Spis treści

[1 Analiza sentymentu 3](#_Toc160745890)

[1.1 Wprowadzenie do Analizy sentymentu 3](#_Toc160745891)

[1.2 Podstawy Przetwarzania Języka Naturalnego (NLP) 10](#_Toc160745892)

[1.3 Narzędzia i Technologie w Analizie Sentymentu 14](#_Toc160745893)

[1.4 Potencjalne Wyzwania i Ograniczenia 17](#_Toc160745894)

[2 Uczenie maszynowe w analizie sentymentu 18](#_Toc160745895)

[2.1 Uczenie nadzorowane 20](#_Toc160745896)

[2.1.1 Regresja Logistyczna 20](#_Toc160745897)

[2.1.2 Metoda k najbliższych sąsiadów 20](#_Toc160745898)

[2.1.3 Drzewo Decyzyjne 20](#_Toc160745899)

[2.2 Uczenie nienadzorowane 21](#_Toc160745900)

[2.2.1 Sztuczne sieci neuronowe 21](#_Toc160745901)

[2.2.2 Deep Learning 21](#_Toc160745902)

[2.2.3 Sieci nauronowe konwolucyjne (CNNs) 22](#_Toc160745903)

[2.2.4 Rekurencyjne sieci neuronowe (RNNs) 22](#_Toc160745904)

[2.3 Regularyzacja 22](#_Toc160745905)

[2.4 Optymalizacja 22](#_Toc160745906)

[3 Analiza sentymentu w mediach społecznościowych 23](#_Toc160745907)

[4 Zakończenie 23](#_Toc160745908)

[5 Literatura 23](#_Toc160745909)

[6 Spis rysunków 24](#_Toc160745910)

[7 Spis tabel 25](#_Toc160745911)

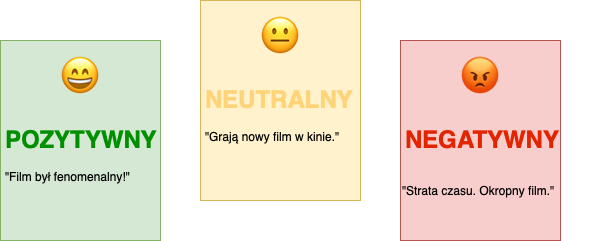
[8 Załączniki 25](#_Toc160745912)

[9 Streszczenie 25](#_Toc160745913)

# 1 Analiza sentymentu

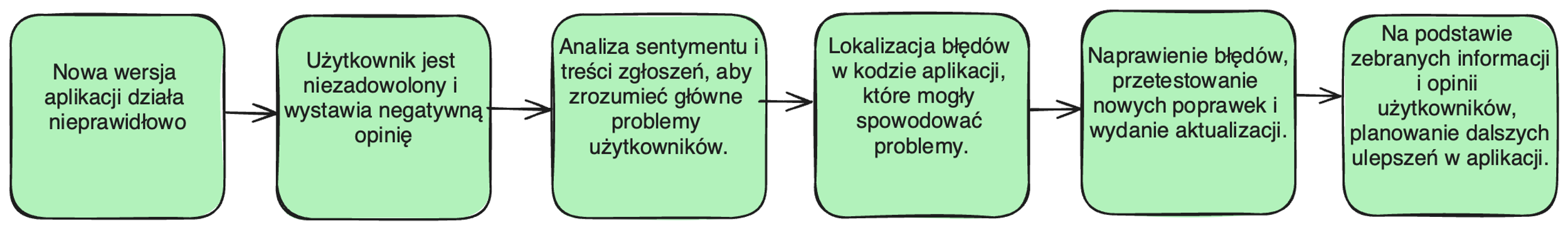
## 1.1 Wprowadzenie do Analizy sentymentu

Analiza sentymentu (SA, od ang. Sentiment Analysis), znana również jako przetwarzanie opinii, jest procesem automatycznego rozpoznawania i klasyfikacji emocji wyrażonych w tekście, stanowiącym kluczowy element w dynamicznie rozwijającym się świecie cyfrowym, gdzie dane tekstowe są nieustannie generowane (Liu, 2012). Głównym celem tego badania jest zrozumienie postawy autora treści, co jest niezmiernie istotne w kontekście szybko zmieniających się trendów i dynamiki społecznej. Technika ta pozwala nie tylko na identyfikowanie ogólnego tonu komunikatu - pozytywnego, negatywnego czy neutralnego (rysunek 1) - ale również na odkrywanie subtelniejszych niuansów emocjonalnych, takich jak złość, radość czy smutek (Pang & Lee, 2008). To umożliwia głębsze zrozumienie ludzkich reakcji i zachowań, co ma istotne znaczenie zarówno w sferze społecznej, jak i biznesowej.



Rysunek 1. Przykłady klasyfikacji sentymentu w opinii o filmach, Źródło: opracowanie własne

Zastosowanie analizy sentymentu może być doskonale zilustrowane na przykładzie prostego, indywidualnego stwierdzenia, które można napotkać w mediach społecznościowych lub recenzjach internetowych. Załóżmy, że użytkownik aplikacji publikuje następujący post: "Jestem rozczarowany nową aktualizacją aplikacji. Spowodowała wiele błędów i spowolniła działanie mojego telefonu". W tym prostym stwierdzeniu, analiza sentymentu pozwala na jednoznaczne sklasyfikowanie ogólnego tonu wypowiedzi jako negatywnego. Słowa takie jak "rozczarowany", "błędy" i "spowolniła" bezpośrednio wskazują na negatywne emocje użytkownika względem ostatniej aktualizacji aplikacji. Tego typu klarowne identyfikowanie znaczenia tekstu jest szczególnie przydatne dla firm technologicznych, które monitorują opinie użytkowników o ich produktach lub usługach. Umożliwia to szybkie reagowanie na problemy i pracę nad ich rozwiązaniem, co jest kluczowe dla utrzymania wysokiej jakości doświadczenia klienta, co ukazano na rysunku 2.



Rysunek 2. Schemat procesu reagowania na problemy z nową wersją aplikacji, Źródło: opracowanie własne

Rozwój technologii informatycznych i cyfryzacja społeczeństwa spowodowały, że analiza sentymentu zyskała na znaczeniu, stając się nieodzownym narzędziem w wielu dziedzinach, od marketingu po monitorowanie opinii publicznej (Bollen, Mao, & Zeng, 2011). W dzisiejszym świecie, gdzie opinie są swobodnie wyrażane na wielu platformach społecznościowych, blogach czy w recenzjach produktów, analiza sentymentu staje się nie tylko narzędziem badawczym, ale także strategicznym zasobem dla spółek i organizacji. Pozwala ona na wychwycenie i zrozumienie subtelnych zmian w percepcji marki, produkcie czy usłudze, co jest nieocenione dla decydentów biznesowych. Wykorzystanie tej technologii do monitorowania i reagowania na opinie klientów w czasie bliskim do rzeczywistego będzie stanowić kluczowy czynnik różnicujący w konkurencyjnym środowisku rynkowym. Zastosowanie analizy sentymentu w biznesie otwiera nowe możliwości w obrębie personalizacji oferty, zwiększenia satysfakcji klientów i optymalizacji strategii marketingowych. Firmy, które potrafią efektywnie interpretować emocje oraz przekonania wyrażane przez swoich konsumentów, zyskują cenne wskazówki, które mogą przekuć w działania prowadzące do wzrostu lojalności kupujących, poprawy reputacji marki, jak również w konsekwencji zwiększenia przychodów. W erze cyfrowej, gdzie dane stanowią nową walutę, analiza sentymentu w mediach społecznościowych, takich jak Twitter, ujawnia się jako fascynujące pole badań naukowych i potężne narzędzie biznesowe. Mikroblogowanie stało się popularnym narzędziem komunikacyjnym wśród użytkowników Internetu, gdzie miliony osób codziennie dzielą się opiniami na różne aspekty życia. Strony mikroblogowe są bogatym źródłem danych do analizy opinii i analizy sentymentu. (Pak & Paroubek, 2010). Portale społecznościowe stały się przez to nieocenionym miejscem pozyskiwania informacji do badania zachowań, reakcji, opinii użytkowników, aby móc im zaoferować doskonalsze, spersonalizowane usługi, które sprostają ich wymaganiom. Warto zauważyć, że z wnioskowania po treści komentarzy korzystają aktualnie nie tylko branża IT. Wiele marek przeniosło swoje sklepy do Internetu, ale także stara się tworzyć sieć społeczną wokół swoich produktów w portalach społecznościowych. Ten nowy kanał informacyjny, zapewnia nieustanny kontakt z nabywcą, aby móc informować go o nowościach, stworzyć więź, ale przede wszystkich umożliwić zbieranie informacji zwrotnej, która jest niezbędna do udoskonalenia usługi, oraz zrozumienia nowych potrzeb.

Warto jednak spojrzeć na analizę sentymentu szerzej - mimo swojej użyteczności napotyka ona na wyzwania związane z interpretacją języka naturalnego (rysunek 3), takie jak rozpoznawanie ironii, sarkazmu oraz różnorodności kontekstów kulturowych (Tsytsarau & Palpanas, 2012). Ponadto, dynamiczna natura języka, wraz z ciągłym rozwojem slangu i nowych form ekspresji w mediach społecznościowych, wymaga od systemów analizy sentymentu ciągłego dostosowywania się i ulepszania.

A screenshot of a computer

Description automatically generated

Rysunek 3. Złożoność analizy sentymentu w różnych typach zdań, Źródło: opracowanie własne

W niniejszej pracy dyplomowej dokonana zostanie dogłębna analiza metod i technik stosowanych w analizie sentymentu, z naciskiem na ich zastosowanie w różnych obszarach biznesowych oraz wyzwania związane z interpretacją danych tekstowych. Praca ta ma na celu nie tylko zbadanie istniejących narzędzi i metod, ale również identyfikację potencjalnych obszarów do dalszych badań i rozwoju w tej dynamicznie rozwijającej się dziedzinie.

Analiza sentymentu, pierwotnie zakorzeniona w lingwistyce i informatyce, zyskała znaczący moment do rozwoju dzięki dynamicznym postępom w dziedzinie nauki o danych. Jako integralna część uczenia maszynowego, przetwarzanie języka naturalnego (NLP) doświadczyło znaczących innowacji i usprawnień. Analiza sentymentu przeszła transformację, stając się bardziej zaawansowaną oraz efektywną technologią. Przez wiele lat była ona blokowana przez brak wystarczającej ilości danych, zbyt słabe rozwiązania sprzętowe, które nie były w stanie wykonywać potrzebnej ilości obliczeń lub trwałyby one nieakceptowanie długo. Początkowe eksploracje w tym obszarze koncentrowały się na elementarnych technikach, bazujących na leksykonach zawierających słowa oznaczone emocjami. Jednakże, to właśnie szybki rozwój w dziedzinie algorytmiki i adaptacyjnych modeli uczenia maszynowego, w tym uczenia głębokiego (Deep Learning), spowodował, że analiza sentymentu przekształciła się w wysoce precyzyjne narzędzie, zdolne do dokładnych i wielowymiarowych analiz emocjonalnych zawartości tekstowej. Zdolność do precyzyjnego rozpoznawania oraz interpretacji złożonych wzorców zachowań w wypowiedziach jest nie tylko triumfem technologicznym, ale również stanowi nieocenioną wartość przez samodzielną umiejętność wykrywania wzorów zachowań. W obecnym trendzie technologicznym, gdzie dane stanowią kluczowy zasób, zastosowanie zaawansowanych technik otwiera przed przedsiębiorstwami nowe horyzonty strategiczne, a przed naukowcami wiele obszarów eksploracji. Patrząc na tę kwestię od strony biznesowej, firmy wykorzystujące te innowacje mogą zyskać głęboki wgląd w percepcje, jak również nastroje klientów, co jest nieocenione w procesie optymalizacji strategii marketingowych, rozwoju produktów czy usług. Inteligentna analiza sentymentu, wspierana przez algorytmy uczenia maszynowego, pozwala na przetwarzanie ogromnych zbiorów danych z różnorodnych źródeł, takich jak media społecznościowe, fora internetowe, czy opinie klientów, dostarczając w ten sposób realnych i mierzalnych wskazówek dla decyzji biznesowych.

A screenshot of a computer

Description automatically generated

Rysunek 4. Porównanie głębokiego uczenia z uczeniem maszynowym, Źródło: opracowanie własne

Zatrudnienie jednostek zajmujących się analizą sentymentu, umożliwi nie tylko usprawnienie dotychczasowych rozwiązań, umożliwiając przedsiębiorstwom reagowanie na bieżące trendy i nastroje. Wdrożenie tego rozwiązania umożliwi także zaoszczędzenie funduszy w procesie budowania nowych rozwiązań, ponieważ będą już znane kierunki warte obrania, a także antycypowanie przyszłych zmian i dostosowywanie się do nich w sposób proaktywny. Dzięki temu, firmy nie tylko reagują na rynek, ale również kształtują go, będąc o krok przed konkurencją i w pełni wykorzystując potencjał ukryty w danych.

Technika analizy sentymentu, przetwarzająca złożone dane tekstowe, jest kluczowa w wykrywaniu subtelnych sygnałów, które mogą być wykorzystane do optymalizacji oferowanych produktów i usług. Rozważmy sektor e-commerce. Badania Pang i Lee (2008) pokazały, jak analiza treści recenzji produktów dostarcza informacji wykraczających poza ogólne tendencje, wskazując na specyficzne cechy produktów cenione przez użytkowników. To umożliwia firmom dostosowanie oferty do precyzyjnie zidentyfikowanych oczekiwań rynku. W branży usługowej, jak wykazało badanie Valdivii, Luzon i Herry (2017), analiza recenzji na platformie *TripAdvisor* ujawnia niespójności między ocenami użytkowników, a treścią ich recenzji, co ma bezpośredni wpływ na zarządzanie reputacją i doskonalenie usług. Zespół badał, czy oceny numeryczne (tzw. „gwiazdki”) są zgodne z sentymentem wyrażonym w opinii. Badacze użyli do tego celu różnych metod analizy sentymentu, w tym opartych na leksykonach i uczeniu maszynowym. Celem było zrozumienie, jak niespójności wpływają na percepcję użytkowników oraz zarządzanie reputacją firm. Praca ta podkreśla, jak ważne jest nie tylko analizowanie ocen numerycznych, ale także głębokie zrozumienie treści tekstowych, aby uzyskać pełniejszy obraz odczuć klientów. Ta perspektywa dodaje nowego wymiaru, pokazując, że analiza sentymentu w e-commerce i branży usługowej wymaga uwzględnienia zarówno liczbowej reprezentacji oceny, jak i subtelności myśli wyrażonej w piśmie. W badaniu stwierdzono, że zgodność między pisemnymi recenzjami użytkowników, a ich ocenami numerycznymi jest niska. Zauważono, że negatywne recenzje często zawierają znaczącą liczbę pozytywnych słów. W badaniu odniesiono się do sytuacji, gdy w negatywnie klasyfikowanych recenzjach pojawiają się wyrażenia, które trudno przypisać do jednej emocji, dobrym przykładem tutaj może być słowo: „niedobrze”. Zauważono, że mimo negatywnej oceny całej wypowiedzi, pozytywny wydźwięk był widoczny - średni procent występowania słowa „dobry” w tych recenzjach wynosi ok. 26%. Dla porównania skali, w recenzjach, które otrzymały wyższe oceny gwiazdkowe, procent ten wynosił ok. 39%. To badanie podkreśla złożoność analizy sentymentu w recenzjach online i wskazuje na potrzebę bardziej zaawansowanych technik analizy, aby lepiej zrozumieć i interpretować opinię. Rozumienie złożoności jak i niespójności w opinii klientów jest niebywale istotne w dążeniu do dostosowywania oferty w oparciu o rzetelne dane z recenzji.

Natomiast badanie przeprowadzone przez Thakora i Sasi (2015) prezentuje innowacyjne podejście do analizy sentymentu opartej na ontologii, które koncentruje się na wykrywaniu negatywnych sentymentów w treściach mediów społecznościowych. Autorzy skupili się na analizie krótkich wypowiedzi użytkowników na portalu Twitter (nowa nazwa marki brzmi: „X”), tzw. „tweedów” dotyczących usług doręczeniowych, takich jak United States Postal Service, Royal Mail z Wielkiej Brytanii i Canada Post analizując kwestie takie jak opóźnienia w dostawach zgubionych paczek czy problem z obsługą klienta. Wykorzystując kombinację technologii do ekstrakcji danych z portalu internetowego Twitter, oczyszczenia pobranych danych, analizy subiektywnej, budowy modelu ontologii oraz analizy sentymentu, naukowcy byli w stanie identyfikować i przetwarzać negatywne sentymenty związane z tweet’ami (Thakor & Sasi, 2015). Jak widać na rysunku 5 w procesie budowy modelu, z danych zostały usunięte hiperłącza do stron internetowych oraz znaki specjalne, które mogłyby utrudniać interpretację. Wykorzystano skrypty napisane w języku programowania *Python* do automatycznego pozyskania i oczyszczenia treści co było kluczowe w dalszym procesie badawczym. Kolejnym krokiem było zastosowanie metod analizy językowej wykorzystując biblioteki NLP w oprogramowaniu GATE, które umożliwiło identyfikację i adnotację rzeczowników oraz czasowników. Korzystając z paradygmatu programowania obiektowego rzeczowniki zdefiniowano jako obiekty, a czasowniki jako właściwości tych obiektów w kontekście ontologii. Informacje te zostały wykorzystane do budowy modelu za pomocą oprogramowania *Protégé*.

A diagram of software

Description automatically generated

Rysunek 5. Proces Analizy Sentymentu Oparty na Ontologii dla Treści Mediów Społecznościowych, Źródło: Thakor, P., & Sasi, S. (2015), "Ontology-based Sentiment Analysis Process for Social Media Content”

W tym modelu, klasy, obiekty i właściwości obiektów zostały określone odpowiednio jako encje, właściwości obiektów. Jako następność firmy oferujące usługi kurierskie mogą wykorzystać opracowane rozwiązanie, wraz z syntezą wyników do podjęcia w odpowiedzi działań korygujących na zidentyfikowane problemy oraz do formowania zautomatyzowanej odpowiedzi w formie online na zgłoszone aspekty, co było decydującym krokiem w administrowanie relacjami z odbiorcą oraz jakością oferowanych usług. Praca ta podkreśla jak istotne jest dogłębne, wielopoziomowe badanie każdej z grup: pozytywne, negatywne, neutralne.

W innej pracy "Sentiment Strength Detection in Short Informal Text" z grudnia 2010 roku autorstwa Mike'a Thelwalla, Kevana Buckleya, Georgiosa Paltoglou, Di Cai i Arvida Kappasa wskazano ponownie analizę sentymentu jako narzędzie do wykrywania krótkich, nieformalnych tekstów, takich jak omawiane wcześniej komentarze w portalach społecznościowych. Praca ta opiera się o badanie przeprowadzone przez zespół naukowców, którzy zwrócili uwagę na fakt, że większość istniejących algorytmów analizy sentymentu była zorientowana na identyfikacji opinii na temat produktów, a nie na wzorcach zachowań użytkowników. Podjęli się, dlatego zadania częściowego wypełnienia tej luki poprzez stworzenie nowego algorytmu *SentiStrength*, który miał za zadanie ekstrahować siłę sentymentu z nieformalnych tekstów w języku angielskim, wykorzystując nowe metody do wykorzystania de facto gramatyki i stylów pisowni. SentiStrength został zastosowany do analizy komentarzy na portalu interentowym „MySpace” i z wykorzystaniem tabeli poszukiwania siły sentymentu terminów zoptymalizowanej przez uczenie maszynowe, był w stanie przewidzieć pozytywne emocje z dokładnością 60% i negatywne emocje z dokładnością 73%, obie na skalach siły od 1 do 5. Podejście to było nowatorskie, ponieważ pozwalało na jednoczesne ocenienie pozytywnego i negatywnego sentymentu, co było nowatorskie w porównaniu do tradycyjnych metod analizy sentymentu, które zazwyczaj koncentrowały się tylko na pozytywnych lub negatywnych emocjach, a nie na ich sile. Przykład tej pracy pokazuje, że jak wiele płaszczyzn należy uwzględnić, aby poprawnie zaadresować przekaz wypowiedzi i emocje mu towarzyszące. Te przykłady ilustrują, że analiza sentymentu wielorako przekształca dane tekstowe w cenne wskazówki, które firmy mogą wykorzystać do przewidywania trendów, personalizacji ofert i kształtowania strategii. W erze cyfrowej, gdzie dane są nowym surowcem, efektywne wykorzystanie tej techniki staje się kluczem do sukcesu.

## 1.2 Podstawy Przetwarzania Języka Naturalnego (NLP)

Analiza sentymentu jest częścią Przetwarzania Języka Naturalnego (NLP, od ang. Natural Language Processing), które skupia się na interakcjach między językiem ludzkim a komputerami. Znajduje się na przecięciu informatyki, sztucznej inteligencji i lingwistyki obliczeniowej (Lopez & Kalita, 2017). NLP jest zbiorczym terminem odnoszącym się do automatycznego przetwarzania komputerowego języków ludzkich. Obejmuje to zarówno algorytmy, które przyjmują tekst wytworzony przez człowieka jako dane wejściowe, jak i algorytmy, które wytwarzają naturalnie wyglądający tekst jako dane wyjściowe (Goldberg, 2015). Ta interdyscyplinarna branża umożliwia systemom komputerowym nie tylko interpretację, ale także zrozumienie i odpowiedź na ludzki język, co otwiera nowe możliwości w analizie dużej ilości danych tekstowych.

Yue Kang, Zhao Cai, Chee-Wee Tan, Qian Huang i Hefu Liu w swoim artykule zatytułowanym: „Natural language processing (NLP) in management research: A literature review” opisują badanie, które polegało na przeprowadzaniu treści artykułów z zakresu wiedzy o zarządzaniu, używając słów kluczowych „Natural language processing” autorzy przeszukali 123 artykuły z dwudziestu czterech czołowych czasopism biznesowych zidentyfikowanych przez Teksański Uniwersytet w Dallas. W wyniku selekcji określono, że 50 artykułów było nie związanych z tą techniką, ponieważ proponowały one NLP jako rozwiązanie przyszłościowe, odrzucały ją lub były elementem w biografii autora. Artykuły, które adresowały zagadnienia z NLP pochodziły z wielu dziedzin nauki (Rysunek 6). Wykazane przez nich wyniki wskazują, że w kolejnych latach technika ta będzie wzrastać na znaczeniu w biznesie i zarządzaniu.

A pie chart with text on it

Description automatically generated

Rysunek 6. Rozkład publikacji dotyczących NLP w prestiżowych periodykach zarządczych, Źródło: Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics, 1–34.

Dzięki rozwojowi technik NLP przedsiębiorcy będą mogli efektywne wydobywać znaczenie, intencje i emocje z tekstu, co jest kluczowe w procesach decyzyjnych opartych na danych. W kontekście biznesowym, NLP rewolucjonizuje sposób, w jaki firmy analizują informacje zawarte w tekście – zaczynając od opinii klientów na portalach społecznościowych przez recenzje produktów po komunikację wewnętrzną. Dzięki NLP, firmy mogą automatycznie analizować duże zbiory danych tekstowych, co przekłada się na lepszą orientację na potrzeby klienta, szybszą reakcję na zmieniające się trendy rynkowe oraz bardziej celowane strategie marketingowe. Analiza sentymentu jawi się jako potężne narzędzie dla użytkowników do wydobywania potrzebnych informacji, jak również do agregowania zbiorowych uczuć wynikających z recenzji. W ostatnich latach na pierwszy plan wysunęły się różne metody realizacji tego zadania (Devika, Sunitha, & Ganesh, 2016). Technologie NLP, wykorzystując zaawansowane modele językowe, pozwalają na głębsze zrozumienie subtelności języka, co ma znaczący wpływ na jakość analizy sentymentu. Za pomocą technik takich jak tokenizacja, stemming i lematyzacja, systemy NLP potrafią przekształcać surowe dane tekstowe w strukturę, którą można łatwiej analizować i interpretować. To pozwala na wyłapywanie nie tylko ogólnych tendencji w wyrażanych opiniach, ale również na identyfikację bardziej złożonych wzorców i niuansów emocjonalnych, co jest nieocenione w tworzeniu strategii biznesowych opartych na danych. Dzięki NLP, przedsiębiorstwa mogą zautomatyzować i usprawnić procesy analizy danych tekstowych, co pozwala na wydobywanie cennych wglądów i informacji strategicznych w znacznie szybszym czasie niż tradycyjne metody. W świecie biznesu, gdzie szybkość i precyzja decyzji mogą być decydujące, zastosowanie NLP w analizie sentymentu stanowi kluczowe narzędzie, umożliwiające firmom pozyskanie przewagi konkurencyjnej.

W ramach przetwarzania języka naturalnego (NLP), wykorzystuje się szereg technik, które mają kluczowe znaczenie dla analizy oraz interpretacji języka, szczególnie w branży technologicznej. Tokenizacja, czyli proces dzielenia tekstu na mniejsze jednostki - tokeny, zwykle słowa, jest niezbędna dla dalszego przetwarzania języka. Na przykład, w analizie danych z mediów społecznościowych, tokenizacja pozwala na efektywne rozpoznanie kluczowych słów i zwrotów, co jest istotne dla algorytmów śledzących trendy i nastroje użytkowników. Jednak warto zauważyć, że istnieje wiele pojęć dotyczących tego, co w NLP (Przetwarzaniu Języka Naturalnego) jest uznawane za token. Różne pojęcia zależą od różnych celów (np. parsowania, MT) i często od różnych tła językowego. Aby dojść do definicji tokena, który jest zarówno lingwistycznie znaczący, jak i metodologicznie użyteczny (Kit & Webster, 1992).

A screenshot of a computer

Description automatically generated

Rysunek 7. Przykład procesu tokenizacji, Źródło: opracowanie własne

Stemming to proces redukcji słów do ich podstawowych form, jest wykorzystywany do usprawnienia procesów wyszukiwania i klasyfikacji tekstu. W procesie stemmizacji 'rdzeń' słowa jest uzyskiwany poprzez zastosowanie zestawu reguł, ale bez uwzględniania części mowy (POS) czy kontekstu wystąpienia słowa. Stemming jest często postrzegany jako narzędzie zwiększające recall, czyli zdolność do odzyskiwania informacji. W językach o stosunkowo prostej morfologii wpływ stemming’u jest mniejszy niż w językach o bardziej złożonej budowie morfologicznej (Jivani, 2011). Na przykład, w wyszukiwarkach internetowych, takich jak Google, stemming pozwala na zwiększenie trafności wyników wyszukiwania, dzięki czemu użytkownicy otrzymują bardziej precyzyjne informacje.

A black and white rectangular object with a green rectangle in the middle

Description automatically generated

Rysunek 8. Przykłady stemmatyzacji dla różnych odmian słowa "nauka", Źródło: opracowanie własne

Lematyzacja, bardziej zaawansowana technika niż stemming, bierze pod uwagę kontekst i znaczenie słowa w zdaniu, a następnie znalezieniu 'lematu', co jest niezwykle cenne w analizie semantycznej tekstu. Algorytm, który przekształca słowo na jego lingwistycznie poprawną formę podstawową, nazywany jest lemmatizer. W morfologii lemma to kanoniczna forma lexeme. Lexeme w tym kontekście odnosi się do zestawu wszystkich form o tym samym znaczeniu, a lemma to szczególna forma wybrana przez konwencję, aby reprezentować lexeme’m. (Jivani, 2011). W branży technologicznej, na przykład w asystentach głosowych typu Siri lub Alexa, lematyzacja pozwala na bardziej precyzyjne rozumienie zapytań użytkowników i dostarczanie odpowiedzi bardziej adekwatnych do ich intencji.

A black and white rectangular object with a green rectangle in the middle

Description automatically generated

Rysunek 9. Przykłady lematyzacji dla różnych odmian słowa "nauka", Źródło: opracowanie własne

Zastosowanie tych technik NLP w branży technologicznej ma szerokie implikacje biznesowe. Umożliwiają one firmom technologicznym analizowanie ogromnych zbiorów danych tekstowych, co jest kluczowe w kontekście Big Data. Dzięki temu mogą one lepiej zrozumieć zachowania i preferencje użytkowników, co pozwala na optymalizację produktów i usług oraz tworzenie bardziej spersonalizowanych doświadczeń dla klientów.

W epoce cyfrowej, gdzie dane tekstowe są nieustannie generowane i przetwarzane, umiejętność efektywnego wykorzystania technik NLP staje się kluczowa dla firm technologicznych. Pozwala im to nie tylko na uzyskanie przewagi konkurencyjnej, ale także na podejmowanie świadomych i strategicznych decyzji biznesowych opartych na dogłębnej analizie danych.

Te techniki są fundamentem dla bardziej zaawansowanych operacji w NLP, umożliwiając efektywne przetwarzanie tekstu, co jest kluczowe w analizie sentymentu. Poprzez ich zastosowanie, systemy NLP mogą lepiej radzić sobie z różnorodnością i złożonością języka naturalnego, co przekłada się na dokładniejszą i bardziej wiarygodną analizę sentymentu. Przyszłość NLP jest ściśle związana z rozwojem sztucznej inteligencji. W miarę jak czytelność języka naturalnego się poprawia, komputery będą mogły uczyć się z informacji dostępnych online i stosować je w rzeczywistości. W połączeniu z generowaniem języka naturalnego, komputery będą coraz bardziej zdolne do odbierania i przekazywania użytecznych i wartościowych informacji lub danych (Chopra, Prashar, i Sain, 2013). Skomplikowane procesy w handlu, takie jak prognozowanie i podejmowanie decyzji w handlu akcjami; upraszczanie interakcji z klientami za pomocą chatbotów na platformach handlowych, czyniąc interakcję bardziej przyjemną; analizowanie problemów obywateli z dużych zbiorów danych w E-administracji; skuteczne zarządzanie operacjami w ochronie zdrowia, takie jak diagnozowanie, dostarczanie usług i zarządzanie dokumentacją; oraz ulepszanie podejść dydaktycznych w sektorze edukacji, to niektóre z korzyści, które mogą przynieść techniki NLP. Ponadto, integracja NLP z zaawansowanymi technologiami, takimi jak ML, AI i głębokie uczenie się, może zapewnić jeszcze dokładniejsze wyniki odpowiadające na potrzeby użytkowników (Wu i Mircea, 2021).

## 1.3 Narzędzia i Technologie w Analizie Sentymentu

W Analizie Sentymentu, kluczową rolę odgrywają zaawansowane narzędzia i technologie, które umożliwiają precyzyjne przetwarzanie i analizę danych tekstowych. Wśród tych narzędzi wyróżnić można Regex, CountVectorizer i Natural Language Toolkit (NLTK), niezbędne w arsenale specjalistów zajmujących się przetwarzaniem języka naturalnego.

Regex, czyli wyrażenia regularne, są potężnym narzędziem do wyszukiwania i manipulowania tekstem, pozwalającym na identyfikację specyficznych wzorców, takich jak słowa kluczowe czy frazy. Wyrażenia regularne oferują ograniczony, ale potężny metalanguage do opisywania wszelkiego rodzaju formatów, protokołów i innych małych języków tekstowych. Wyrażenia regularne pojawiły się w kontekście teorii języków formalnych, a ich podstawowe zastosowanie to część skanerów w kompilatorach (Erwig & Gopinath, 2012). W branży technologicznej, gdzie konieczna jest analiza logów systemowych czy kodów błędów, Regex umożliwia efektywne wyodrębnienie istotnych informacji z dużych zbiorów danych.

CountVectorizer, często stosowany w bezpłatnej bibliotece oprogramowania do języka Python scikit-learn, służy do konwersji kolekcji dokumentów tekstowych na macierz liczb całkowitych, reprezentujących częstotliwość występowania słów. Jest to kluczowe w procesie 'bag-of-words', gdzie tekst jest przekształcany w zestaw jego składników, co jest nieocenione przy identyfikowaniu dominujących trendów i postaw w opinii klientów.

A screenshot of a computer

Description automatically generated

Rysunek 10. Przykłady użycia CountVectorizer dla różnych dokumentów, Źródło: opracowanie własne

NLTK to z kolei wszechstronna platforma do budowania programów języka Python do pracy z danymi języka ludzkiego, oferująca łatwe w użyciu interfejsy do korpusów i zasobów leksykalnych, jak również narzędzia do klasyfikacji, tokenizacji, stemmingu, tagowania i parsowania. W analizie sentymentu, NLTK pozwala na szczegółową i złożoną analizę języka naturalnego, co ma istotne znaczenie dla firm technologicznych w analizie danych klientów, mediów społecznościowych czy treści online.

Wykorzystanie tych narzędzi w analizie sentymentu otwiera przedsiębiorstwom technologicznym możliwość nie tylko efektywnego przetwarzania i analizowania danych tekstowych, ale także przekształcania surowych danych w cenne wglądy biznesowe. W dobie cyfrowej, gdzie analiza danych tekstowych odgrywa kluczową rolę w zrozumieniu rynku i potrzeb klientów, narzędzia takie jak Regex, CountVectorizer i NLTK stanowią niezbędny element w budowaniu przewagi konkurencyjnej na rynku.

W dziedzinie analizy sentymentu, znaczący postęp dokonany został dzięki bogatej literaturze naukowej i innowacyjnym rozwiązaniom technologicznym. Rozważając kluczowe prace i badania, nasuwa się obserwacja, że rozwój tej dziedziny jest ściśle powiązany z postępami w uczeniu maszynowym i przetwarzaniu języka naturalnego.

Wśród pionierskich prac wyróżnia się te, które skupiają się na automatycznym rozpoznawaniu emocji w tekście, wykorzystując zaawansowane algorytmy uczenia maszynowego. Badania prowadzone przez Pang i Lee, ukazują, jak maszyny mogą efektywnie klasyfikować teksty na podstawie ich tonacji emocjonalnej, co ma bezpośrednie zastosowanie w monitorowaniu opinii konsumentów, analizie trendów rynkowych oraz zarządzaniu reputacją marki.

Porównując istniejące rozwiązania i technologie, warto zwrócić uwagę na rozwój narzędzi takich jak IBM Watson, Google Cloud Natural Language API czy Amazon Comprehend. Te systemy, zbudowane przez przodujące firmy technologiczne, wykorzystujące zaawansowane algorytmy i głębokie sieci neuronowe, umożliwiają nie tylko analizę sentymentu na szeroką skalę, ale również pozwalają na dokładne rozpoznawanie subtelnych niuansów emocjonalnych w tekście. Co istotne, ich aplikacja w świecie biznesu otwiera drogę do głębszego zrozumienia potrzeb klientów, umożliwiając firmom tworzenie bardziej spersonalizowanych strategii marketingowych i komunikacyjnych.

Oprócz tego, w literaturze naukowej pojawiają się prace koncentrujące się na specyficznych wyzwaniach związanych z analizą sentymentu, takich jak rozpoznawanie ironii czy sarkazmu, co jest szczególnie istotne w kontekście mediów społecznościowych. Badania te podkreślają potrzebę ciągłego doskonalenia modeli językowych, aby mogły one skutecznie interpretować coraz bardziej złożone i subtelne formy komunikacji ludzkiej.

Podsumowując, przegląd literatury i istniejących rozwiązań w dziedzinie analizy sentymentu wskazuje na dynamiczny rozwój tej branży, który jest napędzany zarówno przez postępy akademickie, jak i innowacje technologiczne. Dla przedsiębiorstw, inwestycja w te technologie to krok ku głębszemu zrozumieniu rynku i większemu dostosowaniu strategii do potrzeb współczesnych konsumentów, co w konsekwencji prowadzi do zwiększenia konkurencyjności i efektywności biznesowej.

## 1.4 Potencjalne Wyzwania i Ograniczenia

W Analizie Sentymentu, napotykamy na szereg wyzwań i ograniczeń, które stanowią przedmiot zainteresowania zarówno w kontekście technologicznym, jak i biznesowym. Jednym z głównych wyzwań jest kompleksowość języka naturalnego, który charakteryzuje się nie tylko bogactwem ekspresji, ale także ironią, sarkazmem i zmiennymi kontekstami kulturowymi. Analiza sentymentu jest zadaniem subiektywnym i dlatego nasze dane mogą nadal cierpieć na uprzedzenia etykietowania, które dotykają większość zbiorów danych (Muhammad et al., 2023). Te aspekty języka mogą prowadzić do ambiwalencji w interpretacji emocji, co staje się problemem szczególnie w automatycznej analizie tekstów pochodzących z mediów społecznościowych.

Dynamiczna ewolucja języka stanowi kolejne wyzwanie. Nowe wyrażenia i ciągłe zmiany w strukturze języka wymagają od systemów analizy sentymentu elastyczności i adaptacyjności. Stały rozwój algorytmów Uczenia Maszynowego, które potrafią śledzić i dostosowywać się do tych zmian, jest kluczowy dla efektywnej Analizy Sentymentu.

Dodatkowo, przetwarzanie i analiza ogromnych ilości danych generowanych przez media społecznościowe wymaga zaawansowanej infrastruktury technologicznej oraz efektywnych metod przetwarzania danych. Wyzwanie to dotyczy zarządzania i analizowania dużych zbiorów danych w sposób wydajny i skuteczny.

Organizacje działają w dynamicznych i ciągle zmieniających się środowiskach, składających się z aktorów takich jak konkurenci i rząd, którzy mają wpływ na to, jak organizacja może i powinna prowadzić biznes. To z kolei wywiera różne rodzaje presji na zdolność i skłonność organizacji do przyjęcia sztucznej inteligencji. (Enholm, Papagiannidis, Mikalef, & Krogstie, 2021). W kontekście biznesowym, te wyzwania przekładają się na potrzebę inwestycji w nowoczesne technologie i specjalistyczną wiedzę. Firmy, które chcą efektywnie wykorzystać analizę sentymentu, muszą podejmować świadome decyzje dotyczące wyboru narzędzi i technologii zdolnych do radzenia sobie z wymienionymi problemami. Sukces w tej dziedzinie może prowadzić do głębszego zrozumienia potrzeb klientów, efektywniejszego reagowania na zmieniające się tendencje rynkowe oraz lepszego zarządzania reputacją marki.

Podsumowując, chociaż analiza sentymentu oferuje obiecujące możliwości dla biznesu, wymaga ona jednocześnie skrupulatnego podejścia do istniejących wyzwań. Rozwiązanie tych problemów za pomocą innowacyjnych metod i technologii jest kluczowe dla efektywnego wykorzystania analizy sentymentu jako narzędzia do uzyskiwania przewagi konkurencyjnej na współczesnym rynku.

# 2 Uczenie maszynowe w analizie sentymentu

Uczenie maszynowe (ML od ang. Machine Learning) jest dziedziną łączącą Sztuczną Inteligencję i informatykę, skupia się na ulepszaniu i badaniu algorytmów systematycznych, które mają na celu uczyć się i generalizować posiadane dane. Algorytmy uczenia maszynowego są szeroko wykorzystywane do analizy sentymentu (Agarwal & Mittal, 2015). Uczenie maszynowe opisuje zdolność systemów do uczenia się na podstawie specyficznych dla problemu danych szkoleniowych, aby automatyzować proces budowania modeli analitycznych i rozwiązywać powiązane zadania (Janiesch, Zschech, & Heinrich, 2021). Freamwork uczenia maszynowego w kontekście ekstrakcji informacji, takim jak wydobywanie opinii, oferuje kilka zalet, w tym mniejszy wysiłek budowania w porównaniu do tworzenia symbolicznych reguł oraz probabilistyczne przypisanie, które jest cenne do oceny niepewności przypisania i stąd potrzeby obliczania kosztownych reprezentacji cech opierających się na przetwarzaniu języka naturalnego. (Boiy & Moens, 2009). Modele uczenia maszynowe są również bardziej adaptacyjne do specyficznych potrzeb branż i tematów. Ważnym aspektem jest również mniejsza subiektywność, która zapewnia większą obiektywność poprzez zastosowanie ustalonych modeli, reguł, niezależnie od osobistych przekonań lub oczekiwań. W analizie tekstu wykonywanej manualnie to od recenzenta zależy interpretacja danego fragmentu, co może prowadzić do niekonsekwencji wśród grupy osób. Dodatkowo modele ML dają wyjątkową możliwość doszkalania w kolejnych iteracjach, poprzez dodawanie nowych danych, ciągłe szkolenie, co wiąże się z możliwością udoskonalania o nowe wzorce, konteksty, bez wprowadzania ludzkiej subiektywności. Ostatecznie, prowadzi to do spójniejszych, powtarzalnych wyników, wszystkie te czynniki prowadzą do zwiększenia wiarygodności, weryfikowalność analizy danych mającego konsekwencje biznesowe.

Jest wiele różnych metod stosowanych analizy sentymentu (Rysunek 11). Wyróżnia się podejścia oparte na uczeniu maszynowym, leksykonie oraz metodach hybrydowych.

A diagram of a learning process

Description automatically generated

Rysunek 11. Taksonomia technik analizy sentymentu, Źródło: Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment Analysis Based on Deep Learning: A Comparative Study. Electronics, 9(3), 483.

W tej pracy skoncentruję się na podejściu ML ze względu na jego adaptacyjność i moce predykcyjne do zmieniających się danych wejściowych lub testowych. Modele stworzone z podejściem uczenia maszynowego, a w jeszcze większym stopniu te zrealizowane w podgrupie uczenia głębokiego, wykazują zdolność do precyzyjniejszego wskazywania subtelnych detali językowych - jest to strategiczne podczas analizy emocjonalnych znaczeń w tekście pisanym. Warto zauważyć, że to właśnie dzięki tej zdolności do treningu na wskazanych przykładach, systemy te cechuje lepsza umiejętność generalizacji, co jest niezwykle wartościowe w szybko zmieniających się kontekstach, takich jak media społecznościowe. Kolejnym aspektem wartym poruszenia, jest kwestia, że uczenie maszynowe daje w tym momencie możliwość automatycznej automatyzacji modeli w momencie, gdy nowa porcja danych jest dostępna. Podejście to ma też możliwość integracji różnych źródeł danych, struktur, formatów – co jest dodatkową korzyścią w multikulturowym środowisku globalnych społeczności internetowych. Kontekst analizy sentymentu wskazuje niejednokrotnie na potrzebę badania dużej ilości danych, gdzie uczenie maszynowe wykazuje możliwość przetwarzania, modelowania, prognozowania danych w skali, a osiągnięcie tego jest bardziej złożone dla metod znaczniej tradycyjnych.

## 2.1 Uczenie nadzorowane

Uczenie nadzorowane (ang. Supervised Learning), czyli innymi słowy scenariusz nadzorowany charakteryzuje się pojęciem nauczyciela lub nadzorcy, którego głównym zadaniem jest dostarczenie agentowi precyzyjnej miary jego błędu (bezpośrednio porównywalnej z wartościami wyjściowymi) (Bonaccorso, 2017). Jest to proces uczenia na podstawie zestawu danych zawierającego obserwacje wraz z określonymi etykietami (w tym przypadku sentymenty), na których model uczy się wykrywać schematy i w ten sposób może przewidzieć sentymenty dla nowych, nieokreślonych danych.

Jest wiele zastosowań, gdzie uczenie nadzorowane umożliwia automatyczna oraz skuteczną analizę dużych ilości tekstów, dostarczając opiniotwórczych informacji dla biznesu. To uczenie jest stosowane w aplikacjach, gdzie dane historyczne przewidują prawdopodobne przyszłe wydarzenia. Istnieje wiele praktycznych przykładów tego uczenia, na przykład aplikacja, która przewiduje gatunek irysa na podstawie zestawu pomiarów jego kwiatu. (Nasteski, 2017). Jako jedne z pierwszych zastosowań uczenia nadzorowanego było klasyfikowanie wiadomości elektronicznych, w celu zidentyfikowania spamu, ale także było używane do interpretacji pisma odręcznego ze zdjęć lub skanu, co miało służyć zautomatyzowaniu obróbki formularzy czy prawidłowemu rozpoznaniu kodu pocztowego na listach. Przez lata zauważono wyjątkowy potencjał tego rozwiązania i poszerzono użytkowania o na przykład, medycynę, gdzie algorytmy pomagają w identyfikacji zmian nowotworowych w obrazach wyników badań medycznych. W finansach znalazło zastosowanie w przewidywaniu trendów na giełdzie papierów wartościowych, analizie ryzyka kredytowego czy automatyzacji doradztwa inwestycyjnego. Jednak ewolucja w tej dziedzinie ciągle postępuje i odkrywane są kolejne przeznaczenia, m. in. w uczeniu głęboki, personalizacji informacji dla klienta czy rozpoznawaniu emocji. Warto zauważyć w jak wielu obszarach analizy sentymentu można zaaplikować uczenie nadzorowane. W tej pracy wielokrotnie wspomniano o analizie emocji w mediach społecznościowych, ocenie recenzji, obsłudze klienta, ale jest też wiele kolejnych aspektów, których ważność tylko wzrasta, na przykład detekcja fałszywych wiadomości i dezinformacji. Klasyfikacja artykułów, tekstów pod kątem ich wiarygodności może obejmować analizę sentymentu w celu zidentyfikowania manipulacji czy fałszywych informacji. Ponieważ szerokie rozprzestrzenianie się fałszywych informacji może mieć poważny negatywny wpływ na jednostki i społeczeństwo, brak skalowalnych strategii weryfikacji faktów jest szczególnie niepokojący. Nic więc dziwnego, że ostatnie wysiłki badawcze są poświęcone nie tylko lepszemu zrozumieniu tego zjawiska, ale również automatyzacji wykrywania fałszywych informacji. (Reis, Correia, Murai, Veloso, & Benevenuto, 2019).

### 2.1.1 Regresja Logistyczna

Regresja logistyczna umożliwia stosowanie ciągłych lub kategorycznych modeli prognozujących oraz zapewnia możliwość dostosowania dla danych z wieloma zmiennymi zależnymi. Sprawia to, że regresja logistyczna jest szczególnie użyteczna do analizy danych obserwacyjnych, gdy konieczne jest dostosowanie w celu zmniejszenia potencjalnego błędu wynikającego z różnic w porównywanych grupach (LaValley, 2008).

Model regresji logistycznej używa funkcji logistycznej (sigmoidalnej) do przewidywania prawdopodobieństwa wystąpienia zdarzenia, że dana obserwacja należy do jednej z uprzednio określonych kategorii. Specyficzną charakterystyką wykresu funkcji jest jej kształt przypominający literę "S" oraz ograniczenie wynik do przedziału między 0, a 1, co sprawia, że ma silne zastosowanie do modelowania prawdopodobieństwa w kontekście klasyfikacji binarnej.

Funkcja sigmoidalna jest określona wzorem:

Gdzie:

e - to podstawa logarytmu naturalnego

x - jest zmienną niezależną

Regresja logistyczna modeluje prawdopodobieństwo, że zmienna zależna Y = 1, na podstawie jednej lub więcej zmiennych niezależnych (oznaczonych zwyczajowo jako X). Wyżej wymienione prawdopodobieństwo jest przekształcone w logarytm szans (logit), który może być modelowany jako liniowa kombinacja zmiennych niezależnych. W regresji logistycznej, proces estymacji parametrów modelu (znanych również jako wagi) odbywa się za pomocą metody maksymalizacji funkcji wiarygodności. Co oznacza, że metoda ta dąży do znalezienia takiego zestawu wartości parametrów, aby prawdopodobieństwo wystąpienia zaobserwowanych wyników było jak największe. Algorytm rozpoczyna działanie z początkowym zestawem wag, które są zazwyczaj losowo wybrane lub inicjowane w inny standardowy sposób. Następnie, używając danych treningowych, algorytm regresji logistycznej iteracyjnie aktualizuje wyżej wymienione wagi. Każdy tego rodzaju powtarzalny krok ma na celu poprawę dokładności działania zbudowanego modelu w przewidywaniu wyników, czyli tym samym zmniejszyć błąd. Metodą najczęściej stosowaną do aktualizacji wag w regresji logistycznej jest gradientowy spadek (ang. gradient descent). Algorytm gradientowego spadku jest jedną z fundamentalny metod optymalizacji służąca do minimalizacji funkcji kosztu (lub funkcji straty) modelu, robi to za pomocą obliczania gradientu (kierunku i wielkości kroku) funkcji wiarygodności względem każdej, poszczególnej wagi, co w kolejnym kroku umożliwia aktualizację o taką wartość, która zwiększa wiarygodność przewidywań modelu. Aby dokonać decyzji o przynależności nowej obserwacji do klasy, porównuje się prawdopodobieństwo z ustalonym progiem (np. 0,5). W takiej sytuacji, jeżeli prawdopodobieństwo prognozowane jest większe niż próg, obserwacja jest klasyfikowana do jednej klasy, w przeciwnym razie do drugiej. Następnym krokiem, po dostosowaniu wag, jest ocena wydajności prognoz na zestawie danych testowych.

Regresja logistyczna ma swoje zastosowanie w wielu dziedzinach, takich jak medycyna (np. przewidywanie prawdopodobieństwa wystąpienia choroby), bankowość (np. ocena ryzyka kredytowego), marketing (np. przewidywanie odpowiedzi klienta na kampanię) i wiele innych.

Regresja logistyczna jest prosta w implementacji i interpretacji wyników, charakteryzuje się dobrą wydajnością w przypadkach, gdy zmienna zależna jest binarna oraz daje możliwość modelowania wpływu zmiennych ciągłych i kategorycznych. Jednak należy uwzględnić, że zakłada ona liniową zależność między logarytmem szans, a zmiennymi niezależnymi, co może nie być zawsze spełnione w rzeczywistych danych. Ma ona również swoje ograniczenia w modelowaniu zależności złożonych lub nieliniowych, co oznacza, że każdego razu, gdy następuje zmiana jednostkowa zmiennej niezależnej będzie mieć to swoje stałe odzwierciedlenie w logarytmie szans zdarzenia, a przy próbie uchwycenia złożonych wzorców w danych będzie to stanowiło wadę.

### 2.1.2 Metoda k najbliższych sąsiadów

Intuicja leżąca u podstaw klasyfikacji Najbliższego Sąsiada (k-NN od ang. k-Nearest Neighbour) jest dość prosta: przykłady są klasyfikowane na podstawie klasy ich najbliższych sąsiadów (Cunningham & Delany, 2021). Algorytm k-najbliższych sąsiadów może przewidywać zarówno atrybuty dyskretne (najczęściej występującą wartość wśród k najbliższych sąsiadów), jak i atrybuty ciągłe (średnią wśród k najbliższych sąsiadów) (Batista & Monard, 2002). Klasyfikacja metodą k najbliższych sąsiadów (k-NN) to proces, który polega na identyfikacji k obserwacji z zestawu treningowego, które są najbliższe nowej, nieznanej obserwacji, na podstawie wybranej metryki odległości. Służą do tego podejścia takie jak odległość euklidesowa, Manhattan czy Minkowski. To podejście opiera się na założeniu, że podobne obserwacje (w sensie cech) będą należały do tej samej klasy. Jako przykład zastosowania, można przyjąć określenie sentymentu nowej opinii o restauracji jako pozytywny czy negatywny, używając metody k najbliższych sąsiadów, dla k =3. Pragniemy analizować, jak bardzo nowa opinia przypomina trzy inne, wcześniej sklasyfikowane opinie, które w tej sytuacji mamy w danych i znamy ich sentyment. Po sprawdzeniu wybranych cech, wybrano trzy najbliższe opinii, które były pozytywne w dwóch przypadkach, a jedna z nich negatywna. Większość z nich była opiniami pozytywnymi, dlatego nową obserwację przypisujemy jako pozytywną. Można tu zauważyć alegorię z życia codziennego, gdy decydowanie o wyborze restauracji jest wykonywane na podstawie rekomendacji większości bliskich nam osób. Jeżeli osoby, które wykazują podobne gusta oceniają nową restaurację dobrze, to istnieje wysokie prawdopodobieństwo, że my też docenimy ich potrawy. Jest to zamysł często używany w rekomendacjach.

Warto tu wspomnieć, że wybór parametru k, jest niebywale istotny, jest on kluczowy dla dokładności k-NN. Zbyt mała wartość parametru k może uczynić model wrażliwym na szum w danych, przez to zbyt dopasować się do danych treningowych (overfitting), podczas gdy zbyt duże k może prowadzić do zbyt ogólnej klasyfikacji, w takim przypadku model będzie niewystarczająco dopasowany (underfitting). Wybór optymalnej wartości k często wymaga eksperymentów i walidacji krzyżowej, a także wiedzy eksperckiej.

Wyraźną zaletą tego algorytmu jest prostota implementacji i intuicyjność działania, wynikająca z podobieństw płynących z życia codziennego. Algorytm ten wykazuje skuteczność w przypadkach, gdzie granica decyzyjna między klasami jest bardzo nieregularna nielinearna, k-NN nie robi założeń a priori o formie granicy decyzyjnej. W przeciwieństwie do wielu innych algorytmów uczenia maszynowego, które zakładają pewną prostotę lub regularność granic decyzyjnych (np. regresja logistyczna zakłada liniową separowalność klas), k-NN może dostosować się do bardzo nieregularnych i złożonych kształtów granic bez konieczności wprowadzania złożonych modyfikacji do modelu. Dodatkowo brak konieczności trenowania modelu, czyni k-NN również leniwym algorytmem uczenia maszynowego (ang. lazy learning algorithm). Klasyfikatory leniwe nie trenują klasyfikatora do momentu przedstawienia próbki testowej (Garcia, Feldman, Gupta, Srivastava, et al., 2009).

### 2.1.3 Drzewo Decyzyjne

Drzewa decyzyjne (ang. Decision Tree) są graficzną metodą wspomagania procesu decyzyjnego. Drzewa decyzyjne wywodzą swoje początki z okresu wczesnego rozwoju zapisu pisemnego. Ta historia ilustruje główną zaletę drzew: wyjątkowo łatwe w interpretacji wyniki, które mają intuicyjny, drzewopodobny wygląd, co z kolei wzmacnia zrozumienie i rozpowszechnianie wyników (De Ville, 2013). Prostota i intuicyjność, prowadzą do łatwości w przedstawieniu nawet dla osób, które nie są ekspertami w dziedzinie analizy danych. Mają one strukturę, które składają się z wierzchołków połączonych krawędziami. Wierzchołki, z których wychodzi co najmniej jedna krawędź, są nazywane węzłami, a pozostałe wierzchołki – liśćmi. To znaczy, że mają one jasno określoną hierarchię – od korzenia poprzez kolejne poziomy decyzji po liście (węzły końcowe bez dzieci). W każdym węźle sprawdzany jest określony warunek, zapytanie na temat danych dotyczący danej obserwacji i na jego podstawie wybierana jest jedna z gałęzi prowadząca do kolejnego wierzchołka. Punktem startowym działania algorytmu jest węzeł startowy, nazwany korzeniem, który jest nominalnym punktem wyjścia dla wszystkich decyzji. Każda gałąź odpowiada na zadane pytanie, prowadząc nad do kolejnych węzłów lub liści, które oznaczają koniec ścieżki oraz podają wyznaczony wynik. Proces decyzyjny w każdym etapie opiera się na ocenie obserwacji wejściowej według określonego kryterium np. czy wartość pewnej cechy jest większa od zadanej liczby. Na podstawie tej oceny, podejmowana jest decyzja o dalszej "ścieżce" – czyli którą gałąź wybrać, aby przejść do następnego węzła. Opisana procedura jest powtarzana, do momentu znalezienia się obserwacji w liściu, który zawiera klasyfikację oraz w ten sposób odpowiada ma pytanie, do jakiej klasy będzie należeć. Procedurę tę można postrzegać jako zachłanne przeszukiwanie przestrzeni wszystkich możliwych drzew decyzyjnych poprzez skanowanie instancji w danym węźle w celu określenia zysku z każdego podziału i wybrania pojedynczego podziału, który zapewnia największy zysk (Kotsiantis, 2013). Drzewa decyzyjne są one faktycznie grafami skierowanymi i acyklicznymi, ponieważ każda krawędzi ma określony kierunek i nie zawiera cykli. Model ten dzieli zbiór danych na mniejsze podzbiory na podstawie prostych zapytań, algorytm wybiera warunki na węzłach, które najefektywniej segregują dane, maksymalizując różnice między kategoriami danych w każdym liściu. To daje możliwość, aby model dokładnie predykował kategorię dla nowej obserwacji, bazując na posiadanych cechach i zestawie danych treningowych. Klasyfikacja nowego przypadku polega na przejściu od korzenia do liścia i przypisaniu do niej klasy zapisanej w danym liściu.

Drzewa decyzyjne są łatwe do interpretacji, ponieważ strukturę drzewa można przedstawić graficznie, a my możemy podążać za gałęziami drzewa zgodnie ze zmiennymi wejściowymi, co wymaga mniej czasu na szkolenie (Sharma & Dey, 2012).

2.2 Uczenie nienadzorowane

Uczenie nienadzorowane (ang. Unsupervised Learning) ułatwia analizę surowych zbiorów danych, pomagając tym samym w generowaniu analitycznych wniosków z danych bez wyznaczonych etykiet (Usama et al., 2019). Jedną z głównych różnic między uczeniem maszynowym nienadzorowanym a uczeniem maszynowym nadzorowanym jest brak zestawu treningowego dla tego pierwszego i w związku z tym, brak oczywistej roli dla walidacji krzyżowej. Drugą ważną różnicą jest to, że chociaż większość algorytmów klastrowania jest formułowana w terminach kryterium optymalności, zazwyczaj nie ma gwarancji, że uzyskano rozwiązanie optymalne globalnie. (Gentleman & Carey, 2008).

### 2.2.1 Sztuczne sieci neuronowe

Sztuczne sieci neuronowe (ANN od ang. Artificial Neural Network) składają się z prostych, silnie połączonych jednostek przetwarzających zwanych neuronami, z których każdy wykonuje dwie funkcje, mianowicie agregację swoich wejść od innych neuronów lub środowiska zewnętrznego oraz generowanie wyjścia z agregowanych wejść (Dongare, Kharde, & Kachare, 2012).

### 2.2.2 Deep Learning

Głębokie uczenie (DL od ang. Deep Learning) to zastosowanie sztucznych sieci neuronowych (ANNs;od ang. Artificial Neural Networks) do zadań uczenia się przy użyciu sieci o wielu warstwach (Zhang, Wang, & Liu, 2018). W tradycyjnych podejściach do uczenia maszynowego cechy są definiowane i ekstrahowane ręcznie lub za pomocą metod selekcji cech. Jednak w modelach głębokiego uczenia cechy są uczone i ekstrahowane automatycznie, co pozwala osiągnąć lepszą dokładność i wydajność (Dang, Moreno-García, & De la Prieta, 2020).

A diagram of machine learning

Description automatically generated

Rysunek 12. Różnice między dwoma podejściami do klasyfikacji polarności sentymentu: uczenie maszynowe (góra) i głębokie uczenie (dół). Część mowy (POS); Rozpoznawanie nazwanych encji (NER); Częstość terminów-Odwrotna częstość dokumentów (TF-IDF). Źródło: Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment Analysis Based on Deep Learning: A Comparative Study. Electronics, 9(3), 483.

### 2.2.3 Sieci nauronowe konwolucyjne (CNNs)

Sieci neuronowe konwolucyjne (CNN od ang. Convolution Neural Networks) są analogiczne do tradycyjnych sieci ANNs, ponieważ składają się z neuronów, które samo-optymalizują się poprzez uczenie. Sieci neuronowe konwolucyjne różnią się od innych form sztucznych sieci neuronowych tym, że zamiast skupiać się na całości dziedziny problemu, wykorzystywana jest wiedza o konkretnym typie wejścia. To z kolei pozwala na ustanowienie znacznie prostszej architektury sieci (O'Shea & Nash, 2015).

### 2.2.4 Rekurencyjne sieci neuronowe (RNNs)

Rekurencyjne sieci neuronowe (RNN od ang. Recurrent Neural Networks) to klasa sieci neuronowych, w których połączenia między neuronami tworzą skierowany cykl, co tworzy pętle sprzężenia zwrotnego wewnątrz RNN (Dang, Moreno-García, & De la Prieta, 2020).

## 2.3 Regularyzacja

## 2.4 Optymalizacja

# 3 Analiza sentymentu w mediach społecznościowych

# 4 Zakończenie

# 5 Literatura

Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Chopra, A., Prashar, A., & Sain, C. (2013). Natural Language Processing. 1(4).

Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment Analysis: A Comparative Study on Different Approaches. Procedia Computer Science, 87, 44–49. https://doi.org/10.1016/j.procs.2016.05.124

Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2022). Artificial Intelligence and Business Value: A Literature Review. Information Systems Frontiers, 24(5), 1709–1734. https://doi.org/10.1007/s10796-021-10186-w

Erwig, M., & Gopinath, R. (2012). Explanations for Regular Expressions. 7212, 394–408. https://doi.org/10.1007/978-3-642-28872-2\_27

Goldberg, Y. (2015). A Primer on Neural Network Models for Natural Language Processing (arXiv:1510.00726). arXiv. https://doi.org/10.48550/arXiv.1510.00726

Jivani, A. G. (2011a). A Comparative Study of Stemming Algorithms. 2.

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics, 7(2), 139–172. https://doi.org/10.1080/23270012.2020.1756939

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers. https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

Lopez, M. M., & Kalita, J. (2017). Deep Learning applied to NLP (arXiv:1703.03091). arXiv. http://arxiv.org/abs/1703.03091

Muhammad, S. H., Abdulmumin, I., Ayele, A. A., Ousidhoum, N., Adelani, D. I., Yimam, S. M., Ahmad, I. S., Beloucif, M., Mohammad, S. M., Ruder, S., Hourrane, O., Brazdil, P., Ali, F. D. M. A., David, D., Osei, S., Bello, B. S., Ibrahim, F., Gwadabe, T., Rutunda, S., … Arthur, S. (2023). AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages (arXiv:2302.08956; Wersja 5). arXiv. http://arxiv.org/abs/2302.08956

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. 10, 1320–1321. https://www.researchgate.net/publication/220746311\_Twitter\_as\_a\_Corpus\_for\_Sentiment\_Analysis\_and\_Opinion\_Mining

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. 2(1–2), 20.

Patodkar, V. N., & I.R, S. (2016). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. IJARCCE, 5(12), 321–322. https://doi.org/10.17148/IJARCCE.2016.51274

Thakor, P., & Sasi, S. (2015). Ontology-based Sentiment Analysis Process for Social Media Content. Procedia Computer Science, 53, 199–207. https://doi.org/10.1016/j.procs.2015.07.295

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment Strength Detection in Short Informal Text. Journal of the American Society for Information Science and Technology, 61, 2544–2558. https://doi.org/10.1002/asi.21416

Valdivia, A., Luzon, M., & Herrera, F. (2017). Sentiment Analysis on TripAdvisor: Are There Inconsistencies in User Reviews? 15–25. https://doi.org/10.1007/978-3-319-59650-1\_2

Webster, J. J., & Kit, C. (1992). Tokenization as the Initial Phase in NLP. COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics. COLING 1992. https://aclanthology.org/C92-4173

Wu, R. M. X., & Mircea, M. (2021). E-Business: Higher Education and Intelligence Applications. BoD – Books on Demand.

# 6 Spis rysunków

Rysunek 1. Przykłady klasyfikacji sentymentu w opinii o filmach, Źródło: opracowanie własne

Rysunek 2. Schemat procesu reagowania na problemy z nową wersją aplikacji, Źródło: opracowanie własne

Rysunek 3. Złożoność analizy sentymentu w różnych typach zdań, Źródło: opracowanie własne

Rysunek 4. Porównanie głębokiego uczenia z uczeniem maszynowym, Źródło: opracowanie własne

Rysunek 5. Thakor, P., & Sasi, S. (2015). Ontology-based Sentiment Analysis Process for Social Media Content.. Procedia Computer Science, 199-207.

Rysunek 6. Rozkład publikacji dotyczących NLP w prestiżowych periodykach zarządczych, Źródło: Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics, 1–34.

Rysunki 7-10. Przykłady narzędzi i algorytmów używanych w przetwarzaniu języka naturalnego Źródło: opracowanie własne

# 7 Spis tabel

# 8 Załączniki

# 9 Streszczenie