnieje potencjał do jego dalszego udoskonalania w celu automatyzacji procesów weryfikacji

zdjęć. Przykładowo, możliwe jest monitorowanie kadrowania, aby zapewnić, że produkt, mimo że niewielki powierzchniowo, jest odpowiednio wyśrodkowany na tle, co dodatkowo zwiększyłoby precyzję oceny zdjęć. Progi algorytmu można również dostosowywać w zależności od kategorii produktów, regulując tolerancję dla udziału tła w zależności od spodziewanego obiektu na zdjęciu. Dzięki temu algorytm może być elastyczny i skutecznie wspierać optymalizację jakości zdjęć w różnych kontekstach i różnych zastosowaniach.

### 7.3.7. Predykcja kategorii GPC

W celu eliminacji problemu jakościowego dotyczącego niepoprawnych lub tymczasowych kategorii produktów [W3.11], opracowano model klasyfikacyjny wykorzystujący techniki uczenia maszynowego, bazując na enkoderach tekstowych opartych na architekturze BERT (Bidirectional Encoder Representations from Transformers). BERT jest modelem językowym, który wykorzystuje dwukierunkowe trenowanie transformatorów w celu lepszego zrozumienia kontekstu w tekście (Devlin i in., 2019).

W początkowej fazie, dla celów porównawczych, na niewielkiej próbce danych (zbiór danych nazw produktowych przypisanych do 96 kategorii GPC podzielony na 7000 danych treningowych i 1500 danych testowych) przetestowano kilka modeli BERT wytrenowanych na polskim korpusie danych. Modele te, wraz z ich źródłami i oceną skuteczności (accuracy), przedstawiono w Tabeli 41. Modele wykorzystane w badaniu zostały pobrane ze strony „<https://huggingface.co/>”, a najlepsze wyniki uzyskano przy użyciu modelu Polbert.

Tabela 41. Wybór modelu do predykcji klasyfikacji

Nazwa Modelu	Model	Skuteczność
<b>deberta-lower</b>	microsoft/deberta-base	0,79867
<b>bert-lower</b>	bert-base-uncased	0,85267
<b>roberta-polish</b>	clarin-pl/roberta-polish-kgr10	0,86667
<b>herbert-lr2e-5-25</b>	allegro/herbert-base-cased	0,86867
<b>polbert-5epochs</b>	dkleczek/bert-base-polish-uncased-v1	0,87133

Źródło: opracowanie własne

Analiza wyników ujawniła, że na jakość modelu pozytywnie (2 punkty procentowe) wpływa obecność nazwy marki jako części nazwy produktu, która, będąc wyuczona przez model, z jednej strony pomaga w przypisywaniu produktów na podstawie podobieństwa z wcześniejszymi wpisami (np. marka „Pepsi” zawsze kojarzona jest z napojami). Paradoksalnie jednak, taki mechanizm może prowadzić do nieprawidłowości, gdy wyuczony

algorytm zbyt mocno przypisze markę do określonego elementu klasyfikacji. Jest to szczególnie problematyczne w sytuacji, gdy dotychczas produkty firmy były kategoryzowane tylko w jednej grupie, co ogranicza model w elastycznym przypisywaniu nowych produktów do innych kategorii. Do trenowania końcowego modelu wykorzystano pełną bazę danych produktowych.

Globalna Klasyfikacja Produktów (GPC) składa się z 5153 elementów (tzw. Bricków), obejmując bardzo szczegółowy podział niektórych produktów (np. pod względem trwałości, stanu przetworzenia czy zastosowania), jednocześnie generalizując pewne obszary. Niektóre kategorie produktów są rzadko stosowane w Polsce (np. GPC Brick nr 10007406 – „*Drzewa baobabu afrykańskiego Adansonia Digitata*”), co utrudnia odpowiednie wytrenowanie modelu ze względu na brak wystarczającej liczby przykładów. Zbiór jest zatem mocno niezbalansowany – pewne kategorie są wykorzystywane częściej, inne natomiast znacznie rzadziej. Jedną z najczęściej stosowanych klasyfikacji jest tzw. klasyfikacja tymczasowa (Brick nr 99999999), która jest często nadużywana w sytuacjach, gdy producenci nie znajdują odpowiedniego elementu na liście, co prowadzi do nieprecyzyjnego przypisywania produktów i obniża jakość klasyfikacji. Są także miksy opakowań, które są używane wówczas, gdy sprzedawany produkt zawiera elementy z różnych kategorii w ramach tej samej klasy (np. Brick nr: 10005186 – „*Opakowania mix zabawek/gier*”, np. lalka z dodatkowymi akcesoriami w jednym opakowaniu sprzedawanymi jako promocja), lub też w ramach różnych rodzin (Brick nr: 10000624 – „*Krzyżowanie się segmentów*” - Opakowania mix, np. piwo z kuflem).

Z drugiej strony jedynym elementem, na podstawie którego algorytm może przypisać kategorię jest nazwa produktu, dlatego tak istotna jest jej czytelna forma opisowa i duża liczba danych treningowych. Przyjęto zasadę, że do modelu klasyfikatora GPC mogą być zakwalifikowane jedynie takie elementy taksonomii, dla których istnieje minimum 100 poprawnie wprowadzonych produktów utworzonych przez 5 różnych firm. Ostatecznie do modelu klasyfikacji włączono 4587 Bricków taksonomii GPC, wyłączając ze zbioru klasyfikację tymczasową. Wiele elementów w klasyfikacji GPC jest tak szczegółowych, że sama nazwa produktu często nie zawiera kluczowych informacji, które umożliwiałyby jednoznaczne przypisanie do właściwej kategorii. W efekcie, zarówno model, jak i człowiek nie są w stanie dokonać poprawnej klasyfikacji bez wcześniejszego zapoznania się ze specyfikacją techniczną lub dokładnym zastosowaniem produktu, którego te informacje w bazie mogą nie zawierać. Przykładem jest klasa „*Samochodowe lampy dekoracyjne*”, która rozróżnia między „*Oświetleniem akcentującym*” a „*Oświetleniem dekoracyjnym – inne*”.



Podobne trudności napotykamy przy klasyfikacji takich produktów jak „Cup&You Cracker z kawą Merry Christmas 400g” – czy jest to kawa, czy produkt okolicznościowy, „Finestra24 Ikona Anioł Stróż 14 x 21 cm P. Chrztu Komunii 50szt” – produkt okolicznościowy czy obraz, oraz „LEDprojektor Rezystor Ceramiczny 10W 39 ohm 2szt” – czy jest to rezystor jako element elektroniczny, czy część układu elektronicznego w samochodzie?

Ostatecznie podjęto decyzję, że wynikiem walidacji nie musi być rekomendacja jednego GPC Brick, lecz kilku. Jeśli wybrany przez użytkownika element nie znajduje się na liście rekomendacyjnej – wówczas walidator powinien wywołać ostrzeżenie i wyświetlić listę rekomendacji, dzięki czemu użytkownik będzie miał możliwość wyboru najlepszej z opcji.

Zbiór danych treningowych składał się z 33 262 287 próbek. Nie był zbalansowany, co oznacza, że liczba produktów w poszczególnych kategoriach nie była równa, co utrudniało poprawność predykcji. Dodatkowo, duża część nazw produktów była identyczna lub bardzo podobna (różniąca się wariantem, czasem jedynie symbolem), co należało wyeliminować przed rozpoczęciem procesu uczenia. Ze względu na potencjalnie dużą liczbę błędów w etykietach danych, badania kontynuowano z zastosowaniem metody usuwania tzw. „brudnych danych” z danych treningowych (tzw. confident learning). W tym celu użyto walidacji krzyżowej, gdzie dane były dzielone na k-krotne części (Bates i in., 2022). W opisywanym przypadku wytrenowano dwa modele na 50% danych w taki sposób, aby jeden model był w stanie zwalidować drugi zbiór i wyeliminować prawdopodobnie błędne dane. Do identyfikacji błędnych danych użyto biblioteki Cleanlab.ai (Northcutt i in., 2022). W wyniku działania biblioteki usunięto 163 660 próbek z danych treningowych, uznanych za błędne.

Końcowy zbiór treningowy modelu obejmował łącznie 33 117 698 próbek, z czego 144 589 stanowiły próbki testowe. Model uzyskał bardzo wysokie wyniki w zakresie oceny skuteczności. Osiągnięta dokładność wyniosła 98%, co wskazuje na ogólną poprawność klasyfikacji. Miara  $F_1$ , będąca średnią harmoniczną precyzji i czułości, wyniosła 84%. Wartość precyzji osiągnęła poziom 85%, co oznacza, że model rzadko popełniał błędy fałszywie pozytywne. Z kolei czułość, również na poziomie 85%, wskazuje, że model z powodzeniem wykrywał większość rzeczywistych przypadków błędnych klasyfikacji. Opis metryk i ich zastosowania szerzej opisano w Rozdziale 6.3. W celu zwiększenia efektywności algorytmu wdrożono dodatkowy mechanizm oceny z wykorzystaniem modelu LLaMA, szczegółowo omówionego w rozdziale 7.3.9. Działanie algorytmu opierało się na analizie rozbieżności pomiędzy najbardziej prawdopodobnym wyborem lub kilkoma wyborami modelu a wyborem użytkownika. W przypadku, gdy wybory modelu o wysokim prawdopodobieństwie nie są

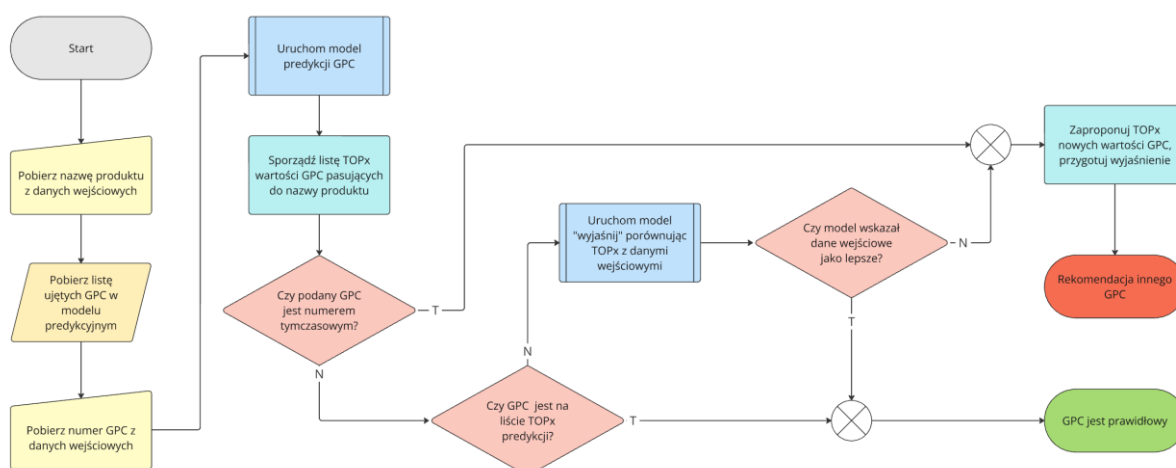
zgodne z wyborem użytkownika, przeprowadzane było dodatkowe porównanie, wykonywane odmienną metodą. Dla każdego z wybranych elementów dołączany był oficjalny opis klasyfikacji, który precyzował, jakie produkty zawierają się w danej kategorii, a także wprowadzania jakich produktów należy unikać. Na podstawie tego opisu tworzony był prompt (zapytanie) dla modelu LLaMA (w wersji 3.1-70B), który jako generatywny model przetwarzania języka naturalnego jest w stanie analizować tekstowe opisy i generować odpowiedzi na ich podstawie. Wynik odpowiedzi modelu LLaMA stanowił dodatkowe uzasadnienie rekomendacyjne, gotowe do przesłania użytkownikowi jako wyjaśnienie wyboru kategorii. Zastosowanie modelu LLaMA w procesie klasyfikacji umożliwiło lepsze dopasowanie wyników oraz zapewnienie jasnych i ustrukturyzowanych uzasadnień dla podejmowanych decyzji. Przykładowe zapytanie wysyłane do modelu LLaMA przedstawiono w Tabeli 42.

Tabela 42. Przykładowe zapytanie do modelu w celu uzyskania rekomendacji dotyczącej właściwej kategorii

Przykładowe zapytanie do modelu (prompt) z odpowiedzią
<p>Podana nazwa produktu to "POP PETS transporter typ L - rozmiar S", użytkownik zaproponował kod klasyfikacji: 10000736. Czy jest właściwy? Czy dałbyś lepszy? Uzasadnij swoją odpowiedź.</p> <p>Kody klasyfikacji i ich definicje są następujące:</p> <p>Kod: 10000736 - Akcesoria dla zwierząt domowych Inne            Obejmuje wszelkie produkty opisane/obserwowane jako akcesoria dla zwierząt domowych, w przypadku, gdy użytkownik schematu nie jest w stanie sklasyfikować produktów w istniejących brickach w schemacie.            Nie zawiera produktów opisanych w istniejących kategoriach akcesoriów dla zwierząt domowych.</p> <p>Kod: 10000662 - Schronienia/posłania dla zwierząt domowych (wielokrotnego użytku)            Obejmuje wszelkie produkty opisane/obserwowane jako produkty wielokrotnego użytku specjalnie zaprojektowane do zapewnienia schronienia i wygody w schronieniu dla zwierząt domowych. Obejmuje produkty takie jak kosze dla zwierząt domowych, klatki, transportery, akwaria, terraria i budki.            Nie zawiera jednorazowych schronień ani posłań.</p> <p>Kod: 10000572 - Schronienia/posłania dla zwierząt domowych (jednorazowe)            Obejmuje wszelkie produkty opisane/obserwowane jako produkty jednorazowe specjalnie zaprojektowane do zapewnienia schronienia i wygody w schronieniu dla zwierząt domowych. Obejmuje produkty takie jak słoma, siano, wióry drzewne oraz jednorazowe gniazda i transportery dla ptaków.            Nie zawiera wielokrotnego użytku produktów ani schronień stałych.</p>
<p>Odpowiedź modelu: Produkt POP PETS transporter typ L - rozmiar S jest schronieniem dla zwierząt domowych, specjalnie zaprojektowanym do zapewnienia wygody i bezpieczeństwa podczas transportu. Kod 10000662 jest bardziej odpowiedni niż kod 10000736, który dotyczy akcesoriów dla zwierząt domowych, a produkt nie jest po prostu akcesorium, ale raczej schronieniem.</p>

Źródło: opracowanie własne

Oba modele sztucznej inteligencji zaimplementowano w algorytmie, którego poglądowy schemat działania zaprezentowano na Rysunku 30.



Rysunek 30. Poglądowy schemat blokowy przebiegu procesu walidacji GPC  
 Źródło: opracowanie własne

Algorytm walidacji GPC rozpoczyna działanie od pobrania nazwy produktu z danych wejściowych. Następnie pobierana jest lista wszystkich numerów Brick GPC ujętych w modelu predykcyjnym oraz numer Brick GPC przypisany do produktu z danych wejściowych. Model predykcji GPC jest uruchamiany w celu zaproponowania alternatywnego numeru /numerów Brick GPC dla produktu na podstawie jego nazwy. Na tej podstawie tworzona jest lista TOPx (maksymalnie 10) wartości Brick GPC pasujących do nazwy produktu. Algorytm następnie sprawdza, czy podany numer Brick GPC jest numerem tymczasowym (tzw. Brick nr: 99999999, który jest często nadużywany w katalogach jak wspomniano w Rozdziale 5.1) i jeśli tak, przechodzi do modułu przygotowania TOPx odpowiedzi. Jeśli numer Brick GPC nie jest tymczasowy, sprawdzane jest czy znajduje się na liście TOPx predykcji modelu. Jeśli tak, to oznacza, że numer Brick ustawiony przez użytkownika jest prawidłowy. Następnie, uruchamiany jest dodatkowy moduł oceny. Ostatecznie na jej podstawie algorytm proponuje nowe TOPx wartości Brick GPC wraz z uzasadnieniem lub zwraca informację o prawidłowości wyboru GPC użytkownika.

W celu weryfikacji działania algorytmu przeprowadzono badanie oparte na specjalnie przygotowanym zbiorze danych, który został zweryfikowany i oczyszczony przez eksperta. Podstawę zbioru do ewaluacji eksperckiej stanowiły dane, które nie były wykorzystywane do trenowania modelu. Z każdej kategorii GPC, dla której w bazie znajdowało się co najmniej 2000 produktów, losowano po dwa produkty: z prawdopodobnie poprawnie wpisaną kategorią GPC przez użytkownika oraz prawdopodobnie niepoprawną, zgodnie z określoną regułą

przydzielania. Każdy z produktów musiał pochodzić od innej firmy, aby wyeliminować powtarzalność danych. Badanie obejmowało 563 najczęściej występujące kategorie GPC, które łącznie reprezentowały 93,7% wszystkich wprowadzonych danych produktowych. Wybrane GPC obejmowały 37% wszystkich kategorii dostępnych dla użytkowników. Reguła przydzielania produktów do zbioru poprawnie i niepoprawnie sklasyfikowanych oparta była na predykcji algorytmu klasyfikacji GPC, według następującego wzoru:

- A.  $S_{\text{correct}} = \{\text{User GPC Data} \mid \text{User GPC Data} \in \text{AI GPC Top5}\}$   
 B.  $S_{\text{incorrect}} = \{\text{User GPC Data} \mid \text{User GPC Data} \notin \text{AI GPC Top5}\}$

Jeśli Brick GPC przypisany przez użytkownika znajdował się wśród pięciu najbardziej prawdopodobnych elementów klasyfikacji przewidywanych przez algorytm (TOP5), produkt uznawano za poprawnie sklasyfikowany. W przeciwnym wypadku przypisywano go do zbioru niepoprawnych. Z początkowo wybranych 1126 danych produktowych usunięto te, których nazwa była niezrozumiała. Łącznie usunięto 140 produktów, z czego 2/3 pochodziło ze zbioru B (zbiór niepoprawnych). Po oczyszczeniu danych, do analizy przyjęto 986 pozycji, jak przedstawiono w Tabeli 43.

Tabela 43. Analiza ekspercka poprawności wprowadzenia klasyfikacji produktów przez użytkownika

	<b>Produkty do analizy</b>	<b>Użytkownik poprawnie sklasyfikował produkty</b>	<b>Użytkownik niepoprawnie sklasyfikował produkty</b>
<b>Dane wybrane jako poprawne:</b>	520	395	125
<b>Dane wybrane jako niepoprawne:</b>	466	110	356
<b>Łącznie</b>	986	505	481

Źródło: opracowanie własne

Łącznie 505 z 986 pozycji testowych zostało poprawnie sklasyfikowanych przez użytkowników, co stanowi około 51,2%. Algorytm zidentyfikował zarówno poprawne, jak i niepoprawne klasyfikacje ze stosunkowo wysoką dokładnością. W Tabeli 44 przeanalizowano różne warianty reguły określającej poprawność klasyfikacji użytkownika. Do analizy włączono także warianty z omawianym modelem LLaMA.

Tabela 44. Warianty reguły poprawności klasyfikacji kategorii GPC

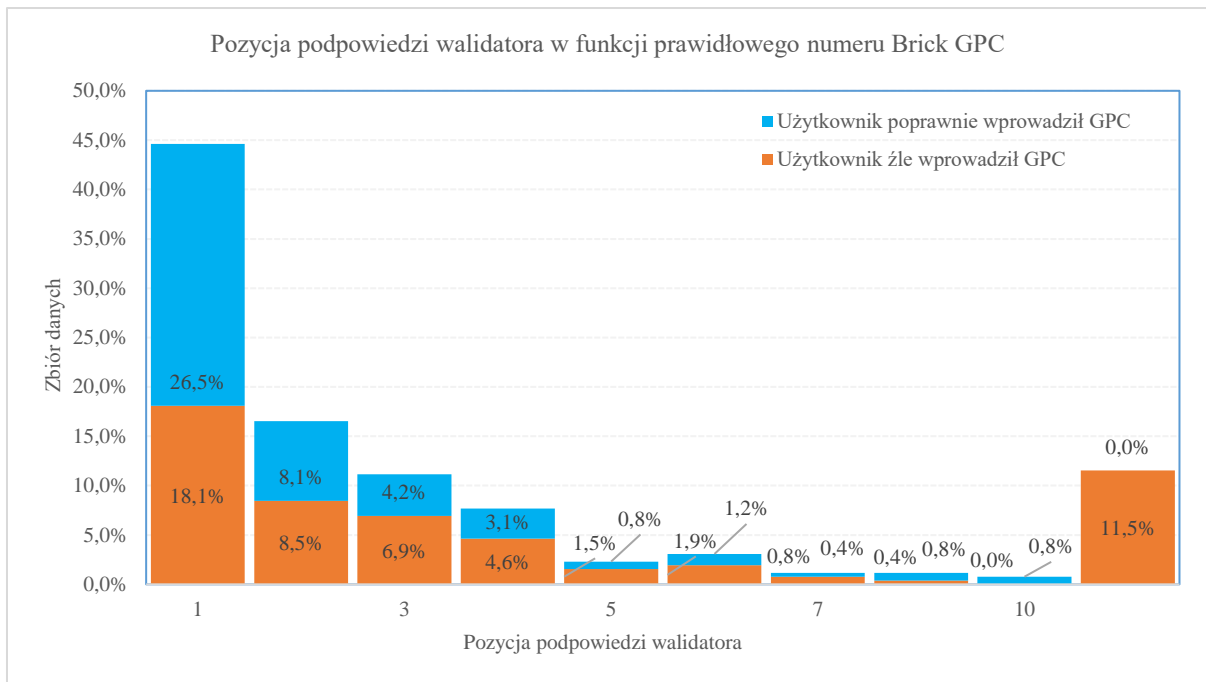
	TOP10	0,10%	TOP5	1%	10%	TOP10+ LLaMA	10%+ LLaMA	TOP5 + LLaMA
<b>Pożądane (True Positive)</b>	260	292	356	330	386	257	358	335
<b>Niepożądane (False Positive)</b>	65	78	110	104	157	58	120	89
<b>Pominięte (False Negative)</b>	221	189	125	151	95	224	123	146
<b>Właściwe (True Negative)</b>	440	427	395	401	348	447	385	416
<b>Precyzja</b>	80%	79%	76%	76%	71%	82%	75%	79%
<b>Czułość</b>	54%	61%	74%	69%	80%	53%	74%	70%
<b>F<sub>1</sub></b>	65%	69%	75%	72%	75%	65%	75%	74%

Źródło: opracowanie własne

Reguła określająca poprawność bazowała na liczbie predykcji modelu, gdzie zbiór posortowany był od najbardziej prawdopodobnego numeru GPC Brick. W ten sposób TOP10 oznaczało, że jeśli w zbiorze 10-ciu najtrafniejszych predykcji znajdzie się GPC Brick wybrany przy użytkownika, wówczas walidacja jest zakończona, a dane uznaje się za poprawne. TOP5 – ogranicza zbiór do 5-ciu najtrafniejszych precyzji. 10% - ogranicza zbiór do predykcji o prawdopodobieństwie równym lub większym niż 10% itd. Przy wykorzystaniu dodatkowego modelu LLaMA, zgodnie ze schematem na Rysunku 30 proces poddany został dodatkowej walidacji wpływającej na wynik, o ile model uznał, że wartość wprowadzona przez użytkownika jest lepsza niż wartości z listy, przy czym do analizy brane były jedynie wartości, których prawdopodobieństwo oceniono na minimum 10%. Każda kolumna zawiera cztery główne kategorie: "Pożądane" (True Positive) – liczba poprawnie sklasyfikowanych pożądaných produktów do zgłoszenia nieprawidłowości; "Niepożądane" (False Positive) – liczba produktów błędnie uznanych za nieprawidłowe; "Pominięte" (False Negative) – liczba pożądaných produktów, które zostały błędnie sklasyfikowane jako poprawne, oraz "Właściwe" (True Negative) – liczba produktów, które były poprawne. Miara precyzji określiła jaka część zaklasyfikowanych jako poprawne produktów rzeczywiście była poprawna. Obliczeniem miary czułości (Recall) zmierzono zdolność algorytmu do poprawnego identyfikowania wszystkich pożądaných produktów. Miara F<sub>1</sub> połączyła te dwa elementy, dzięki której można było ocenić ogólną skuteczność algorytmu. Ponieważ w walidacji typu trzeciego priorytetem była wysoka precyzja, kosztem czułości, najkorzystniejszą opcją okazała się wersja TOP10+LLaMA, która osiągnęła najwyższą precyzję wynoszącą 82%. Model najtrafniej identyfikował poprawne produkty przy minimalnej liczbie błędnie zaklasyfikowanych jako poprawne. Jednakże,

kosztem tej wysokiej precyzji, czułość wynosiła jedynie 53%, co oznacza, że model pominął znaczną część pożądaných produktów (224 przypadki). Wynik  $F_1$  dla tego wariantu wyniósł 65%, co wskazuje na umiarkowaną równowagę między precyzją a czułością.

Walidator nie w każdym przypadku zaproponował lepszą kategorię GPC. Dla 11,5% przypadków właściwa kategoria znajdowała się poza 10 wybranymi propozycjami. Szczegółowy rozkład przedstawiono na Rysunku 31. Na pierwszej pozycji walidator prawidłowo zaproponował kategorię dla 44,6% przypadków, dla drugiej pozycji odsetek wyniósł 16,5%, a dla trzeciej 11,2%. Pozostałe pozycje miały niewielki udział procentowy. Charakterystycznym elementem tego wykresu jest podział na bloki w zależności od tego, czy użytkownik poprawnie, czy błędnie wprowadził dane. Należy pamiętać, że model został wytrenowany na nieoczyszczonym zbiorze danych, co mogło wpłynąć na jakość odpowiedzi. W sytuacjach, gdy użytkownik prawidłowo wprowadził dane, walidator częściej podawał właściwą kategorię na wcześniejszych pozycjach. Natomiast w przypadku błędnych danych wejściowych, prawidłowa kategoria była rzadziej proponowana w czołowych pozycjach. W ramach oczyszczania zbioru czułość i precyzja modelu powinna ulec znaczącej poprawie. Redukcja błędnych i nieadekwatnych wpisów w danych treningowych powinna zwiększyć trafność proponowanych przez walidator kategorii GPC, szczególnie na wcześniejszych pozycjach listy. Oczyszczenie zbioru danych może również zredukować liczbę przypadków, w których właściwa kategoria znajduje się poza 10 wybranymi propozycjami, co z kolei przełoży się na bardziej precyzyjne dopasowanie w sytuacjach, gdy użytkownik błędnie wprowadzi dane.



Rysunek 31. Pozycja podpowiedzi walidatora w funkcji prawidłowego numeru Brick GPC

Źródło: opracowanie własne

W ramach działania walidatora przeanalizowano również zbiór danych sklasyfikowanych przez użytkownika za pomocą numeru tymczasowego. Do analizy wybrano 143 produkty, a po eliminacji nieprawidłowych nazw oceniono możliwość przypisania ich do właściwych kategorii. W przypadku 15 produktów zarekomendowano pozostawienie klasyfikacji tymczasowej. Dla 106 produktów zaproponowano właściwą klasyfikację, z czego dla 96 z nich prawidłowa kategoria znalazła się w TOP10 predykcji, co stanowiło 90,6%.

Analizując zbiór ewaluacyjny, zestawiono poprawne i niepoprawne rekomendacje modelu LLaMA z oceną ekspercką. W 119 przypadkach model LLaMA zaproponował kategorię zgodną z wyborem eksperta, natomiast w 129 przypadkach rekomendacja różniła się. Pomimo że zastosowanie modelu LLaMA w procesie rekomendacji poprawiło działanie walidatora, ogólna trafność rekomendacji budzi uzasadnione wątpliwości. Dodatkowym problemem były opisowe wyjaśnienia modelu. Choć sugestie LLaMA często wydawały się logiczne, to z natury w 52% były błędne. Pomimo tych niedoskonałości, wartością dodaną rekomendacji opisowych generowanych przez LLaMA jest fakt, że proces dedukcyjny oparty na nazwie produktu, zwłaszcza gdy nazwa jest nieprecyzyjna, często zwraca uwagę na potencjalne problemy z jej jakością, co może stanowić impuls do poprawy. Niemniej jednak, z uwagi na niewystarczająco wysoką skuteczność wyników, opisy generowane przez LLaMA przy projektowaniu graficznym narzędzia powinny być dostępne jedynie na żądanie, aż do

momentu poprawy jakości opisowej predykcji. Błędy predykcji w 12,5% przypadków wynikały z faktu, że model miał dostęp jedynie do fragmentu najlepszych kategorii GPC (zbiór wyników modelu BERT z minimum 10% prawdopodobieństwa), co nie zawsze pokrywało się z kategorią wybraną przez eksperta. Z drugiej strony, zwiększenie zakresu do TOP10 predykcji skutkowało generowaniem przez model tzw. "halucynacji" (fałszywych predykcji), co wymusiło ograniczenie liczby sugestii. W 8,8% przypadków odpowiednie dane nie znajdowały się na liście, co powodowało, że model naturalnie wskazywał błędne kategorie. Pozostały procent błędów ma prawdopodobnie inną przyczynę. Zidentyfikowano kilka kluczowych problemów. Pierwszym z nich była jakość nazw produktów, które często nie oddają w pełni charakterystyki i przeznaczenia danego produktu. Drugim problemem był formalny, urzędowy charakter opisów kategorii GPC, który nie jest dostosowany do potrzeb przeciętnego użytkownika, a także nie jest dostatecznie dobrze zrozumiały dla modelu, który wymaga bardziej logicznej i uproszczonej formy opisu. Trzecim problemem był brak przykładów produktów w opisie kategorii, co prawdopodobnie znacznie utrudniało modelowi LLaMA prawidłową ocenę i klasyfikację. Przykładowo, w opisach GPC Brick brakowało nawiązania do takich produktów jak: kurtka męska, taśma elektroizolacyjna, torba na zakupy, pościel, obrus ceratowy, pisaki, flamastry, maty silikonowe pod meble, znicze, haft diamentowy czy słomki do picia. Nawet po przeczytaniu pełnych opisów tych kategorii nie wiadomo było do końca czy są one odpowiednie dla tych produktów. Inne produkty to „pidżama”, z klasyfikacją „10001358 opakowania mix odzieży do spania” (ponieważ góra i dół piżamy są różnymi kategoriami), „dres sportowy”, będący miksem odzieży sportowej z tego samego powodu, „łóżko” jako „10002213 domowe łóżka/materace – opakowania mix” ponieważ osobną kategorią są materace i ramy łóżek. Te przykłady również wskazują, że brak jednoznacznych przykładów stanowił problem nie tylko dla modelu, ale także dla przeciętnego użytkownika, przyzwyczajonego do innych kategorii w e-commerce. Dlatego, aby poprawić działanie modelu przy wdrożeniu produkcyjnym algorytmu walidatora, konieczne będzie uzupełnienie opisów kategorii GPC o odpowiednie przykłady produktów. Dodatkowo, ograniczenia wynikające z mocy obliczeniowej zastosowanej wersji modelu LLaMA stanowiły kolejny czynnik wpływający na jakość wyników. Użycie bardziej zaawansowanej wersji modelu mogłoby znacząco wspomóc główny proces walidacji.

Zgodnie z zaproponowaną metodyką oceny skuteczności zestawiono potencjalny wpływ na podmiot badawczy (Tabela 45). Badanie przeprowadzono na stratyfikowanej próbie, obejmującej 1% bazy danych (0,5 mln produktów), przy zachowaniu proporcji produktów



w ramach poszczególnych kategorii GPC. Następnie, na podstawie wyników dokonano proporcjonalnego przeliczenia wpływu na całą bazę. Należy jednak podkreślić, że badanie przeprowadzono na danych, które były używane do trenowania modelu. W związku z tym w rzeczywistości liczba poprawnych wykryć może być większa, natomiast dla zbioru nowych produktów można założyć większą proporcję wykrytych problemów.

Tabela 45. Wpływ algorytmu walidacji GPC na całą bazę danych

Wyniki na całej bazie	Dane
Liczba wykrytych problemów jakościowych w całej bazie	2 531 250
Liczba potwierdzeń braku problemów jakościowych	42 428 050
Zakładana liczba poprawnych wykryć problemów jakościowych	2 075 625

Źródło: opracowanie własne

Dla produktów z tymczasowym GPC Brick 99999999 dokonano obliczeń wpływu walidatora i jego rekomendacji na podstawie rzeczywistej liczby danych. Liczba znalezionych rekordów w bazie to 624 934, zatem realny wpływ będzie wynosił potencjalnie 566 190 poprawnych propozycji. Łącznie wpływ walidatora na doskonalenie bazy to 2 641 815 poprawionych rekordów, czyli 6% całej bazy.

### 7.3.8. Predykcja jednostki i zawartości netto

Nieadekwatna jednostka oraz nieprawidłowa zawartość netto w odniesieniu do produktów [W3.12], [W3.13] stanowią istotne problemy jakościowe, szczególnie w sytuacji, gdy zawartość netto jest elementem wyróżniającym w tej samej linii produktowej lub grupie produktów. Pomimo że jest to cenna informacja dla konsumenta, zawartość netto często nie jest właściwie wprowadzana lub importowana z danych źródłowych. Błędy te wynikają głównie z braku odpowiedniego przeszkolenia osób wprowadzających dane, pomyłek edycyjnych oraz prób omijania walidacji obligatoryjności pola w przypadku braku informacji. Często problemem są także niewłaściwie przeliczenia zawartości netto w stosunku do wybranej pochodnej jednostki. W trakcie realizacji badań okazało się również, że problemem jest przede wszystkim mylenie pojęcia zawartości netto z atrybutami logistycznymi, jak waga produktu, czy jego długość, pojemność (np. pojemność zbiornika) itp.

Analizę możliwości poprawy pól jednostki i zawartości netto rozpoczęto od procesu „Data Profiling”, czyli zebrania i wizualizacji obecnie wprowadzonych danych. Na podstawie zebranych danych dokonano pomiarów zależności między GPC a jednostką miary, z uwzględnieniem liczby firm wykorzystujących te pary i liczby przypisanych produktów.

Poprzez analizę rzadko występujących połączeń GPC – jednostka, oznaczono 2736 połączeń jednostek z GPC jako prawdopodobnie niepoprawne, co objęło 36 918 aktywnych produktów w bazie. Dodatkowo, 15 912 par GPC – jednostka nigdy nie zostało użytych. Choć podejście statystyczne do analizy zależności między kategorią GPC a jednostką miary pozwala na wykrycie pewnych nieprawidłowości, nie jest ono wystarczające. Powodem jest specyficzna struktura GPC, która w zależności od obszaru obejmuje w niektórych przypadkach ogólne kategorie dla całej gamy produktów o różnych atrybutach. W takim przypadku produkty mogą być opisywane w różnych jednostkach miary – od litrów i kilogramów, przez metry, aż po sztuki. Tego typu różnorodność utrudnia precyzyjną ocenę poprawności przypisanych jednostek, wyłącznie na podstawie częstotliwości statystycznych powiązań GPC – jednostka miary. Przykładem może być kategoria "*Środki/akcesoria kosmetyczne*", która obejmuje produkty o szerokim zakresie cech fizycznych. W tej kategorii jednocześnie mogą występować jednostki wagowe (gramy, dla produktów takich jak proszki kosmetyczne), objętościowe (litry, dla płynów do demakijażu), ilościowe (sztuki, dla pędzli do makijażu) oraz długości (metry, dla taśm kosmetycznych). W tym przypadku każda z jednostek jest potencjalnie właściwa w zależności od specyfiki produktu. Lepszym rozwiązaniem w takiej sytuacji jest zastosowanie modelu opartego na zaawansowanej analizie semantycznej nazw produktów. Model, opierający się na rozumieniu znaczenia i kontekstu nazw, w połączeniu z kategorią produktu jako dodatkowym elementem, powinien znacznie dokładniej rekomendować prawidłową jednostkę miary.

W celu rozwiązania problemu opracowano model klasyfikacyjny, wykorzystujący sztuczną inteligencję, techniki uczenia maszynowego oparte na enkoderach tekstowych z architekturą BERT, podobne do tych, które opisano w rozdziale 7.3.7. Model został wytrenowany na danych zastanych, obejmujących nazwy produktów, odpowiadające im numery klasyfikacji GPC Brick i znormalizowane jednostki zawartości netto do postaci podstawowych miar: wagi, długości, objętości, powierzchni, liczby sztuk. Połączenie nazwy produktu i numeru GPC Brick pozwoliło dokładniej ocenić, która jednostka miary zawartości netto jest właściwa. Było to szczególnie przydatne, gdy nazwa nie wskazywała wyraźnie charakteru lub zastosowania produktu. Dodatkowo, analizowanie tekstu nazwy umożliwiło lepszą predykcję odpowiedniej jednostki niż opieranie się jedynie na kategorii. Normalizacja jednostek miar była konieczna ze względu na dużą liczbę produktów operujących w granicach pochodnych jednostek miary (np. napoje sprzedawane w butelkach o pojemności od 250 ml do 1,5 l). Zachowanie jednostek z pochodnymi mogłoby skutkować tym, że algorytm

przewidywałyby nie tylko samą jednostkę, ale również przedział zawartości, czego starano się w tym walidatorze uniknąć.

Model trenowany analogicznie jak w rozwiązaniu dla predykcji kategorii GPC, uzyskał wysokie wskaźniki precyzji i czułości we wszystkich kategoriach jednostek miary. Średnia precyzja modelu wyniosła 95,9%, a średnia wartość czułości: 93,9%. Wysoka wartość miary precyzji oznaczała, że model na danych testowych przewidywał odpowiednią jednostkę miary z niewielką liczbą błędnych klasyfikacji (tj. minimalna liczba przypadków, w których model błędnie przypisał jednostkę). Natomiast wysoka czułość wskazywała, że model potrafił poprawnie zidentyfikować większość poprawnych jednostek miary spośród wszystkich rzeczywiście poprawnych przykładów. Dodatkowo obliczono miarę  $F_1$ , stanowiącą harmoniczną średnią precyzji i czułości. W Tabeli 46 zaprezentowano wyniki modeli dla poszczególnych jednostek miary, obejmujące miarę  $F_1$ , precyzję i czułość.

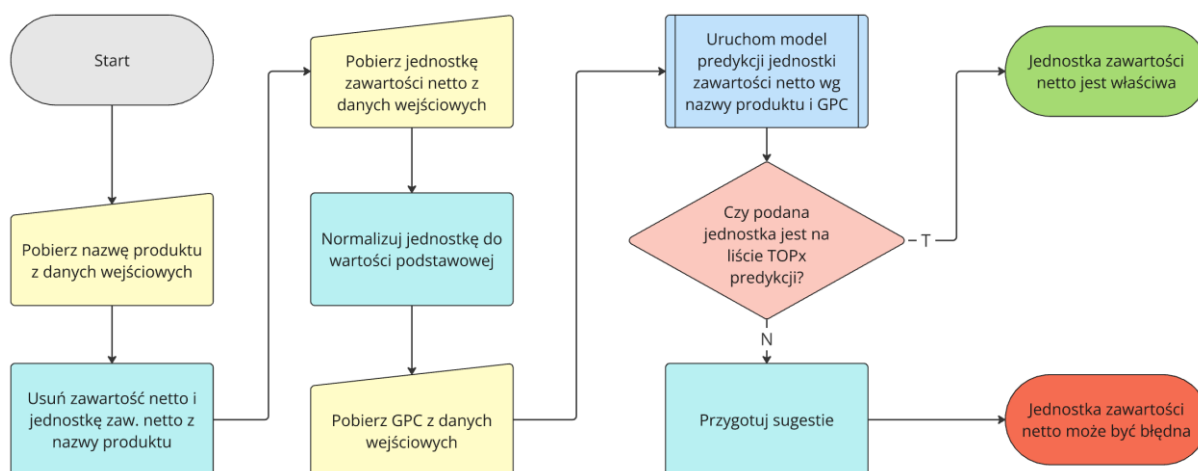
Tabela 46: Zestawienie metryk dla modelu oceny jednostki netto

Jednostka miary	$F_1$	Precyzja	Czułość
Gramy (g)	94,2%	94,9%	93,5%
Litry (l)	91,7%	92,0%	91,4%
Metry (m)	94,5%	95,9%	93,2%
Metry kwadratowe (m <sup>2</sup> )	94,6%	97,4%	92,0%
Sztuki (szt)	99,5%	99,4%	99,5%

Źródło: opracowanie własne

Dane użyte do treningu i ewaluacji modelu niestety nie były zbalansowane pod względem liczby przykładów na daną jednostkę miary. Brak zbalansowania danych wynikał z naturalnego rozkładu jednostek miary w rzeczywistych danych produktowych i niekorzystnie wpłynął na wydajność modelu, faworyzując te klasy, które są lepiej reprezentowane. Wiele kategorii produktowych zawiera jedynie produkty opisywane jako jedna sztuka. Produkty o jednostkach miary takich jak gramy, czy litry są powszechne głównie w branży spożywczej z racji konieczności stosowania odpowiednich przepisów unijnych w tym zakresie. Tymczasem jednostki miary, takie jak metry czy metry kwadratowe, są charakterystyczne dla niewielu produktów, co sprawia, że ich reprezentacja w katalogu jest ograniczona. Zdecydowano jednak nie tworzyć sztucznie dodatkowych danych dla wyrównania, aby uniknąć wprowadzenia nieautentycznych przykładów, ale bazować na danych zastanych. Wykorzystując zbudowany model utworzono na jego podstawie walidator jednostek zawartości netto.

Na Rysunku 32 przedstawiono poglądowy schemat blokowy walidatora.



Rysunek 32. Schemat blokowy walidatora jednostki zawartości netto  
 Źródło: opracowanie własne

Proces rozpoczyna się od pobrania nazwy produktu z danych wejściowych. Następnie, usuwana jest zawartość netto oraz jednostka miary z nazwy produktu, po czym następuje pobranie jednostki zawartości netto z danych wejściowych i normalizacja do wartości podstawowej. W kolejnym kroku system pobiera z danych wejściowych numer klasyfikacji GPC dla produktu. Gdy wszystkie dane są przetworzone, uruchamiany jest model predykcji jednostki zawartości netto, który opiera się na analizie nazwy produktu i klasyfikacji GPC. Algorytm generuje rekomendacje najbardziej prawdopodobnych jednostek miary, analogicznie do podejścia zastosowanego przy klasyfikacji GPC. Liczba rekomendacji zależy od ustalonej wartości prawdopodobieństwa i/lub innego dobranego progu, np. TOPx liczby rekomendacji. Jeżeli wśród dopuszczonych rekomendacji znajduje się jednostka wpisana przez użytkownika – walidacja przechodzi pomyślnie. Jeśli podana jednostka nie znajduje się na liście, system generuje sugestie poprawy, informując, że jednostka zawartości netto może być błędna.

Działanie zaprezentowanego walidatora zostało przetestowane metodą ekspercką na bazie odpowiednio przygotowanego zbioru danych, który nie był wykorzystywany do trenowania modelu. Do analizy wybrano 702 produkty różnych firm, przypisane do najpopularniejszych klas GPC. W każdej klasie mogły znaleźć się maksymalnie dwa produkty: potencjalnie poprawny (A) i niepoprawny (B), pod warunkiem, że nie zostały wykluczone z powodu wadliwej nazwy lub niepoprawnej klasyfikacji. Początkowo zbalansowany zbiór A i B, po procesie oczyszczania, stracił swoją równowagę, jednak mimo to nadal umożliwił precyzyjną analizę danych. Podzbiory A i B zdefiniowano w następujący sposób:

- A.  $S_{correct} = \{ \text{User UoM Data} \mid \text{User UoM Data} \in \{ \text{AI UoM Data} \mid \text{AI UoM Prediction} > 0,1 \} \wedge \text{User GPC Data} \mid \text{User GPC Data} \in \{ \text{AI GPC Data} \mid \text{AI GPC Prediction} > 0,1 \} \}$
- B.  $S_{incorrect} = \{ \text{User UoM Data} \mid \text{User UoM Data} \notin \{ \text{AI UoM Data} \mid \text{AI UoM Prediction} > 0,1 \} \wedge \text{User GPC Data} \mid \text{User GPC Data} \in \{ \text{AI GPC Data} \mid \text{AI GPC Prediction} > 0,1 \} \}$

Oba zbiory zostały ograniczone do produktów, dla których predykcja modelu GPC była zgodna z danymi wprowadzonymi przez użytkownika na poziomie co najmniej 10% prawdopodobieństwa. Podział na zbiory A i B wynikał ze wstępnej oceny walidatora, gdzie podzbiór "A" zawierał produkty, dla których jednostki miary zawartości netto wprowadzone przez użytkownika były zgodne z predykcją modelu na poziomie minimum 10% prawdopodobieństwa. Natomiast zbiór "B" obejmował dane, dla których to prawdopodobieństwo było niższe. Przygotowany zbiór poddano analizie eksperckiej pod kątem prawidłowości wprowadzenia zawartości netto. Wyniki zaprezentowano w Tabeli 47.

Tabela 47. Liczba poprawnych i niepoprawnych danych w zbiorze ewaluacyjnym walidatora jednostki netto

	Produkty do analizy	Użytkownik poprawnie wprowadził zaw. netto	Użytkownik niepoprawnie wprowadził zaw. netto
<b>Poprawne (&gt;= 10%):</b>	451	376	85
<b>Niepoprawne (&lt; 10%):</b>	251	81	160
<b>Łącznie:</b>	702	457	245

Źródło: opracowanie własne

Zgodnie z założeniami, w algorytmie walidacji wykorzystującym zbudowany model przyjęto, że próg zadziałania walidatora nie musi być równoznaczny z najlepszą predykcją wartości. Po procesie analizy eksperckiej przeprowadzono analizę poprawności działania algorytmu i jego skuteczności dla różnych progów zadziałania. Szczegóły różnic przedstawiono w Tabeli 48.

Tabela 48. Analiza progu zadziałania walidatora jednostki zawartości netto

	TOP1	TOP2	TOP3	10%	1%	0,1%
<b>Pożądane (True Positive)</b>	171	50	15	160	135	99
<b>Niepożądane (False Positive)</b>	109	58	27	81	56	33
<b>Pominięte (False Negative)</b>	74	195	230	85	110	146
<b>Właściwe (True Negative)</b>	348	399	430	376	401	424
<b>Precyzja (Precision)</b>	61%	46%	36%	66%	71%	75%
<b>Czułość (Recall)</b>	70%	20%	6%	65%	55%	40%
<b>F1</b>	65%	28%	10%	66%	62%	53%

Źródło: opracowanie własne

Skuteczność walidatora zależała od przyjętego progu zadziałania. W przypadku TOP1 brana była pod uwagę wartość z najwyższym prawdopodobieństwem, tak jak w modelu predykcyjnym. TOP2 obejmowała dwie pierwsze najlepsze predykcje, a TOP3 – trzy. Dla progów procentowych, takich jak 10%, klasyfikacja uwzględniła jedynie predykcje z prawdopodobieństwem równym lub wyższym niż dany próg (np. 10%, 1%, 0,1%). Z perspektywy miary  $F_1$ , która łączy precyzję i czułość, najlepsze wyniki uzyskano przy progu 10%. Z kolei najwyższą precyzję osiągnięto przy progu 0,1%, co ostatecznie zadecydowało o jego wyborze do wdrożenia. W miarę postępów w procesie poprawy jakości danych oraz ponownego trenowania modelu sztucznej inteligencji, można oczekiwać lepszych wyników, co pozwoli na podniesienie progu i dalszą optymalizację działania algorytmu walidatora.

Wyniki precyzji i skuteczności algorytmu różnią się od zaprezentowanych wyników modelu predykcyjnego z kilku powodów. Po pierwsze, różny rozkład klas w zbiorach danych miał istotny wpływ na rezultaty. Zbiór testowy, na którym uzyskano wyniki precyzji 95,9% i recall 93,9%, charakteryzował się bardziej zrównoważonym rozkładem danych, co oznacza, że proporcje między danymi poprawnymi a błędnymi były bliższe rzeczywistej dystrybucji, jaką model przetwarzał podczas trenowania. Modele klasyfikacyjne, takie jak BERT, działają zazwyczaj najlepiej, gdy dane testowe są zbliżone do danych treningowych. Kolejny zbiór, w którym precyzja wynosiła 61%, a czułość 70%, miał inny balans między danymi poprawnymi i niepoprawnymi. Zwiększono w nim liczbę błędnych przykładów, aby ocenić odporność modelu na trudniejsze przypadki, co w efekcie wpłynęło na obniżenie precyzji i czułości. Dodatkowo, drugi zbiór zawierał bardziej zróżnicowane i trudne przypadki do klasyfikacji, takie jak produkty o niejednoznacznych nazwach lub nietypowych jednostkach miary, co przyczyniło się do zwiększenia liczby błędnych klasyfikacji. Warto również zauważyć, że standardowe katalogi produktowe często zawierają powtarzalne pozycje, takie jak klamki czy śruby, różniące się jedynie wariantem. W takim kontekście model może osiągać lepsze wyniki, gdyż klasyfikacja staje się łatwiejsza. W przypadku drugiego zbioru celowo każdy produkt był unikalny, co znacząco utrudniło klasyfikację. Dodatkowym czynnikiem był fakt, że model był trenowany na "brudnych" danych, co potwierdza analiza nowego zbioru, z którego początkowo usunięto 168 przykładów, a z pozostałych 702-óch aż 245 okazało się błędnych. Mimo to, wynik modelu nadal można uznać za wystarczająco dobry i wskazać na jego potencjał w sukcesywnej poprawie jakości danych.

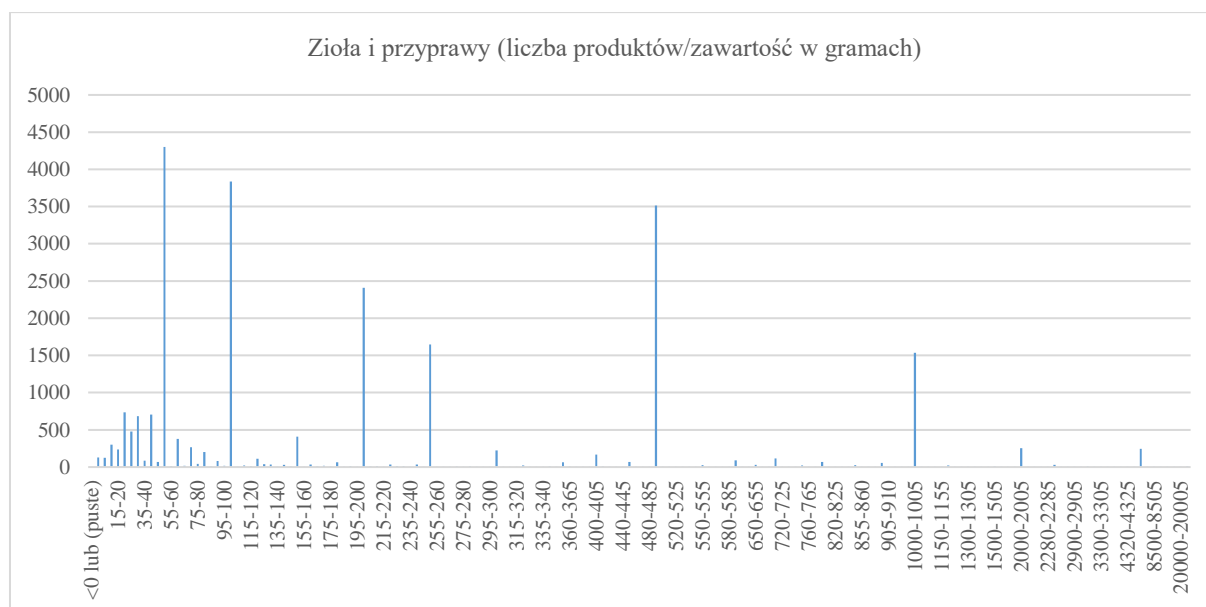
Potencjalny wpływ walidatora na całą bazę danych przedstawiono w Tabeli 49.

Tabela 49. Wpływ algorytmu walidacji jednostki zawartości netto na całą bazę danych

Wyniki na całej bazie	Dane
Liczba wykrytych problemów jakościowych w całej bazie	2 217 428
Liczba potwierdzeń braku problemów jakościowych	42 741 872
Zakładana liczba poprawnych wykryć problemów jakościowych	1 663 071

Źródło: opracowanie własne

Kolejnym etapem w procesie analizy, po predykcji jednostek miary, było określenie wartości granicznych zawartości netto dla każdej jednostki w powiązaniu z klasyfikacją GPC. To podejście umożliwiło dokładniejsze zrozumienie rozkładu danych i zidentyfikowanie typowych wartości dla danej kategorii produktowej. W celach poglądowych, zaprezentowano wybrane wykresy, które ilustrują omawiane zależności. Na ich podstawie możliwe było zaproponowanie modelu predykcyjnego do szacowania zawartości netto dla odpowiedniego walidatora danych.



Rysunek 33: Wykres liczby produktów dla zawartości netto w kategorii "Zioła i przyprawy"

Źródło: opracowanie własne

W kategorii "Zioła i przyprawy" (Rysunek 33) wyraźnie widoczne są wzrosty liczby produktów dla kilku wiodących wartości zawartości netto. Te wartości wskazują na standardowe rozmiary opakowań, które są najpopularniejsze w sprzedaży detalicznej ziół i przypraw. Ponadto, widać

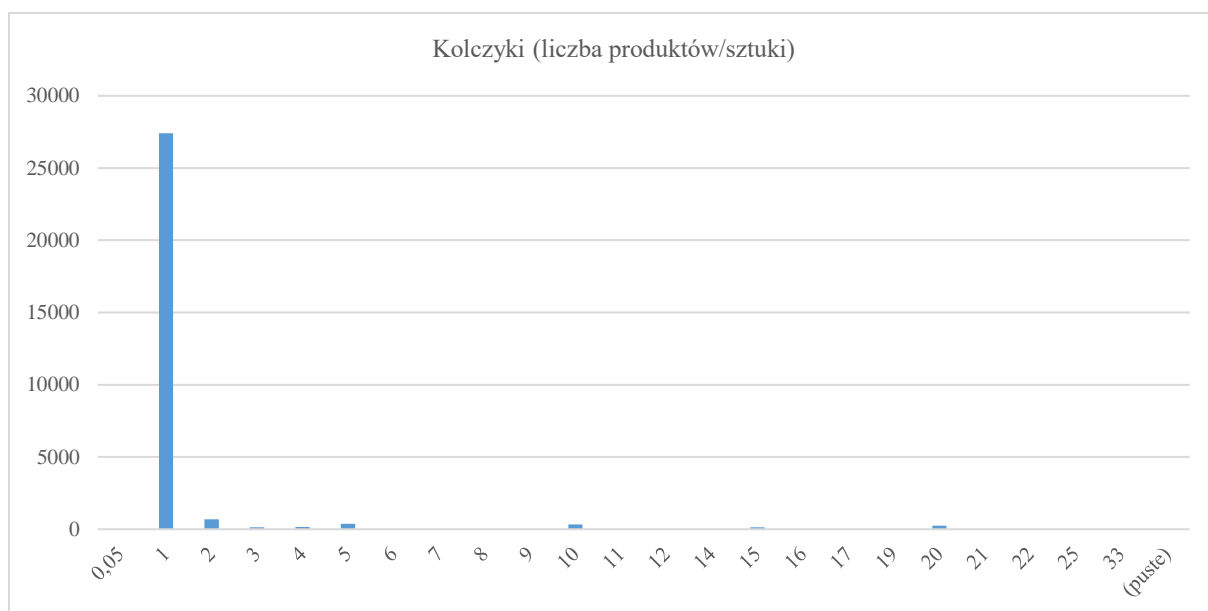
regularne odstępy między szczytami, co może sugerować pewną harmoniczną w wyborze opakowań wagowych przez producentów. Harmoniczność ta może wynikać z praktyk przemysłowych lub preferencji konsumentów, którzy kupują zioła i przyprawy w określonych ilościach. Zakres wartości granicznych jest wyjątkowo szeroki, od 0,1 grama („*Szafran nitki 0,1 g DOY100*”) do 25000 gram („*NASIONA GORCZYCY BIO (SUROWIEC) (25 kg)*”).



Rysunek 34: Wykres liczby produktów dla zawartości netto w kategorii "Wkręty"  
 Źródło: opracowanie własne

W kategorii "Wkręty" (Rysunek 34) liczba sztuk wyraźnie zgrupowana jest na wartościach do 300, z wyraźnymi harmonicznymi wzrostami dla wartości 1000, 2000 itd.

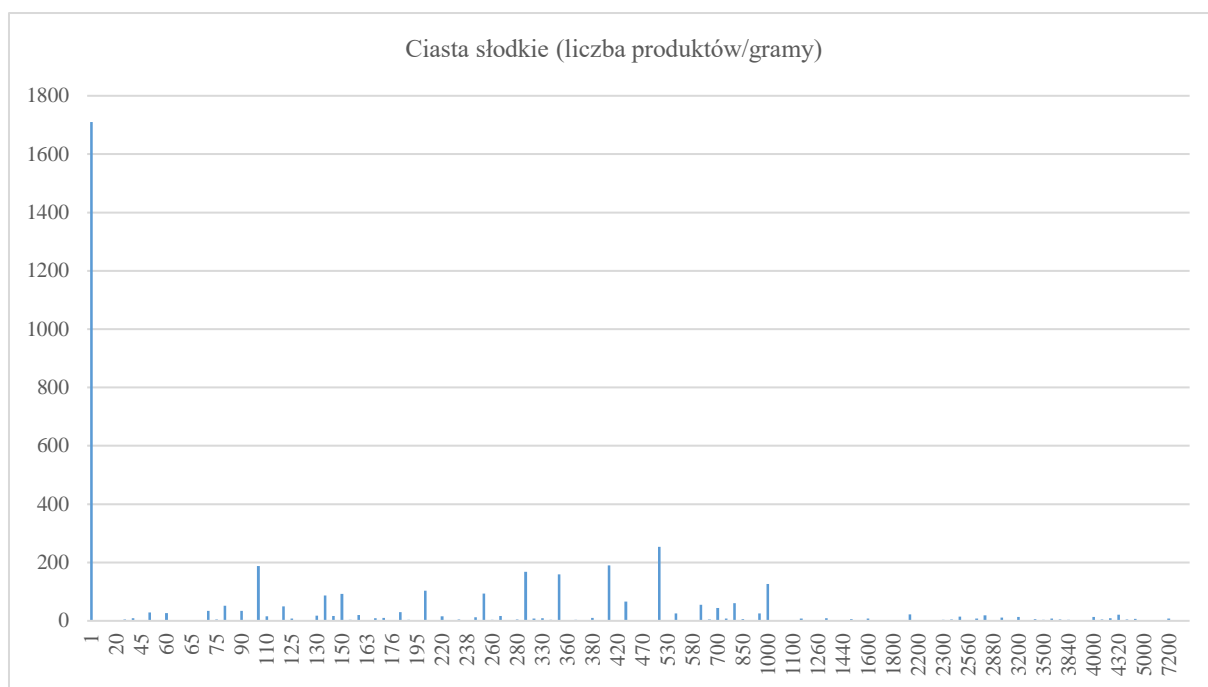




Rysunek 35: Wykres liczby produktów dla zawartości netto w kategorii "Kolczyki"

Źródło: opracowanie własne

Na podstawie wykresu liczby produktów dla zawartości netto w kategorii "Kolczyki" (Rysunek 35) zaobserwowano kilka kluczowych aspektów. Przede wszystkim, występują początkowe błędne wartości poniżej 1 sztuki, co wskazuje na błędy wprowadzania danych. Największa liczba produktów przypada na wartość 1 sztuka, co jest wartością poprawną i oczekiwaną dla tej kategorii. Z kolei wartość 2 sztuk może częściowo wynikać z błędnej interpretacji jednostki jako pary kolczyków zamiast pojedynczej sztuki. Dalsze wartości, choć mniej liczne, układają się w nieco rozproszony ciąg, co może sugerować błędy wprowadzania danych lub różnorodność pakowania kolczyków w zestawy. Układ wartości liczbowych sugeruje pewien ciąg harmoniczny, co może być wynikiem standardowych zestawów sprzedażowych, takich jak zestawy po 10 czy 20 sztuk.



Rysunek 36: Wykres liczby produktów dla zawartości netto w kategorii "Ciasta słodkie"  
 Źródło: opracowanie własne

Analizując wykres liczby produktów dla kategorii "Ciasta słodkie" (Rysunek 36), można dostrzec pewne wzorce, które mogą sugerować kolejną harmoniczną w rozkładzie danych. Oprócz znaczącego wzrostu liczby produktów o wartości netto 1 gram, który jest wynikiem błędu – importu, który dokonano omijając walidator obligatoryjności pola dla zawartości netto, można zauważyć regularne pojawianie się pików w określonych odstępach. Przykładowo, większa liczba produktów pojawia się dla wartości takich jak 50, 100, 200, 300, 500 gramów, co może sugerować standardowe wartości wagowe dla produktów w tej kategorii. Wykres kończą częściowo błędy, takie jak „Mazurek staropolski z kajmakiem 450g” z zawartością netto 27000 gram, ale także najprawdopodobniej poprawne zbiorcze kartony z produktami, np.: „Croissant VIGEN z wiśnią i czekoladą (karton)” z zawartością 7200 gram.

Na podstawie wykresów przedstawiających liczbę produktów względem zawartości netto dla danej kategorii i danej jednostki zawartości netto zaobserwowano, że:

- A. W wielu kategoriach widoczne są wyraźne wzrosty liczby produktów przy pewnych wartościach netto, co sugeruje istnienie standardowych wielkości opakowań, jednak jest to różne dla kategorii i nie zawsze można dostrzec taką regułę.
- B. Wzorce harmonicznego są widoczne w regularnych odstępach między szczytami na wykresach. To sugeruje, że producenci mogą stosować określone standardy

wagowe, które są preferowane przez konsumentów lub wynikać z praktyk przemysłowych.

- C. Wartości graniczne nie zawsze oznaczają, że występują w nich błędy, ale jest to znacznie częstsze niż w wewnętrznej przestrzeni liczbowej.
- D. Zawartości netto zależą nie tylko od kategorii, ale od konkretnego produktu w kategorii (przykład: szafran zawsze będzie miał wartość netto mniejszą niż inne przyprawy w kategorii).
- E. Jeden import dużej ilości błędnych danych może całkowicie zaburzyć statystykę bazującą na liczbach.

Na podstawie wyników zaprojektowano i przebadano kilka rozwiązań wspierających walidację danych. Jednym z pierwszych było podejście statystyczne, które polegało na ustaleniu granic wartości zawartości netto dla każdej kategorii produktów. Proces ten obejmował zliczanie minimalnych i maksymalnych wartości netto dla produktów każdej firmy, zgrupowanych według danej jednostki miary w danej kategorii GPC. Następnie z tych wartości obliczano mediany, które stanowiły granicę pomiędzy poprawnymi i niepoprawnymi – odstającymi wartościami netto, dla każdej jednostki w kategorii GPC. Zaletą implementacji takiego rozwiązania była jego prostota i łatwość adaptacji do nowych danych. Wadą – ograniczenie jedynie do wartości skrajnych w danej kategorii. Innym przebadanym rozwiązaniem było zastosowanie algorytmów wykrywających anomalie w celu odzwierciedlenia charakterystyki wykresów. Jednak ze względu na niedokładność GPC i różnorodność produktów, metoda ta okazała się mało precyzyjna. Przeanalizowano również algorytmy do analizy anomalii, gdzie danymi wejściowymi był numer Brick GPC, jednostka i zawartość netto. Niestety, nieoczyszczone dane i brak zbalansowania spowodował, że wyniki nie były zadowalające. Interesujące wyniki uzyskano poprzez zastosowanie algorytmu klasyfikacyjnego opartego na modelu BERT, gdzie danymi wejściowymi były nazwy produktów oraz numery Brick klasyfikacji GPC, a elementami wyjściowymi określone przedziały zawartości netto. W ten sposób wytrenowano modele dla każdej jednostki miary i jej pochodnych ( $10^3$ ,  $10^{-3}$ ). Przedziały zostały wyznaczone na podstawie danych zgromadzonych w bazie. W obliczeniach uwzględniono jedynie takie wartości, które zostały wprowadzone przez co najmniej pięć różnych firm. Dzięki temu procesowi dane zostały oczyszczone z wartości skrajnych i potencjalnie błędnych. Wyznaczono po pięć przedziałów zawartości netto (kwantyle) osobno dla każdej jednostki zawartości netto, które nazwano xs, s, m, l oraz xl.

Badanie objęło analizę skuteczności modelu klasyfikacyjnego dla różnych jednostek miary: metrów, milimetrów, litrów, mililitrów, kilogramów, gramów oraz sztuk. W każdym przypadku użyto pretrenowanego modelu „Polbert-5epoch”. W przypadku każdej jednostki miary badano metryki  $F_1$ , precyzji (precision) i czułości (recall), z podziałem na zdefiniowane zakresy wielkości ([XS], [S], [M], [L] oraz [XL]). Wyniki metryk umieszczono w Tabeli 50.

Tabela 50. Metryki predykcji zawartości netto

	Metr	Milimetr	Litr	Mililitr	Kilogram	Gram	Sztuka
<b>Zbiór</b>	537 538	512 599	124 431	317 870	1 165 693	1 129 288	34 916 240
<b>Dokładność</b>	89%	98%	84%	85%	94%	81%	99%
<b><math>F_1</math></b>	81%	94%	88%	86%	93%	83%	94%
<b>Precyzja</b>	81%	95%	87%	86%	93%	76%	94%
<b>Czułość</b>	81%	94%	90%	85%	93%	86%	94%

Źródło: opracowanie własne

Rozmiar zbioru treningowego wahał się od około 100 tysięcy do 34 milionów, w zależności od rodzaju jednostki miary, co wynikało z dużego niezbalansowania bazy produktów. Mimo satysfakcjonujących wyników miary  $F_1$ , ręczna ewaluacja na zbiorze przygotowanym analogicznie jak dla jednostek zawartości netto, wykazała znaczną nieodporność modeli na produkty nietypowe. To wpłynęło na wyniki precyzji i czułości walidacji poniżej 20% dla badanego zbioru. Poprawne wskazania głównie dotyczyły wartości skrajnie wysokich, co można było wykryć za pomocą rozwiązania opartego na statystyce.

Problemem okazała się jakość danych - brudne dane znacząco wpłynęły na jakość predykcji. Na model wpłynęły zarówno niedokładna nazwa produktu, niepoprawna klasyfikacja GPC, jak i źle dobrana jednostka zawartości netto. Błędy nakładały się, co jeszcze bardziej utrudniało precyzyjny trening. Poza istotnymi błędami wynikającymi z „brudnych danych” kluczową trudność stanowiła nazwa produktu, która zazwyczaj nie zawierała informacji o formie opakowaniowej, opakowaniu zbiorczym czy specyficznych cechach wpływających na zawartość netto. Brak tych istotnych danych powodował trudności modelu w prawidłowej klasyfikacji zawartości netto do najbardziej prawdopodobnego przedziału. Na przykład w branży HoReCa (Hotele, Restauracje, Catering) majonez jest zazwyczaj sprzedawany w dużych opakowaniach, takich jak 5 litrów, a nie w standardowych 500 ml. Brak takich informacji w nazwie produktu lub w oddzielnych polach bazy danych prowadzi do błędnych wyników walidatora, szczególnie jeśli baza częściej zawiera produkty przeznaczone do sprzedaży detalicznej. Problemem było również prawdopodobnie nadmierne uwzględnianie

marki produktu, będącej częścią nazwy. Marka, choć pomocna przy określaniu kategorii produktów, w tym przypadku prawdopodobnie zbyt mocno wpływała na prognozowaną zawartość netto, co prowadziło do błędnej oceny przedziału.

Podsumowując, analiza wyników wykazała, że budowa takiego walidatora na podstawie dostępnych danych, opierających się na klasyfikacji GPC oraz opisowych nazwach produktów zawierających m.in. markę w obliczu silnego zaszumienia danych, nie przynosi wymiernych korzyści. Poprawa skuteczności takiego walidatora przy obecnych danych bez dodatkowej anotacji nie jest możliwa. Perspektywicznym rozwiązaniem mogłoby być wytrenowanie modeli do ekstrakcji nazw powszechnych z nazw produktów, głębsza klasyfikacja z wykorzystaniem dodatkowego, niższego poziomu klasyfikacji GPC (tzw. atrybutów klasyfikacji GPC (GS1, 2024), obecnie nie używanych w przedmiotowym katalogu) lub transformacja nazw przy użyciu modeli takich jak T5 (Text-to-Text Transfer Transformer) (Roberts i in., 2019) do wspólnego formatu. Tak „otagowane” dane mogłyby następnie stanowić solidną podstawę do dalszego uczenia proponowanego modelu lub, co byłoby korzystniejsze, powrotu do analizy wzorców, wykorzystując właściwości przedstawione na Rysunkach 33, 34, 35, 36. Realizacja takiego rozwiązania wymagałaby jednak treningu modelu na bardzo dużym zbiorze czystych i odpowiednio przygotowanych danych, co przy obecnej skali i generalizacji badanego katalogu byłoby procesem wysoce czasochłonnym. Ponieważ z dużym prawdopodobieństwem ekstrakcja nazw powszechnych z produktów wystarczająco efektywna do takiej analizy będzie możliwa w kolejnych wersjach modeli typu LLM, np. LLaMA czy GPT, będzie to przedmiotem dalszych badań. W realizacji walidacji powrócono jednak do zaproponowanej analizy statystycznej. W jej wyniku otrzymano 3 903 393 przypadków przekroczeń ustalonych zawartości netto, w tym wychwycono produkty takie jak: „*Pestilence Levels of Perception Sinister DIGITAL SINGLE 9999999999 szt*”, „*ELHANDEL Foremki na muffinki 50×25 mm kwiatki 23-61 TAMIPOL 60 szt 10005430 szt*”, „*Goldkids Majewski Piórnik bez wyposażenia Avengers 2023 plus dzienniczek 500 kg*”.

### **7.3.9. Ocena poprawności nazwy produktu**

Analiza poprawności nazwy produktu [W3.14] stanowi złożony problem, wymagający opracowania skutecznych algorytmów walidacyjnych oraz generujących sugestie. Wiele z dotychczasowych rozwiązań koncentrowało się na walidacji składniowej, której celem była ocena formalnej poprawności nazw produktów. Pomijały one jednak inne aspekty, takie jak

znaczenie nazwy czy jej „rozumiałość” z punktu widzenia konsumenta, co stanowi kluczowy element przy ocenie adekwatności nazw w kontekście rynkowym.

Walidację „rozumiałości” rozpoczęto od zbudowania i uruchomienia walidatora, który analizował, czy nazwa produktu zawiera powszechnie stosowane słowa w języku polskim, wiodącym języku opisu nazw produktów w bazie. W tym celu wykorzystano bibliotekę Morfeusz2, opisaną w rozdziale 7.3.3, To rozwiązanie efektywnie eliminowało nazwy składające się z samych skrótów, numerów seryjnych, symboli producentów, a także literówek i błędów ortograficznych. Przeprowadzone testy na 6240 nazwach produktów wykazały, że algorytm działał skutecznie, choć czasem pojawiały się braki słownikowe dla niektórych rzadziej używanych słów, takich jak: granola, rawierka, bulionówka, posypywaczka, mufinka, czubryca, ulotkownik, ekspozytor, trybunka, śpiworek, fałdownik, woskowijka, cenówka, flizelina, penseta co doprowadziło do 444 wyników fałszywie negatywnych na 6240 sprawdzanych. Precyzja algorytmu wynosiła więc 93,36%. Wynik walidacji w ujęciu całej bazy podmiotu badań przedstawiono w Tabeli 51. Morfeusz2 wykrył nazwy powszechnie dla 37 710 554 produktów z oznaczonym polskim językiem opisu, natomiast 4 498 037 nazw oznaczono jako błędne (bez naw powszechnych).

Tabela 51. Wpływ walidatora nazw powszechnych w nazwach produktu na całą bazę danych

Wyniki na całej bazie	Dane
Liczba wykrytych problemów jakościowych w całej bazie	4 498 037
Liczba potwierdzeń braku problemów jakościowych	37 710 554
Zakładana liczba poprawnych wykryć problemów jakościowych	4 199 367

Źródło: opracowanie własne

Opisany walidator nie dostarczył jednak wystarczających informacji, na podstawie których można byłoby jednoznacznie ocenić poprawność nazwy produktu. Jego zastosowanie miało charakter wstępny, umożliwiając szybkie wychwycenie oczywistych błędów. Nazwy produktów mogą zawierać nazwę pospolitą, ale nadal nie przekazywać wystarczających informacji o samym produkcie. Wykryta nazwa powszechna może też odnosić się jedynie do nazwy marki, będącej częścią nazwy produktu. Przykładami mogą być nazwy marek składające się z nazw pospolitych, jak „Cykoria”, „Motyl”, „Biały Jeleń” czy „Malwa”, które mogą być błędnie sklasyfikowane jako poprawne nazwy produktów. Z tego względu do oceny poprawności konieczne było zastosowanie bardziej zaawansowanych rozwiązań. Korzystając z wyników badań (Niemir i Mrugalska, 2023a; Plaskowski i in., 2024), przeanalizowano możliwość wykorzystania sztucznej inteligencji, a konkretnie dużego modelu językowego

(LLM – Large Language Model), do opracowania walidatora, który zweryfikuje poprawność nazw od strony semantycznej, o ile zawierają one przynajmniej jedną nazwę powszechną.

Do realizacji zadania zbadano użycie modeli GPT - „Generative Pre-trained Transformer” (Brown i in., 2020) oraz LLaMA - „Large Language Model Meta AI (Touvron i in., 2023). Oba modele stanowią zaawansowane architektury językowe, przeszkolone na obszernych zbiorach danych tekstowych, co umożliwia im efektywne przetwarzanie naturalnego języka. Zarówno GPT, jak i LLaMA charakteryzują się zdolnością do generowania odpowiedzi, które są semantycznie adekwatne oraz zgodne z kontekstem. Choć ich wewnętrzne architektury różnią się pod pewnymi względami, oba modele wykorzystują mechanizmy transformacyjne do modelowania zależności między słowami, umożliwiając uzyskiwanie wysokiej jakości wyników w zadaniach związanych z przetwarzaniem tekstu. Tego typu modele wykorzystywane są głównie do generowania tekstu na podstawie wprowadzonego wejścia tekstowego, zwanego „promptem”. Prompt to zapytanie lub instrukcja, którą użytkownik podaje modelowi, w celu uzyskania odpowiedzi. W zaproponowanym rozwiązaniu, poprzez prompty, model najpierw uczył się poprawnej konstrukcji nazwy produktu, a następnie analizował badaną nazwę produktu i oceniał jej poprawność. Innymi słowy, modele wykorzystano do pełnienia roli przeszkolonego eksperta, analizującego jakość nazwy produktu.

Przykład konwersacji z wykorzystaniem modelu GPT-4 przedstawiono w Tabeli 52.

Tabela 52. Przykład konwersacji służącej do walidacji i oceny poprawności nazw produktu

<p><b>Zapytanie:</b>  <i>Nauczę cię walidować pełne nazwy produktów w języku polskim dla e-commerce, aby były zrozumiałe dla klientów, minimalizowały pomyłki przy zakupie oraz były łatwe do wyszukania. Pełna nazwa produktu, przedstawiana zwykle na etykiecie, powinna składać się z: marki, ogólnej nazwy określającej przeznaczenie produktu oraz ewentualnie wariantu i zawartości netto dla odróżnienia od innych produktów tego samego typu. Twoim zadaniem jest ocena nazwy w formacie [OK], [WARNING], [ERROR], z uzasadnieniem i propozycją lepszej nazwy w przypadku [WARNING]. [ERROR] jest przyznawane, gdy nazwa jest niejednoznaczna.</i>  <i>Przykład:</i>  <i>Zwaliduj nazwę: 639531 Emolium Krem ochronny na wiatr i mróz 75ml</i></p> <p><b>Odpowiedź:</b>  Ocena: [ERROR]  Uzasadnienie: Numer SKU producenta nie powinien być częścią pełnej nazwy produktu.  Propozycja: Emolium Krem ochronny na wiatr i mróz 75ml</p>
<p><b>Zapytanie:</b>  <i>Zwaliduj nazwę: Karta graficzna Gigabyte GeForce RTX 3060 Gaming OC</i></p> <p><b>Odpowiedź modelu:</b>  Ocena: [OK]  Uzasadnienie: Pełna nazwa produktu zawiera markę (Gigabyte), ogólną nazwę określającą przeznaczenie produktu (Karta graficzna), model (GeForce RTX 3060), wariant (Gaming OC)  Propozycja: Brak.</p>
<p><b>Zapytanie:</b>  <i>Zwaliduj nazwę: EKLAND nap herb wiś-jab MV203 10szt*1kgT</i></p>

<p><b>Odpowiedź modelu:</b>  Ocena: [WARNING]  Uzasadnienie: Nazwa produktu jest niejasna, trudno zrozumieć co oznacza "nap herb wiś-jab" i "10szt*1kgT".  Propozycja: EKLAND Napój herbaciany Wiśnia-Jabłko 10 sztuk × 1 kg.</p>
<p><b>Zapytanie:</b>  Zwaliduj nazwę: EWA 500G (12)</p> <p><b>Odpowiedź modelu:</b>  Ocena: [ERROR]  Uzasadnienie: Nazwa produktu jest zbyt ogólna, brak informacji o produkcie, jego przeznaczeniu i brak marki (EWA jest marką, ale nie wystarczy).  Propozycja: Brak, potrzeba więcej informacji o produkcie.</p>

*Źródło: opracowanie własne*

Jednym z głównych ograniczeń rozwiązania jest czas generowania odpowiedzi związany z długością generowanego tekstu w odpowiedzi oraz moc obliczeniowa lub koszty w przypadku modeli płatnych uruchamianych w chmurze. W toku prac badawczych uproszczono zadanie, opracowując krótkie, generyczne pytania oraz żądając uzasadnienia tylko w przypadkach wykrywania nieprawidłowości, co pozwoliło na przeprowadzenie eksperymentów przy niższych kosztach obliczeniowych. Zdecydowano się także na użycie modelu LLaMA 3.1:70b ze względu na ograniczenie kosztów mając jednocześnie świadomość gorszych wyników w porównaniu z modelem GPT-4 (Wang i in., 2024).

W celu przeprowadzenia ewaluacji, z bazy danych wybrano 500 nazw produktów różnych producentów, które ekspercko oceniono jako poprawne, i 500 nazw, które oceniono negatywnie, a które pozytywnie przeszły walidację zarówno pierwszego jak i drugiego poziomu, a także walidację z wykorzystaniem biblioteki Morfeusz2, czyli posiadały wyrazy uznane za powszechne. Wśród nazw ocenionych negatywnie znalazły się takie, które były złożone ze skrótów, z przewagą symboli, z niezrozumiałymi słowami itd. Kategorie nazw produktów wybierano losowo spośród najczęściej występujących w bazie danych. Utworzony w ten sposób zbiór testowy poddano walidacji, testując różne warianty promptów, w tym zmieniając styl, język oraz formułowanie zapytań. Celem tych działań było opracowanie optymalnego promptu, który umożliwiłby modelowi dokładną analizę nazw produktowych przy minimalizacji liczby fałszywie pozytywnych wyników (False Positives). Zadanie to okazało się wysoce złożone ze względu na specyfikę różnych branż, która przekładała się na sposób tworzenia nazw produktów. Wiele zapytań (promptów) wykazywało tendencję do wysokiej czułości, ale także niskiej precyzji. W konsekwencji, przy wdrożeniu takiego walidatora skutkowałoby to wieloma pytaniami ze strony zdezorientowanych użytkowników którym wyświetlałby się nieuzasadniony komunikat o potencjalnym błędzie. W rezultacie wypracowano prompt (Próba 4 w Tabeli 53) dla którego model precyzyjnie wykrywał



niepoprawne nazwy produktów, niestety kosztem ogólnej czułości klasyfikatora. Prompt złożony został z części kontekstowej i głównego pytania. Głównym pytaniem było „Zweryfikuj nazwę produktu: {product\_name}”, gdzie „product\_name” to zmienna, do której podstawiana jest rzeczywista nazwa produktu. Różne, przykładowe wersje kontekstu przedstawiono w Tabeli 53.

Tabela 53. Przykłady promptów walidatora oceny poprawności nazwy produktu

Próba	Kontekst pytania
1	<p>„Zweryfikuj, na podstawie nazwy produktu, czy kupujący zrozumie, co kupuje. Dla jasności, przyjmijmy następujące założenia: - kupujący posługuje się wyłącznie językiem polskim; - kupujący nie jest specjalistą w dziedzinie, do której odnosi się produkt. Dodatkowo, nazwa produktu, aby była zrozumiała, nie powinna zawierać powtarzającej się nazwy marki ani powtarzającej się zawartości netto. Jeśli uważasz, że kupujący nie będzie wiedział, co kupuje, napisz krótkie uzasadnienie.”</p>
2	<p>„Zweryfikuj, na podstawie nazwy produktu, czy kupujący zrozumie, co kupuje. Dla jasności, przyjmijmy następujące założenia: - językiem ojczystym kupującego jest język Polski; - kupujący zna szerszy kontekst, ponieważ kupuje produkt w sklepie o określonej branży; Jeśli uważasz, że kupujący nie będzie wiedział, co kupuje, napisz krótkie uzasadnienie.”</p>
3	<p>„Verify, based on the product name, that the buyer will understand what he is buying. For the sake of clarity, make the following assumptions: - the buyer's native language is Polish; - the buyer is familiar with the broader context because he is buying the product from a store with a specific industry; If you think the buyer won't know what he or she is buying, write a brief justification. Translate the reasoning into Polish”</p>
4 (wybrana)	<p>„Zweryfikuj, na podstawie nazwy produktu, czy kupujący zrozumie, co kupuje. Założenia do weryfikacji: - językiem ojczystym kupującego jest język Polski; - po nazwie można określić kategorię produktu; - kupujący zna trudne i specjalistyczne słowa branżowe; - nazwa może zawierać dodatkowe symbole i specyficzne oznaczenia; - nazwa może być opisowa; - nazwa może być bardzo długa; - nazwa zawiera markę produktu, zwykle umieszczoną na początku. Marka nie musi pomagać w identyfikacji produktu; Jeśli uważasz, że kupujący nie będzie wiedział, co kupuje, napisz krótkie uzasadnienie. Zwróć odpowiedź w języku polskim.”</p>

Źródło: opracowanie własne

Podczas analizy wyników badania zidentyfikowano kilka przypadków zasługujących na szczególną uwagę. Jeden z nich dotyczył produktu „Goldkids Majewski Piornik bez wyposażenia avengers 2023 plus dzienniczek 500kg”. Algorytm ocenił: „Nazwa zawiera wiele niejasnych elementów, takich jak '500kg', co może sugerować wagę produktu, ale w kontekście piornika i dzienniczka wydaje się to nielogiczne. Dodatkowo, obecność słowa 'Avengers' może sugerować związek z marką lub licencją, ale nie jest to jasne.” Algorytm dodatkowo

zweryfikował błędnie wprowadzoną zawartość netto. Nie zwrócił jednak uwagi na błąd pisowni słowa „piórnik”. Inny przypadek dotyczył produktu „*vasopro askorbinian sodu 50% odczynnik chemiczny 100ml*”. Model uznał nazwę za nieprecyzyjną, argumentując: „*Kupujący może nie wiedzieć, co to jest i do czego służy, ponieważ nazwa zawiera specyficzne terminy chemiczne (askorbinian sodu), które mogą być nieznane przeciętnemu konsumentowi*”. Dopiero zmiana nazwy na „*vasopro odczynnik chemiczny askorbinian sodu 50% 100ml*” lub „*vasopro odczynnik chemiczny askorbinian sodowy roztwór wodny 50% 100ml*” spowodowała akceptację produktu przez walidator. Ten przypadek pokazuje jak ważny jest kontekst i jego szyk w nazwie / zdaniu. Kluczowym wnioskiem z badania jest więc konieczność opracowania użytecznego interfejsu, który w połączeniu z wdrożeniem walidatora zapewni użytkownikowi nie tylko automatyczne rekomendacje, ale także jasne wytyczne dotyczące poprawy nazwy produktu najlepiej z indywidualnymi przykładami dostosowanymi do branży.

Wyniki badań przedstawiono w formie macierzy pomyłek (Tabela 56).

Tabela 54. Macierz pomyłek i metryki dla predykcji poprawności nazwy

		Predykcja	
		Pozytywne	Negatywne
Rzeczywistość	Pozytywne	TP 698	FN 2
	Negatywne	FP 323	TN 377

Swoistość (TNR) **53,86%**  
 Ujemna wartość predykcyjna (NPV) **99,47%**  
 Dokładność (ACC) **76,79%**

Źródło: opracowanie własne

Analiza wyników dostarcza szczegółowego obrazu skuteczności walidatora w rozpoznawaniu poprawnych i błędnych nazw produktów. Model poprawnie sklasyfikował 698 nazw jako właściwe (True Positive, TP). Tylko w dwóch przypadkach błędnie ocenił poprawne nazwy jako niepoprawne (False Negative, FN). Z kolei 323 przypadki zostały zaklasyfikowane jako False Positive (FP), co oznacza, że model błędnie uznał niepoprawne nazwy za poprawne. To obszar wymagający dalszej optymalizacji, ponieważ błędne nazwy produktów mogą przechodzić niezauważone i być dopuszczone do użytku. Walidator dobrze rozpoznał 377 nazwy do poprawy (True Negative, TN) ze swoistością na poziomie 53,86%.

Ujemna wartość predykcyjna wyniosła 99,47%, wskazując, że prawie wszystkie sklasyfikowane przypadki rzeczywiście wymagały poprawy. Dokładność modelu osiągnęła 76,79%.

Podsumowując, walidator wykazał wysoką skuteczność w identyfikacji niepoprawnych nazw produktów. Dalsza optymalizacja, w celu zmniejszenia liczby ocen fałszywie pozytywnych jest możliwa przy zmianie modelu na dokładniejszy (nowa wersja LLaMA, lub GPT-4) oraz w przypadku utworzenia dedykowanych zapytań pod konkretną branżę lub grupę asortymentową.

Badania na rzeczywistej bazie danych przeprowadzono na stratyfikowanej części liczącej 0,5 miliona produktów, z odzwierciedleniem proporcji Global Product Classification (GPC). Następnie wyniki z tej próbki zostały ekstrapolowane na całą bazę danych, co pozwoliło oszacować pełen potencjał wykorzystania walidatora w skali całego systemu. Po proporcjonalnym przeliczeniu założono, że walidator może zgłosić uwagi dla około 5 310 422 nazw produktów (Tabela 55) przyjmując, że podstawą do obliczeń była baza jedynie polskich nazw produktów, które zostały zwalidowane na poziomie 1 i 2, oraz wykryto w nich nazwy powszechne biblioteką Morfeusz2.

Tabela 55. Wpływ walidatora jakości nazwy produktu w ujęciu semantycznym na całą bazę danych

Wyniki na całej bazie	Dane
Liczba wykrytych problemów jakościowych w całej bazie	5 338 595
Liczba potwierdzeń braku problemów jakościowych	32 371 958
Zakładana liczba poprawnych wykryć problemów jakościowych	5 310 422

Źródło: opracowanie własne

Łącząc wyniki Morfeusz2 i LLaMA zakładany wpływ walidatora na katalog produktów dla produktów w języku polskim wyniósł 9 509 789 na 42 208 591 produktów, co daje potencjalnie 23% poprawy danych.

### 7.3.10. Ocena przydatności zdjęcia do zastosowań e-commerce

W handlu elektronicznym jakość zdjęć produktowych odgrywa kluczową rolę w procesie podejmowania decyzji zakupowych przez klientów. Ze względu na brak możliwości fizycznego obejrzenia produktu, konsumenci opierają swoje oceny głównie na podstawie

dostarczonych obrazów, co czyni je jednym z najważniejszych elementów prezentacji produktu w e-commerce. Dlatego właściwa walidacja przydatności zdjęć pod względem ich jakości informacyjnej, staje się istotnym elementem doskonalenia jakości danych produktowych w katalogach elektronicznych. W niniejszym rozdziale zostaną omówione możliwości wykorzystania sztucznej inteligencji w tym zakresie, przybliżając je do wytycznych omówionych w Rozdziale 3.2.4, a także potrzeb będących wynikiem audytu opisanego w Rozdziale 5.1.

Wykrywanie i rozpoznawanie obiektów to tematy od ponad dwóch dekad szeroko omawiane w światowej literaturze naukowej. Postęp w tym obszarze zawdzięcza się między innymi odkryciu Konwolucyjnych (Splotowych) Sieci Neuronowych (CNN) (Lecun i in., 1998). CNN, dzięki swojej zdolności do automatycznego wykrywania hierarchicznych wzorców przestrzennych w danych wejściowych, stały się podstawą metod rozpoznawania obrazów. Architektury te efektywnie uczą się reprezentacji danych z obrazów poprzez stosowanie wielowarstwowych operacji splotowych, które sukcesywnie ekstrahują coraz bardziej złożone cechy obiektów, począwszy od prostych krawędzi i kształtów, po bardziej skomplikowane wzorce. W rozwoju CNN kluczową rolę odegrał model ResNet (He i in., 2015), w którym wprowadzono innowacyjną koncepcję połączeń resztkowych (residual connections), umożliwiającą skuteczne trenowanie znacznie głębszych sieci bez napotykania problemu zanikającego gradientu. Problem zanikającego gradientu polega na tym, że w bardzo głębokich sieciach neuronowych sygnał błędu (gradient) stawał się coraz słabszy podczas propagacji wstecznej, co utrudniało uczenie się początkowych warstw sieci.


Implementacja modeli do klasyfikacji obrazów zazwyczaj przebiega według standardowego procesu, w którym model musi najpierw zostać wytrenowany na określonym zbiorze zdjęć, aby był gotowy do użycia i rozpoznawania podobnych obiektów. Jego dokładność i precyzja zależą od stopnia wytrenowania. W przypadku budowy walidatora badającego jakość zdjęcia założono, że otrzymanie predykcji wiążącej informację o kształcie opakowania z daną kategorią GPC będzie wystarczające do akceptacji lub zgłoszenia problemu jakościowego. W koncepcji przyjęto zatem, że np. dla kategorii produktów mlecznych zdjęcie można będzie uznać za prawidłowe, jeżeli będzie widnieć na nim produkt o kształcie kartonika lub butelki. Trenując model będzie można otrzymać stosowne powiązanie kształtów i kategorii, a więc potwierdzenie, lub zaprzeczenie zgodności zdjęcie – klasyfikacja GPC. Wyniki świadczące o słuszności takiej koncepcji przedstawiono na konferencji naukowej (Szymkowski i Niemir, 2024). Mimo pierwszych pozytywnych wyników, późniejsze rozszerzenie modelu na




cały zakres danych, zwiększenie elementów klasyfikujących do całej klasyfikacji GPC i analiza ekspercka wykazały ograniczenia takiego rozwiązania. Koncepcję oceny zdjęcia zmieniono, wykorzystując technikę zbliżoną do rozwiązania opisanego w Rozdziale 7.3.9, czyli opartego o duże modele sterowane zapytaniami. W analizowanym przypadku użyto rozwiązania wykorzystującego duży model multimodalny LLM (Large Multimodal Model) oraz mniejszy - VLM (Vision-Language Model), specyficzny dla kombinacji wizji i języka. Wybrano: multimodalny model LLaVA - Large Language and Vision AI (Liu i in., 2024), oraz mniejszy liczbą parametrów, dostosowany do małych urządzeń mobilnych - model Moondream2 (M87 Labs, 2024).

LLaVA i Moondream2 to modele integrujące dane wizualne i tekstowe, umożliwiające precyzyjną analizę oraz zrozumienie obrazów w kontekście informacji tekstowej. Łączą technologie głębokiego uczenia, konwolucyjnych sieci neuronowych (CNN) oraz transformerów. Przeszkolone na globalnych, różnorodnych zbiorach danych, oferują szeroki zakres możliwości zastosowania – od analizy kształtów i struktur, przez ocenę jakości zdjęć, po generowanie odpowiedzi na pytania dotyczące treści wizualnych. Dzięki zastosowaniu tych modeli możliwe jest nie tylko rozpoznawanie i klasyfikacja obiektów, ale również interpretowanie bardziej złożonych relacji między elementami obrazu i tekstu. Z tego powodu zrezygnowano z połączenia GPC – kształt opakowania, na rzecz analizy przydatności zdjęcia sensu stricto – czy zdjęcie nadaje się do sprzedaży on-line.

Podobnie jak w modelu do oceny poprawności nazwy produktu, walidator wchodzi w interakcję z modelem, aby w postaci tzw. „promptów” czyli zapytań do modelu, uzyskać informacje o tym co znajduje się na zdjęciu i zadać pytanie o jego jakość. W Tabeli 56 przedstawiono przykładową konwersację i ocenę na przykładzie modelu LLaVA.

Tabela 56. Przykład konwersacji służącej do walidacji jakości zdjęcia

Zdjęcie	Komunikacja z modelem
	<p><u>Pytanie do modelu:</u>            Czy to może być zdjęcie do sprzedaży online?            Klient musi wiedzieć, co kupuje, patrząc na zdjęcie.            Odpowiedz [TAK] lub [NIE].</p> <hr/> <p>[TAK]. Zdjęcie pokazuje butelkę syropu z etykietą zawierającą nazwę produktu, ilość oraz inne istotne informacje. To wystarcza, aby klient mógł zidentyfikować produkt, który zamierza kupić. Etykieta jest czytelna, a produkt jest zaprezentowany w sposób umożliwiający klientowi zobaczenie opakowania i zawartości.</p>

Zdjęcie	Komunikacja z modelem
	<p>[NIE]. Zdjęcie nie nadaje się do sprzedaży online, ponieważ nie pokazuje produktu wyraźnie. Koszulka jest pokazana na tle teksturowanym, co utrudnia dostrzeżenie szczegółów, takich jak krój, faktura materiału i ewentualne elementy wzornictwa. Jasne i dobrze oświetlone zdjęcie produktu jest niezbędne do sprzedaży online, aby zapewnić klientom możliwość podjęcia świadomej decyzji zakupowej.</p> <p><u>Pytanie dodatkowe:</u> Czy to produkt ogólny? Odpowiedz [TAK] lub [NIE].</p> <p>[TAK]. Zdjęcie przedstawia zwykłą białą koszulkę, która jest produktem ogólnym. Nie ma na niej widocznych logotypów, wzorów ani unikalnych cech, które odróżniałyby ją od innych zwykłych białych koszulek.</p>
	<p>[TAK]. Zdjęcie pokazuje butelkę soku jabłkowego marki Cook's Pure Apple Juice z czytelną etykietą zawierającą nazwę marki, nazwę produktu oraz wizualne przedstawienie jabłka. Etykieta również wskazuje, że sok jest z koncentratu i nie zawiera sztucznych słodzików. Butelka jest umieszczona na drewnianej powierzchni, a tło jest rozmyte, co pomaga skupić uwagę na produkcie. Tego typu zdjęcie byłoby odpowiednie do sprzedaży online, ponieważ dostarcza niezbędnych informacji, aby klient mógł zidentyfikować produkt, który rozważa zakup.</p>
	<p>[NIE]. Zdjęcie przedstawia trzy butelki produktu, ale nie ma wyraźnego wskazania, czym jest ten produkt. Etykiety są nieczytelne i brak jest widocznej marki lub nazwy produktu, która pozwoliłaby klientowi zidentyfikować zawartość. Ponadto zdjęcie nie dostarcza informacji na temat rozmiaru, ilości lub ceny produktu. Aby zdjęcie mogło być wykorzystane do sprzedaży online, konieczne byłoby uwzględnienie tych informacji i upewnienie się, że produkt jest wyraźnie identyfikowalny dla klienta.</p>

Źródło: opracowanie własne

Zapytania do modelu przetłumaczono w celu łatwiejszej analizy. Oryginalnie konwersacja z modelem odbywała się w języku angielskim, ze względu na ograniczenia modelu i różnicę w działaniu na korzyść rozwiązania angielskojęzycznego.

W procesie oceny przydatności zdjęć do zastosowań e-commerce zdecydowano się na wyeliminowanie kategorii produktów, dla których zdjęcia charakteryzowały się innymi wymaganiami w kontekście prezentacji. Były to:

- **Książki, czasopisma, CD, DVD, płyty winylowe:** O ile z założenia zdjęcie produktu powinno przedstawiać produkt – w tej kategorii książkę czy płytę, to w przeważającej części w tej kategorii znajdują się zdjęcia przedstawiające jedynie pełnowymiarowe okładki, etykiety lub płyty.
- **Pokrycia ścian, wykładziny dywanowe:** Zdjęcia tych produktów zazwyczaj prezentują jedynie wzory lub faktury, zamiast samego produktu w całości, co

komplikuje ocenę ich przydatności do sprzedaży online. Kluczowym elementem jest tutaj brak możliwości rozpoznania całego produktu (np. rolki tapety) na zdjęciu, co prowadzi do trudności w ocenie, czy zdjęcie adekwatnie prezentuje produkt, jaki klient zamierza zakupić.

- **Dekoracyjne magnesy, naklejki, obrazy, plakaty:** W przypadku tych kategorii zdjęcia często przedstawiają elementy dekoracyjne na różnych powierzchniach, co może prowadzić do błędów w rozpoznawaniu, czy dany obiekt jest rzeczywiście produktem przeznaczonym do sprzedaży. Model mógłby zacząć wyszukiwać na zdjęciu elementy, które nie są faktycznym produktem (np. scenerie, obiekty tła), co wprowadza niepożądane zamieszanie w kontekście oceny przydatności zdjęcia.
- **Rośliny żywe:** Zdjęcia przedstawiające rośliny są zazwyczaj wykonane w plenerze i mogą nie spełniać standardowego wymogu prezentacji produktów na jednolitym, neutralnym tle, co jest istotne w kontekście e-commerce. Przykładowo zdjęcie kwitnącego fragmentu rośliny może nie być sklasyfikowane jako produkt.
- **Usługi:** Zdjęcia związane z usługami mogą prezentować logo, reklamy lub inne elementy graficzne, które nie są produktami fizycznymi. W związku z tym, standardowe podejście do oceny zdjęć nie jest adekwatne dla tej kategorii.

Wymienione kategorie wymagały zastosowania odmiennych strategii analizy zdjęć. Eliminacja tych kategorii z procesu ogólnej oceny zdjęć była zatem uzasadniona w celu uniknięcia błędów klasyfikacyjnych oraz zwiększenia ogólnej efektywności algorytmu w zakresie przydatności do zastosowań e-commerce w sensie uniwersalnym.

Po wyeliminowaniu wyżej wymienionych kategorii oraz zdjęć z prawdopodobnie niepoprawnym tłem (Rozdział 7.3.6), przeanalizowano łącznie 1663 różniących się 64 bitowym haszem percepcyjnym zdjęć z losowych kategorii. Oznaczono je jako poprawne i niepoprawne (174 zdjęć niepoprawnych i 1489 zdjęć poprawnych). Ocena "niepoprawnego" zdjęcia przez eksperta oznaczała, że zdjęcie nie spełniało podstawowych kryteriów przydatności do sprzedaży online. Kryteria te obejmowały takie elementy, jak czytelność przedstawianych na zdjęciu informacji, występowanie głównego obiektu sprzedaży, forma prezentacji, sceneria (zgodnie z wymaganiami opisanymi w Rozdziale 3.2.4). Innymi słowy, zdjęcie uznane za niepoprawne nie prezentowało produktu w sposób wystarczająco klarowny lub zgodny z oczekiwaniami dla sprzedaży internetowej, co mogło powodować trudności w zrozumieniu, co jest przedmiotem oferty.

Do badania skuteczności wykrywania prawidłowych i nieprawidłowych zdjęć wykorzystano dwa modele: LLaVA oraz Moondream2. Przeanalizowano kilkadziesiąt zapytań, aby uzyskać dokładne wyniki dotyczące wykrywania zdjęć nieprawidłowych, bez konieczności szczegółowej analizy zgodności fotografii z konkretną kategorią produktu. Model LLaVA skutecznie odpowiadał na pytanie: „Czy zdjęcie może być użyte do sprzedaży online?”, podczas gdy model Moondream2 okazał się lepszy w udzielaniu odpowiedzi na bardziej ogólne pytanie: „Czy zdjęcie przedstawia produkt?” („Does the photo show the product?”), wykonując je 10 razy szybciej. W przypadku zdjęć produktów odzieżowych, dodatkowo zadawano pytanie, czy zdjęcie przedstawia modelkę lub modela prezentującego odzież na białym tle („Is there a model with a white background in the picture?”). Wyniki porównawcze modeli zaprezentowano w Tabeli 57.

Tabela 57. Porównanie dwóch modeli detekcji obiektów na obrazie

Model	Dokładność	Precyzja (dobre zdjęcie)	Czułość (dobre zdjęcie)	NPV (nieodpowiednie zdjęcie)	Swoistość (nieodpowiednie zdjęcie)
<b>LLaVA</b>	89,84%	90,34%	99,26%	59,26%	9,20%
<b>Moondream2</b>	93,87%	95,06%	98,25%	79,03%	56,32%

Źródło: opracowanie własne

Moondream2 wykazał większą skuteczność w wykrywaniu zdjęć nieodpowiednich, co znalazło odzwierciedlenie w wyższej wartości NPV (79,03%) i lepszym wyniku swoistości (56,32%). To oznacza, że Moondream2 znacznie częściej poprawnie klasyfikował zdjęcia jako błędne. Model LLaVA rzadko wykrywał nieodpowiednie zdjęcia. Miał niższą wartość NPV (59,26%) i szczególnie niską swoistość (9,20%). W kontekście fałszywej klasyfikacji poprawnych zdjęć jako niepoprawne Moondream2 uzyskał wysoki wynik czułości (98,25%), co oznacza, że rzadziej popełniał błędy polegające na oznaczaniu poprawnych zdjęć jako niepoprawne, natomiast Model LLaVA był wyjątkowo czuły (99,26%) na poprawne zdjęcia, co zapewniło bardzo niską liczbę fałszywie pozytywnych wyników w tej kategorii, ale kosztem wykrywania błędnych zdjęć. Podsumowując, mimo mniejszej czułości, wyniki modelu Moondream2 okazały się lepsze dla tego zdania, choć nadal mało precyzyjne i czułe w kontekście wykrywalności wad. W Tabeli 58 zaprezentowano podsumowanie wyników w ujęciu całej bazy danych z pominięciem zdjęć zbyt małych, na niejednorodnym lub kolorowanym tle, oraz na zbyt dużej powierzchni tła. W Tabeli 59 zaprezentowano przykłady wykrytych problemów jakościowych wraz z uzasadnieniem generowanym przez rozwiązanie.



Tabela 58. Wpływ walidatora jakości zdjęcia na całą bazę danych

Wyniki na całej bazie	Dane
Liczba wykrytych problemów jakościowych w całej bazie	277 646
Liczba potwierdzeń braku problemów jakościowych	1 122 394
Zakładana liczba poprawnych wykryć problemów jakościowych	219 424

Źródło: opracowanie własne

Tabela 59. Przykłady wykrytych problemów jakościowych zdjęć

	Zdjęcie nie prezentuje produktu, lecz minimalistyczny symbol – zielone koło z białym sercem w środku, symbolizującym miłość, troskę i uczucie. Brak dodatkowych szczegółów pozwala skupić się na symbolicznym znaczeniu serca, co często jest wykorzystywane w materiałach marketingowych i komunikacji wizualnej.
	Zdjęcie nie ukazuje produktu efektywnie, ponieważ jest czarno-białe, co utrudnia odwzorowanie kolorów i szczegółów. W tym przypadku, tortilla oraz tortilla pszenna są kluczowymi elementami, a obraz mógłby lepiej oddać ich wygląd w formie fotografii kolorowej.
	Produkt nie jest widoczny, ponieważ zdjęcie skupia się na zbliżeniu cukierków owiniętych w plastik. Główna uwaga skierowana jest na cukierki oraz ich opakowanie, pomijając otoczenie i inne elementy.
	Zdjęcie nie przedstawia produktu, ponieważ ukazuje jedynie jego etykietę, skupiając się na informacji o składnikach, wartościach odżywczych oraz cechach charakterystycznych. Brak rzeczywistego obrazu produktu sprawia, że nie jest on widoczny na zdjęciu.
	Produkt nie jest widoczny na zdjęciu, ponieważ kobieta trzyma go przed twarzą, co powoduje, że głównym elementem obrazu staje się ona oraz jej działania. Tło, na którym widać ocean, nadaje zdjęciu estetyczny walor, jednak nie jest istotne dla przedstawienia produktu, jakim jest herbata KABISA.
	Produkt nie jest przedstawiony bezpośrednio, ponieważ zdjęcie ukazuje jego odbicie – stoik z kokosem i połówki kokosa umieszczone na refleksyjnej powierzchni. Efekt odbicia tworzy artystyczną kompozycję, podkreślając produkt oraz zwracając na niego uwagę.

Źródło: opracowanie własne

Po przeanalizowaniu wyników zauważono, że trudnością w ocenie jakości danych produktowych były takie zdjęcia, które przedstawiały produkty, na których były inne zdjęcia. Problem stanowiło także rozpoznanie zdjęć przedstawiających sprzedawane etykiety, a także

takie, które przedstawiały produkty złożone z prostych kształtów. Takimi przykładami były tablice ostrzegawcze, wycięte z blachy metalowe przedmioty, kartki świąteczne oraz szare kartony bez logo. Głównym powodem problemów z poprawną identyfikacją takich zdjęć była ich złożoność semantyczna. Modele, jak wybrany Moondream2, zazwyczaj opierają się na jednoznacznych cechach wizualnych do identyfikacji kategorii. Kiedy obrazy nie posiadają wyraźnych wskazówek semantycznych lub zawierają elementy wspólne dla różnych kategorii, modele mogą mieć trudności z ich poprawnym sklasyfikowaniem. Dodatkowo, podobieństwo wizualne może wprowadzać modele w błąd. Na przykład, metalowe przedmioty o prostych, wyciętych kształtach mogą przypominać kliparty, co utrudnia ich identyfikację jako produktów. W przypadku kartek świątecznych z elementami pejzaży, modele mogą skupić się na dominujących cechach wizualnych, takich jak krajobrazy, pomijając kontekst samej kartki, szczególnie jeśli jest biała na białym tle. Szare kartony bez logo stanowią wyzwanie, ponieważ brakuje na nich charakterystycznych elementów, takich jak tekst czy symbole, które modele mogłyby wykorzystać do prawidłowej identyfikacji. W celu poprawy skuteczności modelu, konieczne byłoby nie tylko rozszerzenie i urozmaicenie zbioru treningowego, ale również zastosowanie bardziej zróżnicowanych podejść w generowaniu promptów, dostosowanych indywidualnie do każdego rodzaju produktu / branży / GPC. Warto jednak zaznaczyć, że to, co w klasycznym rozumieniu mogłoby być uznawane za wadę modelu, może paradoksalnie stanowić jego zaletę w przypadku walidacji. Jeśli walidator zbudowany w oparciu o taki model sygnalizuje, że zdjęcie nie jest wystarczająco poprawne, warto rozważyć jego zmianę – na przykład poprawić kąt fotografowania, oświetlenie lub inne aspekty techniczne. Wynika to z faktu, że skoro algorytm zgłosił uwagę do zdjęcia, trudności w jego poprawnej ocenie mogą mieć również inne algorytmy, boty wpływające na wyszukiwanie produktu w Internecie oraz użytkownicy końcowi, szczególnie w aplikacjach mobilnych, gdzie zdjęcie stanowi jeden z najważniejszych atrybutów produktu (Di i in., 2014). Wskazanie przez model na potencjalne nieprawidłowości może zatem pełnić funkcję ostrzegawczą, sugerując konieczność weryfikacji jakości obrazu przed jego ostatecznym wykorzystaniem w Internecie. Rozszerzając w przyszłości walidator, można także zwrócić uwagę na aspekty wizualne opakowania produktu prezentowanego na zdjęciu w wybranych kategoriach. Można rozpocząć od kontroli, czy produkt jest w opakowaniu, co obecnie nie jest standardem, a przez to ogranicza możliwość zastosowania walidacji dotyczącej procentu pokrycia tła. Następnie można włączyć analizę estetyki opakowania. Jak udowadniają Autorzy badań (Michalski, 2024; Michalski i Grobelny, 2016), zarówno kształt, jak i kolor opakowania mają istotny wpływ na percepcję produktu przez konsumentów oraz ich skłonność do zakupu.

## 7.4. Ocena wpływu walidatorów na dane w zaproponowanym modelu

W ramach niniejszej rozprawy dokonano analizy wpływu wybranych walidatorów opracowanych w modelu wielopoziomowej walidacji danych produktowych, realizując zapisy Zadania 5 oraz postępując zgodnie z proponowaną metodyką opisaną w Rozdziale 6.3. Ocenę przeprowadzono w oparciu o rzeczywiste wyniki zebrane z uruchomienia walidatorów na danych pochodzących z elektronicznego katalogu produktów będącego podmiotem badań. Proces oceny obejmował trzy główne proponowane poziomy walidacji: normalizację danych, walidację techniczno-logiczną oraz sugerowanie zmian. Analizę zbiorczą w ujęciu wpływu na dane produktów, uwzględniającą wszystkie proponowane walidatory, przedstawiono w Tabeli 60. Zaprezentowano w niej informacje o wykrytych problemach jakościowych, a także procentowy udział każdego z nich w poprawie jakości lub w przypadku walidatorów 3-go poziomu, w potencjalnej poprawie jakości. Podstawą do liczenia wartości procentowych była baza produktów licząca 44 959 300 rekordów.

Tabela 60. Zestawienie walidatorów i ich wpływu na polepszenie jakości danych

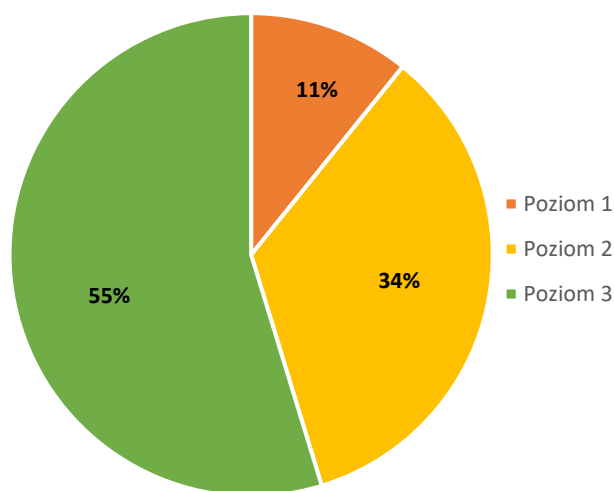
Problem jakości	Wpływ na bazę produktów	Procent bazy
[W1.1] Encja HTML zamiast znaku specjalnego w tekście	29 116	0,065%
[W1.1] Encja HTML zamiast znaku specjalnego w tekście (w nazwie marki)	3 464 676	7,71%
[W1.2] Znak tabulacji w tekście jednowierszowym	16 646	0,037%
[W1.2] Znak tabulacji w tekście jednowierszowym (w nazwie marki)	63 828	0,14%
[W1.3] Kod sterujący w tekście	181 956	0,40%
[W1.3] Kod sterujący w tekście (w nazwie marki)	930	0,002%
[W1.4] Znak nowej linii w tekście jednolinijkowym	67 520	0,15%
[W1.4] Znak nowej linii w tekście jednolinijkowym (w nazwie marki)	328 435	0,73%
[W1.5] Nadmiarowe spacje w tekście	1 934 469	4,30%
[W1.5] Nadmiarowe spacje w tekście (w nazwie marki)	11 018 054	24,5%
[W1.6] Podwójny znak cudzysłowu w tekście	7 407	0,016%
[W1.7] Spacja przed tekstem właściwym	4 413	0,010%
[W1.7] Spacja przed tekstem właściwym (w nazwie marki)	189 005	0,42%
[W1.8] Spacja po tekście właściwym	38 578	0,086%
[W1.8] Spacja po tekście właściwym	2 858 311	6,36%
[W2.1] Brak wskazania zawartości netto w nazwie produktu	5 160 051	11,5%
[W2.2] Brak marki w nazwie produktu	34 219 091	76,1%
[W2.3] Powtarzające się nazwy produktów dla różnych GTIN	15 321 846	34,1%
[W2.4] Produkt z nazwą jednowyrazową	3 399 820	7,56%
[W2.5] Brak słów w nazwie produktu	1 920 134	4,27%
[W2.6] Marka produktu z nadmierną ilością słów	1 123 461	2,50%
[W2.7] Niecałkowita liczba sztuk, par, egzemplarzy	687 517	1,53%
[W2.8] Początek tekstu to nieprawidłowy znak	320 302	0,71%

<b>Problem jakości</b>	<b>Wpływ na bazę produktów</b>	<b>Procent bazy</b>
[W2.9] Niepoprawne nazwy marek (lista)	162 760	0,36%
[W2.10] Nadmiarowe podkreślenia w tekście	150 949	0,34%
[W2.11] Niedozwolone znaki w tekście wynikające z błędnej konwersji znaków (w nazwie marki)	310	0,001%
[W2.11] Niedozwolone znaki w tekście wynikające z błędnej konwersji znaków (lista)	2 666	0,006%
[W2.12] Nazwa produktu z nadmierną ilością słów	1 348 850	3,00%
[W2.13] Zawartość netto $\leq 0$	33 030	0,073%
[W2.14] Identyczna nazwa produktu i marki	37 132	0,083%
[W2.15] Nazwa marki bez liter	5 702	0,013%
[W2.16] Formuła Excela zamiast treści	5 429	0,012%
[W2.16] Formuła Excela zamiast treści (w nazwie marki)	428 824	0,95%
[W2.17] Zbyt niska rozdzielczość zdjęcia	5 548	0,21%
[W2.18] Nadmiernie długa nazwa marki	74 982	0,17%
[W2.19] Nadmiernie długa nazwa produktu	10	0,00002%
[W2.20] Zbyt duża rozdzielczość zdjęcia	158	0,0004%
[W2.21] Przekroczona wielkość pliku	15	0,00003%
[W2.22] Przekroczona wielkość zawartości netto	10	0,00002%
[W3.1] Nazwa produktu pisana wyłącznie wielkimi literami	11 401 826	25,4%
[W3.2] Nadmiernie długa nazwa produktu do niektórych zastosowań e-commerce	5 271 882	11,7%
[W3.3] Początek tekstu to cudzysłów lub apostrof	29 759	0,066%
[W3.4] Różne zapisy tej samej marki	3 228 885	7,18%
[W3.5] Nazwa produktu w języku nieodpowiadającym wskazanemu językowi	4 787 554	10,6%
[W3.6] Zbyt niska rozdzielczość zdjęcia dla niektórych zastosowań e-commerce	978 926	36,3%
[W3.7] Zbyt niska rozdzielczość zdjęcia dla większości zastosowań e-commerce	17 440	0,65%
[W3.8] To samo zdjęcie dla różnych produktów	1 465 797	54,4%
[W3.9] Zdjęcie produktu na niejednorodnym lub kolorowym tle	559 740	20,8%
[W3.10] Zdjęcie produktu na zbyt dużej powierzchni tła	7 252	0,27%
[W3.11] Nieadekwatna jednostka zawartości netto w odniesieniu do kategorii/grupy produktów	1 663 071	3,70%
[W3.12] Nieadekwatna zawartość netto w odniesieniu do kategorii/grupy produktów	3 903 393	8,68%
[W3.13] Niepoprawna lub tymczasowa kategoria produktu	2 641 815	5,88%
[W3.14] Nazwa jest niezrozumiała	10 129 536	22,5%
[W3.15] Zdjęcie nieprzedstawiające produktu do sprzedaży	219 424	19,5%

Źródło: opracowanie własne

Na poziomie pierwszym (walidatory [W1.x]) przeprowadzono normalizację danych produktowych, koncentrując się głównie na formatowaniu pól tekstowych nazwy produktu i marki. Dzięki temu procesowi wykryto i skorygowano 3 256 100 problemów w atrybutach produktowych. Mnożąc markę przez liczbę produktów, na których występowała, wpływ na pełną informację o produktach wyniósł aż 20 203 343. Na poziomie drugim (walidatory

[W2.x]) zrealizowano walidację techniczno-logiczną, polegającą na sprawdzeniu zgodności danych z ustalonymi normami oraz detekcji błędów logicznych. Ten etap umożliwił wykrycie 63 906 108 problemów jakościowych (licząc z wpływem marek 64 408 597), które wymagały odrzucenia produktów. Poziom trzeci, zgodnie z założeniami, obejmował nie tylko wykrywanie błędów, ale także sugerowanie możliwych zmian i poprawek, co doprowadziło do identyfikacji 42 468 089 problemów jakościowych (uwzględniając wpływ marek, to 46 306 300). Wartość ta jest szacowana ze względu na zastosowanie stratyfikacji dla niektórych walidatorów, związane z ograniczeniami zasobów obliczeniowych niezbędnych do przetworzenia całej bazy danych oraz na dokładność walidatorów obliczoną na podstawie wycinków bazy danych. Przy obliczaniu wartości proporcjonalnie odjęto fałszywie pozytywne wyniki. Łączna szacowana liczba wszystkich wykrytych rzeczywistych problemów w atrybutach wyniosła 108 712 148 (130 918 241 w odniesieniu wpływu marek na pełną listę atrybutów produktu). Zbiorcze wyniki z podziałem na poziomy walidacji zaprezentowano na Rysunku 37.



Rysunek 37. Wpływ walidatorów na bazę produktów w podziale na poziomy walidacji  
Źródło: opracowanie własne

Z przedstawionego wykresu kołowego wynika, że dla wszystkich zaproponowanych walidatorów w modelu, walidatory 3-go poziomu, umożliwiające korzystanie m in. z algorytmów sztucznej inteligencji bez konsekwencji ryzyka wprowadzenia błędów do bazy danych, stanowią 55% wpływu na poprawę danych wszystkich walidatorów. Ze względu na fakt, że jedynie 6% danych produktowych zawierało zdjęcia, walidatory dotyczące tego obszaru zostały przedstawione osobno, w Tabeli 61, w celu możliwości obserwacji wpływu walidatorów wyłącznie na dane obrazowe wraz z procentowym rozkładem.

Tabela 61. Zestawienie walidatorów zdjęć i ich wpływu na polepszenie jakości danych obrazowych

<b>Problem jakości</b>	<b>Wpływ na bazę zdjęć</b>	<b>% bazy zdjęć</b>
[W2.17] Zbyt niska rozdzielczość zdjęcia	5 548	0,206%
[W3.6] Zbyt niska rozdzielczość zdjęcia dla niektórych zastosowań e-commerce	978 926	36,319%
[W3.7] Zbyt niska rozdzielczość zdjęcia dla większości zastosowań e-commerce	17 440	0,647%
[W3.8] To samo zdjęcie dla różnych produktów	1 465 797	54,382%
[W3.9] Zdjęcie produktu na niejednorodnym lub kolorowym tle	559 740	20,767%
[W3.10] Zdjęcie produktu na zbyt dużej powierzchni tła	7 252	0,269%
[W3.15] Zdjęcie nieprzedstawiające produktu do sprzedaży	219 424	8,141%

Źródło: opracowanie własne

Analiza przedstawionych tabel pozwoliła sformułować kilka istotnych wniosków:

- A. Zastosowanie sztucznej inteligencji nie jest konieczne we wszystkich przypadkach, szczególnie tam, gdzie otoczenie biznesowe narzuca konkretne reguły i wytyczne dotyczące jakości danych.
- B. Nawet w dobrze zarządzanym katalogu danych można zaproponować dodatkowe, klasyczne walidatory do istniejących, które udoskonalą dane zarówno na poziomie pierwszym, jak i drugim.
- C. Poziom trzeci otwiera zupełnie nowe możliwości, ukazując potencjał rozwojowy oraz przyszłość interakcji z wykorzystaniem najnowocześniejszych rozwiązań, głównie z zakresu sztucznej inteligencji.

## 8. Podsumowanie i kierunki dalszych badań

Niniejsza rozprawa doktorska o charakterze teoretyczno-empirycznym miała na celu opracowanie i adaptację modelu walidacji do zarządzania danymi produktowymi w katalogach elektronicznych, który uwzględni możliwość korzystania z nowoczesnych rozwiązań z zakresu sztucznej inteligencji.

Przeprowadzone badania i analizy ukazały złożoność problematyki jakości danych produktowych i bardzo duży potencjał w rozwoju walidatorów opartych o sztuczną inteligencję. Badania nad standardami i praktykami w zakresie wprowadzania danych o produktach do katalogów elektronicznych pomogły odkryć przede wszystkim brak spójności w nazewnictwie produktów między platformami sprzedażowymi, rozbieżności wytycznych dotyczących atrybutów tekstowych jak i zdjęć produktów, problemy jakościowe w zapisie zawartości netto oraz kategoryzacji produktu. Przeprowadzona analiza danych produktowych pochodzących z wytypowanych baz, sklepów internetowych oraz platform internetowych ujawniła, że problem jakości występuje powszechnie. Przedstawiono również, że wyzwania związane z poprawą danych tekstowych i graficznych wymagają zaawansowanych metod analizy semantycznej i zrozumienia kontekstu, czego nie można wykonać za pomocą algorytmów regułowych. Analiza danych podmiotu badań ujawniła szereg problemów różnej charakterystyki, które pogrupowano względem możliwości poprawy, uwzględniając również interakcję z osobami wprowadzającymi dane (tzw. społeczność wprowadzającą dane). W ten sposób utworzono model wielopoziomowej walidacji, w którym pierwszy poziom uwzględnił automatyczną poprawę danych, drugi – odrzucenie nieprawidłowych danych, a trzeci – warunkową akceptację z oznaczeniem błędów i sugestiami poprawek. Taka systematyka z kolei otworzyła możliwości zbadania rozwiązań – w postaci konkretnych walidatorów, które pomogłyby w wykrywaniu problemów jakościowych z uwzględnieniem zastosowania najnowszych algorytmów sztucznej inteligencji. Wykorzystanie nowoczesnych algorytmów otworzyło nowe możliwości w obszarze doskonalenia jakości danych, choć z uwzględnieniem konieczności monitorowania wyników.

W procesie walidacji jakości danych produktowych zastosowano szereg algorytmów sztucznej inteligencji oraz narzędzi wspomagających analizę tekstu i obrazu. Kluczowe walidatory w których użyto sztuczną inteligencję to:

- A. **Biblioteka SpaCy** - wykorzystana do analizy nazw produktów ([W3.1]). Biblioteka wykorzystuje uczenie maszynowe do wykrywania części zdania (parsing składniowy).

- B. **LLaMA** - użyta w połączeniu z operacjami na ciągach tekstu do sugestii krótszej nazwy produktu, przy zbyt długich nazwach produktów oraz oczyszczonej nazwy dla nazw z niepoprawnymi znakami ([W3.2], [W3.3]). LLaMA to generatywny, zaawansowany duży model językowy (LLM) opracowany przez Meta AI. Umożliwia generowanie i rozumienie tekstu, optymalizując zadania językowe.
- C. **LLaMA** – użyta z algorytmem Jaro-Winklera do detekcji różnych zapisów tej samej marki i sugestii najlepszej ([W3.4]).
- D. **FastText** - do analizy językowej wraz z **Morfeusz2**, co pozwoliło na identyfikację produktów w języku niezgodnym z oczekiwanym ([W3.5]). FastText to algorytm uczenia maszynowego, stworzony przez Facebook AI Research, który umożliwia szybkie i efektywne tworzenie wektorowych reprezentacji słów oraz klasyfikację tekstu, z uwzględnieniem struktury wewnętrznej słów.
- E. **RemBG** wraz z Pillow - użyte do przetwarzania obrazu, usuwania i analizy tła ([W3.6], [W3.7], [W3.9], [W3.10]). RemBG to narzędzie wykorzystujące sieci neuronowe do segmentacji obrazów, umożliwiające usunięcie tła, co ułatwia dalszą analizę wizualną.
- F. **BERT**, wyuczony do klasyfikacji jednostek zawartości netto, zastosowany w celu walidacji i predykcji najlepszej jednostki po nazwie i klasyfikacji produktu ([W3.12]). BERT to zaawansowany model językowy oparty na architekturze Transformer, który umożliwia zrozumienie kontekstu dwukierunkowego w analizie tekstu.
- G. **BERT i LLaMA**, w hybrydowym połączeniu do klasyfikacji kategorii produktów GPC wraz z jej sprawdzeniem i uzasadnieniem ([W3.11]).
- H. **LLaMA** - zastosowana do identyfikacji niezrozumiałych nazw produktów wraz z uzasadnieniem. ([W3.14]).
- I. **Moondream2** – do identyfikacji zdjęć nieprzedstawiających produktu ([W3.15]) i do tworzenia uzasadnienia oceny. Moondream2 to kompaktowy model wizji językowej (VLM), przystosowany do zadań związanych z rozumieniem i klasyfikacją obrazów.

Przeprowadzono szczegółową analizę działania każdego z proponowanych walidatorów, a uzyskane wyniki, w postaci liczby wykrytych problemów jakościowych, przedstawiono w celu ukazania wpływu oraz potencjału zarówno proponowanego rozwiązania, jak i całej koncepcji.



## **8.1. Zrealizowane zadania badawcze**

### **Zadanie 1: Analiza literatury przedmiotu.**

Przeprowadzono przegląd literatury dotyczącej różnych aspektów jakości danych produktowych, w tym metod walidacji oraz problemów występujących w zarządzaniu danymi w katalogach elektronicznych. Szczególną uwagę zwrócono na możliwość zastosowania sztucznej inteligencji oraz na specyficzne wyzwania związane z otwartymi polami tekstowymi oraz zdjęciami produktów, które stanowią istotny element jakości danych podstawowych w katalogu produktowym. Analiza literatury pomogła zidentyfikować istniejące luki poznawcze w obszarze walidacji danych produktowych i sformułować kluczowy problem badawczy, związany z doskonaleniem metod automatycznej walidacji przy użyciu sztucznej inteligencji.

### **Zadanie 2: Eksploracja standardów i analiza sektora e-commerce w kontekście przyjętych norm i praktyk w procesie wprowadzania danych o produktach.**

W ramach tego zadania przeprowadzono analizę obowiązujących standardów oraz praktyk stosowanych w sektorze e-commerce w zakresie wprowadzania danych produktowych. Szczególną uwagę zwrócono na różnorodność podejść stosowanych przez platformy e-commerce do podstawowych atrybutów produktowych, takich jak nazwy produktów, identyfikatory GTIN, marki oraz zdjęcia. Badania wykazały, że brak spójności w interpretacji tych atrybutów oraz różnice w implementacji standardów prowadzą do problemów z jakością danych. Analiza ujawniła, że choć niektóre standardy, jak GTIN, są powszechnie stosowane, nie zawsze są odpowiednio walidowane, co utrudnia wymianę informacji między różnymi systemami. Wykazano potrzebę standaryzacji globalnych norm oraz opracowania uniwersalnych wytycznych, które mogłyby zostać zastosowane w szerokim zakresie platform e-commerce.

### **Zadanie 3: Identyfikacja wyzwań związanych z walidacją danych na podstawie analizy bazy danych będącej przedmiotem badań.**

Analiza bazy danych produktowych, będącej przedmiotem badań, pozwoliła zidentyfikować kluczowe wyzwania i problemy jakościowe. W wyniku analizy stwierdzono występowanie 45 specyficznych problemów jakościowych, związanych z różnymi atrybutami, takimi jak nazwa, marka, klasyfikacja produktu, zdjęcia produktów czy zawartość netto. Zauważono problemy związane z zależnościami pomiędzy danymi, między nieprawidłową kategoryzacją produktów, a także dużą liczbę problemów z jakością nazwy produktu, która polega nie tylko na interpretacji syntaktycznej, ale semantycznej, związanej z rozumieniem tekstu.

### **Zadanie 4: Opracowanie struktury wielopoziomowego modelu walidacji.**

Opracowano wielopoziomowy model walidacji, który uwzględnia różne aspekty jakości danych produktowych. Model złożono z trzech głównych poziomów: Pierwszy poziom obejmuje normalizację danych, czyli automatyczne czyszczenie, korekcję formatowania i harmonizację terminologii bez zmiany znaczenia danych. Drugi poziom to walidacja techniczno-logiczna, która polega na sprawdzeniu zgodności danych z przyjętymi standardami oraz analizie logicznej spójności pomiędzy różnymi polami danych. Trzeci poziom obejmuje sugestie dotyczące poprawek i ulepszeń, które wspierają proces poprawy jakości danych, jednak nie blokują ich wprowadzania. Taki model umożliwił otwarcie możliwości skorzystania walidacji opartej na sztucznej inteligencji, której rozwiązania mogą nie dawać dokładnych wyników, a także która może nie być deterministyczna.

### **Zadanie 5: Selekcja, adaptacja i ocena wpływu algorytmów walidacyjnych do zastosowania w zaproponowanym modelu walidacji.**

W ramach tego zadania wybrano i zaadaptowano algorytmy walidacyjne na każdym poziomie walidacji zgodnie z modelem opracowanym w Zadaniu 4. Postanowiono uwzględnić wszystkie problemy zidentyfikowane w Zadaniu 3, co wymagało analizy możliwości zastosowania algorytmów zarówno opartych na regułach, jak i na sztucznej inteligencji, w zależności od potrzeb oraz z uwzględnieniem efektywnego zarządzania zasobami obliczeniowymi. Dzięki temu przedstawiono potencjał walidacji na trzecim poziomie

w porównaniu do pozostałych dwóch poziomów. Następnie przeprowadzono szczegółową analizę działania każdego z proponowanych walidatorów. Uzyskane wyniki, w postaci liczby wykrytych problemów jakościowych, przedstawiono w celu ukazania wpływu oraz potencjału zarówno proponowanego rozwiązania, jak i całej koncepcji trypoziomowego modelu.

## **8.2. Odpowiedzi na pytania badawcze**

Przeprowadzone badania własne oraz analiza uzyskanych wyników pozwoliły na udzielenie odpowiedzi na przyjęte w rozprawie pytania badawcze.

### **Pytanie 1. Jakie standardy i praktyki są stosowane w sektorze e-commerce w kontekście procesu wprowadzania danych o produktach?**

Badania przeprowadzone na 10-ciu platformach e-commerce oraz analiza baz produktowych z 25-ciu sklepów internetowych, platformy zakupowej z 771 ofertami tego samego produktu o różnych nazwach i 3-ech katalogów danych ujawniły kluczowe problemy jakości danych w e-commerce. Sektor e-commerce korzysta z własnych, zdefiniowanych wytycznych w procesie wprowadzania danych o produktach, które są zbliżone wymaganiami, ale różnią się w szczegółach. Jedynie numer GTIN jako standard identyfikacyjny jest powszechnie stosowany, ale nie jest obowiązkowy i nie zawsze jest poprawnie walidowany. Wytyczne sieci globalnej wymiany danych produktowych GS1 GDSN, z których korzysta również Globalny Rejestr Produktów GS1 GRP, nie definiują dokładnego schematu nazwy produktu, co utrudnia odniesienie się do wzorca przy analizie semantycznej. Istnieje jednak ogólna zasada, że pole to powinno zawierać dane z etykiety produktu. Badania na rzeczywistych bazach danych produktowych pochodzących z różnych źródeł dostępnych online ugruntowały tezę o globalnym problemie braku spójności nazw produktów.

### **Pytanie 2. Jakie problemy jakościowe można znaleźć w danych produktowych w elektronicznych katalogach?**

Na podstawie przeprowadzonej analizy bazy danych będącej przedmiotem badań zidentyfikowano 45 specyficznych problemów jakościowych w danych produktowych. Najczęściej występujące problemy dotyczyły nazwy produktu (23 przypadki), marki

(17 przypadków), następnie zdjęć produktów oraz zawartości netto. Analiza ujawniła również częste zależności pomiędzy polami danych, takie jak nieprawidłowa kategoryzacja produktów, nieadekwatne opisy czy zdjęcia nieprzedstawiające produktu. Zidentyfikowano m.in. błędy pisowni, problemy z formatowaniem tekstu, naruszenia zasad biznesowych organizacji, naruszenia ograniczeń integralności oraz zduplikowane dane. Szczególnie problematyczne okazały się naruszenia zasad biznesowych (16 przypadków) oraz formatowanie tekstu (11 przypadków). Algorytmy o niskiej złożoności, rozwiązujące 34 problemy, mogły być skutecznie zastosowane do prostych walidacji, takich jak kontrola długości pól tekstowych czy zgodności z predefiniowaną listą wartości. Algorytmy wysokiej złożoności, stosowane do niektórych problemów, takich jak analiza semantyczna nazw produktów czy wizualna analiza zdjęć, wymagały użycia zaawansowanych technik sztucznej inteligencji.

### **Pytanie 3. Jaka struktura modelu walidacji danych produktowych zapewni skorzystanie z metod sztucznej inteligencji w elektronicznych katalogach?**

Struktura modelu walidacji danych produktowych, który skutecznie wykorzystuje metody sztucznej inteligencji w elektronicznych katalogach, powinna być wielopoziomowa, obejmując trzy kluczowe komponenty.

1. **Poziom normalizacji danych** – na tym etapie dane powinny być wstępnie przetworzone i uporządkowane. Automatyczne mechanizmy oczyszczają dane z błędów formatowania, usuwają nadmiarowe spacje, tabulacje oraz standaryzują wartości (np. formatowanie jednostek miary lub numerów GTIN).
2. **Poziom walidacji techniczno-logicznej** – ten poziom bazuje na weryfikacji spójności danych pod względem reguł i norm branżowych, np. czy dane atrybuty produktów są zgodne z wymogami formalnymi.
3. **Poziom sugestii poprawy danych** – na tym etapie wykorzystuje się zaawansowane algorytmy sztucznej inteligencji, takie jak modele generatywne (np. ChatGPT), które potrafią wspierać użytkowników w poprawie jakości danych. Sztuczna inteligencja może pełnić rolę asystenta ("co-pilota") poprzez analizę semantyczną danych, co umożliwia sugerowanie zmian lub poprawnych wartości, nawet jeśli nie są one jednoznacznie błędne. W tym przypadku SI może wykrywać subtelne różnice w nazewnictwie produktów lub wspierać tworzenie bardziej spójnych opisów, co

przyczynia się do lepszej jakości danych. Sztuczna inteligencja może zostać także użyta do analizy zależności między różnymi atrybutami produktu oraz do wykrywania anomalii w danych. Modele predykcyjne, jak te bazujące na transformerach, mogą identyfikować potencjalnie nieprawidłowe wartości (np. nieadekwatne jednostki netto w odniesieniu do kategorii produktów) i proponować alternatywy.

Taka struktura modelu walidacji zapewnia elastyczne połączenie metod regułowych i zaawansowanych algorytmów sztucznej inteligencji, co umożliwi automatyczną detekcję błędów oraz poprawę jakości danych w elektronicznych katalogach, z minimalnym udziałem człowieka.

#### **Pytanie 4. Jakie algorytmy walidacyjne mogą pozytywnie wpłynąć na doskonalenie jakości danych produktów w elektronicznych katalogach?**

W badaniach zaproponowano i zweryfikowano 45 walidatorów, które wykorzystują zarówno reguły walidacyjne do walidacji syntaktycznej, logiki warunkowej, algorytmy klasyfikacyjne z analizą semantyczną, algorytmy dopasowania i integracji danych, jak i multimodalne algorytmy generatywne do analizy tekstu i obrazu, jednak szczególną uwagę poświęcono możliwościom wykorzystania sztucznej inteligencji do ich budowy, co miało zasadniczy wpływ na ostateczne wyniki. Zaproponowane algorytmy zostały dobrane i zaadaptowane do specyficznych wymagań walidacji danych produktowych podmiotu badań na każdym z 3-ech poziomów walidacji.

### **8.3. Kierunki dalszych badań**

Opracowany model wielopoziomowej walidacji danych otwiera liczne możliwości rozwoju i stanowi punkt wyjścia do przyszłych badań nad doskonaleniem jakości danych w katalogach elektronicznych. Dalsze badania powinny koncentrować się na rozwijaniu nowych technologii w obszarze walidacji oraz zautomatyzowanych metod doskonalenia jakości danych zgodnie z koncepcją Jakości 4.0, co może przynieść wymierne korzyści zarówno dla przedsiębiorstw, jak i konsumentów.

Kluczowe kierunki dalszych badań mogą obejmować:

- A. Rozwój i uniwersalizację modelu walidacji - adaptację do innych obszarów zastosowań.

- B. Rozwój algorytmów walidacyjnych, szczególnie poziomu trzeciego, w zakresie sztucznej inteligencji.
- C. Dalsze doskonalenie zapytań (promptów) sterujących dla modeli LLM, VLM i LMM, z uwzględnieniem kontekstu kategorii produktów, co umożliwi zwiększenie precyzji i efektywności walidacji.
- D. Opracowanie mechanizmów umożliwiających dynamiczne dostosowywanie do nowych danych i wzorców bez potrzeby ręcznego przeuczania, tzw. Continuous Learning.

#### **8.4. Wdrożenie wyników prac**

W ramach realizowanego doktoratu wdrożeniowego podjęto szereg działań mających na celu wdrożenie opracowanych koncepcji i rozwiązań w środowisku biznesowym.

Wiedza, koncepcje i zdobyte umiejętności w trakcie realizacji doktoratu wdrożeniowego, obejmujące tematykę jakości danych, katalogów elektronicznych, standardów danych, walidatorów oraz sztuczną inteligencję, umożliwiła wsparcie wielu projektów badawczych realizowanych w Łukasiewicz – Poznańskim Instytucie Technologicznym. Instytut prowadzi zarówno badania naukowe, jak i projekty rynkowe z wieloma partnerami biznesowymi w obszarze transformacji cyfrowej, gdzie taka wiedza jest przydatna w rozwijaniu innowacyjnych rozwiązań, optymalizacji procesów oraz podnoszeniu jakości danych. Dzięki temu możliwe było efektywne wdrażanie nowych technologii, które przyczyniają się do zwiększenia konkurencyjności przedsiębiorstw oraz wspierają rozwój gospodarki.

Jednym z głównych partnerów biznesowych Instytutu jest Fundacja GS1 Polska, która zarządza danymi katalogu produktowego szeroko omawianego w niniejszej dysertacji. Dzięki współpracy możliwe było przetestowanie walidatorów na rzeczywistych danych obejmujących kilkadziesiąt milionów rekordów produktowych. W trakcie trwania programu doktoratu wdrożeniowego do omawianego katalogu wprowadzono zarówno modelową koncepcję wielopoziomowej walidacji, jak i wybraną część z przedstawionych w niniejszej pracy walidatorów. Planowane są kolejne wdrożenia, które pozwolą na dalszą poprawę jakości danych.

Niezależnie od działań na rzecz doskonalenia danych badanego katalogu, nawiązano współpracę z firmami działającymi w sektorze e-commerce oraz zarządzającymi dużymi

bazami danych produktowych. Analizowano możliwości zastosowania opracowanych walidatorów oraz rozwój kolejnych narzędzi w ramach wielowarstwowego modelu walidacji. Działania te miały na celu integrację narzędzi walidacyjnych z istniejącymi systemami zarządzania danymi, poprawę jakości danych produktowych oraz standaryzację procesów weryfikacji informacji. Ponadto nawiązano kontakt z przedsiębiorstwem technologicznym zainteresowanym wdrożeniem sztucznej inteligencji do własnego katalogu produktowego, z którym zaplanowano współpracę w ramach funduszy europejskich wspierających rozwój nowoczesnej gospodarki. Wdrożenie wyników prac planowane jest także w ramach ostatniej fazy przedwdrożeniowej projektu realizowanego na zlecenie instytucji publicznej. Projekt dotyczy monitorowania produktów dostępnych na rynkach europejskich. Wypracowane koncepcje doskonale sprawdzą się w kontroli i poprawie jakości zbieranych danych. Aby rozszerzyć zasięg i wpływ opracowanych rozwiązań, podjęto także działania zmierzające do umiędzynarodowienia badań. Przeprowadzono spotkania z przedstawicielami właścicieli zagranicznych rejestrów danych w celu omówienia możliwości adaptacji narzędzi walidacyjnych. Zaplanowano prace nad uniwersalizacją algorytmów, tak aby mogły być zastosowane w różnych krajach, uwzględniając specyfikę lokalnych rynków. Wymiana doświadczeń i najlepszych praktyk może przyczynić się do globalnego podniesienia standardów jakości danych produktowych.

## Bibliografia

1. Abedjan, Z., Golab, L., i Naumann, F. (2015). Profiling relational data: A survey. *The VLDB Journal — The International Journal on Very Large Data Bases*, 24(4), 557–581. <https://doi.org/10.1007/s00778-015-0389-y>
2. Abraham, J. (2014). Product information management. *Springer International Publishing*, 10, 978–3.
3. Allegro.pl. (2021). *Allegro.pl—Wystawianie i edycja oferty*. <https://allegro.pl/pomoc/dla-sprzedajacych/wystawianie-i-edycja-oferty>
4. Amazon.com. (2021). *Amazon seller central*. <https://sellercentral.amazon.com/>
5. Ao, J., i Chirkova, R. (2019). Effective and Efficient Data Cleaning for Entity Matching. *Proceedings of the Workshop on Human-In-the-Loop ...*, Query date: 2023-10-25 03:57:50. <https://doi.org/10.1145/3328519.3329127>
6. Appelbaum, D., Kogan, A., Vasarhelyi, M., i Yan, Z. (2017). Impact of business analytics and enterprise systems on managerial accounting. *International Journal of Accounting Information Systems*, 25, 29–44. <https://doi.org/10.1016/j.accinf.2017.03.003>
7. Arasu, A., Chaudhuri, S., Chen, Z., Ganjam, K., i ... (2011). Towards a domain independent platform for data cleaning. *Data Engineering ...*, Query date: 2023-10-25 03:57:50. <https://www.microsoft.com/en-us/research/publication/towards-a-domain-independent-platform-for-data-cleaning/>
8. Ashraf, A., Ilyas, S., ur Rehman, K. U., i Ahmad, S. (2020). *Algorithms for Data Cleaning in Knowledge Bases*.
9. Ballou, D. P., Madnick, S. E., i Wang, R. Y. (2004). Special Section: Assuring Information Quality. *J. Manag. Inf. Syst.*, 20, 9–11.
10. Ballou, D. P., i Pazer, H. L. (1985). Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science*, 31(2), 150–162. <https://doi.org/10.1287/mnsc.31.2.150>
11. Bates, S., Hastie, T., i Tibshirani, R. (2022). *Cross-validation: What does it estimate and how well does it do it?* (arXiv:2104.00673). arXiv. <http://arxiv.org/abs/2104.00673>
12. Batini, C., Cappiello, C., Francalanci, C., i Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3), 1–52. <https://doi.org/10.1145/1541880.1541883>
13. Battistello, L., Haug, A., Suzic, N., i Hvam, L. (2021). Implementation of product information management systems: Identifying the challenges of the scoping phase. *Computers in Industry*, 133, 103533. <https://doi.org/10.1016/j.compind.2021.103533>
14. Beskales, G., Ilyas, I. F., i Golab, L. (2010). Sampling the repairs of functional dependency violations under hard constraints. *Proceedings of the VLDB Endowment*, 3(1–2), 197–207. <https://doi.org/10.14778/1920841.1920870>
15. Beskales, G., Ilyas, I. F., Golab, L., i Galiullin, A. (2013). *On the relative trust between inconsistent data and inaccurate constraints*. 541–552.
16. Birch-Jensen, A., Gremyr, I., i Halldórsson, Á. (2020). Digitally connected services: Improvements through customer-initiated feedback. *European Management Journal*, 38(5), 814–825. <https://doi.org/10.1016/j.emj.2020.03.008>
17. Blikle, A. J. (2009). *Doktryna Jakosci. Rzecz o Turkusowej Samoorganizacji*.
18. Bohannon, P., Fan, W., Flaster, M., i Rastogi, R. (2005). *A cost-based model and effective heuristic for repairing constraints by value modification*. 143–154.
19. Borek, A., Woodall, P., Oberhofer, M., i Parlikad, A. (2015). A Classification of Data Quality Assessment Methods. *researchgate.net*, Query date: 2023-10-25 03:57:50.



- [https://www.researchgate.net/profile/Ajith-Kumar-Parlikad/publication/259081866\\_A\\_Classification\\_of\\_Data\\_Quality\\_Assessment\\_Methods/links/54cfc2510cf298d656658ac4/A-Classification-of-Data-Quality-Assessment-Methods.pdf](https://www.researchgate.net/profile/Ajith-Kumar-Parlikad/publication/259081866_A_Classification_of_Data_Quality_Assessment_Methods/links/54cfc2510cf298d656658ac4/A-Classification-of-Data-Quality-Assessment-Methods.pdf)
20. Boyd, M. (2006). Product Information Management-Forcing the Second Wave of Data Quality; The only way to reliably understand and describe a product is to isolate and extract its attributes. *Information Management*, 16(10), 38.
  21. Brabham, D. C. (2008). Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence*, 14(1), 75–90. <https://doi.org/10.1177/1354856507084420>
  22. Bradji, L., i Boufaida, M. (2011). *A rule management system for knowledge based data cleaning*. Query date: 2023-10-25 03:57:50. [https://www.scirp.org/html/3-8701125\\_8238.htm](https://www.scirp.org/html/3-8701125_8238.htm)
  23. Brechan, I. (2006). The different effect of primary and secondary product attributes on customer satisfaction. *Journal of Economic Psychology*, 27(3), 441–458. <https://doi.org/10.1016/j.joep.2005.10.003>
  24. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. <https://doi.org/10.48550/arXiv.2005.14165>
  25. Brunner, U., i Stockinger, K. (2020). Entity matching with transformer architectures-a step forward in data integration. *23rd International Conference on ...*, Query date: 2023-10-25 03:57:50. <https://digitalcollection.zhaw.ch/handle/11475/19637>
  26. BuiltWith.com. (2021). *Distribution for websites using e-commerce technologies*. <https://trends.builtwith.com/shop>
  27. Caballero, I., i Piattini, M. (2003). CALDEA: A data quality model based on maturity levels. *Third International Conference on Quality Software, 2003. Proceedings.*, 380–387. <https://doi.org/10.1109/QSIC.2003.1319125>
  28. Cai, L., i Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data science journal*, Query date: 2023-10-25 03:57:50. <https://account.datascience.codata.org/index.php/up-j-dsj/article/view/dsj-2015-002>
  29. Cao, J. (2023). *E-Commerce Big Data Mining and Analytics* (1st ed. 2023.). Singapore : Springer Nature Singapore : Imprint: Springer, 2023. <https://doi.org/10.1007/978-981-99-3588-8>
  30. Cao, M., i Zhang, Q. (2011). Supply chain collaboration: Impact on collaborative advantage and firm performance. *Journal of operations management*, 29(3), 163–180. <https://doi.org/10.1016/j.jom.2010.12.008>
  31. Cappiello, C., Caro, A., Rodríguez, A., i Caballero, I. (2013). *An Approach To Design Business Processes Addressing Data Quality Issues*. European Conference on Information Systems. <https://www.semanticscholar.org/paper/An-Approach-To-Design-Business-Processes-Addressing-Cappiello-Caro/851b9f39797c6248d5dc33b080eae24faec49188>
  32. Ceneo.pl. (2021). *Instrukcja tworzenia pliku XML*. Ceneo.pl. <https://www.ceneo.pl/poradniki/Instrukcja-tworzenia-pliku-XML>
  33. Chalamalla, A., Ilyas, I. F., Ouzzani, M., i Papotti, P. (2014). *Descriptive and prescriptive data cleaning*. 445–456.
  34. Chapman, P. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*.
  35. Charantimath, P. M. (2011). *Total Quality Management*. Pearson Education India.
  36. Chen, M., Song, M., Han, J., i Haihong, E. (2012). Survey on data quality. *2012 World Congress on Information and Communication Technologies*, 1009–1013. <https://doi.org/10.1109/WICT.2012.6409222>

37. Cheung, W. M., i Schaefer, D. (2010). Product lifecycle management: State-of-the-art and future perspectives. *W Enterprise information systems for business integration in SMEs: Technological, organizational, and social dimensions* (s. 37–55). IGI Global.
38. Chu, X., Ilyas, I. F., Krishnan, S., i Wang, J. (2016). Data Cleaning: Overview and Emerging Challenges. *Proceedings of the 2016 International Conference on Management of Data*, 2201–2206. <https://doi.org/10.1145/2882903.2912574>
39. Chu, X., Ilyas, I. F., i Papotti, P. (2013a). Discovering denial constraints. *Proceedings of the VLDB Endowment*, 6(13), 1498–1509. <https://doi.org/10.14778/2536258.2536262>
40. Chu, X., Ilyas, I. F., i Papotti, P. (2013b). *Holistic data cleaning: Putting violations into context*. 458–469.
41. Chung, Y., Krishnan, S., i Kraska, T. (2016). A data quality metric (DQM): How to estimate the number of undetected errors in data sets. *arXiv preprint arXiv:1611.04878*, *Query date: 2023-10-25 03:57:50*. <https://doi.org/10.14778/3115404.3115414>
42. Cichy, C., i Rass, S. (2019). An overview of data quality frameworks. *IEEE Access*, 7, 24634–24648. <https://doi.org/10.1109/ACCESS.2019.2899751>
43. Cohen, W., Ravikumar, P., i Fienberg, S. (2003). A comparison of string DISTANCE metrics for name-matching tasks. *IWeb*, 2003.
44. Cong, G., Fan, W., Geerts, F., Jia, X., i Ma, S. (2007). *Improving data quality: Consistency and accuracy*. 315–326.
45. Corbett, L. M., i Cutler, D. J. (2000). Environmental management systems in the New Zealand plastics industry. *International Journal of Operations & Production Management*, 20(2), 204–224. <https://doi.org/10.1108/01443570010304260>
46. Corrales, D. C., Corrales, J. C., i Ledezma, A. (2018). How to Address the Data Quality Issues in Regression Models: A Guided Process for Data Cleaning. *Symmetry*, 10(4), Article 4. <https://doi.org/10.3390/sym10040099>
47. Craig, J. H. S., i Lemon, M. (2008). Perceptions and reality in quality and environmental management systems: A research survey in China and Poland. *The TQM Journal*, 20(3), 196–208. <https://doi.org/10.1108/17542730810867227>
48. Crosby, P. B. (1967). *Cutting the cost of quality: The defect prevention workbook for managers*. (No Title).
49. Czakon, W. (2011). *Podstawy metodologii badań w naukach o zarządzaniu*. Warszawa: Oficyna a Wolters Kluwer business.
50. Dani, M. N., Faruquie, T. A., Garg, R., Kothari, G., Mohania, M. K., Prasad, K. H., Subramaniam, L. V., i Swamy, V. N. (2010). *A knowledge acquisition method for improving data quality in services engagements*. 346–353.
51. Dasu, T., i Johnson, T. (1999). *Hunting of the Snark: Finding Data Glitches using Data Mining Methods*. IQ. <https://www.semanticscholar.org/paper/Hunting-of-the-Snark%3A-Finding-Data-Glitches-using-Dasu-Johnson/e19dc0b172574a3ae034e3e33af31274f617e518>
52. David, M., i Rowe, F. (2016). What does PLMS (product lifecycle management systems) manage: Data or documents? Complementarity and contingency for SMEs. *Computers in Industry*, 75, 140–150. <https://doi.org/10.1016/j.compind.2015.05.005>
53. de Almeida, W. G., de Sousa, R. T., de Deus, F. E., Daniel Amvame Nze, G., i de Mendonça, F. L. L. (2013). Taxonomy of data quality problems in multidimensional Data Warehouse models. *2013 8th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–7. <https://ieeexplore.ieee.org/abstract/document/6615784>
54. Deming, W. E. (2018). *Out of the Crisis, reissue*. MIT press.

55. Devlin, J., Chang, M.-W., Lee, K., i Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. <http://arxiv.org/abs/1810.04805>
56. Di, W., Sundaresan, N., Piramuthu, R., i Bhardwaj, A. (2014). Is a picture really worth a thousand words? - On the role of images in e-commerce. *Proceedings of the 7th ACM international conference on Web search and data mining*, 633–642. <https://doi.org/10.1145/2556195.2556226>
57. Do, N. (2018). Identifying experts for engineering changes using product data analytics. *Computers in Industry*, 95, 81–92. <https://doi.org/10.1016/j.compind.2017.12.004>
58. Drd'gan, M., i Metz, D. (2017). Analysis and improvement of data quality. Improvement of data quality for business purposes. *Review of Economic Studies and Research Virgil ...*, Query date: 2023-10-25 03:57:50. <https://doi.org/10.24193/RVM.2017.10.09>
59. Dziuba, D. T. (2012). Crowdsourcing internetowy i jego ekonomiczno-społeczne implikacje. Segment rynku odpłatnego crowdsourcingu. *Informatyka Ekonomiczna*, 26, 64–80.
60. eBay Inc. (2018). *Catalog best practices guide*. <https://developer.ebay.com/devzone/merchant-products/catalog-best-practices/content/index.html>
61. Eckerson, W. W. (2002a). Achieving business success through a commitment to high quality data. *TDWI Report Series*, 3.
62. Eckerson, W. W. (2002b). Data quality and the bottom line. *TDWI Report, The Data Warehouse Institute*, 1–32.
63. English, L. (1998). The high costs of low quality data. *DM REVIEW*, 8, 38–39.
64. English, L. P. (1999, marzec 11). *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. <https://www.semanticscholar.org/paper/Improving-Data-Warehouse-and-Business-Information-English/3bd1e135c28a5e61edff068d2304ac93920a69a9>
65. Feltham, G. A. (1968). The value of information. *The accounting review*, 684–696.
66. Fürber, C., i Hepp, M. (2011). Towards a vocabulary for data quality management in semantic web architectures. *Proceedings of the 1st International Workshop on Linked Web Data Management*, 1–8. <https://doi.org/10.1145/1966901.1966903>
67. Gandomi, A., i Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144. <https://doi.org/10.1016/j.ijinfomgt.2014.10.007>
68. Gargova, S., Temnikova, I., Dzhumerov, I., i Nikolaeva, H. (2022). Evaluation of Off-the-Shelf Language Identification Tools on Bulgarian Social Media Posts. *Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria (CLIB 2022)*, 152–161. <https://aclanthology.org/2022.clib-1.18>
69. Gartner. (2021). *Magic Quadrant for Data Quality Solutions*. [resource.itbusinessplus.com. https://resource.itbusinessplus.com/whitepapers/12252-IBM-CPL-Q1-AI-US-1.pdf](https://resource.itbusinessplus.com/whitepapers/12252-IBM-CPL-Q1-AI-US-1.pdf)
70. Ge, M., i Helfert, M. (2007). *A Review of Information Quality Research—Develop a Research Agenda*. MIT International Conference on Information Quality. <https://www.semanticscholar.org/paper/A-Review-of-Information-Quality-Research-Develop-a-Ge-Helfert/eeb8fe6942820bf85c014cefc5fe065658561b96>
71. Glowalla, P., Balazy, P., Basten, D., i Sunyaev, A. (2014). Process-Driven Data Quality Management – An Application of the Combined Conceptual Life Cycle Model. *2014 47th Hawaii International Conference on System Sciences*, 4700–4709. <https://doi.org/10.1109/HICSS.2014.575>

72. Golab, L., Karloff, H., Korn, F., Srivastava, D., i Yu, B. (2008). On generating near-optimal tableaux for conditional functional dependencies. *Proceedings of the VLDB Endowment*, 1(1), 376–390. <https://doi.org/10.14778/1453856.1453900>
73. Gölzer, P., i Fritzsche, A. (2017). Data-driven operations management: Organisational implications of the digital transformation in industrial practice. *Production Planning & Control*, 28(16), 1332–1343. <https://doi.org/10.1080/09537287.2017.1375148>
74. Google.com. (2021). *Specyfikacja danych produktów—Google Merchant Center—Pomoc*. <https://support.google.com/merchants/answer/7052112>
75. Grönroos, C. (1984). A Service Quality Model and its Marketing Implications. *European Journal of Marketing*, 18(4), 36–44. <https://doi.org/10.1108/EUM0000000004784>
76. GS1. (2015). *GS1 Web Vocabulary Standard* (Wersja 1.6.1). [https://www.gs1.org/docs/gsl-smartsearch/GS1\\_Vocabulary\\_Standard.pdf](https://www.gs1.org/docs/gsl-smartsearch/GS1_Vocabulary_Standard.pdf)
77. GS1. (2021). *GS1 General Specification* (Wersja Release 21.0.1, Ratified, Jan 2021).
78. GS1. (2022a). *Global Data Dictionary (GDD)*. <https://apps.gs1.org/>
79. GS1. (2022b). *GS1 Product Image Specification Standard* (Wersja 4.1). GS1. [https://www.gs1.org/sites/default/files/docs/gdsn/Product\\_Image\\_Specification.pdf](https://www.gs1.org/sites/default/files/docs/gdsn/Product_Image_Specification.pdf)
80. GS1. (2022c). *GS1 Package and Product Measurement Standard* (Wersja 3.1). [https://www.gs1.org/docs/gdsn/3.1/GS1\\_Package\\_Product\\_Measurement\\_Standard.pdf](https://www.gs1.org/docs/gdsn/3.1/GS1_Package_Product_Measurement_Standard.pdf)
81. GS1. (2023a). *GS1 GDSN standards | GS1* (Wersja 3.1.25). <https://gs1.org/standards/gdsn>
82. GS1. (2023b, wrzesień 18). *GS1 The Global Language of Business Homepage*. <https://www.gs1.org/>
83. GS1. (2023c). *GS1 Global Data Model (GDM)* (Wersja 2.8). <https://www.gs1.org/standards/gsl-global-data-model>
84. GS1. (2024). *How GPC works*. <https://www.gs1.org/standards/gpc/how-gpc-works>
85. GS1.pl. (2023). *Verified by GS1 | GS1.PL*. Verified by GS1 - GS1 Polska. <https://gs1pl.org/verified-by-gs1/>
86. Guoling, L., i Qinyun, W. (2008). *Research on e-business model of distance education*. 400–403.
87. Hamrol, A. (2005). *Zarządzanie jakością z przykładami*. Wydawnictwo naukowe PWN.
88. Han, J., Kamber, M., i Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
89. Haug, A., i Stentoft Arlbjørn, J. (2011). Barriers to master data quality. *Journal of Enterprise Information Management*, 24(3), 288–303. <https://doi.org/10.1108/17410391111122862>
90. Haug, A., Stentoft Arlbjørn, J., i Pedersen, A. (2009). A classification model of ERP system data quality. *Industrial Management & Data Systems*, 109(8), 1053–1068. <https://doi.org/10.1108/02635570910991292>
91. Haug, A., Zachariassen, F., i Van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management*, 168–193. <https://doi.org/10.3926/jiem.2011.v4n2.p168-193>
92. He, K., Zhang, X., Ren, S., i Sun, J. (2015). *Deep Residual Learning for Image Recognition* (arXiv:1512.03385). arXiv. <https://doi.org/10.48550/arXiv.1512.03385>
93. Heinrich, B., i Klier, M. (2009). A novel data quality metric for timeliness considering supplemental data. *ECIS 2009 Proceedings*, 2651–2662. <https://aisel.aisnet.org/ecis2009/14>
94. Hernández, M. A., i Stolfo, S. J. (1998). Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. *Data Mining and Knowledge Discovery*, 2(1), 9–37. <https://doi.org/10.1023/A:1009761603038>
95. Hipp, J., Güntzer, U., i Grimmer, U. (2001). *Data Quality Mining—Making a Virtue of Necessity*. Workshop on Research Issues on Data Mining and Knowledge Discovery.

- <https://www.semanticscholar.org/paper/Data-Quality-Mining-Making-a-Virute-of-Necessity-Hipp-G%C3%BCntzer/2a81e1f1e4cdd42c47bfb9db27f7d35319b3f20d>
96. Hirth, M., Hoßfeld, T., i Tran-Gia, P. (2011). Anatomy of a Crowdsourcing Platform—Using the Example of Microworkers.com. *2011 Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, 322–329. <https://doi.org/10.1109/IMIS.2011.89>
  97. Hole, Y., Pawar, S., i Bhaskar, M. P. (2018). Service marketing and quality strategies. *Periodicals of Engineering and Natural Sciences (PEN)*, 6(1), 182–196. <https://doi.org/10.21533/pen.v6i1.291>
  98. Howard, P. (2011, maj 23). *Data Migration White Paper*. Bloor Research. <https://www.bloorresearch.com/research/data-migration-white-paper/>
  99. Hüner, K., Schierning, A., Otto, B., i Österle, H. (2011). Product data quality in supply chains: The case of Beiersdorf. *Electronic Markets*, 21(2), 141–154. <https://doi.org/10.1007/s12525-011-0059-x>
  100. Ilyas, I. F., i Chu, X. (2019). *Data Cleaning*. Morgan & Claypool.
  101. Informatica. (2019). *Product Information Management? Product MDM? Or Both?*
  102. International Organization for Standardization. (2008). *ISO/IEC 25012:2008(en) Software engineering—Software product Quality Requirements and Evaluation (SQuaRE)—Data quality model*. <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:25012:ed-1:v1:en>
  103. International Organization for Standardization. (2015). *ISO 9000:2015(en) Quality management systems—Fundamentals and vocabulary*. <https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-4:v1:en>
  104. Isaksson, R., i Garvare, R. (2003). Measuring sustainable development using process models. *Managerial Auditing Journal*, 18(8), 649–656. <https://doi.org/10.1108/02686900310495142>
  105. Jelonek, D. (2023). Environmental uncertainty and changes in digital innovation strategy. *Procedia Computer Science*, 225, 1468–1477.
  106. Joulin, A., Grave, E., Bojanowski, P., i Mikolov, T. (2016). *Bag of Tricks for Efficient Text Classification* (arXiv:1607.01759). arXiv. <https://doi.org/10.48550/arXiv.1607.01759>
  107. Juran, J. M. (2003). *Juran on leadership for quality*. Simon and Schuster.
  108. Karpischek, S., Michahelles, F., i Fleisch, E. (2011). The not so unique Global Trade Identification Number—Exploring inconsistencies in online product information sources. *RFID SysTech 2011 7th European Workshop on Smart Objects: Systems, Technologies and Applications*, 1–4. <https://ieeexplore.ieee.org/document/6138701>
  109. Karpischek, S., Michahelles, F., i Fleisch, E. (2012). The not so unique global trade identification number: Product master data quality in publicly available sources. *Proceedings of the 14th Annual ...*, Query date: 2023-10-25 03:57:50. <https://doi.org/10.1145/2346536.2346543>
  110. Karpischek, S., Michahelles, F., i Fleisch, E. (2014). Detecting incorrect product names in online sources for product master data. *Electronic Markets*, 24(2), 151–160. <https://doi.org/10.1007/s12525-013-0136-4>
  111. Kawa, A., i Pierański, B. (2021). Green logistics in E-commerce. *Logforum*, 17(2), 1. <https://doi.org/10.17270/J.LOG.2021.588>
  112. Kenyon, H., Cassela, J., Lambert, R., i Jordaan, W. (2001). *Global data management survey*.
  113. Kieraś, W., i Woliński, M. (2017). Morfeusz 2—analizator i generator fleksyjny dla języka polskiego. *Język Polski*, 75–83.
  114. Kim, Y., Lee, H., i Shim, J. (2021). Developing data-conscious deep learning models for product classification. *Applied Sciences*, Query date: 2023-10-25 03:57:50. <https://www.mdpi.com/2076-3417/11/12/5694>

115. Kindlarski, E. (1988). *Jakość wyrobów, Państwowe Wydawnictwo Naukowe*.
116. Kolman, R., Grudowski, P., Meller, A., i Preihs, J. (1996). Wybrane zagadnienia zarządzania jakością. *Wyższa Szkoła Administracji i Biznesu w Gdyni, Gdynia*, 59.
117. Krieger, R., i Schorr, C. (2019). A Reference Model for Product Data Profiling in Retail ERP Systems. W S. Hammoudi, C. Quix, i J. Bernardino (Red.), *PROCEEDINGS OF THE 8TH INTERNATIONAL CONFERENCE ON DATA SCIENCE, TECHNOLOGY AND APPLICATIONS (DATA)* (s. 317–324). SCITEPRESS. <https://doi.org/10.5220/0007953303170324>
118. Krima, S., Feeney, A., i Fougou, S. (2014). Dynamic customisation, validation and integration of product data models using semantic web tools. *International Journal of ...*, *Query date: 2023-10-25 03:57:50*. <https://doi.org/10.1504/IJPLM.2014.065463>
119. Kropsu-Vehkaperä, H., i Haapasalo, H. (2011). Defining Product Data Views for Different Stakeholders. *Journal of Computer Information Systems*, 52(2), 61–72.
120. Kruse, F., Hassan, A., Awick, J., Gómez, J., i Bui, T. (2020). A Qualitative Literature Review on Linkage Techniques for Data Integration. *HICSS*, *Query date: 2023-10-25 03:57:50*. <https://doi.org/10.24251/HICSS.2020.132>
121. Küpper, D., Knizek, C., Ryeson, D., i Noecker, J. (2019). Quality 4.0 takes more than technology. *Boston Consulting Group (BCG)*, 1–14.
122. Laranjeiro, N., Soydemir, S. N., i Bernardino, J. (2015). A Survey on Data Quality: Classifying Poor Data. *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)*, 179–188. <https://doi.org/10.1109/PRDC.2015.41>
123. Lecun, Y., Bottou, L., Bengio, Y., i Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. *Proceedings of the IEEE*. <https://doi.org/10.1109/5.726791>
124. Lederman, R., Shanks, G., i Gibbs, M. R. (2003). *Meeting privacy obligations: The implications for information systems development*.
125. Lee, M. L., Lu, H., Ling, T. W., i Ko, Y. T. (1999). Cleansing Data for Mining and Warehousing. W T. J. M. Bench-Capon, G. Soda, i A. M. Tjoa (Red.), *Database and Expert Systems Applications* (s. 751–760). Springer. [https://doi.org/10.1007/3-540-48309-8\\_70](https://doi.org/10.1007/3-540-48309-8_70)
126. Levitin, A. V., i Redman, T. C. (1998). Data as a resource: Properties, implications, and prescriptions. *MIT Sloan Management Review*.
127. Li, B., Wang, W., Sun, Y., Zhang, L., Ali, M., i ... (2020). GraphER: token-centric entity resolution with graph convolutional neural networks. *Proceedings of the AAAI ...*, *Query date: 2023-10-25 03:57:50*. <https://ojs.aaai.org/index.php/AAAI/article/view/6330>
128. Liong, V. E., Lu, J., Wang, G., Moulin, P., i Zhou, J. (2015). Deep hashing for compact binary codes learning. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2475–2483. <https://doi.org/10.1109/CVPR.2015.7298862>
129. Liu, A., Zhang, Y., Lu, H., Tsai, S.-B., Hsu, C.-F., i Lee, C.-H. (2019). An Innovative Model to Choose E-Commerce Suppliers. *IEEE Access*, 7, 53956–53976. <https://doi.org/10.1109/ACCESS.2019.2908393>
130. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., i Lee, Y. J. (2024). *Llava-next: Improved reasoning, ocr, and world knowledge*.
131. Liu, X., Cheng, G., i Wu, J. X. (2002). Analyzing outliers cautiously. *IEEE Transactions on Knowledge and Data Engineering*, 14(2), 432–437. *IEEE Transactions on Knowledge and Data Engineering*. <https://doi.org/10.1109/69.991726>
132. Liu, Z., Chen, Q., i Cai, L. (2018). Application of requirement-oriented data quality evaluation method. *2018 19th IEEE/ACIS International ...*, *Query date: 2023-10-25 03:57:50*. <https://doi.org/10.1109/SNPD.2018.8441103>

133. M87 Labs. (2024). *Vikhyatk/moondream2*. Hugging Face. <https://huggingface.co/vikhyatk/moondream2>
134. Madlberger, M. (2011). *Can data quality help overcome the penguin effect? The case of item master data pools*. 13. <https://aisel.aisnet.org/ecis2011/271/>
135. Magento.com. (2020). *Magento user guide*. <https://docs.magento.com/user-guide/catalog/product-create.html>
136. Maier, M. (2013). *Towards a big data reference architecture*. <https://www.semanticscholar.org/paper/Towards-a-big-data-reference-architecture-Maier/b7fc1205337145f314cab67f162238a11157139e>
137. Mantura, W. (2012). Wybrane zastosowania kwalitologii. *Zarządzanie i Finanse*, 1(3), 37–24.
138. Mao, X., Gong, B., Su, F., Xu, K., Xian, K., Liu, D., i ... (2019). Data quality management and measurement. *Signal and Information ...*, Query date: 2023-10-25 03:57:50. [https://doi.org/10.1007/978-981-13-1733-0\\_28](https://doi.org/10.1007/978-981-13-1733-0_28)
139. Marsh, R. (2005). Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management. *Journal of Database Marketing & Customer Strategy Management*, 12(2), 105–112. <https://doi.org/10.1057/palgrave.dbm.3240247>
140. Martin, J., Elg, M., i Gremyr, I. (2020). The Many Meanings of Quality: Towards a Definition in Support of Sustainable Operations. *Total Quality Management & Business Excellence*, 0(0), 1–14. <https://doi.org/10.1080/14783363.2020.1844564>
141. Michalski, R. (2024). The influence of product digital visual presentation on purchase willingness: Effects of roundedness axes and degree. *Multimedia Tools and Applications*, 83(1), 2173–2202. <https://doi.org/10.1007/s11042-023-15786-z>
142. Michalski, R., i Grobelny, J. (2016). *The Effects of Background Color, Shape and Dimensionality on Purchase Intentions in a Digital Product Presentation* (T. 9739, s. 479). [https://doi.org/10.1007/978-3-319-40238-3\\_45](https://doi.org/10.1007/978-3-319-40238-3_45)
143. Monge, A. (2000). Matching Algorithms within a Duplicate Detection System. *IEEE Data(base) Engineering Bulletin*, 23, 14–20.
144. Mroczo, F. (2012). *Zarządzanie jakością*. Prace Naukowe Wałbrzyskiej Wyższej Szkoły Zarządzania i Przedsiębiorczości. [https://wwszip.pl/wp-content/uploads/2017/07/25\\_Zarzadzanie\\_jakoscia\\_Mroczo\\_2012\\_wyd\\_spec.pdf](https://wwszip.pl/wp-content/uploads/2017/07/25_Zarzadzanie_jakoscia_Mroczo_2012_wyd_spec.pdf)
145. Nakatani, K., Chuang, T.-T., i Zhou, D. (2006). Data Synchronization Technology: Standards, Business Values and Implementation Challenges. *Communication of Association for Information Systems*, 17. <https://doi.org/10.17705/1CAIS.01744>
146. Naumann, F. (2014). Data profiling revisited. *Rekord ACM SIGMOD*, 42(4), 40–49. <https://doi.org/10.1145/2590989.2590995>
147. Niemir, M., i Mrugalska, B. (2021). Basic Product Data in E-Commerce: Specifications and Problems of Data Exchange. *European Research Studies Journal*, XXIV(Special Issue 5), 317–329. <https://doi.org/10.35808/ersj/2735>
148. Niemir, M., i Mrugalska, B. (2022a). Identifying the cognitive gap in the causes of product name ambiguity in e-commerce. *Logforum*, 18(3), 9. <https://doi.org/10.17270/J.LOG.2022.738>
149. Niemir, M., i Mrugalska, B. (2022b). Product Data Quality in e-Commerce: Key Success Factors and Challenges. *Production Management and Process Control*. 13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022). <https://doi.org/10.54941/ahfe1001626>
150. Niemir, M., i Mrugalska, B. (2023a). Data Science Challenges of Automated Quality Verification Process in Product Data Catalogues. *Materials Research Proceedings*, 34, 390–399. <https://doi.org/10.21741/9781644902691-45>

151. Niemir, M., i Mrugalska, B. (2023b). Monitoring and Improvement of Data Quality in Product Catalogs Using Defined Normalizers and Validation Patterns. W B. Mrugalska, T. Ahram, i W. Karwowski (Red.), *Human Factors in Engineering: Manufacturing Systems, Automation, and Interactions* (s. 173–187). CRC Press. <https://doi.org/10.1201/9781003383444>
152. Niemir, M., Muszyński, K., i Skwarek, S. (2022). Searching For Ai Solutions To Improve The Quality Of Master Data Affecting Consumer Safety. *Business Logistics in Modern Management*, 121–140. [http://blmm-conference.com/wp-content/uploads/BLMM2022\\_Conference\\_Proceedings.pdf](http://blmm-conference.com/wp-content/uploads/BLMM2022_Conference_Proceedings.pdf)
153. Northcutt, C. G., Jiang, L., i Chuang, I. L. (2022). *Confident Learning: Estimating Uncertainty in Dataset Labels* (arXiv:1911.00068). arXiv. <http://arxiv.org/abs/1911.00068>
154. Oliveira, P., Rodrigues, F., i Rangel Henriques, P. (2005, styczeń 1). A Formal Definition of Data Quality Problems. *Proceedings of the 2005 International Conference on Information Quality (MIT IQ Conference)*.
155. Onyeabor, G., i Ta'a, A. (2019). A model for addressing quality issues in big data. *Recent Trends in Data Science and Soft Computing ...*, Query date: 2023-10-25 03:57:50. [https://doi.org/10.1007/978-3-319-99007-1\\_7](https://doi.org/10.1007/978-3-319-99007-1_7)
156. Osmólski, W., i Muszyński, K. (2020). Monitoring Of Goods-Documentation Flows In Modern Logistic Supply Chain, Based On Blockchain Technology. *Business Logistics in Modern Management*.
157. Otto, B. (2015). Quality and value of the data resource in large enterprises. *Information Systems Management*, Query date: 2023-10-25 03:57:50. <https://doi.org/10.1080/10580530.2015.1044344>
158. Oyrzanowski, B. (1984). *Mała encyklopedia jakości*. Polskie Towarzystwo Ekonomiczne.
159. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
160. Pantziarka, P. (2005). *Machine Learning And Data Validation*. University of Surrey.
161. Parasuraman, A., Zeithaml, V. A., i Berry, L. L. (1985). A Conceptual Model of Service Quality and Its Implications for Future Research. *Journal of Marketing*, 49(4), 41–50. <https://doi.org/10.1177/002224298504900403>
162. Parviainen, P., Tihinen, M., Kääriäinen, J., i Teppola, S. (2017). Tackling the digitalization challenge: How to benefit from digitalization in practice. *International Journal of Information Systems and Project Management*, 5(1), Article 1. <https://doi.org/10.12821/ijispm050104>
163. Peralta, V. (2006). *Data quality evaluation in data integration systems* [Université de Versailles-Saint Quentin en Yvelines; Université de la ...]. <https://theses.hal.science/tel-00325139>
164. Ponsignon, F., Kleinhans, S., i Bressolles, G. (2019). The contribution of quality management to an organisation's digital transformation: A qualitative study. *Total Quality Management & Business Excellence*, 30(sup1), S17–S34. <https://doi.org/10.1080/14783363.2019.1665770>
165. Popa, V., i Duica, M. (2011). *Supply Chain Information Alignment in the Consumer Goods and Retail Industry: Global Standards and Best Practices*. 14(1), 16.
166. Prasad, K., Faruque, T., Joshi, S., i ... (2011). Data cleansing techniques for large enterprise datasets. *2011 Annual SRII ...*, Query date: 2023-10-25 03:57:50. <https://doi.org/10.1109/SRII.2011.26>
167. Project Management Institute. (2021). *A Guide to the Project Management Body of Knowledge: PMBOK (®) Guide*.



168. Prpic, J., P, i Shukla. (2017). *Crowd Science: Measurements, Models, and Methods* (arXiv:1702.04221). arXiv. <https://doi.org/10.48550/arXiv.1702.04221>
169. Qalati, S. A., Vela, E. G., Li, W., Dakhan, S. A., Hong Thuy, T. T., i Merani, S. H. (2021). Effects of perceived service quality, website quality, and reputation on purchase intention: The mediating and moderating roles of trust and perceived risk in online shopping. *Cogent Business & Management*, 8(1), 1869363. <https://doi.org/10.1080/23311975.2020.1869363>
170. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., i Jagersand, M. (2020). U2-Net: Going Deeper with Nested U-Structure for Salient Object Detection. *Pattern Recognition*, 106, 107404. <https://doi.org/10.1016/j.patcog.2020.107404>
171. Rahm, E., i Do, H. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Data Eng. Bull.*, 23, 3–13.
172. Raman, V., i Hellerstein, J. M. (2001). Potter’s Wheel: An Interactive Data Cleaning System. *Proceedings of the 27th International Conference on Very Large Data Bases*, 381–390.
173. Redman, T. (2004). Data: An unfolding quality disaster. *Dm Review*, 14(8), 21–23.
174. Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2), 79–82. <https://doi.org/10.1145/269012.269025>
175. Redyuk, S., Kaoudi, Z., Markl, V., i Schelter, S. (2021). Automating Data Quality Validation for Dynamic Data Ingestion. *EDBT, Query date: 2023-10-25 03:57:50*. <https://sergred.github.io/files/edbt.reds.pdf>
176. Reeves, C., i Bednar, D. (1994). Defining quality: Alternatives and implications. *Academy of management Review*, 419–445. <https://doi.org/10.2307/258934>
177. Richter, G., Raban, D. R., i Rafaeli, S. (2015). Studying Gamification: The Effect of Rewards and Incentives on Motivation. W T. Reiners i L. C. Wood (Red.), *Gamification in Education and Business* (s. 21–46). Springer International Publishing. [https://doi.org/10.1007/978-3-319-10208-5\\_2](https://doi.org/10.1007/978-3-319-10208-5_2)
178. Roberts, A., Raffel, C., Lee, K., Matena, M., Shazeer, N., Liu, P. J., Narang, S., Li, W., i Zhou, Y. (2019). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Google.
179. Russom, P. (2011). Big data analytics. *TDWI best practices report, fourth quarter*, 19(4), 1–34.
180. Saaksvuori, A., i Immonen, A. (2008). *Product lifecycle management systems*. Springer.
181. Sage, A. P., i Rouse, W. B. (2011). *Handbook of Systems Engineering and Management*. John Wiley & Sons.
182. Saha, B., i Srivastava, D. (2014). Data quality: The other face of Big Data. *2014 IEEE 30th International Conference on Data Engineering*, 1294–1297. <https://doi.org/10.1109/ICDE.2014.6816764>
183. Salem, A., i Boufares, F. (2019). *IDQMS: An Intelligent Data Quality Management System Tool*. 16th International Conference on Applied Computing 2019. [https://scholar.archive.org/work/6o47tf7jcrdexjjrf4jv6rabwi/access/wayback/http://www.iadisportal.org/components/com\\_booklibrary/ebooks/201912L001.pdf](https://scholar.archive.org/work/6o47tf7jcrdexjjrf4jv6rabwi/access/wayback/http://www.iadisportal.org/components/com_booklibrary/ebooks/201912L001.pdf)
184. Schäffer, T., i Stelzer, D. (2017, grudzień 2). Assessing tools for coordinating quality of master data in inter-organizational product information sharing. *Wirtschaftsinformatik 2017 proceedings*. 13th International Conference on Wirtschaftsinformatik, St. Gallen, Switzerland.
185. Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F., i Grafberger, A. (2018). Automating large-scale data quality verification. *Proceedings of the VLDB Endowment*, 11(12), 1781–1794. <https://doi.org/10.14778/3229863.3229867>
186. Schema.org Community Group. (2015). *Product—Schema.org Type*. <https://schema.org/Product>

187. Schemm, J., i Legner, C. (2008). Toward the Inter-organizational Product Information Supply Chain – Evidence from the Retail and Consumer Goods Industries. *Journal of the Association for Information Systems*, 9(4), 119–150. <https://doi.org/10.17705/1jais.00156>
188. Schemm, J., Legner, C., i Otto, B. (2007). Global Data Synchronization—Current Status and Future Trends. W [Http://www.alexandria.unisg.ch/Publikationen/67245](http://www.alexandria.unisg.ch/Publikationen/67245).
189. Schmidts, O., Kraft, B., Winkens, M., i Zündorf, A. (2020). Catalog Integration of Low-quality Product Data by Attribute Label Ranking: *Proceedings of the 9th International Conference on Data Science, Technology and Applications*, 90–101. <https://doi.org/10.5220/0009831000900101>
190. Schmidts, O., Kraft, B., Winkens, M., i Zündorf, A. (2021). Catalog Integration of Heterogeneous and Volatile Product Data. W S. Hammoudi, C. Quix, i J. Bernardino (Red.), *Data Management Technologies and Applications* (s. 134–153). Springer International Publishing.
191. Semenov, V., Ilyin, D., Morozov, S., i Tarlapan, O. (2019). Effective consistency management for large-scale product data. W *JOURNAL OF INDUSTRIAL INFORMATION INTEGRATION* (T. 13, s. 13–21). ELSEVIER. <https://doi.org/10.1016/j.jii.2018.11.006>
192. Shewhart, W. A. (1931). *Economic control of manufactured product*. van Nostrand.
193. Shopify.com. (2021). *Shopify help center*. <https://help.shopify.com/en/manual/products/details>
194. Sidi, F., Shariat Panahy, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H., i Mustapha, A. (2012). Data quality: A survey of data quality dimensions. *2012 International Conference on Information Retrieval & Knowledge Management*, 300–304. <https://doi.org/10.1109/InfRKM.2012.6204995>
195. Silvola, R., Tolonen, A., Harkonen, J., Haapasalo, H., i Mannisto, T. (2019). Defining one product data for a product. *International Journal of Business Information Systems*, 30(4), 489. <https://doi.org/10.1504/IJBIS.2019.099308>
196. SimilarWeb. (2021, marzec 1). *Top sites ranking for e-commerce and shopping in the world*. Similarweb. <https://www.similarweb.com/top-websites/e-commerce-and-shopping/>
197. Singh, J., i Vashishtha, S. (2015). *Data quality tools for datawarehouse models*.
198. Singh, S. K., i Dwivedi, D. R. K. (2020). Data mining: Dirty data and data cleaning. *Available at SSRN 3610772*.
199. Skrzypek, E. (2003). Koszty jakości jako narzędzie oceny efektywności organizacji. *Annales Universitatis Mariae Curie-Skłodowska. Sectio H. Oeconomia*, 37, 415–439.
200. Sony, M., Antony, J., i Douglas, J. A. (2020). Essential ingredients for the implementation of Quality 4.0: A narrative review of literature and future directions for research. *The TQM Journal*, 32(4), 779–793. <https://doi.org/10.1108/TQM-12-2019-0275>
201. Stubblebine, T. (2008). *Wyrażenia regularne. Leksykon kieszonkowy. Wydanie II*. Helion.
202. Sung, S. Y., Li, Z., i Sun, P. (2002). *A fast filtering scheme for large database cleansing*. 76–83.
203. Szymkowski, M., i Niemir, M. (2024). Convolutional Neural Networks and Vision Transformers in Product GS1 GPC Brick Code Recognition. W B. Sheng, L. Bi, J. Kim, N. Magnenat-Thalmann, i D. Thalmann (Red.), *Advances in Computer Graphics* (s. 440–450). Springer Nature Switzerland. [https://doi.org/10.1007/978-3-031-50069-5\\_36](https://doi.org/10.1007/978-3-031-50069-5_36)
204. Taguchi, G. (1986). *Introduction to quality engineering: Designing quality into products and processes*.
205. Taleb, I., Serhani, M. A., Bouhaddioui, C., i Dssouli, R. (2021). Big data quality framework: A holistic approach to continuous quality management. *Journal of Big Data*, 8(1), 76. <https://doi.org/10.1186/s40537-021-00468-0>

206. Taleb, I., Serhani, M., i Dssouli, R. (2018, lipiec 2). *Big Data Quality: A Survey*. <https://doi.org/10.1109/bigdatacongress.2018.00029>
207. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., i Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models* (arXiv:2302.13971). arXiv. <https://doi.org/10.48550/arXiv.2302.13971>
208. Unicode. (2023). *UAX #44: Unicode Character Database* (Wersja 15.1.0). [https://www.unicode.org/reports/tr44/#General\\_Category\\_Values](https://www.unicode.org/reports/tr44/#General_Category_Values)
209. Vasiliev, Y. (2020). *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press.
210. Wach, K., Ejdyś, J., Duong, C. D., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiwicz, J., i Ziemia, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, 11(2), 7–30. <https://doi.org/10.15678/eber.2023.110201>
211. Wand, Y., i Wang, R. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), 86–95. <https://doi.org/10.1145/240455.240479>
212. Wang, J., Kraska, T., Franklin, M. J., i Feng, J. (2012). Crowder: Crowdsourcing entity resolution. *arXiv preprint arXiv:1208.1927*.
213. Wang, R. Y., i Strong, D. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4), 5–33. <https://doi.org/10.1080/07421222.1996.11518099>
214. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., i Tang, X. (2018, wrzesień 1). *ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks*. arXiv.Org. <https://arxiv.org/abs/1809.00219v2>
215. Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., i Chen, W. (2024, czerwiec 3). *MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark*. arXiv.Org. <https://arxiv.org/abs/2406.01574v4>
216. Wang, Y., Xu, Y. E., Li, X., Dong, X. L., i Gao, J. (2020). Automatic Validation of Textual Attribute Values in E-commerce Catalog by Learning with Limited Labeled Data. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2533–2541. <https://doi.org/10.1145/3394486.3403303>
217. Watts, S., Shankaranarayanan, G., i Even, A. (2009). Data quality assessment in context: A cognitive perspective. *Decision support systems*, 48(1), 202–211. <https://doi.org/10.1016/j.dss.2009.07.012>
218. Wawak, S. (2022). Korzyści i bariery wdrażania jakości 4.0 w polskich przedsiębiorstwach – badania wstępne. *Problemy Jakości, R. 54, nr 5*. <https://doi.org/10.15199/46.2022.5.1>
219. Wawak, T. (1989). *Makroekonomiczne problemy jakości produktów przemysłowych w Polsce* (Numer 170). Nakł. Uniwersytetu Jagiellońskiego.
220. Wilda Kurnia Putri i Vera Pujani. (2019). The influence of system quality, information quality, e-service quality and perceived value on Shopee consumer loyalty in Padang City. *The International Technology Management Review*, 8(1), 10–15. <https://doi.org/10.2991/itmr.b.190417.002>
221. Winkler, W. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods*.
222. Witten, I. H., Frank, E., Hall, M. A., i Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

223. WooCommerce.com. (2021). *Adding and managing products—Woocommerce docs*. <https://docs.woocommerce.com/document/managing-products/>
224. Yakout, M., Elmagarmid, A. K., Neville, J., Ouzzani, M., i Ilyas, I. F. (2011). Guided data repair. *arXiv preprint arXiv:1103.3103*.
225. Yang, Y., Yuan, Y., i Li, B. (2018). Data Quality Evaluation: Methodology and Key Factors. ... , *SmartCom 2017, Shenzhen, China, December 10 ...*, Query date: 2023-10-25 03:57:50. [https://doi.org/10.1007/978-3-319-73830-7\\_22](https://doi.org/10.1007/978-3-319-73830-7_22)
226. Zahari, N., Hussin, W., Yussof, M., i ... (2015). Data quality issues in data migration. *Soft Computing in Data ...*, Query date: 2023-10-25 03:57:50. [https://doi.org/10.1007/978-981-287-936-3\\_4](https://doi.org/10.1007/978-981-287-936-3_4)
227. Zhang, Q. (2001). Quality dimensions, perspectives and practices: A mapping analysis. *International Journal of Quality & Reliability Management*, 18(7), 708–722. <https://doi.org/10.1108/EUM000000005777>
228. Zheng, H. (2014). *ANTECEDENTS OF SOLUTION QUALITY IN CROWDSOURCING: THE SPONSOR'S PERSPECTIVE*. 15(3).
229. Zhou, C. H., Chen, B., Gao, Y., Zhang, C., i Guo, Z. J. (2011). *A technique of filtering dirty data based on temporal-spatial correlation in wireless sensor network*. 511–516.
230. Zmud, R. (1978). An empirical investigation of the dimensionality of the concept of information. *Decision sciences*, 187–195. <https://doi.org/10.1111/j.1540-5915.1978.tb01378.x>
231. Zonnenshain, A., i Kenett, R. S. (2020). Quality 4.0—The challenging future of quality engineering. *Quality Engineering*, 32(4), 614–626. <https://doi.org/10.1080/08982112.2019.1706744>
232. Zymonik, Z. (2003). *Koszty jakości w zarządzaniu przedsiębiorstwem*. Oficyna Wydawnicza Politechniki Wrocławskiej Wrocław.
233. Żuchowski, W. (2022). The smart warehouse trend: Actual level of technology availability. *Logforum*, 18(2), 7. <https://doi.org/10.17270/J.LOG.2022.702>

## Spis rysunków

Rysunek 1. Wizualizacja słów kluczowych i powiązań publikacji .....	13
Rysunek 2. Proces selekcji publikacji zgodnie z metodą PRISMA .....	15
Rysunek 3. Schemat realizacji pracy .....	19
Rysunek 4. Oś czasu kluczowych zmian w definicji jakości .....	21
Rysunek 5. Wymiary jakości danych – podział na 4 kategorie.....	28
Rysunek 6. Diagram wymiarów jakości.....	30
Rysunek 7. Podstawowe atrybuty produktu na etykiecie produktu.....	45
Rysunek 8. Struktura sieci GDSN .....	46
Rysunek 9. Struktura numeru GTIN-13 .....	90
Rysunek 10. Zgodność nazwy produktu .....	116
Rysunek 11. Przykładowe zdjęcia produktów dla wyszukiwanej frazy .....	117
Rysunek 12. Przykłady nazw kolorów z przypisanymi numerami wzornika Pantone.....	117
Rysunek 13. Procentowy rozkład wykrytych problemów .....	133
Rysunek 14. Klasyfikacja wykrytych problemów.....	134
Rysunek 15. Diagram podziału zidentyfikowanych problemów z jakością danych produktowych ...	140
Rysunek 16. Struktura wielopoziomowego modelu walidacji danych produktowych.....	145
Rysunek 17. Schemat blokowy procesu ewaluacji walidatorów .....	151
Rysunek 18. Procentowy udział dużych liter w nazwie produktu.....	164
Rysunek 19. Liczba produktów w stosunku do długości nazw w bazie danych .....	166
Rysunek 20. Poglądowy przebieg walidacji języka dla treści nazwy produktu .....	172
Rysunek 21. Liczba pikseli mniejszego z boków zdjęcia produktu .....	178
Rysunek 22. Liczba megapikseli zdjęć produktów .....	179
Rysunek 23. Porównanie algorytmów do wykrywania identycznych i podobnych zdjęć .....	181
Rysunek 24. Przykładowe podobne zdjęcia z tym samym 64-bitowym odciskiem palca .....	182
Rysunek 25. Schemat blokowy procesu walidacji kolorystyki tła. ....	186
Rysunek 26. Schemat blokowy procesu walidacji powierzchni tła.....	186
Rysunek 27. Liczba zdjęć produktów w funkcji poziomu jakości tła zdjęcia.....	188
Rysunek 28. Liczba zdjęć w funkcji procentu powierzchni tła .....	190
Rysunek 29. Przykłady zdjęć ze zbyt dużą powierzchnią tła .....	190
Rysunek 30. Poglądowy schemat blokowy przebiegu procesu walidacji GPC.....	195
Rysunek 31. Pozycja podpowiedzi walidatora w funkcji prawidłowego numeru Brick GPC .....	199
Rysunek 32. Schemat blokowy walidatora jednostki zawartości netto .....	204
Rysunek 33: Wykres liczby produktów dla zawartości netto w kategorii "Zioła i przyprawy".....	207
Rysunek 34: Wykres liczby produktów dla zawartości netto w kategorii "Wkręty" .....	208
Rysunek 35: Wykres liczby produktów dla zawartości netto w kategorii "Kolczyki".....	209
Rysunek 36: Wykres liczby produktów dla zawartości netto w kategorii "Ciasta słodkie" .....	210
Rysunek 37. Wpływ walidatorów na bazę produktów w podziale na poziomy walidacji .....	229

## Spis tabel

Tabela 1. Wyniki wyszukiwania poszczególnych haseł w bazach (stan na dzień 2022-01-02).....	13
Tabela 2. Wymiary wraz z tłumaczeniem i omówieniem ich znaczenia.....	29
Tabela 3. Omówienie diagramu wymiarów .....	31
Tabela 4. Zadania wykonywane poprzez crowdsourcing.....	39
Tabela 5. Źródła problemów z jakością danych.....	49
Tabela 6. Problemy z jakością informacji w danych produktowych.....	51
Tabela 7. Rodzaje algorytmów możliwe do wykorzystania do walidacji danych.....	64
Tabela 8. Przegląd rozwiązań teoretycznych .....	72
Tabela 9. Przegląd rozwiązań praktycznych .....	82
Tabela 10. Przedstawiciele rynku, platformy i powszechnie używana narzędzia w e-commerce .....	88
Tabela 11. Porównanie identyfikatorów produktów .....	92
Tabela 12. Porównanie różnych atrybutów "nazw produktów" w GDD.....	94
Tabela 13. Porównanie nazw produktów .....	95
Tabela 14. Porównanie interpretacji nazw marek .....	97
Tabela 15. Porównanie technicznych aspektów atrybutu "Zdjęcie produktu" .....	99
Tabela 16. Porównanie jakościowych aspektów atrybutu "Zdjęcie Produktu" .....	100
Tabela 17. Porównanie atrybutu zawartości netto produktu .....	103
Tabela 18. Porównanie klasyfikatorów produktu.....	104
Tabela 19. Różnice w nazwach produktów.....	109
Tabela 20. Różne nazwy przykładowego produktu.....	112
Tabela 21. Przykład różnych nazw produktów na rynku brytyjskim .....	113
Tabela 22. Przykład różnych nazw produktów na rynku niemieckim .....	113
Tabela 23. Problemy z jakością danych zgłoszone przez ankietowanych.....	122
Tabela 24. Zestawienie znalezionych problemów jakościowych w katalogu .....	124
Tabela 25. Zidentyfikowane zależności między atrybutami danych.....	133
Tabela 26. Ocena możliwości utworzenia algorytmów w celu eliminacji wykrytych problemów .....	136
Tabela 27. Zaimplementowane normalizatory danych.....	154
Tabela 28. Lista walidatorów drugiego poziomu .....	157
Tabela 29. Lista walidatorów trzeciego poziomu.....	162
Tabela 30. Przykłady proponowanych nazw produktów.....	165
Tabela 31. Długie nazwy produktów.....	166
Tabela 32. Przykłady nazw produktów rozpoczynających się od cudzysłowu .....	168
Tabela 33. Weryfikacja ekspercka algorytmu proponowania lepszej marki .....	170
Tabela 34. Wpływ walidatora proponowania lepszej marki na całą bazę danych .....	170
Tabela 35. Wyniki działania algorytmu do wykrywania języka .....	173
Tabela 36. Przykłady specyficznych nazw produktów .....	174
Tabela 37. Weryfikacja ekspercka algorytmu walidacji języka .....	176
Tabela 38. Wpływ walidatora języka na całą zawartość bazy danych.....	176
Tabela 39. Przykłady wyodrębnienia tła z wykorzystaniem różnych metod .....	184
Tabela 40. Przykłady zdjęć niespełniających kryteriów jakościowych tła.....	188
Tabela 41. Wybór modelu do predykcji klasyfikacji .....	191
Tabela 42. Przykładowe zapytanie do modelu w celu uzyskania rekomendacji dotyczącej właściwej kategorii.....	194
Tabela 43. Analiza ekspercka poprawności wprowadzenia klasyfikacji produktów przez użytkownika .....	196
Tabela 44. Warianty reguły poprawności klasyfikacji kategorii GPC .....	197

Tabela 45. Wpływ algorytmu walidacji GPC na całą bazę danych.....	201
Tabela 46: Zestawienie metryk dla modelu oceny jednostki netto .....	203
Tabela 47. Liczba poprawnych i niepoprawnych danych w zbiorze ewaluacyjnym walidatora jednostki netto .....	205
Tabela 48. Analiza progu zadziałania walidatora jednostki zawartości netto .....	205
Tabela 49. Wpływ algorytmu walidacji jednostki zawartości netto na całą bazę danych.....	207
Tabela 50. Metryki predykcji zawartości netto .....	212
Tabela 51. Wpływ walidatora nazw powszechnych w nazwach produktu na całą bazę danych .....	214
Tabela 52. Przykład konwersacji służącej do walidacji i oceny poprawności nazw produktu .....	215
Tabela 53. Przykłady promptów walidatora oceny poprawności nazwy produktu .....	217
Tabela 54. Macierz pomyłek i metryki dla predykcji poprawności nazwy .....	218
Tabela 55. Wpływ walidatora jakości nazwy produktu w ujęciu semantycznym na całą bazę danych .....	219
Tabela 56. Przykład konwersacji służącej do walidacji jakości zdjęcia.....	221
Tabela 57. Porównanie dwóch modeli detekcji obiektów na obrazie .....	224
Tabela 58. Wpływ walidatora jakości zdjęcia na całą bazę danych.....	225
Tabela 59. Przykłady wykrytych problemów jakościowych zdjęć .....	225
Tabela 60. Zestawienie walidatorów i ich wpływu na polepszenie jakości danych.....	227
Tabela 61. Zestawienie walidatorów zdjęć i ich wpływu na polepszenie jakości danych obrazowych .....	230