



Studium Magisterskie

Kierunek Analiza danych - big data

Specjalność -

Imię i nazwisko autora

Maciej Sadkowski

Nr albumu ms107984

Wykorzystanie analizy sentymentu do analizy leków na podstawie komentarzy pacjentów

Praca magisterska

pod kierunkiem naukowym

Mariusza Rafała

Instytut

Informatyki i Gospodarki Cyfrowej

Warszawa 2022

Spis treści

Wstęp	3
1. Text mining oraz algorytmy i struktury w NLP	
1.1 Text mining.....	7
1.2 Definicja NLP.....	9
1.3 Analiza sentymentu oraz jej rodzaje.....	12
1.4 Definicja macierzy Document-Term i Term-Document.....	16
1.5 N-gramy.....	17
1.6 Algorytm LDA.....	18
2. Uczenie maszynowe w analizie tekstowej	
2.1 Uczenie maszynowe.....	20
2.2 Sieci neuronowe.....	22
2.3 Diagnostyka sieci neuronowych.....	27
3. Analiza sentymentu w badaniach rynkowych	
3.1 Zastosowanie biznesowe analizy sentymentu.....	29
3.2 Źródła danych.....	38
3.3 Technologie wykorzystujące analizę sentymentu.....	40
3.4 Analiza sentymentu w produkcji leków.....	44
4. Zastosowanie NLP w badaniu ocen leków	
4.1 Opis zbioru danych i przedstawienie schematu działania systemu.....	48
4.2 Analiza eksploracyjna.....	49
4.3 Podział na n-gramy.....	52
4.4 Budowa sieci neuronowej do analizy sentymentu.....	58
4.5 Ocena jakości modelu.....	63
4.6 Użyte biblioteki oraz pakiety.....	66
Zakończenie	67
Spis tabel i wykresów	71
Bibliografia	72

Wstęp

Początki medycyny nowoczesnej na przełomie XIX i XX wieku dokonały przełomu w historii ludzkości. Jeszcze na początku XX wieku dostępnymi lekami były tylko: naparstnica (środek pobudzający pracę mięśni sercowych), chinina (środek używany do leczenia malarii), ipekakuana (wykorzystywana do leczenia dyzenterii), aspiryna oraz rtęć (używano jej do leczenia kiły). W 1928 roku Alexander Fleming odkrył działanie penicyliny przeciwko gronkowcom. W 1944 roku, dzięki działaniom Howarda Floreya oraz Ernsta Chaina, umożliwiona została produkcja na dużą skalę penicyliny, która stała się pierwszym antybiotykiem (Ng, 2009). W 1966 roku Monroe E. Wall oraz Mansukh C. Wani odkryli, że kamptotecyna, substancja występująca w korze oraz łodydze drzewa *Camptotheca acuminata*, niszczy komórki rakowe (Kohn, 2020). Odkrycie tych leków oraz stworzenie wielu innych środków doprowadziło do tego, że przeciętna długość życia w Stanach Zjednoczonych w 1998 roku wynosiła 74 lat dla mężczyzn oraz 80 lat dla kobiet (dane pochodzące z Uniwersytetu Berkeley). Dla porównania, w 1900 roku oczekiwana długość życia mężczyzn była na poziomie 46 lat oraz 48 lat dla kobiet. Rozwój medycyny oraz powszechny dostęp do leków wpłynęły również na światową populację. W 1900 roku liczyła ona 1 656 000 000 ludzi. W roku 1950 liczba ta wzrosła do 2 516 000 000 (pomiędzy tymi dwoma okresami doszło do wybuchu I wojny światowej i II wojny światowej oraz pandemii grypy hiszpanki). W 1995 roku światowa populacja liczyła 5 576 000 000 ludzi, a w 2020 osiągnęła poziom 7 772 850 162 ludzi (Kaneda, 2021). Ponadto w poszczególnych okresach liczba urodzeń na 1000 osób wynosiła na świecie odpowiednio: 40, 38, 31 i 19. Oznacza to, że w coraz późniejszych okresach rodziło się coraz mniej dzieci, lecz populacja zwiększała się mimo tego faktu (Kaneda, 2021).

Do zagrożeń wynikających z zażywania leków należą tzw. efekty uboczne. W latach pięćdziesiątych XX wieku zachodnioniemiecki koncern farmaceutyczny Chemie Grünenthal GmbH opracował lek o nazwie Talidomid. Środek ten był pierwotnie stosowany jako lek usypiający. W latach 1957–1961 używano go powszechnie jako środka przeciwbólowego dla kobiet w ciąży, a także przeciwko przeziębieniu czy grypie. Lek ten jednak okazał się wykazywać silne działania teratogenne (tzn. uszkodzające płód) w pierwszych fazach okresu prenatalnego, co doprowadziło do narodzin ponad 10 tysięcy dzieci z poważnymi wadami, takimi jak deformacje stawów czy kończyn. Według szacowań ok. 50% spośród tych dzieci nie przeżyło pierwszego roku (sciencemuseum.org.uk, 2019). W czasach współczesnych powszechne są już portale oraz strony, na których pacjenci mogą opisać lek oraz jego działanie, a także wskazać na ewentualne skutki uboczne. Zaletą dostępu do ocen i recenzji leków są:

- możliwość oszacowania opinii o różnych lekach przeciwko konkretnej dolegliwości,
- znalezienie informacji o efektach ubocznych,
- oszacowanie zadowolenia pacjentów,

- redukcja kosztów czasowych i finansowych (mniejsze zużycie czasu niezbędnego do znalezienia odpowiedniego leku, wybranie odpowiedniego leku pozwala na pozbycie się kosztów kupna nieodpowiedniego środka).

Przeczytanie wszystkich komentarzy i oszacowanie wszystkich ocen jest jednak procesem wymagającym od użytkownika dużych zasobów czasowych. Rozwiązaniem, które mogłoby wpłynąć na poprawę doświadczenia użytkownika oraz skrócenie czasu potrzebnego do zrealizowania jego potrzeb, jest zbudowanie systemu opartego na analizie sentymentu. Analiza sentymentu jest metodą analizy tekstu pozwalającą na szybkie i zautomatyzowane uzyskanie wglądu w to, czy zdanie bądź dany komentarz jest opinią pozytywną, czy negatywną. Znaczenie analizy sentymentu nie jest duże tylko dla autorów organizacji (czyli dla podmiotu udostępniającego analizę), ale również dla użytkowników. W ten sposób organizacja ma wgląd w takie czynniki, jak: reputacja firmy, opinia klientów o produkcie czy też zadowolenie pracowników. Ponadto w zależności od rodzaju implementacji analizy sentymentu, można uzyskać informację o przyczynach satysfakcji lub jej braku ze strony klienta (Robinson, 2021). Użytkownicy mogą natomiast uzyskać lepszą informację o produkcie czy organizacji, co może wpłynąć na ich lepszy wybór oraz lepsze doświadczenie klienta (Robinson, 2021). Doświadczenie klienta (po angielsku: „customer experience”) jest efektem interakcji pomiędzy organizacją a klientem w trakcie trwania ich relacji (Zalewska, 2018). Interakcja między firmą a klientem może być wygenerowana na trzy sposoby: poprzez drogę, którą przebywa klient, punkty styczności z marką oraz rzeczywistości, z którymi styka się klient (włączając rzeczywistość wirtualną) (Zalewska, 2018). Doświadczenie klienta to dokładnie generowanie przyjemnych konotacji z marką, dostarczanie mu pozytywnych wrażeń i w ten sposób budowanie jego lojalności, wynikającej z zadowolenia z interakcji z organizacją (Zalewska, 2018).

Celem pracy jest stworzenie systemu opartego na analizie sentymentu do oceny leków na podstawie opinii pacjentów. Pierwszą grupą docelową wśród użytkowników tego systemu są pacjenci. W systemie tym pacjenci mogliby sprawdzić każdy komentarz na temat danego leku, jego uśrednioną ocenę oraz sprawdzić, czy komentarze są pozytywne bądź negatywne. W ten sposób ułatwiłoby to pacjentom wybór odpowiedniego leku na daną przypadłość. Drugą grupę użytkowników systemu stanowiliby producenci leków, którzy mogą w ten sposób badać swoje produkty. Dzięki temu mogą w krótkim czasie monitorować opinie pacjentów na temat leków oraz zbadać w ten sposób przyczynę, dlaczego opinia pacjentów jest właśnie taka. Oprócz tego, zarówno pacjenci, jak i producenci mogą w tym systemie zobaczyć słowa kluczowe w danej grupie leków oraz przeanalizować w ten sposób, które wyrazy powtarzały się najczęściej i wykazywały najwyższą kontrybucję do komentarzy. W pracy zostało przedstawione, czym jest text mining oraz przetwarzanie języka naturalnego. Ponadto opisano również użyte struktury i algorytmy, takie jak

sieci neuronowe. Opisana została też szczegółowo analiza sentymentu, jej zastosowania w biznesie oraz różne jej rodzaje i implementacje. Analiza została zaimplementowana w języku R.

Do analizy danych tekstowych powszechnie stosowane są algorytmy uczenia maszynowego oraz algorytmy NLP (ang. *natural language processing*) (Stedman, 2020). Przetwarzanie języka naturalnego jest szybko rozwijającą się dziedziną służącą do m.in. filtrowania spamu, wykrywania mowy nienawiści w tekście czy też badania, czy dany komentarz był nacechowany pozytywnie czy negatywnie (Stedman, 2020). Uczenia maszynowego używa się do m.in. predykcji cen akcji na rynku, określenia, czy klient zrezygnuje z usług, czy też przypisania klienta do właściwej grupy konsumenckiej.

Analiza danych oraz tworzenie modeli predykcyjnych wiązą się również z przetwarzaniem dużej ilości danych. Oprócz używania typowych, ustrukturyzowanych danych znajdujących się w bazach relacyjnych coraz częściej wymagane jest przetwarzanie danych nieustrukturyzowanych w postaci plików, zdjęć, filmów, rejestrów sensorów, danych tekstowych czy też logów systemowych. Rozwiązania big data w dzisiejszych czasach najczęściej opierają się o obliczenia chmurowe (Berisha, Meziu i Shabani, 2022). Platformy chmurowe zwalniają organizacje z odpowiedzialności za fizyczne utrzymywanie centrów danych (ang. *data centres*) (Schmelzer, 2022). Ponadto organizacje takie jak Amazon czy Microsoft pozwalają firmom na dostosowanie opłaty za udostępnianie zasobów do ich potrzeb. Platformy takie jak AWS czy Azure pozwalają też na przechowanie różnych danych w tzw. jeziorach danych (ang. *data lake*). Jest to rozwiązanie pozwalające na przechowywanie danych zarówno ustrukturyzowanych, jak i nieustrukturyzowanych (takich jak: pliki, tabele NoSQL czy też kolejki) w swojej pierwotnej postaci (Schmelzer, 2022). Wykorzystanie platformy chmurowej mogłoby zostać użyte w przyszłości, gdyby zdecydowano się na stworzenie systemu opartego o zaimplementowaną w pracy analizę sentymentu.

W systemie wykorzystującym np. platformę Azure (i jego infrastrukturę, jak np. jezioro danych Azure Data Lake Storage Gen2), odpowiedzialność za utrzymanie infrastruktury zostałaby przerzucona na dostawcę usług. Do budowy tego systemu składającego się z interfejsu graficznego użytkownika, silnika przetwarzającego dane oraz modelu sieci neuronowej zdecydowano się na użycie języka programowania R. Istotną kwestią jest też, by system ten był w stanie przetwarzać duże ilości danych, tak, aby móc w stanie poddać analizie jak najwięcej komentarzy w jak najkrótszym czasie.

W pierwszym rozdziale pracy opisano teoretyczne podstawy pracy, wyjaśniono, czym jest NLP oraz text mining. Drugi rozdział jest natomiast poświęcony uczeniu maszynowemu i sieciom neuronowym. W rozdziale tym zostały opisane zagadnienia uczenia maszynowego, metryki, których używa się do badania dopasowania modelu. Trzeci rozdział jest poświęcony analizie

sentymetu. W tej części pracy przedstawione są: definicja, zastosowanie oraz rodzaje analizy sentymetu. Czwarty rozdział jest poświęcony implementacji analizy sentymetu, przedstawione zostały jej kroki. Ponadto rozdział dotyczy również modelu analizy sentymetu bazującego na sieci neuronowej, pokazano w tej części pracy architekturę sieci oraz przedstawione zostały jej metryki i statystyki dopasowania. W rozdziale tym podana jest także szczegółowa lista bibliotek i pakietów użytych do implementacji analizy sentymetu. Adres repozytorium projektu znajduje się pod adresem: <https://github.com/HomeSeeker88/MastersThesis>

1. Text mining oraz algorytmy i struktury w NLP

1.1 Text mining

Text mining jest to transformacja nieustrukturyzowanych danych tekstowych w uporządkowany format w celu otrzymania wartościowych informacji, które pozwalają na zrozumienie tekstu (Stedman, 2020). Ponadto przy użyciu różnych narzędzi statystycznych czy algorytmów uczenia maszynowego, jak maszyna wektorów nośnych czy deep learning, można badać związek między danymi w tekście. Same dane mogą przyjmować format:

- ustrukturyzowany,

Dane są w ustandaryzowanej formie tabelarycznej, składającej się z wielu wiersów i kolumn. Wówczas dane tekstowe stanowią wartość jednej z kolumn. W takiej formie łatwiej przechowywać i procesować dane, można dokonywać analizy czy też budować modele predykcyjne (Stedman, 2020).

- nieustrukturyzowany,

Takie dane nie mają z góry zdefiniowanego formatu, może to być tekst z różnych źródeł, jak media społecznościowe recenzje produktów czy formaty multimedialne, np. pliki audio czy wideo (Stedman, 2020).

- pół-ustrukturyzowany.

Jest to pewne połączenie danych ustrukturyzowanych i nieustrukturyzowanych, dane są w pewny sposób uporządkowane, ale nie spełniają kryteriów relacyjnej bazy danych. Przykładem takiego formatu są pliki XML, JSON czy też HTML, dane tekstowe stanowią wówczas wartość jednego z pól w takim pliku (Stedman, 2020).

Szacuję się, że ok. 80% danych na świecie jest w formie nieustrukturyzowanej, co czyni text mining wyjątkowo cennym narzędziem do analizy danych tekstowych (Stedman, 2020). Oprócz badania korelacji między słowami w tekście, text mining skupia się też na częstotliwości ich występowania oraz powtarzalnych wzorcach, jakie występują w tekstowym zbiorze danych (Stedman, 2020). Text mining wyróżnia algorytmy, które nie wymagają interakcji człowieka czy też znajomości semantyki języka ze strony analityka. Składa się on z metod, wśród których można wyróżnić następujące pojęcia:

- klasyfikacja tekstu,
- ekstrakcja tekstu,
- klasteryzacja tekstu.

Klasyfikacja tekstu jest procesem polegającym na przypisaniu danym tekstowym odpowiedniej kategorii (tzw. tagu). W ten sposób umożliwia się analizowanie różnych źródeł danych i uzyskiwanie wartościowych wniosków w szybki i mało kosztowny sposób

(MonkeyLearn, 2022). Do najważniejszych czynności klasyfikacji tekstu należą m.in. analiza sentymentu oraz rozpoznanie języka. Analiza sentymentu jest procesem, w którym rozpoznaje się emocjonalny wydźwięk testu. W ten sposób istnieje możliwość zautomatyzowanego sklasyfikowania, czy opinia jest pozytywna bądź negatywna (MonkeyLearn, 2022). Rozpoznanie języka polega na odpowiednim rozpoznaniu języka, w którym tekst został napisany. Zastosowanie rozpoznania języka można znaleźć w centrach obsługi, gdzie zgłoszenie wysłane przez użytkownika, które dotyczyło jakiejś usterki lub prośby o nadanie dostępu, może zostać szybko skierowane do odpowiedniego regionu.

Ekstrakcja tekstu to technika, w której ze źródła danych tekstowych są wyciągane specyficzne elementy: słowa kluczowe, nazwy własne, adresy zamieszkania, adresy e-mail itp. Dzięki temu czasochłonne i monotonne manualne wyszukiwanie takich informacji może zostać w pełni zautomatyzowane. Ekstrakcja tekstu występuje najczęściej w połączeniu z jego klasyfikacją. Do głównych czynności ekstrakcji tekstu należą: wyszukiwanie słów kluczowych, rozpoznawanie nazw własnych oraz wykrywanie cech (MonkeyLearn, 2022). Słowa kluczowe są to najbardziej istotne wyrazy w tekście, które mogą zostać użyte do podsumowania zbioru tekstowego. Formą podsumowania tekstu jest np. chmura słów, czyli graficzne zobrazowanie zawartości tekstu, w którym największe znaczniki oznaczają najistotniejsze wyrazy. Identyfikacja nazw własnych jest to metoda służąca do rozpoznawania imion, nazw firm czy organizacji w tekście. Wykrywanie cech jest to technika służąca do rozpoznawania określonych charakterystyk w tekście. Jest ona często używana do analizy opisu produktów, gdzie za pomocą tej techniki można z tekstu wyciągnąć informacje o takich atrybutach, jak: kolor, marka, model czy też cena (MonkeyLearn, 2022).

Klasteryzacja tekstu jest to proces służący do pogrupowania nieprzypisanych źródeł danych w taki sposób, by w jednej grupie znajdowały się teksty podobne do siebie w większym stopniu niż zbiory tekstowe z innych grup (Kunwar, 2013). W tej implementacji dokumenty mogą być reprezentowane jako wektory cech. Podobieństwo między tekstami jest obliczanie poprzez mierzenie odległości między tymi cechami. Obiekty będące blisko siebie powinny wówczas należeć do jednego klastra, w przypadku gdy ta odległość jest większa – wówczas te źródła tekstowe powinny należeć do dwóch innych grup (Kunwar, 2013). Klasteryzacja tekstu uwzględnia trzy aspekty:

- wybór odpowiedniej miary dystansu do identyfikacji bliskości pomiędzy dwoma wektorami cech;
- funkcje kryterium, która pozwala obliczyć odpowiednie dopasowanie klastrów;
- algorytm optymalizujący funkcję kryterium.

Klasteryzacja tekstu znajduje zastosowanie w:

- identyfikacji fake newsów (rozpoznanie, czy informacja jest prawdziwa bądź fałszywa),
- filtrowaniu spamu (rozpoznawanie niechcianych wiadomości wysłanych pocztą elektroniczną),
- tłumaczeniu tekstu (translacja tekstu z jednego języka na drugi),
- generowaniu taksonomii,
- analizie zgłoszeń w centrum wsparcia (identyfikacja i przypisanie zgłoszeń do odpowiedniej grupy przyjętych wcześniej raportów na podstawie podobieństwa w tekście) (Kunwar, 2013).

W ten sposób organizacje mogą znaleźć potencjalnie cenne spostrzeżenia w źródłach danych: firmowych dokumentach, wiadomościach od klientów, rejestrach rozmów z centrali rozmów, ankietach, recenzjach, wpisach użytkowników na portalach mediów społecznościowych, zapisach z placówek medycznych dotyczące czy innych źródłach danych tekstowych (Stedman, 2020).

1.2 Definicja NLP

Przetwarzanie języka naturalnego (ang. *natural language processing* – NLP) jest to zbiór technik pozwalających na zrozumienie języka przez komputer. W przetwarzaniu tym wyróżnia się dwa etapy (Sreemany, 2021):

- preprocessing danych,
- rozwój algorytmu.

Preprocessing danych to zbiór metod służących do przygotowania oraz oczyszczenia danych dla komputera w celu przygotowania ich formatu w taki sposób, by można było na nich pracować. Wśród metod składających się na preprocessing można wyróżnić (Sreemany, 2021):

- tokenizację,
- eliminowanie słów stop listą,
- stemming,
- dzielenie na części zdania.

Metoda tokenizacji polega na dzieleniu tekstu na mniejsze części i zostanie szerzej opisana w dalszej części pracy. Powstałe w wyniku podziału tzw. tokeny służą do zbudowania słownika i mogą to być odpowiednio: słowa, znaki lub fragmenty słów, przy czym słownik jest to zbiór wszystkich unikatowych tokenów. Dzielenie całego tekstu na słowa jest najpowszechniej stosowanym algorytmem tokenizacji. Problemem takiego rozwiązania jest natomiast tzw. problem słów spoza słownika (ang. *OOV words* – Out Of Vocabulary) (Pai, 2020). Problem ten dotyczy przypadku, gdy w zbiorze testowym znajdują się słowa spoza słownika powstałego w wyniku tokenizacji zbioru treningowego. Do możliwych rozwiązań należy zebranie ze zbioru testowego

tw. nieznanych tokenów (ang. *unknown tokens* – UNK), czyli tokenów brakujących w zbiorze treningowym, które występują w zbiorze testowym (Pai, 2020). W następnym kroku dokonuje się selekcji najczęściej występujących tokenów, natomiast rzadziej występujące słowa są zastępowane nieznanymi tokenami. W ten sposób eliminowany jest problem z nieznanymi wyrazami podczas przetwarzania zbioru testowego. Ograniczeniem tego rozwiązania jest natomiast częściowa utrata informacji podczas odrzucania rzadziej występujących słów, które mogą mieć wysoki stopień istotności zgodnie z prawem Zipfa. Według prawa Zipfa, częstotliwość występowania termu jest odwrotnie proporcjonalna do jej rangi statystycznej (Silge i Robinson, 2017). Innym ograniczeniem dzielenia całego tekstu na słowa jest złożoność obliczeniowa algorytmu. Zbiory treningowe, jeszcze nie przetworzone, mogą być bardzo obszernymi korpusami (ang. *corpus*). W rezultacie obliczenie częstotliwości każdego unikatowego tokenu dla tak dużego korpusu może obciążyć pracę komputera. Oba ograniczenia tokenizacji według wyrazów mogą zostać rozwiązane przy pomocy tokenizacji znakowej. Wówczas zdanie „Nauka uczenia maszynowego jest interesująca” zostanie rozbite podczas tokenizacji znakowej na sekwencję: [„N”, „a”, „u”, „k”, „a”, „u”, „c”, „z”, „e”, „n”, „i”, „a”, „m”, „a”, „s”, „z”, „y”, „n”, „o”, „w”, „e”, „g”, „o”, „j”, „e”, „s”, „t”, „i”, „n”, „t”, „e”, „r”, „e”, „s”, „u”, „j”, „a”, „c”, „a”]. Każdemu unikatowemu tokenowi (w tym przypadku są to znaki) w następnej kolejności jest przypisany identyfikator (Kanna, 2021). Zaletą tego rozwiązania jest znacznie mniejszy słownik, alfabet w języku angielskim czy łacińskim liczy 26 liter, co pozwala na dużą oszczędność zasobów pamięciowych komputera (Kanna, 2021). Ponadto kwestia nieznanych tokenów w ten sposób też jest rozwiązana przez zachowanie informacji pochodzącej ze słowa. Dzieje się tak poprzez rozbicie takiego słowa na znane już tokeny, przez co nie dochodzi do utraty informacji. Do ograniczeń takiego rozwiązania należy natomiast fakt, że rozbudowane zdania czy fragmenty tekstu wpływają na powstawanie bardzo dużych sekwencji znaków, co utrudnia zbadanie relacji pomiędzy znakami w celu skonstruowania poprawnych słów (Pai, 2020). Ostatnim typem tokenizacji jest dzielenie tekstu na podśłowa, np. wówczas słowo „smartest” zostałoby podzielone na dwa tokeny „smart” i „est” (Pai, 2020).

Eliminacja słów stop listą polega na usunięciu ze źródła danych wyrazów, które znajdują się w tzw. stop liście (ang. *stop words*) (Silge i Robinson, 2017). Najczęściej takimi słowami są takie części mowy, jak: rodzajniki, przyimki, zaimki czy też spójniki. W języku angielskim przykładami takich słów będą takie wyrazy, jak: „the”, „an”, „where”, „on”, „because” – nie przenoszą one dużo informacji w zdaniu, dlatego warto je eliminować w celu pozostawienia słów, które zawierają dużo informacji (Kanna, 2021). Kolejną zaletą tego rozwiązania jest mniejszy słownik, co pozwala na oszczędność zasobów obliczeniowych komputera. Ograniczeniem tego rozwiązania natomiast jest fakt, iż różne stop listy mogą zawierać inne słowa, zbiory takich słów

są udostępniane przez takie biblioteki programistyczne, jak *tidytext*, *scikit-learn* czy też *nlk*. Wpłynąć to może na różniące się od siebie wyniki analizy tekstowej.

Stemming polega zaś na usuwaniu formantów ze słów w celu uproszczenia zbioru tekstowego, dzięki temu pobrany jest sam rdzeń słowa (jest to najmniejsza część wyrazu niosąca ze sobą znaczenie i której nie można podzielić na mniejsze jednostki znaczeniowe). Jako przykład można podać słowa „programming” i „programmer” (Kibble, 2013). Słowa te nie mają takiego samego znaczenia, lecz wskazują na tę samą dziedzinę, więc można je uprościć poprzez stemming. Wówczas otrzymane w rezultacie jest słowo „programm”, które będzie częściej się powtarzać oraz, z racji swojej krótszej nazwy, będzie też mniej kosztowne pamięciowo i obliczeniowo dla komputera (Kibble, 2013). Najczęściej stosowaną implementacją stemmingu jest algorytm Portera (Manning, Schütze i Raghavan, 2009). Algorytm ten składa się z 5 kroków upraszczania danego słowa, które następują bezpośrednio po sobie. W każdym kroku należy wybrać zasadę uproszczenia końcówki słowa dla najdłuższego przyrostka. Jako przykład pierwszego kroku można podać następującą grupę zasad (Manning, Schütze i Raghavan, 2009):

$$SSES \rightarrow SS$$
$$IES \rightarrow I$$
$$SS \rightarrow SS$$
$$S \rightarrow S$$

oraz grupę słów, które zostaną poddane transformacji według tych zasad:

$$caresses \rightarrow caress$$
$$ladies \rightarrow ladi$$
$$princess \rightarrow princess$$
$$dogs \rightarrow dog$$

W późniejszych krokach algorytmu często stosowana jest tzw. miara słowa. Koncept ten polega na sprawdzaniu liczby sylab w słowie w celu określenia, czy końcówka słowa jest faktycznie sufiksem, który należy usunąć (Manning, Schütze i Raghavan, 2009). Przykładem zastosowania zasady z użyciem miary słowa może być następująca reguła:

$$(m > 1) EMENT \rightarrow$$

oraz następujące słowa poddane transformacji:

$$achievement \rightarrow achiev$$
$$cement \rightarrow cement$$

Słowo „cement” nie zostało poddane skróceniu ze względu na to, że słowo to składa się z jednej sylaby, gdyby nie uwzględniono miary słowa w zasadzie transformacji, wynikiem zostałoby „c”, które doprowadziłoby do utraty ważnych informacji (Manning, Schütze i Raghavan, 2009).

Dzielenie na części zdania polega natomiast na przypisaniu słowom w tekście odpowiednich poziomów będącymi takimi częściami zdania, jak np. podmiot, orzeczenie czy dopełnienie. Zaletą tego rozwiązania jest łatwiejsza identyfikacja logiki zdania oraz przechwycenie większej ilości informacji. Ograniczeniem tej metody jest natomiast większa złożoność obliczeniowa, w szczególności dotyczy to źródeł tekstu ze zdaniami wielokrotnie złożonymi.

Następnym etapem przetwarzania języka naturalnego po preprocessingu danych jest rozwój i opracowanie algorytmu. Do systemów analizujących tekst należą (Dorash, 2017):

1. Systemy oparte na regułach

Ten rodzaj klasyfikacji opiera się na opracowanych przez człowieka lingwistycznych regułach pomiędzy specyficznym wzorcem lingwistycznym a klasą odpowiedzi. W momencie gdy algorytm zostanie zaimplementowany, system będzie w stanie zaklasyfikować różne struktury językowe i przypisać je do odpowiedniej grupy (Dorash, 2017). Reguły bazują na wzorcach składniowych, morfologicznych oraz leksykalnych. Zasady te mogą też być związane z aspektami semantycznymi oraz fonologicznymi. Przykładową regułą może być następujący przypadek:

(Czarny | Biały | Niebieski | Zielony | Czerwony) → Kolor

Na podstawie tej reguły system przypisze kategorię „Kolor” w momencie, gdy jakikolwiek z podanych w regule wyrazów pojawi się w tekście. Systemy oparte na regułach są łatwe do zrozumienia, ponieważ są opracowane przez człowieka. Aczkolwiek dodawanie nowych reguł do działającego systemu wymaga przeprowadzenia dużej ilości testów w celu sprawdzenia, czy nowa reguła nie wpływa na działanie predykcji pozostałych zasad (Dorash, 2017). W rezultacie wadą systemów opartych na regułach jest problem ze skalowalnością rozwiązania. Kolejnym ograniczeniem tych systemów jest również wymaganie posiadania konkretnej znajomości lingwistyki i danych, które należy analizować.

2. Systemy oparte na uczeniu maszynowym

Te systemy bazują na algorytmach uczenia maszynowego. Uczenie maszynowe zostanie objaśnione dokładniej w dalszej części pracy.

3. Systemy hybrydowe

Są to systemy decyzyjne wykorzystujące zarazem reguły, jak i algorytmy uczenia maszynowego. Pozwala to na poprawę osiąganych rezultatów.

1.3 Analiza sentymentu oraz jej rodzaje

Analiza sentymentu polega na wykorzystaniu przetworzenia języka naturalnego, analizy tekstu, lingwistyki komputerowej do systematycznego identyfikowania, wyodrębniania, określania emocjonalnego wydźwięku źródła tekstowego (tj. czy wydźwięk ten jest pozytywny, negatywny

bądź neutralny). Znaczenie analizy sentymentu jest tak duże ze względu na to, że może ona być w pełni zautomatyzowana, dzięki czemu oszczędza to dużo pracy i czasu człowiekowi. Ponadto jest to narzędzie o rosnącej popularności, które używane jest w takich dziedzinach, jak: e-commerce, marketing (badanie satysfakcji klientów z produktu), polityka (badanie nastrojów społecznych) i badania rynku. Ogromnym źródłem takich danych są różne platformy mediów społecznościowych: Twitter, Facebook, Reddit, gdzie codziennie miliony użytkowników udzielają wpisów na różne tematy (Silge i Robinson, 2017).

Przeprowadzanie analizy sentymentu należy zawsze zacząć od stworzenia lub pobrania leksykonu (czyli specjalnego zbioru danych) zawierającego poszczególne wyrażenia i ich scoring sentymentu (jest to metryka określająca emocjonalny wydźwięk słowa). W następnym kroku algorytm sam zaczyna po podziale tekstu na określone części (takie jak np. n-gramy) klasyfikować wydźwięk emocjonalny danego fragmentu tekstu. Algorytm wówczas dokonuje bilansu kontrybucji na podstawie każdego poszczególnego słowa w analizowanym fragmencie i podaje wynik sentymentu. Zakres wyników może przyjmować różny charakter: od ciągłego, gdzie wartości poniżej 0 odpowiadają negatywnemu wydźwiękowi. Wyniki powyżej 0 natomiast wskazują na coraz bardziej pozytywny ton wypowiedzi. Wyniki też mogą być w postaci dyskretnej; wówczas wszystkie negatywne słowa są w kategorii negatywnej (lub -1), pozytywne zaś w kategorii pozytywnej (lub 1).

Głównymi wyzwaniem i trudnościami podczas przeprowadzania analizy sentymentu są:

- zaprzeczenia,
- złożone zdania o kontrastującym wydźwięku,
- sarkazm,
- rozpoznanie nazw własnych (ang. *named-entity recognition*).

Zaprzeczeniami mogą być różne zdania z takim wyrazami, jak „nie”, „niezbyt” czy „wcale”. Przykładem tego może być następujące zdanie złożone: „Biorąc pod uwagę wszelkie wydarzenia, należy uznać dzisiejszy dzień za udany”. Wskazuje ono na pozytywny wydźwięk wypowiedzi, która powinna być bez problemu poprawnie sklasyfikowana. Dla zdania: „Biorąc pod uwagę wszelkie wydarzenia, wcale nie należy uznać dzisiejszego dnia za udany”, z kolei już nie można być pewnym poprawności wyniku analizy. Przyczyną tego jest fakt, iż przy sumowaniu lub uśrednieniu wyniku sentymentu poszczególnych słów, wynik może wskazywać na pozytywny ton wypowiedzi pomimo tego, że w rzeczywistości to zdanie miało negatywny wydźwięk emocjonalny.

Innym ograniczeniem analizy sentymentu są sytuacje, gdy w zdaniu występują słowa o kontrastującym wydźwięku, co może skutkować w błędnym wyniku analizy. Przykładem może być następujące zdanie: „Dzisiaj rano czułem się okropnie, ale mimo to na przyjęciu bawiłem się

przednio”. Takie zdanie stanowi problem dla modelu ze względu na obecność słów o bardzo rozbieżnym wydźwięku. Występują tu słowa oraz człony o bardzo pozytywnym wydźwięku, jak: „bawiłem się” i „przednio”, ale też jest słowo „okropnie”, które ma bardzo negatywny wynik sentymentu.

Dużym problemem w analizie sentymentu są również opinie zawierające ironię bądź sarkazm. Dzieje się tak, ponieważ wydźwięk emocjonalny nie jest przenoszony dosłownie w wyrazach, dlatego też zdanie: „Bardzo się cieszę, że mój pociąg przyjedzie z dwugodzinnym opóźnieniem”, może zostać z dużym prawdopodobieństwem źle sklasyfikowane, w szczególności przez prostsze implementacje analizy sentymentu.

Ze względu na czynnik wykrywany w analizie sentymentu można wyróżnić jej poszczególne rodzaje, takie jak:

- standardowa analiza sentymentu,
- szczegółowa analiza sentymentu,
- wykrywanie emocji,
- analiza sentymentu bazująca na aspekcie,
- wykrywanie zamiaru.

Standardowa analiza sentymentu identyfikuje szczegóły w opinii i dokonuje jej klasyfikacji. Jest to najpowszechniejsza odmiana analizy sentymentu. Dla przykładowego zdania: „Produkt x jest niezawodny, bardzo polecam stosowanie go”, standardowa analiza sentymentu powinna zwrócić wynik: „Opinia pozytywna”. Natomiast dla zdania: „Produkt y jest wadliwy oraz bardzo drogi, oceniam go źle”, analiza ta powinna zwrócić wynik: „Opinia negatywna”.

Szczegółowa analiza sentymentu (ang. *fine-grained sentiment analysis*) jest rozszerzeniem standardowej analizy sentymentu. W przeciwieństwie do standardowej analizy sentymentu szczegółowa analiza sentymentu dokonuje też analizy polaryzacji opinii. W rezultacie wynikiem analizy danego tekstu może być 5 klas takich jak: „Opinia bardzo pozytywna”, „Opinia pozytywna”, „Opinia neutralna”, „Opinia negatywna” czy „Opinia bardzo negatywna”. Szczegółowa analiza sentymentu znajduje zastosowanie w badaniu ocen produktów oraz recenzji (Roldós, 2020).

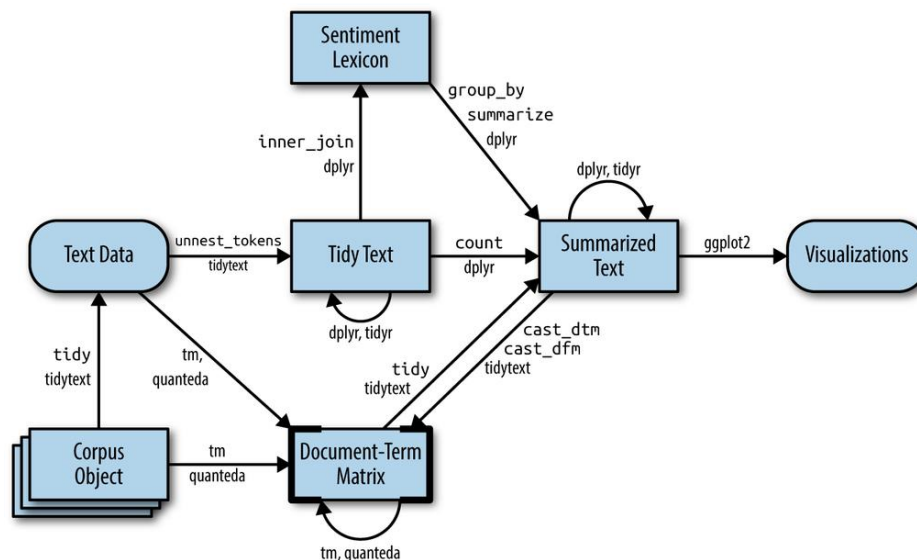
Analizy sentymentu, która opiera się na wykrywaniu emocji używa się do rozpoznania emocji autora tekstu. W przeciwieństwie do standardowej i szczegółowej analizy sentymentu wykrywanie emocji jako rezultaty wyjściowe przyjmie wartości takie jak: smutek, radość, złość, strach czy zmartwienie. Systemy oparte na wykrywaniu emocji wykorzystują do tego specjalne leksykony – czyli zbiory określonych słów, które mają pewny wydźwięk emocjonalny. Innym rozwiązaniem stosowanym do wykrycia określonych emocji w tekście są algorytmy uczenia maszynowego (Roldós, 2020). Uczenie maszynowe w wykryciu emocji może być skuteczniejsze

od korzystania z leksykonów ze względu na fakt, iż autor tekstu może przekazać swoje emocje w sposób niejednoznaczny. Przykładowe zdanie: „Ten lek mnie wykańcza, stanowczo go nie polecam”, może przenosić takie emocje, jak: ból, złość bądź smutek. Słowo „wykańcza” może zostać zinterpretowane w różny sposób, dlatego skorzystanie z leksykonu w tym przypadku może doprowadzić do błędnego wykrycia emocji autora (Roldós, 2020).

Analiza sentymentu bazująca na aspekcie polega na rozszerzeniu analizy sentymentu o dokonanie oceny przedmiotu opinii autora. Jest to szczególnie użyteczne w przypadku badania opinii klienta o produkcie, gdyż można w ten sposób określić, co jest przyczyną satysfakcji klienta lub z czego był on niezadowolony. Zaletą tego rozwiązania jest również jego skalowalność: analiza sentymentu bazująca na aspekcie pozwala na zautomatyzowaną analizę dużej ilości danych tekstowych, co pozwala na oszczędność zasobów finansowych oraz czasowych. Inną korzyścią analizy sentymentu bazującej na aspekcie jest możliwość dokonywania analizy w czasie rzeczywistym. Dzięki temu można w znacznie szybszy sposób wykryć niepożądane efekty i podjąć kroki w celu ich eliminacji. Oprócz możliwości szybszej reakcji na zdanie klienta analiza sentymentu bazująca na aspekcie pozwala na lepsze zrozumienie produktu oraz na lepszą analizę potrzeb klienta (Pascual, 2019).

Wykrywanie zamiaru w analizie sentymentu polega na wykryciu zamiaru lub planów autora tekstu oraz poprawnej klasyfikacji tego celu. Rozwiązanie to jest stosowane w oprogramowaniu wykorzystującym chatbot oraz systemach inteligentnego dialogu, gdzie bardzo istotna jest poprawna interpretacja potrzeb i zamiarów użytkownika. Do analizy zamiaru używa się transformerów, czyli modeli sieci neuronowych pozwalających na wyliczenie istotności poszczególnych elementów wejściowych w zbiorze danych. Inną strukturą wykorzystywaną do analizy zamiaru jest kapsułowa sieć neuronowa. Ten model jest używany do modelowania hierarchicznych związków, neurony w tej sieci to kapsuły, które są wektorami prawdopodobieństw wystąpienia określonej cechy (Pascual, 2019).

1.4 Definicja macierzy Document-Term i Term-Document



Wykres 1: Schemat działania metod biblioteki tidytext przy pracy z macierzami DTM i TDM
Źródło: (Silge i Robinson, 2017)

W eksploracji tekstu do jednych z najczęściej używanych struktur analitycznych należą macierze Document-Term i Term-Document. Macierz Document-Term jest to macierz, w której:

- każdy wiersz reprezentuje dokument (czyli źródło tekstowe takie jak książka czy artykuł);
- każda kolumna reprezentuje term (czyli słowo);
- pojedyncza wartość w macierzy reprezentuje liczbę wystąpień danego termu (czyli wyrazu) w danym dokumencie.

Ponieważ taka macierz przeważnie składa się z zer, macierze DTM są głównie implementowane w postaci rzadkiej (ang. *sparse matrix*). Macierze rzadkie są często używaną strukturą w obliczeniach i informatyce ze względu na łatwość w ich skompresowaniu, co skutkuje mniejszym obciążeniem pamięciowym. Mierzenie rzadkości (ang. *sparsity*) takiej macierzy jest dane wzorem:

$$Rzadkość = \frac{\text{Liczba zer w macierzy}}{\text{Liczba wszystkich wartości w macierzy}} \times 100\%$$

Transponowana macierz DTM jest nazywaną macierzą Term-Document. Wtedy w to w danej kolumnie znajdują się licznosci poszczególnych termów w danym dokumencie (Kibble, 2013). Wówczas w bardziej wygodny sposób można wyliczać takie metryki, jak: częstotliwość występowania termu *tf* (term-frequency), *idf* (inverse document frequency) czy też współczynnik *tf-idf*.

Macierze DTM czy TDM znajdują zastosowanie w realizacji różnych zadań w przetwarzaniu języka naturalnego. Do takich czynności należą:

- rozbiecie tekstu i dokonanie atomizacji, w ten sposób dzięki macierzom DTM i TDM można

poprawić działanie silników wyszukiwania wyrazów poprzez ujednoznaczenie słów o różnym znaczeniu i wyszukiwanie synonimów (Verma, 2021);

- ekstrakcja i analiza danych behawioralnych, w ten sposób poprzez przeprowadzanie wielowymiarowej analizy macierzy DTM można określić wiele tematów poruszonych w danych (Verma, 2021).

1.5 N-gramy

Jedną z technik działania na ogromnych i skomplikowanych zbiorach tekstowych jest dzielenie tekstu na tzw. n-gramy. N-gramy są to sekwencje składające się z n ilości słów. Pozwala to na łatwiejsze operowanie na tekście i ograniczenie czasowe działania algorytmów. Do uproszczenia tekstu używa się także stop listy (Silge i Robinson, 2017). Stop lista jest to zbiór słów, które nie mają same w sobie znaczenia, więc nie są potrzebne w analizie znaczenia tekstu (Silge i Robinson, 2017). Są to różnego rodzaju spójniki i przyimki, które służą do nadania logicznego sensu wypowiedzi, lecz nie nadają jej żadnego znaczenia merytorycznego, tonu lub wydźwięku. Liczba n-gramów w jednym zdaniu lub fragmencie tekstu K można określić wzorem:

$$Ngramy_K = X - (N - 1)$$

gdzie X oznacza liczbę słów w zdaniu K .

W modelach i analizie NLP najczęściej używa się bigramów (sekwencji dwuwyrazowych) oraz trigramów (sekwencji trójwyrazowych). N-gramy są używane m.in. do budowania języków, gdzie sprawdzana jest poprawność pisowni, czy też skracania źródeł tekstowych w celu pozbycia się niepotrzebnych wyrazów. Za przykład można podać zdanie: „Dzisiaj przewidywane są przelotne opady”. Gdy rozbije się to zdanie na bigramy, otrzyma się wówczas:

- „Dzisiaj przewidywane”
- „przewidywane są”
- „są przelotne”
- „przelotne opady.”

Zgodnie ze wzorem ze zdania składającego się z 5 słów otrzymano 4 bigramy (Silge i Robinson, 2017). Z kolei gdy to zdanie rozbije się na trigramy, powstaną wówczas poszczególne człony:

- „Dzisiaj przewidywane są”
- „przewidywane są przelotne”
- „są przelotne opady”.

Zgodnie ze wzorem ze zdania składającego się z 5 słów otrzymano 3 trigramy (Silge i Robinson, 2017).

1.6 Algorytm LDA

Jednym z zagadnień dziedziny, jaką jest text mining, jest modelowanie tematyki (ang. *topic modeling*). Polega ono na znajdowaniu wśród dokumentów grup o zbliżonej tematyce. Jedną z najbardziej powszechnie stosowanych metod w modelowaniu tematyki, jest algorytm LDA (*Latent Dirichlet Allocation*) (Silge i Robinson, 2017). W algorytmie tym występują dwie zasady:

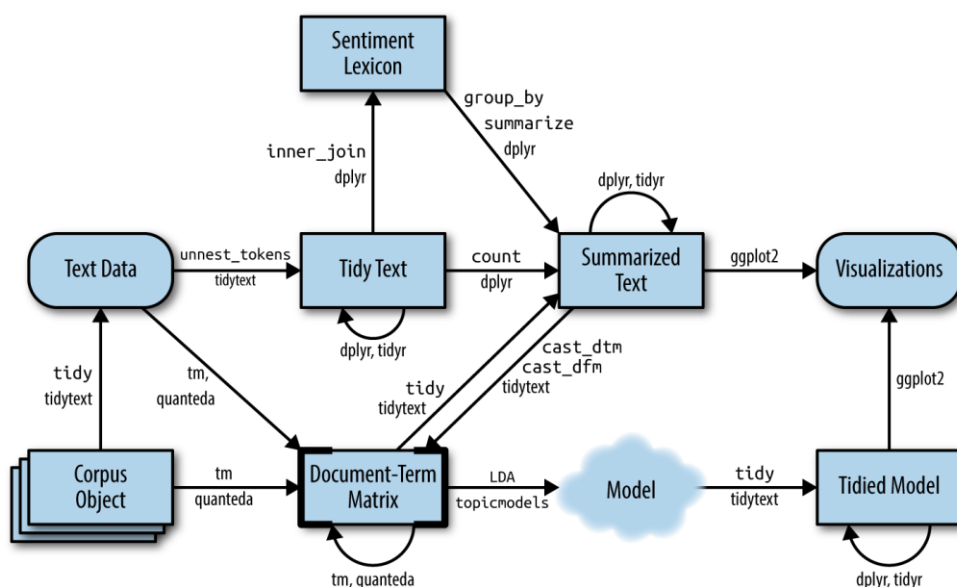
1. Każdy dokument jest zbiorem tematów.

Należy przyjąć, że każdy zbiór tekstowy może zawierać słowa o różnej tematyce w szczególnych proporcjach. Przykładem może być dwutematyczny model, wówczas zbiory tekstowe mogą zostać sklasyfikowane w następujący sposób (Silge i Robinson, 2017):

- Dokument 1 jest w 80% poświęcony tematyce A oraz w 20% tematyce B.
- Dokument 2 jest w 30% poświęcony tematyce A oraz w 70% tematyce B.

2. Każdy temat jest zbiorem słów.

Przykładem może być dwutematyczny model do badania wiadomości w portalu informacyjnym, który dzieli wiadomości na dwie kategorie: „polityka” i „rozrywka”. Do najczęściej pojawiających się słów w informacjach poświęconych polityce będą takie wyrazy, jak: „prezydent”, „parlament”, „rząd”, „ministrowie”. Z kolei wśród wiadomości pochodzących ze świata rozrywki słowami występującymi bardzo często będą: „aktor”, „gwiazda”, „telewizja”, „serial”, „piosenkarz”. Bardzo istotnym faktem jest to, że słowa mogą występować w dwóch tematach jednocześnie, np. wyraz „pieniądze” może należeć do obu kategorii (Silge i Robinson, 2017).



Wykres 2: Schemat działania metod biblioteki tidytext w języku R przy działaniu z algorytmem LDA

Źródło: (Silge i Robinson, 2017)

LDA jest matematyczną metodą, która jednocześnie określa grupę słów powiązaną z danym tematem oraz określa zbiór tematów, które opisują dany dokument. Do poddania dokumentów analizie należy najpierw je poddać preprocessingowi oraz tokenizacji, a następnie przekształcić w macierz DTM składającą się z m dokumentów i n unikatowych wyrazów. W następnym kroku macierz DTM zostaje przekształcona w dwie macierze: macierz DocumentTopic oraz macierz TopicWord. Macierz DocumentTopic zawiera możliwe k tematów jako kolumny, które dany dokument może zawierać. Macierz TopicWord składa się ze słów, które mogą występować w danej tematyce. Algorytm wykorzystuje również dwa parametry do kontroli rozkładu (Seth, 2021):

- α (alfa), prawdopodobieństwo przynależności tematu do dokumentu;
- β (beta), prawdopodobieństwo przynależności słowa do tematu.

Końcowym celem algorytmu LDA jest znalezienie najbardziej optymalnej reprezentacji macierzy DocumentTopic oraz TopicWord, w której współczynniki prawdopodobieństwa o najwyższej wartości determinują, do jakiego tematu należy przypisać słowa oraz jakie tematy należy przypisać do danego dokumentu (Seth, 2021).

2. Uczenie maszynowe w analizie tekstowej

2.1 Uczenie maszynowe

Systemy oparte na uczeniu maszynowym bazują na modelach, które zostały nauczone poprzez przetwarzanie prawidłowo sklasyfikowanych danych. W tym celu zbiory treningowe muszą być spójne oraz reprezentatywne, aby osiągnąć model, który będzie dokonywać poprawnej klasyfikacji. Na początku tego procesu dane tekstowe są poddawane tzw. wektoryzacji (Neppali, Caragea i Caragea, 2018). Ze zdefiniowanego zbioru wyrazów dla każdego słowa jest obliczona liczba jego wystąpień w tekście. Przetransformowane informacje wraz z oczekiwanymi predykcjami są przetwarzane przez algorytm uczenia maszynowego (Neppali, Caragea i Caragea, 2018). W ten sposób model poddany etapowi treningu może dokonywać predykcji danych, które nie zostały sklasyfikowane. Do wykorzystywanych algorytmów uczenia maszynowego należą:

- naiwne klasyfikatory bayesowskie,

Klasyfikatory te bazują na teorii prawdopodobieństwa oraz teorii bayesowskiej do zaklasyfikowania tekstu. Dane wówczas są zwektoryzowane do wektorów prawdopodobieństw tekstu należącego do określonej kategorii. Klasyfikatory te są w stanie osiągnąć dobre dopasowanie nawet w przypadku, gdy zbiór treningowy zawiera niewiele danych (Neppali, Caragea i Caragea, 2018).

- maszyna wektorów nośnych,

Algorytm ten dzieli zaklasyfikowane wektory na dwie grupy: grupę, w której znajdują się wektory należące do określonej klasy, i drugą grupę zawierającą te wektory, które do tej klasy nie należą. Algorytm ten osiąga lepsze wyniki od klasyfikatorów bayesowskich, lecz są znacznie trudniejsze w zaimplementowaniu (Dorash, 2017)

- deep learning.

Systemy oparte na sieciach neuronowych – ich działanie zostanie opisane w dalszej części pracy (Dorash, 2017).

Kończącą fazą opracowania, trenowania i testowania modelu jest ewaluacja jego dopasowania. Do oceny modelu klasyfikacji tekstu wykorzystuje się różnego rodzaju metryki:

1. Skuteczność (ang. *accuracy*) to najprostsza i najbardziej podstawowa metryka stosowana do mierzenia dopasowania modelu. Skuteczność jest dana wzorem:

$$Acc = \frac{TP + TN}{TP + TN + FN + FP}$$

gdzie *TP* (ang. *True Positive*) oznacza liczbę poprawnych klasyfikacji pozytywnych (czyli odpowiedzi twierdzącej), a *TN* (ang. *True Negative*) to liczba poprawnych klasyfikacji negatywnych (czyli odpowiedzi przeczącej). *FN* (ang. *False Negative*) jest za to liczbą

błędnych klasyfikacji negatywnych (czyli błędne zaklasyfikowanie obserwacji jako nienależącej do danej klasy), a *FP* (ang. *False Positive*) natomiast jest liczbą błędnych klasyfikacji pozytywnych (czyli będących zaklasyfikowanymi jako należące do danej klasy niezgodnie z prawdą) (Malik, 2020). Oznacza to, że skuteczność jest stosunkiem poprawnych odpowiedzi do wszystkich dokonanych predykcji. Wiarygodność skuteczności modelu ma jedno zasadnicze ograniczenie: w przypadkach, gdy rozkład przynależności obserwacji do klas jest nierównomierny (tj. gdy w jednej klasie może znajdować się o wiele więcej obserwacji niż w innej), może dojść do tzw. paradoksu skuteczności. Paradoks ten polega na sytuacji, w której wysoka skuteczność modelu może nie oznaczać, że model jest dobrze dopasowany (Malik, 2020). Przykładowo, jedna klasa odpowiedzi może zawierać 99% wszystkich obserwacji, wówczas model może klasyfikować te dane poprawnie, ale istnieje zagrożenie błędnej klasyfikacji pozostałych obserwacji. Stanowi to poważne zagrożenie w sytuacjach, gdzie poprawne wykrycie klasy rzadziej występującej jest o wiele istotniejsze. Metrykami, które w sposób bardziej wiarygodny opisują model, są: czułość, swoistość i precyzja.

2. Czułość (ang. *sensitivity* lub *recall*) to metryka, która jest również znana jako współczynnik poprawnych pozytywnych klasyfikacji. Czułość jest dana wzorem:

$$Recall = \frac{TP}{TP + FN}$$

Oznacza to, że czułość jest stosunkiem poprawnych klasyfikacji pozytywnych wskazanych przez model do wszystkich pozytywnych obserwacji. Im wyższa wartość współczynnika czułości, tym model w lepszym stopniu klasyfikuje pozytywne obserwacje (Malik, 2020).

3. Swoistość (ang. *specificity*) to statystyka znana też jako współczynnik poprawnych negatywnych klasyfikacji. Swoistość jest dana wzorem:

$$Specificity = \frac{TN}{TN + FP}$$

Oznacza to, że swoistość jest stosunkiem poprawnych klasyfikacji negatywnych wskazanych przez model do wszystkich negatywnych obserwacji. Wysoka wartość współczynnika swoistości oznacza, że model rozpoznaje negatywne obserwacje na wysokim poziomie (Malik, 2020).

4. Precyzja (ang. *precision*) jest miarą służącą do zbadania, ile pozytywnych klasyfikacji spośród wszystkich pozytywnych klasyfikacji modelu było poprawnych. Precyzja jest dana wzorem:

$$Precision = \frac{TP}{TP + FP}$$

Miary te w bardziej miarodajny sposób opisują dopasowanie modelu, ponieważ oprócz badania poprawności odpowiedzi, uwzględniają również rozkład zmiennej objaśnianej. Dzięki temu można zbadać, czy model np. poprawnie diagnozuje choroby nowotworowe. Wówczas wysoka skuteczność modelu nie jest aż tak istotna jak precyzja, która informuje o tym, ile procent diagnoz choroby okazało się być prawidłowe. W tym przypadku bardzo ważna jest też wysoka czułość modelu, w ten sposób prawie wszyscy pacjenci z chorobą zostaliby poprawnie zdiagnozowani.

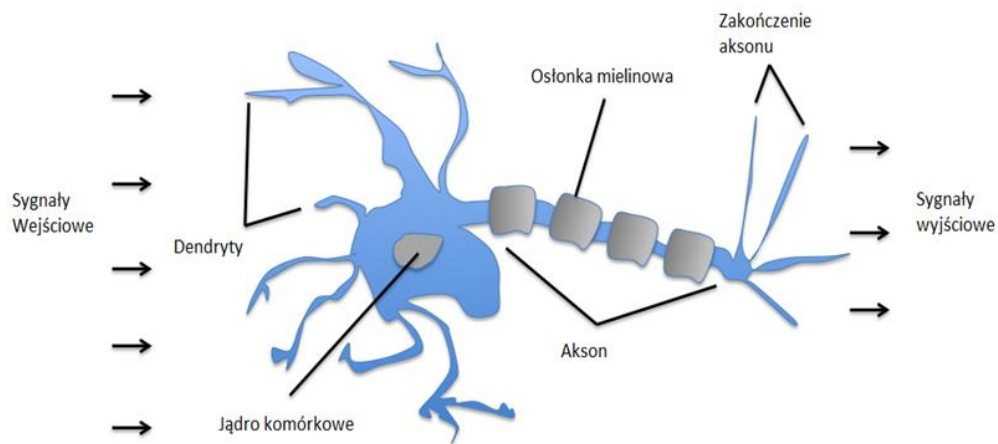
5. Wskaźnik F1 (ang. *F1 score*) jest dany wzorem:

$$F = \frac{2 \times Precision \times Recall}{(Precision + Recall)}$$

Wskaźnik F1 uwzględnia zarówno czułość, jak i precyzję. Jest to metryka znacznie bardziej miarodajna od czułości, ponieważ uwzględnia poprawność klasyfikacji modelu dla każdej z kategorii.

2.2 Sieci neuronowe

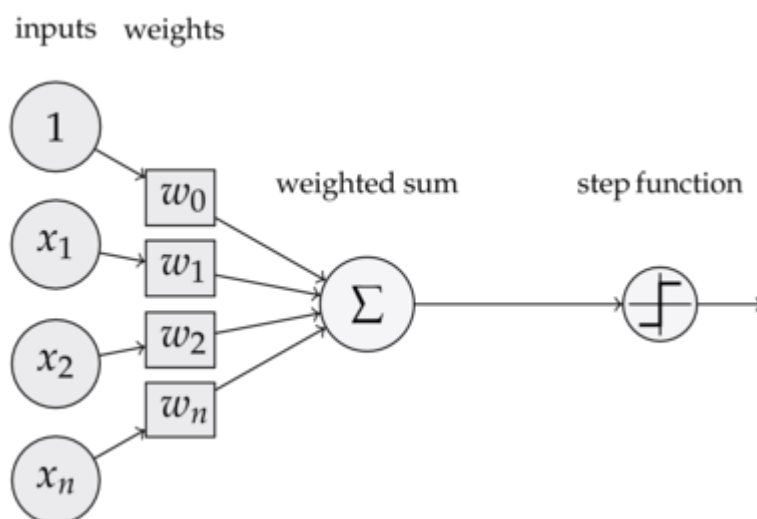
Do klasyfikacji komentarzy użyto metody uczenia głębokiego (ang. *deep learning*). Opiera się ona na strukturach matematycznych zwanych sieciami neuronowymi. Nazwa tych struktur nie jest przypadkowa, ich działanie jest inspirowane naturą, a konkretnie działaniem mózgu ludzi i innych zwierząt. Sam neuron jest komórką składającą się z ciała komórkowego (perikarionu), czyli wypustek określanych mianem dendrytów i aksonów. Neurony potrafią odbierać i przysyłać sygnały elektryczne oraz różnego rodzaju informacje (Cherry, 2021). Dendryt to wypustka stanowiąca przedłużenie komórki nerwowej, która odpowiada za odbieranie impulsów oraz przesyłanie ich do ciała komórki w celu ich integracji. Dendryt to największa część neuronów – stanowi do 90% powierzchni wielu z nich. Akson z kolei służy do przesyłania informacji z ciała komórkowego do reszty komórek nerwowych. Proces ten znany jest jako neurotransmisja (Cherry, 2021).



Wykres 3: Schemat budowy neuronu

Źródło: www.healthline.com

W 1958 roku Frank Rosenblatt opracował i zbudował najprostszy model sieci neuronowej zwanej perceptronem. W swojej najprostszej wersji perceptron był zbudowany z dwóch warstw neuronów reprezentujących odpowiednio wejście i wyjście. Rosenblatt odkrył też ważną właściwość perceptronu, którą przedstawił w swoim twierdzeniu: „Jeżeli tylko istnieje taki wektor wag w , przy pomocy którego element perceptronowy odwzorowuje w sposób zbiór oczekiwanych wartości wyjściowych, to istnieje metoda uczenia tego elementu gwarantująca zbieżność do wektora w (Loiseau, 2019).



Wykres 4: Schemat perceptronu Rosenblatta

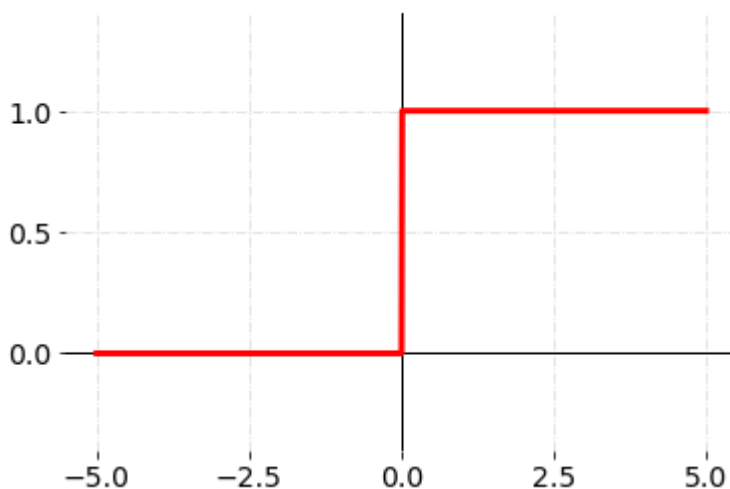
Źródło: www.statworx.com

Na wykresie 4 przedstawiono perceptron z tzw. uprzedzeniem (ang. *bias*), czyli do

parametrów wejściowych dodano również wartość równą 1 razem ze swoją wagą w celu lepszego dopasowania modelu do zmieniających się danych wejściowych (Gebel, 2020). Obliczenie sumy ważonej jest dane wzorem:

$$\sum_i^n w_i \times x_i$$

gdzie x oznacza i -ty parametr wejściowy, a w jego wagę. W kolejnym kroku suma ważona trafia do funkcji aktywacji, gdzie wartość sumy jest przekształcona na wartość wyjściową. Pierwotnie funkcjami aktywacji były binarne funkcje dyskretne, które w zależności od tego, czy suma była większa od wartości progowej, zwracały 1 lub 0 bądź 1 lub -1 (Loiseau, 2019).



Wykres 5: Przykład binarnej funkcji aktywacji

Źródło: www.towardsdatascience.com

Taka bardzo prosta funkcja aktywacji ma dwie duże wady. Z racji tego, że jest binarna, oznacza to, że nie nadaje się do rozwiązywania problemów klasyfikacji, gdzie występują więcej niż 2 klasy rozpoznania. Innym ograniczeniem jest fakt, iż gradient takiej funkcji wynosi 0, co poważnie utrudnia przeprowadzenie korekcji wag algorytmem propagacji wstecznej.

Rozwiązaniem tego problemu są nieliniowe funkcje aktywacji, ponieważ:

- pozwalają na przeprowadzenie algorytmu propagacji wstecznej ze względu na powiązanie pochodnej funkcji aktywacji z wejściem, co umożliwia lepsze zrozumienie, które wagi w neuronie są w stanie dokonać lepszej predykcji;
- pozwalają na składanie wielu warstw neuronów, co umożliwia uzyskanie wartości wyjściowej w postaci nieliniowej kombinacji parametrów wejściowych przepuszczonych przez wiele warstw.

Modele sieci neuronowych, podobnie jak inne modele predykcyjne, mogą być

nadzorowane i trenowane. Taka optymalizacja sieci odbywa się poprzez korekcję wartości wag w neuronach. Do najpopularniejszego algorytmu korekcji wag należy tzw. algorytm propagacji wstecznej. Jest to algorytm używany w sieciach wielowarstwowych, czyli takich sieciach neuronowych, które oprócz warstwy wejściowej i wyjściowej mają jeszcze warstwy ukryte. Algorytm ten opiera się na minimalizacji sumy kwadratów błędu uczenia z wykorzystaniem optymalizacji metody największego spadku (Korbicz, Obuchowicz i Uciński, 1994). Należy najpierw rozważyć błąd sieci ξ :

$$\xi = \sum_{\mu=1}^P \xi_{\mu} = \frac{1}{2} \sum_{\mu=1}^P \sum_{j=1}^n (y_j^{z\mu} - \varphi_j^{\mu})^2$$

gdzie:

$$\xi_{\mu} = \frac{1}{2} \sum_{j=1}^n (\delta_j^{\mu})^2$$

przy czym n oznacz liczbę elementów w warstwie wyjściowej, $y_j^{z\mu}$ to oczekiwana wartość wyjścia j -tego elementu, a φ_j^{μ} jest ważoną sumą wejść wyznaczoną w sumatorze według μ -tego wzorca.

Problemem przy uczeniu sieci neuronowych jest znalezienie globalnego minimum funkcji błędu ξ . Bardzo często używaną do tego metodą jest gradientowa metoda największego spadku. Polega ona na iteracyjnym poszukiwaniu kolejnego lepszego punktu w kierunku przeciwnym do gradientu funkcji celu w danym punkcie. Stosując tę metodę, zmiana wagi połączenia w_{ji} powinna spełniać relację:

$$\Delta w_{ji} = -\eta \frac{\partial \xi}{\partial w_{ji}} = -\eta \sum_{\mu=1}^P \frac{\partial \xi_{\mu}}{\partial w_{ji}} = -\eta \sum_{\mu=1}^P \frac{\partial \xi_{\mu}}{\partial y_j^{\mu}} \frac{\partial y_j^{\mu}}{\partial w_{ji}}$$

gdzie η oznacza współczynnik proporcjonalności.

W przypadku liniowych elementów przetwarzających zachodzą równości:

$$\frac{\partial \xi_{\mu}}{\partial \xi_j^{\mu}} = -(y_j^{z\mu} - y_j^{\mu}) = -\delta_j^{\mu}$$

$$\frac{\partial y_j^{\mu}}{\partial w_{ji}} = \frac{\partial \varphi_j^{\mu}}{\partial w_{ji}} = u_i^{\mu}$$

Stąd zostanie otrzymana korekta wagi:

$$\Delta w_{ji} = \eta \sum_{\mu=1}^P \delta_j^{\mu} u_i^{\mu}$$

Po czym ostatecznie otrzymana jest tzw. „reguła delty”, która jest dana wzorem:

$$w_{ji}^n = w_{ji}^s + \Delta w_{ji}$$

gdzie górny indeks n oznacza nową, a s starą wagę. Reguła delty przy dostatecznie małym współczynniku proporcjonalności uczenia η poszukuje zbioru wag minimalizującego funkcję błędu sieci liniowej. Należy pamiętać, że metoda propagacji wstecznej działa tylko dla wielowarstwowych sieci jednokierunkowych, nie zadziała ona np. w modelach generatywnych, które opierają się na sieciach rekurencyjnych. Są to mianowicie takie sieci, które posiadają sprzężenie zwrotne (Korbicz, Obuchowicz i Uciński, 1994).

Wśród warstw ukrytych możemy rozróżnić wiele ich rodzajów: są to na przykład sieci gęste, czyli składające się z od kilkunastu do kilkudziesięciu neuronów (w jednej warstwie może być 16, 32 czy nawet 64 neuronów). Najczęściej używaną funkcją aktywacji w tego rodzaju warstwach jest funkcja ReLU. Warstwę gęstą można również użyć jako warstwę końcową (np. w postaci neuronów posiadających sigmoidalną funkcję aktywacji). Innymi, szeroko stosowanymi warstwami, są sieci konwolucyjne. W sieciach tych używa się filtra nazywanego jądrem (ang. *kernel*). Jednym z hiperparametrów sieci konwolucyjnej jest rozmiar jądra. Mniejszy rozmiar jądra wpływa na lepszą dokładność w uzyskaniu informacji kluczowych z danych wejściowych. Wpływa to również na zmniejszenie dalszych warstw, co w rezultacie daje głębszą architekturę (tj. więcej warstw w sieci neuronowej). Większy rozmiar jądra odbija się na mniejszej dokładności modelu, lecz lepszej generalizacji problemu, gdy np. w danych wejściowych nie występuje dużo wartości szczegółowych. Podczas uczenia sieci neuronowej dokonywana jest korekta wag w filtrze. Następnie uogólniony wynik jest poddawany funkcji aktywacji – najczęściej używana jest wówczas funkcja ReLU. Sieci CNN (ang. *Convolutional Neural Networks*) są powszechnie używane w rozpoznawaniu i klasyfikacji obrazów (filtry 2D) oraz w analizie tekstowej (filtr 1D). Po przejściu wartości przez filtr wyniki przechodzą przez pooling (Mamczur, 2021). Pooling służy do upraszczania obrazu bądź tekstu, co wpływa pozytywnie na wydajność modelu ze względu na mniejszą liczbę parametrów do przetworzenia. Wyróżnia się trzy rodzaje tej warstwy:

- max pooling (z danych elementów wybiera się ten o największej wartości),
- min pooling (z danych elementów wybiera się ten o najmniejszej wartości),
- average pooling (wynik się uśrednia).

Jako przykład można podać macierz pikseli:

$$\begin{bmatrix} 129 & 243 \\ 85 & 100 \end{bmatrix}$$

Przy zastosowaniu max pooling wynikiem będzie piksel o wartości 243. Jeżeli użyty zostanie min pooling, rezultatem będzie 85. Z kolei average pooling zwróci średnią wartość, czyli 139. Max pooling jest powszechnie stosowany przy analizie sentymentu ze względu na potrzebę wyróżniania wyrazów o mocnym wydźwięku (Mamczur, 2021). Do zapobiegania nadmiernemu przeuczeniu się sieci używa się warstw, w których dokonywany jest tzw. dropout (odrzućenie). Polega on na losowym wyłączaniu niektórych połączeń między neuronami. Sprawia to, że sieć nie przyjmuje zbyt szybko wyuczonego wzorca, przez co problem przeuczenia jest znacznie zredukowany.

Oprócz rozpoznawania i klasyfikacji obrazów czy tekstów sieci neuronowe można wykorzystywać również do analizy wideo czy też zwykłych danych tabelarycznych. Ponadto można też wyróżnić sieci generatywne, są to specjalne sieci rekurencyjne (czyli takie, które w odróżnieniu do jednokierunkowych sieci neuronowych, posiadają sprzężenie zwrotne w swojej topologii), które potrafią generować dane oraz takie obiekty, jak: zdjęcia, dźwięk czy tekst. Sieci generatywne można zastosować np. w odtworzeniu bardzo starych nagrań czy przy odnowieniu starych zdjęć z brakującymi elementami (Mamczur, 2021).

2.3 Diagnostyka sieci neuronowych

Do oceniania dopasowania modeli sieci neuronowych używa się różnych metryk. Oprócz wcześniej wspomnianych statystyk, jak skuteczność, precyzja czy współczynnik F1, do oceny sieci neuronowych używa się również (Furbush, 2021):

1. współczynnika ucieczki (ang. *escape rate*),

Jest to współczynnik określający stosunek błędnych klasyfikacji negatywnych do wszystkich klasyfikacji.

$$Escape = \frac{FN}{TP + TN + FN + FP}$$

W ten sposób można określić, jaki np. procent wadliwych produktów (u których nie wykryto defektów) „ucieknie” na rynek. Duży współczynnik ucieczki może doprowadzić do tego, że ucierpi na tym wizerunek firmy, co może ją kosztować w rezultacie bardzo dużą ilość pieniędzy (Furbush, 2021).

2. współczynnika overkill (ang. *overkill rate*).

Jest to metryka służąca do określania stosunku błędnych klasyfikacji pozytywnych do wszystkich klasyfikacji.

$$Overkill = \frac{FP}{TP + TN + FN + FP}$$

Dzięki tej statystyce można określić np. jaki procent produktów poprawnie działających zostało mylnie wyłączonych z linii produkcji. Takie produkty mogą zostać wyrzucone bądź zostaną ręcznie wykonane po raz kolejny. Obie te decyzje generują dla producenta dodatkowe koszty zarówno finansowe, jak i robocze. (Furbush, 2021)

Miarami służącymi natomiast do optymalizacji sieci w celu jej najlepszego dopasowania są funkcje straty. Wśród nich można wyróżnić:

1. średni błąd kwadratowy (ang. *mean square error*),

Jest to jedna z funkcji straty, która jest średnią kwadratów różnic pomiędzy wartościami rzeczywistymi a tymi podanymi przez model sieci neuronowej. Wartość tej funkcji będzie zawsze dodatnia, w idealnej sytuacji powinna ona wynieść 0 (Brownlee, 2019).

2. entropia krzyżowa (ang. *cross-entropy loss*).

Jest to inny rodzaj funkcji straty, w tym przypadku każde prawdopodobieństwo przewidziane przez model jest najpierw porównane do faktycznej wartości (wynoszącej 0 lub 1), a wartość straty bazuje na różnicy między wynikiem podanym przez model a wynikiem prawdziwym. Wartość kary jest logarytmiczna (Brownlee, 2019). Wartość tego rodzaju funkcji straty jest minimalizowana, gdzie mniejsze wartości wskazują na lepszy model niż w przypadku, gdy wartości kar są większe. Użycie odpowiedniego rodzaju funkcji straty zależy w dużym stopniu od problemu, którego dany model dotyczy. W problemach regresyjnych, gdzie wartością wyjściową jest ilość bądź inna wartość z ciągłego przedziału, lepszym wyborem będzie użycie średniego błędu kwadratowego. Natomiast w przypadku klasyfikacji binarnych oraz klasyfikacji wieloklasowych należy użyć entropii krzyżowej (przy czym konieczne jest przekształcenie zmiennej objaśnianej na zmienną binarną) (Brownlee, 2019).

3. Analiza sentymentu w badaniach rynkowych

3.1 Zastosowanie biznesowe analizy sentymentu

Do najczęściej realizowanych celów biznesowych przez analizę sentymentu należą takie zagadnienia, jak (Patel, 2022):

- monitorowanie marki,
- poprawa obsługi klienta,
- badanie satysfakcji pracowników,
- zapewnienie lepszej analizy produktu,
- monitorowanie rynku,
- śledzenie konkurencji,
- obserwacja mediów społecznościowych.

1. Monitorowanie marki

Monitoring marki oraz zarządzanie reputacją firmy to jedno z najpowszechniejszych rozwiązań, w których stosuje się analizę sentymentu. Użycie narzędzi wykorzystujących analizę sentymentu umożliwia szybkie wykrycie negatywnych opinii na temat firmy i podjęcie kroków w celu szybkiego poprawienia reputacji przedsiębiorstwa. Inną zaletą tego rozwiązania jest możliwość śledzenia zmian reputacji marki w czasie, co pozwala na śledzenie postępu w budowie dobrej opinii o firmie. Kolejną możliwością, jaką analiza sentymentu oferuje w tym przypadku, jest użycie uczenia maszynowego do wykrycia trendów oraz dokonanie estymacji potencjalnej reakcji klientów na określoną decyzję ze strony firmy (Roldós, 2020). Użycie analizy sentymentu do stworzenia raportu dotyczącego reputacji marki powinno również zawierać następujące informacje:

- łączna liczba dokonanych transakcji oraz kampanii przez firmę w określonym czasie,
- łączna liczba wspomnień na temat marki,
- procent pozytywnych opinii,
- procent negatywnych opinii,
- obliczenie wyniku sentymentu społecznego,
- graficzne przedstawienie wyniku sentymentu społecznego w czasie (w celu zobrazowania jak reputacja zmieniała się w czasie) (Newberry, 2020).

Obliczenie wyniku sentymentu społecznego dokonuje się na kilka sposobów:

- obliczenie stosunku pozytywnych opinii do wszystkich komentarzy,
- obliczenie stosunku pozytywnych opinii do komentarzy zawierających sentyment (wówczas odrzuca się opinie neutralne) (Newberry, 2020).

2. Poprawa obsługi klienta

Według badań przeprowadzanych przez firmę McKinsey & Company więcej niż 25% klientów rezygnuje z usług bądź produktów firmy po jednym złym doświadczeniu w związku z obsługą (McKinsey, 2016). Ponadto wzrost popularności mediów społecznościowych oraz forów powoduje, że jedno złe doświadczenie może wywoływać straty w firmie wiele razy. Analiza sentymentu na danych opisujących takie zdarzenia może przygotować zespół odpowiedzialny za doświadczenie klienta na potencjalne trudności oraz pomóc mu lepiej zrozumieć odczucia klienta podczas całego procesu biznesowego (Roldós, 2020). W analizie sentymentu służącej do badania doświadczenia klienta używane są dwa czynniki:

- biegunowość (tj. czy wydzźwięk emocjonalny w tekście jest pozytywny bądź negatywny),
- stopień (tj. jak silne są emocje opisane przez użytkownika) (Patel, 2022).

Dane wykorzystywane do analizy sentymentu w badaniu obsługi klienta mogą pochodzić z różnych źródeł. Feedback od klienta może zostać zebrany w sytuacjach takich jak:

- rozwiązania zgłoszenia o obsługę (wówczas można zebrać poszczególne wiadomości od użytkownika),
- uruchomienie sprzedaży nowego produktu,
- zakończenie procesu wdrażania klienta w system i ogólną organizację firmy.

W celu uzyskania jak najwartościowszych informacji w danych organizacja może dokonać podziału klientów na 3 grupy (Patel, 2022). Są to:

- promotorzy (ang. *promoters*) – najbardziej pożądana przez firmę grupa klientów. Spośród grup klientów promotorzy mają zdecydowanie najwyższy współczynnik LTV (ang. *lifetime value*) i to oni są źródłem największych dochodów dla firmy;
- grupa pasywnych klientów (ang. *passives*) – klienci o neutralnym stosunku do firmy, u których prawdopodobieństwo na zrezygnowanie z usług organizacji na rzecz konkurenta jest wyższe;
- krytycy (ang. *detractors*) – niezadowoleni klienci, którzy krytykują produkt lub obsługę organizacji. Wśród tych 3 grup krytycy mają najniższy współczynnik LTV (Patel, 2022).

Zebranie wszystkich opinii i podział na te 3 grupy pozwala na dokonanie analizy sentymentu. Wówczas można określić, jakie scenariusze bądź rozwiązania są skorelowane i związane z opiniami wyrażanymi przez grupę promotorów (Patel, 2022). W ten sposób można też wyróżnić, jakie błędy oraz nieprawidłowości występują w działaniu firmy, na które grupa krytyków zwraca dużą uwagę. W rezultacie organizacja jest w stanie zmodyfikować swoje produkty, dokonać zmian w ich cechach oraz poprawić jakość usług w celu osiągnięcia jak największej liczby zadowolonych oraz lojalnych klientów przy jednoczesnej minimalizacji liczby krytyków firmy (Patel, 2022).

Kolejnym źródłem danych, które można wykorzystać do poprawy obsługi klienta, są media społecznościowe. Platformy takie, jak: Facebook, Twitter czy Instagram, stanowią duże i istotne źródło danych, gdyż klienci na tych portalach mogą w sposób bardziej ekspresywny wyrażać swoją opinię (Patel, 2022). Wśród słów używanych przez użytkowników mediów społecznościowych mogą znajdować się takie pozytywne wyrazy, jak: „good”, „excellent”, „fantastic”, albo takie negatywne wyrażenia, jak: „awful”, „hate” czy też „horrible”. Zaletą kolekcji danych z mediów społecznościowych jest np. udostępnienie przez portale takie jak Twitter czy Facebook otwartego API. W ten sposób można kolekcjonować i analizować dane w czasie oraz śledzić zadowolenie klientów z obsługi (Patel, 2022).

3. Badanie satysfakcji pracowników

Źródłami danych, które informują o satysfakcji pracowników, mogą być np. ankiety czy opinie użytkowników na portalach takich jak Glassdoor czy GoWork oraz wiadomości wysyłane pocztą elektroniczną. Pozwala to na lepsze zrozumienie potrzeb pracowników, a co za tym idzie, poprawę ich satysfakcji. Dzięki temu można poprawić takie wskaźniki, jak produktywność czy wskaźnik rotacji, czyli procent pracowników, którzy odeszli w określonym przedziale czasowym (Roldós, 2020). Do analizy dużej ilości danych tekstowych dotyczących satysfakcji zatrudnionych wykorzystuje się uczenie maszynowe w celu zautomatyzowania procesu. Dzięki temu można oszacować na podstawie informacji uzyskanej w komentarzach bądź ankietach, którzy konkretnie pracownicy są zadowoleni, a którzy zatrudnieni nie czują satysfakcji ze swojego miejsca pracy (Rosencrance, 2020). Po oszacowaniu wyników i ich agregacji dane należy zaprezentować zarządowi bądź menedżerom w organizacji. Zdolność wizualizacji sentymentu wśród pracowników pozwala również na zwiększenie zaangażowania pracowników. Zarząd organizacji może też poprawić dzięki temu proces zarządzania pracownikami czy lepsze zrozumienie doświadczenia pracownika (Rosencrance, 2020). Firmy mogą też przeprowadzać analizę sentymentu komentarzy pracowników na wiele sposobów, np. na podstawie wiadomości e-mail i poprzez wyszukanie kluczowych słów dla określonych departamentów (Rosencrance, 2020). W ten sposób firma może określić potrzeby pracowników i dostosować swoje zarządzanie do opinii działu bądź indywidualnych osób. Korzyści z rozwiązania, jakim jest analiza sentymentu, czerpią również dział HR (ang. *Human Resources*). Departament ten w dużym stopniu opiera się głównie na jakościowych i wysoko ustrukturyzowanych danych pochodzących z list płac i innych źródeł dedykowanych dla HR (Rosencrance, 2020). Analiza sentymentu pozwala na przetwarzanie danych pochodzących z innych, często nieustrukturyzowanych źródeł oraz umożliwia wykrycie informacji o pozytywnym bądź negatywnym wydźwięku. Inną zaletą tego rozwiązania jest uzyskanie lepszego i efektywniejszego wglądu w opinie pracowników o firmie poprzez analizę ich komunikacji wewnątrz organizacji (Rosencrance, 2020). Umożliwia to również pracownikom

działu HR obserwację, jakim tonem czy językiem posługuje się pracownik w swoich wiadomościach e-mail, przez co można zidentyfikować satysfakcję zatrudnionego z firmy oraz swojej pozycji w niej (Rosencrance, 2020).

4. Zapewnienie lepszej analizy produktu

Poprzez analizę sentymentu opinii użytkowników na temat produktu można w lepszy sposób analizować produkt. W ten sposób można określić oczekiwania klienta od produktu oraz stwierdzić, jakie zmiany wpłynęłyby pozytywnie na jego satysfakcję. Ponadto wykorzystanie analizy sentymentu bazującej na aspekcie umożliwia ocenę potencjalnych obszarów do zmian w konkretnej cesze produktu takiej jak np. funkcjonalność, interfejs czy też doświadczenie użytkownika (ang. *user experience*, UX) (Roldós, 2020). Analiza nieustrukturyzowanych danych może pomóc w:

- porównaniu ocen i recenzji produktu z towarami stworzonymi przez konkurencję,
- osiągnięciu wglądu w najnowsze produkty w czasie rzeczywistym, 24 godziny na dobę,
- oszczędzeniu dużej ilości czasu na ręczne przetwarzanie danych.

W ten sposób, poprzez zautomatyzowany processing i analizę, recenzje produktów zostają sklasyfikowane jako pozytywne bądź negatywne. W pierwszym kroku należy zebrać dane zawierające recenzje oraz opinie klientów o danym produkcie. Cennymi źródłami takich informacji są portale: Yelp, Amazon, Google Play czy też Capterra (Pascual, 2019). Do zebrania informacji ze stron internetowych powszechnie używa się techniki, która nazywa się web scraping. Jest to metoda służąca do wyodrębnienia danych ze stron internetowych zastępująca ręczne wpisywanie bądź kopiowanie i wklejanie tekstu. Pozyskane w ten sposób dane często są ustrukturyzowane (Pascual, 2019). Web scraping ze względu na sposób przeprowadzenia można podzielić na dwie kategorie:

- narzędzia wizualne (wyspecjalizowane programy z graficznym interfejsem użytkownika, do ich użycia niepotrzebna jest znajomość programowania – dzięki temu ich użycie jest o wiele prostsze),
- biblioteki programistyczne (różnego rodzaju pakiety programistyczne pozwalające programiście na stworzenie narzędzia dokonującego web scraping według potrzeb programisty. Do bibliotek tych można zaliczyć: Scrapy, BeautifulSoup bądź Pyspider dla języka Python, Simplecrawler lub Nodecrawler w języku JavaScript czy też Rvest dla języka R) (Pascual, 2019).

Po zebraniu danych należy je ustrukturyzować w odpowiednią formę. Do kluczowych kolumn należą:

- data (kiedy tekst został napisany bądź wysłany),
- tekst (czyli treść opinii lub recenzji),

- wynik (ocena wystawiona przez użytkownika).

Rodzajem analizy sentymentu używanej do analizy produktów jest analiza sentymentu bazująca na aspekcie ze względu na fakt, iż oprócz samego wydźwięku emocjonalnego w tekście bardzo istotne też jest wykrycie, który element bądź cecha produktu jest przyczyną takiego wyniku (Pascual, 2019). Rozbudowany tekst, jak np. recenzja, może mieć w sobie więcej niż jeden sentyment oraz aspekt, np. użytkownik może w recenzji okazać zadowolenie z dobrej jakości zdjęć, jaką aparat wywołuje, ale może też jednocześnie wyrazić rozczarowanie szybkim zużyciem baterii. W tym celu dzieli się recenzję na zdania, żeby z każdego z nich można było wyróżnić dokładnie jeden aspekt i dokładnie jeden sentyment (Pascual, 2019).

5. Monitorowanie rynku

Serwisy informacyjne, blogi, fora lub media społecznościowe stanowią bardzo bogate źródło informacyjne w kontekście badania nastrojów na rynku, oczekiwań klientów bądź konwersacji na temat określonej kampanii marketingowej czy też produktu, który został wprowadzony na rynek. W tym obszarze również analiza sentymentu bazująca na aspekcie pozwala na wskazanie konkretnych elementów, które należy uznać za szczególnie istotne (Wonderflow, 2018). Do monitoringu rynku za pomocą analizy sentymentu używa się słów kluczowych (ang. *keywords*) (Pinkowska, 2020). Słowa kluczowe służą do wyszukania wśród danych tekstowych tych fragmentów, które są dla organizacji istotne, jak np. firmy będące konkurencją, kraje będące potencjalnymi rynkami czy też obszary, w których firma chce uruchomić inwestycje. Oprócz słów kluczowych należy też wybrać język, w którym dane informacje zostały napisane. W analizie rynku używana jest rozwinięta analiza sentymentu (Pinkowska, 2020). Zastosowanie klasycznej, binarnej analizy sentymentu klasyfikującej pozytywne oraz negatywne opinie posiada wiele ograniczeń. Nowoczesne narzędzia do analizy sentymentu w badaniu rynku wykorzystują sieci neuronowe. W ten sposób można zapewnić poprawne określenie sentymentu dla określonego słowa i jego synonimów (Pinkowska, 2020). Dodatkową zaletą tego rozwiązania jest fakt, iż badany jest sentyment słów na podstawie rodziny języków. W ten sposób algorytm, oprócz osiągnięcia większej precyzji, jest też w stanie działać poprawnie dla większej liczby języków (Pinkowska, 2020).

Jednym z zagrożeń wynikających z analizy sentymentu w celu badania rynku jest wiarygodność danej informacji (Pinkowska, 2020). Zadaniem stojącym przed analitykiem oraz firmą jest określenie, czy informacja pochodzi z dobrze zbadanego źródła, czy też jest to niezweryfikowana wieść. Problemem powszechnie badanym jest również korelacja między zmianą cen a sentymentem. Cena oraz sentyment powinny być ze sobą skorelowane, lecz nie zawsze wiadomym jest, która z tych rzeczy ma pierwotny wpływ (tj. czy negatywne opinie miały wpływ na obniżenie kursu bądź akcji lub na odwrót – czy wzrost cen akcji wpływa na zwiększenie

ilości pozytywnych opinii oraz informacji) (Pinkowska, 2020). Jest to nie tylko część technicznej analizy, ale również kluczowa informacja dla inwestorów oraz handlowców. Stosowanie analizy sentymentu do predykcji cen oraz akcji było badane przez wielu uczonych z takich uczelni, jak Uniwersytet Stanforda czy Uniwersytet Londyński (Pinkowska, 2020).

W czasach współczesnych, przy bardzo dużej ilości informacji, zmienność na rynku jest jeszcze większa. Każda negatywna bądź pozytywna opinia może mieć duży wpływ na sytuację na rynku i wpłynąć na jej zmianę. Celem analizy sentymentu jest jak najszybsze wykrycie wszelkich nowych pozytywnych lub negatywnych opinii, by mieć możliwość zareagowania na zmiany na rynku. Ponadto analiza sentymentu pozwala na przetwarzanie informacji w czasie rzeczywistym, dzięki czemu można badać nowo przychodzące informacje (Pinkowska, 2020). W ten sposób firma może podjąć kroki w celu uniknięcia negatywnych sytuacji w przyszłości i zwiększenia szans na wystąpienie tych dobrych. Do zagrożeń wynikających z analizy sentymentu należą zaś niezweryfikowane informacje oraz tzw. fake newsy. Poprzez zastosowanie się do nich firma może narazić się na niepożądane konsekwencje oraz błędne podjęcie kroków (Pinkowska, 2020).

6. Śledzenie konkurencji

Analiza sentymentu może zostać również użyta do badania sytuacji na rynku u firm konkurencyjnych czy też do obserwacji ich reputacji. Pozwala to na wykrycie potencjalnych rozwiązań oraz zagrożeń dla firmy. Można w ten sposób też określić swoją pozycję na rynku w tej chwili. Ponadto dzięki analizie sentymentu bazującej na aspekcie można porównać sytuację firmy z konkurencją na różnych szczeblach takich jak: obsługa klienta, jakość produktów oraz doświadczenie użytkownika. Kolejną zaletą tego rozwiązania jest możliwość śledzenia postępów oraz zmian pomiędzy przedsiębiorstwem a jego konkurencją (Wonderflow, 2018). Najczęstszym źródłem danych do śledzenia konkurencji są platformy mediów społecznościowych lub portale z recenzjami. Rodzajem analizy sentymentu, który jest używany w monitoringu konkurencji, jest analiza sentymentu bazująca na aspekcie. Do filtrowania danych też używa się słów kluczowych, czyli np. kategorii, w której firma chce znaleźć informacje o konkurencji, lub nazwy firm, z którą organizacja bezpośrednio rywalizuje na rynku albo która ma konkurencyjne produkty na rynku. Dzięki temu można oszacować, w których szczegółach produktu konkurencja mogła osiągnąć przewagę lub które elementy produktu wzbudziły brak satysfakcji wśród klientów.

7. Obserwacja mediów społecznościowych

Codziennie na portalach mediów społecznościowych publikowane są wiele tysięcy wpisów na różne tematy. Analiza sentymentu pozwala na zautomatyzowane klasyfikowanie tychże wpisów i uzyskanie wglądu w takie obszary, jak: zadowolenie społeczeństwa z pełniącego władzę aktualnego rządu, reakcja na ostatnie wydarzenia polityczne, wydarzenia sportowe bądź muzyczne czy też ocena najnowszych produktów. Dodatkowym atutem obserwacji mediów

społecznościowych jest fakt, iż serwisy takie jak Twitter czy YouTube udostępniają otwarte API, co pozwala na analizę danych w czasie rzeczywistym (Wonderflow, 2018). Analizę sentymentu danych pochodzących z mediów społecznościowych można użyć do badania reputacji oraz opinii użytkowników o organizacji w czasie rzeczywistym (Gavran, 2022). Systemy oparte na analizie sentymentu dokonują również analizy porównawczej w opinii klientów między organizacją a firmami konkurencyjnymi. Dodatkowo, użytkownicy różnych platform mediów społecznościowych mogą przejawiać różne gusta, dlatego w systemach opartych na analizie sentymentu stosuje się również analizę porównawczą dla wielu platform mediów społecznościowych w celu sprawdzenia, na której stronie użytkownicy mają korzystniejsze zdanie bądź na której platformie znajduje się większa ilość krytycznych komentarzy (Gavran, 2022). W analizie sentymentu mediów społecznościowych powszechnie stosowane są słowa kluczowe w celu wyszukania wpisów na określony temat (dana firma, technologia bądź model produktu). Ponadto w celu określenia przyczyny określonego wydźwięku emocjonalnego do systemów analizy sentymentu w mediach społecznościowych używa się analizy sentymentu bazującego na aspekcie. Jest to rozwiązanie używane szczególnie przy monitorowaniu zdań użytkowników o jakości obsługi klienta bądź ich doświadczenia (Gavran, 2022). W ten sposób organizacja może określić, które czynniki w obsłudze należy poprawić, a które elementy stanowią przewagę biznesową nad konkurencją. Inną zaletą monitorowania mediów społecznościowych jest możliwość szybkiego wykrycia kryzysowej sytuacji w czasie (Gavran, 2022). Jeżeli w krótkim czasie doszło do nagłego wzrostu liczby negatywnych informacji – to dzięki analizie sentymentu bazującej na aspekcie i słowach kluczowych organizacja jest w stanie szybko wykryć nagły kryzys wizerunkowy, określić przyczynę tego zdarzenia i podjąć decyzje mające na celu poprawę sytuacji wizerunkowej (Gavran, 2022). Do zagrożeń występujących w analizie sentymentu w mediach społecznościowych należą np. opinie sarkastyczne, które mogą mieć często odwrotne przesłanie emocjonalne w porównaniu do tłumaczenia dosłownego. W źródłach tekstowych znajdujących się na platformach mediów społecznościowych przeważa liczba zdań pisanych językiem nieformalnym. Często wśród tych zdań można znaleźć piktogramy służące do wyrażania emocji, czyli tzw. emoji (Pozzi, 2018). Piktogramy te można podzielić na dwie kategorie: te o pozytywnym wydźwięku emocjonalnym oraz te o negatywnym sentymencie. Analiza sentymentu uwzględniająca emoji pozwala na osiągnięcie większej dokładności w ocenie ze względu na dodatkowy czynnik zawierający sentyment (Pozzi, 2018). Jest to w szczególności istotne, ponieważ duży procent wpisów w mediach społecznościowych zawiera te piktogramy (Pozzi, 2018).

1. Zastosowanie analizy sentymentu do wykrycia mowy nienawiści we wpisach na platformie Twitter

Wykrycie zjawiska mowy nienawiści jest niezbędne dla wykrycia zdarzeń kontrowersyjnych, zbudowania chatbotów czy też systemu rekomendacji treści. Wpisy zawierające mowę nienawiści to komentarze atakujące pojedynczą osobę (mogą to być treści rasistowskie, dyskryminujące na tle płciowym) lub jakąś grupę (np. mniejszość etniczną, wspólnotę religijną czy organizację). Wykrycie takiego zjawiska jest trudnym zadaniem ze względu na złożoność języka naturalnego – różne formy dyskryminacji czy też cele ataku. Ponadto można to samo znaczenie wypowiedzi wyrazić na wiele sposobów, co stanowi kolejne wyzwanie (Badjatiya, Gupta, Gupta i Varma, 2017). Rozwój systemów wykrywających mowę nienawiści zawdzięczamy wykorzystaniu metod uczenia maszynowego i klasyfikatorów takich jak: lasy losowe, drzewa decyzyjne wzmocnione o gradient (ang. *gradient boosted decision trees*) czy też głębokie sieci neuronowe (Badjatiya, Gupta, Gupta i Varma, 2017). Szczególnie dobre wyniki te systemy osiągają przy użyciu takich rodzajów sieci neuronowych, jak sieci konwolucyjne czy sieci LSTM (ang. *Long Short-Term Memory Network*). Oprócz tego ważną rolę w wykrywaniu mowy nienawiści za pomocą analizy sentymentu pełnią n-gramy, wektory TF-IDF czy też worki słów (ang. *bag-of-words*). Ta ostatnia struktura jest uproszczoną (np. o pominięcie gramatyki) reprezentacją pierwotnego tekstu. Worki słów często stanowią wejście dla modeli sieci neuronowych (Badjatiya, Gupta, Gupta i Varma, 2017). Istotną kwestią jest też rozmiar zbioru treningowego – powinien on liczyć minimum 10 tysięcy wpisów (Badjatiya, Gupta, Gupta i Varma, 2017). W momencie gdy model jest nauczony, potrafi on wówczas dokonywać klasyfikacji tekstu i klasyfikować dany komentarz jako rasistowski, dyskryminujący na tle płciowy czy też będący zaklasyfikowany jako niezawierający obraźliwych treści. Rozwiązanie to pozwala na zautomatyzowane wykrywanie mowy nienawiści, zaklasyfikowanie jaki rodzaj dyskryminacji zawierał dany komentarz. Ponadto pozwala to na skalowalne przetwarzanie danych w przeciwieństwie do ręcznego monitoringu wpisów (Badjatiya, Gupta, Gupta i Varma, 2017).

2. Określenie sentymentu w wypowiedziach wulgarnych

Wyrazy wulgarne są dość powszechne. Szacuje się, że stanowią od 0,5% do 0,7% wyrazów używanych w mowie codziennej oraz 1,15% wśród wpisów udostępnianych na portalu Twitter (Cachola, Holgate, Preotiuc-Pietro i Li, 2018). Wyrazy wulgarne mogą zostać użyte w wielu celach. Jednym z zastosowań jest wzmocnienie wyrażenia w komentarzu. Przekleństw używa się również w celu ataku bądź obrazu czy też jako sygnał, że konwersacja jest nieformalna. Badanie kontekstu wypowiedzi wulgarnych jest istotne, gdyż pozwala zrozumieć pragmatykę przekleństw, a modele przetwarzające wypowiedzi wulgarne stanowią cenne źródło dla psychologów badających czynniki społeczne wpływające na użycie wulgaryzmów. Do identyfikacji wyrazów wulgarnych należy użyć specjalnego leksykonu wyrazów wulgarnych. Jeden z takich leksykonów znajduje się pod adresem www.noswearing.com. (Cachola, Holgate, Preotiuc-Pietro i Li, 2018).

W dalszej kolejności, po zebraniu danych i identyfikacji przekleństw, bada się korelację częstotliwości ich występowania z takimi czynnikami, jak: wiek autora, jego płeć, wykształcenie, zarobki czy też wyznanie. Słowa wulgarne stosowane w kontekście wzmocnienia opinii mogą być użyte zarówno do wyrażenia opinii pozytywnej, jak i negatywnej (Cachola, Holgate, Preotiuc-Pietro i Li, 2018). Natomiast jeśli służą do wyrażenia emocji, wówczas przekleństwa zdecydowanie częściej występują w komentarzach o ujemnym sentymencie. Do badania komentarzy zawierających opinie wulgarne analizą sentymentu opartą na deep learning często wykorzystuje się sieci LSTM (Cachola, Holgate, Preotiuc-Pietro i Li, 2018). Do zliczenia wszystkich wulgarnych słów w komentarzu używa się tzw. maskowania. Polega to na tym, że dla każdego znalezionej wulgarnego słowa w jego miejscu wstawiany jest token. W ten sposób można w szybki sposób wyeliminować gramatyczne różnice w słowie, które nie mają wpływu na sam wydźwięk emocjonalny w tekście (Cachola, Holgate, Preotiuc-Pietro i Li, 2018). Następnie, po zliczeniu tokenów dokonywana jest konkatencja łącznej liczby tokenów z reprezentacją wpisu w celu przetworzenia tego wejścia przez sieć LSTM. Zmienna objaśniana jest 5-stopniowa, gdzie dopasowanie modelu może zostać wówczas określone poprzez obliczenie współczynnika średniego błędu bezwzględnego (Cachola, Holgate, Preotiuc-Pietro i Li, 2018).

3. Wykrycie emocji

Analizę sentymentu można również wykorzystać do sklasyfikowania emocji autora wypowiedzi. Wśród komercyjnych zastosowań analizy sentymentu można wyróżnić takie produkty, jak: wyraz twarzy, ton głosu, język oraz gestykulacja interlokutora. Rozpoznanie konkretnej emocji na podstawie samej treści wypowiedzi jest trudnym zadaniem obejmującym wiele wyzwań. Jednym z nich jest użycie zdawkowych wypowiedzi, jak np. „Tak”, „W porządku”, „Nie”. Ponadto zbiór danych MELD (ang. *Multimodal Emotion-Lines Dataset*), wykorzystywany do klasyfikacji emocji, składa się w 42% z wypowiedzi liczącej mniej niż 5 słów (Poria i inni, 2019). Ten zbiór danych zawiera m.in. dialogi z popularnego serialu komediowego pt. „Przyjaciele” (ang. „Friends”), w dialogach tych występuje kilku interlokutorów. Zbiór danych oprócz tekstu dialogów zawiera również znaczniki czasu w celu określenia kolejności wypowiedzi. Ponadto wszystkie wypowiedzi muszą należeć do tego samego dialogu bądź sceny i muszą być posortowane względem znacznika czasu w kolejności rosnącej (czyli od najwcześniejszej wypowiedzi do najpóźniejszej). Zmienna objaśniana będzie składać się z 7 klas odpowiedzi:

- radość,
- strach,
- złość,
- smutek,
- zaskoczenie,

- obrzydzenie,
- neutralność.

Przy czym klasy takie jak: „Strach”, „Złość”, „Smutek”, „Obrzydzenie” należą do klasy negatywnej, klasa „Neutralność” należą do klasy neutralnej, a „Radość” do klasy pozytywnej. Klasa „Zaskoczenie” jest nieco bardziej skomplikowana i może należeć zarówno do pozytywnej, jak i negatywnej klasyfikacji (Poria i inni, 2019). Do klasyfikowania wypowiedzi w dialogach używa się głównie modeli analizy sentymentu opartej na sieciach neuronowych. Jedną z przykładowych sieci, używaną do klasyfikacji emocji, jest sieć text-CNN (Poria i inni, 2019). Jest to model konwolucyjnej sieci neuronowej, która za wejście przyjmuje poszczególne wypowiedzi w dialogu. Model ten reprezentuje najprostsze podejście do klasyfikacji problemu, gdyż nie uwzględnia kontekstu dialogu. Innym przykładem sieci używanej do klasyfikacji emocji jest sieć bcLSTM, która jest dwukierunkową siecią rekurencyjną. Model ten uwzględnia kontekst wypowiedzi, lecz nie rozróżnia rozmówców i przetwarza cały dialog jako pojedynczą sekwencję (Poria i inni, 2019). Najbardziej rozbudowanym modelem sieci neuronowej do klasyfikacji emocji jest sieć DialogueRNN. Model ten jest w stanie uwzględnić autorów wypowiedzi w konwersacji oraz kontekst i emocje towarzyszące podczas dyskusji. Model ten wykorzystuje sieć GRU (ang. *Gated recurrent unit*) – jest to sieć rekurencyjna, przypominająca sieci LSTM, różniącą się jednak tym, iż ma ona mniej parametrów oraz brakuje w niej warstwy wyjściowej (Poria i inni, 2019). Model ten osiąga najlepsze rezultaty spośród 3 wymienionych. Systemy rozpoznania emocji znajdują zastosowania w aplikacjach służących do prowadzenia konwersacji, jak np. Siri czy Google Assistant (Poria i inni, 2019).

3.2 Źródła danych

Wśród najczęściej stosowanych źródeł danych tekstowych używanych w analizie sentymentu wyróżnia się następujące obszary:

- media społecznościowe,
- opinie klientów i konsumentów,
- artykuły oraz filmy z portali informacyjnych,
- ankiety.

1. Media społecznościowe

Według szacowań portalu statista, z serwisów mediów społecznościowych korzysta więcej niż 3,6 miliarda ludzi. Serwisy takie jak: TikTok, Twitter, Facebook, YouTube czy Instagram, posiadają kilkaset milionów aktywnych użytkowników, co pozwala na uzyskanie ogromnej ilości informacji (Dixon, 2022). Ponadto wśród serwisów takich jak TikTok czy Tumblr dominującą grupą użytkowników są ludzie w przedziale wiekowym od 16 do 24 lat, podczas gdy wśród serwisu

Facebook są to ludzie w grupie wiekowej od 25 do 34 lat (Repustate, 2021). Najbardziej zróżnicowanym wiekowo portalem jest Twitter, zrzesza on 350 milionów użytkowników, a wśród nich są również najważniejsi przedstawiciele państw, właściciele i prezesi największych korporacji na świecie czy też sportowcy oraz artyści (Repustate, 2021). Następnym krokiem po wyborze danej platformy jest zebranie danych. Serwisy takie jak: Twitter, Facebook czy YouTube, udostępniają otwarte API pozwalające na dostęp do komentarzy bądź wpisów. Innym sposobem na zebranie danych tekstowych jest korzystanie z takich metod, jak web scraping w przypadku, gdy portal nie udostępnia prostszej metody dostępu do informacji (Repustate, 2021). Analiza sentymentu danych pochodzących z tych portali jest bardzo wartościowa, gdyż informacje o liczbie „polubieni”, udostępnień czy komentarzy pod wpisem bądź filmem nie są zawsze w stanie przedstawić wiarygodnej informacji o zadowoleniu użytkowników oraz ich opinii na dany temat. Analizy sentymentu w mediach społecznościowych używa się do:

- odkrywania nowych trendów na rynku,
- śledzenia świadomości marki,
- badania potencjalnych nowych produktów,
- analizy działalności konkurencji,
- badania możliwości rozwoju kampanii reklamowych online.

2. Opinie klientów i konsumentów

Wartościowym źródłem danych do analizy sentymentu są również opinie klientów. Serwisy takie jak: Google, TripAdvisor, Yelp, Imdb, zawierają bardzo dużą ilość komentarzy na temat filmów, restauracji, kawiarni czy też placówek medycznych, jak zakłady stomatologiczne, czy też opinie o usługach lekarskich. Wśród portali, które zawierają opinie na temat leków razem z ich oceną, możemy wyróżnić m.in. stronę: www.drugs.com oraz stronę: www.webmd.com.

W tym przypadku analiza sentymentu pozwala firmom na badanie satysfakcji klientów z usług oraz produktów, jak również na uzyskanie cennych informacji o jakości produktów, takich jak np. wady w produkcie czy też efekty uboczne w przypadku leków (Repustate, 2021).

3. Artykuły oraz filmy z portali informacyjnych

Analiza sentymentu dla danych pochodzących z portali informacyjnych pozwala na monitorowanie reputacji marki oraz sytuacji na rynku. Uzyskanie takiej wiedzy jest bardzo wartościowe, gdyż można w ten sposób przewidzieć zmianę cen akcji danej firmy na rynku czy też ceny surowców w zależności od obecnej sytuacji. W ten sposób firma może uzyskać przewagę na rynku oraz wiedzę o potencjalnych możliwościach inwestycyjnych (Repustate, 2021).

4. Ankiety

Ankiety są jednym z najlepszych źródeł danych do analizy sentymentu, jeżeli celem biznesowym

jest zbadanie klientów czy pracowników. Otwarte pytania czyli takie, w których użytkownik może udzielić obszernej wypowiedzi, stanowią bardzo bogate źródło informacji takich jak zadowolenie klienta czy też poznanie jego potrzeb. W opiece zdrowotnej bardzo cennym źródłem takich informacji jest system Patient Voice, w którym pacjenci mogą udzielić opinii zarówno o wizycie u lekarza, jak również o całej kuracji czy obsłudze placówki medycznej (Repustate, 2021). Stanowi to w rezultacie ogromnie ważne źródło dla szpitali czy klinik, gdyż w ten sposób można przy pomocy analizy sentymentu uzyskać informacje o satysfakcji pacjentów oraz o tym, czy personel szpitala, kliniki bądź przychodni odpowiednio wykonuje swoje obowiązki (Repustate, 2021).

3.3 Technologie wykorzystujące analizę sentymentu

1. Talkwalker (analiza trendów, reputacji oraz produktu)

Jest to narzędzie wykorzystujące NLP, w tym analizę sentymentu, do badania opinii i odczuć klientów (Poria i inni, 2019). Można w ten sposób określić reputację firmy oraz satysfakcję klientów z produktu. Dzięki temu można wykryć panujące trendy na rynku, zbadać dotychczasową sytuację firmy na rynku oraz porównać pozycję i reputację firmy z konkurencją. Inną zaletą tego produktu jest jego wrażliwość na emocjonalny wydźwięk tekstu; model analizy sentymentu jest w stanie np. wykryć sarkazm w zdaniu, co jest jednym z największych wyzwań przy analizie sentymentu (Opitz, 2017). System ten ponadto jest w stanie dokonać analizy dla ponad 90 języków. Model analizy sentymentu wykorzystany w tym systemie jest oparty na nowym podejściu (Opitz, 2017). W pierwszej kolejności sieć neuronowa uczy się znaczenia danego zdania. W ten sposób sieć neuronowa symuluje kognitywne funkcje mózgu, a model jest w stanie poprawnie klasyfikować bardzo złożone zdania, w tym również te zawierające ironię lub sarkazm. Skuteczność tego modelu jest wprost proporcjonalna do ilości danych tekstowych przetworzonych przez sieć neuronową. W celu osiągnięcia 90% skuteczności, model powinien przetworzyć ok. 10 milionów obserwacji w zbiorze treningowym (Opitz, 2017). Dla porównania model analizy sentymentu bazujący tylko na słowach kluczowych jest w stanie osiągnąć skuteczność na poziomie wynoszącym tylko od 50% do 80%, taki model wówczas nie jest dobrze dopasowany oraz nie nadaje się do spełnienia określonego celu biznesowego (Opitz, 2017). Skuteczność modelu opartego na sieci neuronowej wykorzystuje się do poprawy doświadczenia klienta. System ten, będąc rozszerzonym o oparcie na aspekcie, jest w stanie wykryć wzrost negatywnych opinii i ich przyczynę. W ten sposób organizacja jest w stanie podjąć odpowiednie kroki w celu rozwiązania kryzysowej sytuacji. Innym zastosowaniem tego rozwiązania jest monitorowanie trendów na rynku oraz opinii o produktach, dzięki czemu można dokonać wyróżnienia cech produktów wzbudzających zadowolenie. Monitoring ten pozwala również na wykrycie krytyki wśród użytkowników bądź wyróżnienie elementów wzbudzających dużą uwagę w danym przedziale

czasowym (Opitz, 2017). Z usług Talkwalkera korzystają takie firmy oraz instytucje, jak: Goodwill, Spotify, Orange, Dentsu International czy Europejski Bank Inwestycyjny (Opitz, 2017). Strona produktu jest dostępna pod adresem: www.talkwalker.com.

2. Clarabridge

Produkt ten wykorzystuje różne źródła danych, jak: ankiety, recenzje, wiadomości wysłane pocztą elektroniczną czy wpisy z mediów społecznościowych – aplikacje, z których pobierane są dane, to m.in. Yelp, Glassdoor, Slack, Twitter, Facebook, WhatsApp czy też TripAdvisor (Clarabridge, 2019). Narzędzie to jest w stanie rozpoznać temat wypowiedzi oraz wyróżnić kluczowe informacje. System używa swojej własnej implementacji analizy sentymentu wykorzystującej podejście uwzględniające gramatykę oraz leksykalność w tekście. Główną funkcjonalnością tego produktu jest badanie zachowania klientów i predykcja przyszłego zachowania (Clarabridge, 2019). W ten sposób model jest w stanie dokonać poprawnej klasyfikacji źródeł tekstowych, w których występują m.in. zaprzeczenia, zdania warunkowe, sentyment w specyficznym kontekście w zdaniu oraz wyjątki. Przykładem wyjątku w źródle tekstowym jest zdanie zawierające słowo zmieniające wartość sentymentu, jak: „Te spodnie są zbyt pomarańczowe”. Słowo „pomarańczowe” jest neutralne w skali sentymentu, aczkolwiek, słowo to w połączeniu z wyrazem „zbyt” staje się wyrażeniem negatywnym (Clarabridge, 2019). Dlatego też można wyróżnić regułę określającą zależność, że słowa takie jak „zbyt” czy „za” (ang. *too*) w połączeniu ze słowem neutralnym tworzą w rezultacie wyrażenie negatywne. Aplikacja Clarabridge wykorzystuje w analizie sentymentu ponad 500 wyjątków dla procesowania danych napisanych w języku angielskim w celu poprawnej analizy danych tekstowych (Clarabridge, 2019). Ponadto produkt ten wykorzystuje szczegółową analizę sentymentu – model korzysta ze skali o wartościach od -5 do 5. Pozwala to na rozróżnienie stopnia wydźwięku emocjonalnego np. zdanie „Ten posiłek był dobry” uzyska ocenę 1, wypowiedź „To było najlepsze danie, jakie jadłem w życiu” osiągnie zaś ocenę 5. Dzięki temu zmienna wynikowa nie jest uproszczona do postaci binarnej, co pozwala na uzyskanie większej wiedzy o danej opinii. Do użytkowników tej platformy należą takie korporacje, jak: BMW, Under Armour, Accenture, Deloitte czy też Ernst & Young. Strona produktu znajduje się pod adresem: www.qualtrics.com/clarabridge.

3. MeaningCloud

Narzędzie to pozwala na integrację nowych danych z obszerną bazą wiedzy i dokonanie analizy sentymentu. Ponadto można użyć go do analizy:

- dokumentów, wraz z integracją z systemami CMS (ang. *Content Management System* – System Zarządzania Treścią) oraz RPA (ang. *Robotic Process Automation* – Zrobotyzowana Automatyzacja Procesów),
- VoC (ang. *Voice of the Customer*), czyli analizowania informacji zwrotnej od klienta,

- pracowników – wykrycia mocnych i słabych stron pracowników w organizacji w celu poprawy ich satysfakcji oraz osiągnięcia lepszej produktywności,
- mediów społecznościowych,
- serwisów kontaktowych w celu lepszej klasyfikacji incydentów i poprawy satysfakcji klienta.

Produkt ten oferuje również możliwość integracji z takimi aplikacjami, jak Microsoft Excel czy też ze stronami internetowymi w postaci wtyczek. Kolejną zaletą tej platformy jest opcja analizy sentymentu dla ponad 50 języków, w tym również języka polskiego. Oprócz tego MeaningCloud udostępnia również API, co pozwala na wykorzystanie go w aplikacjach organizacji i brak potrzeby wdrażania go w system wewnętrzny firmy. Do klientów tego rozwiązania należą takie firmy i organizacje, jak: Pfizer, ING, World Bank Group, Vocento oraz Unidad Editorial. Strona tego produktu znajdują się pod następującym adresem: www.meaningcloud.com.

4. Mention

Do rozwiązań tego systemu należy monitorowanie sytuacji, w których zostanie wspomniana nazwa organizacji (ang. *mention* – wzmianka) w mediach społecznościowych, portalach informacyjnych czy też wyszukiwarkach internetowych (Mention, 2021). Głównym zastosowaniem analizy sentymentu w tej platformie jest monitoring reputacji organizacji na rynku oraz porównanie jej z firmami konkurencyjnymi. Model analizy sentymentu używany przez Mention oprócz klasyfikowania opinii pozytywnych bądź negatywnych może również określić niektóre komentarze jako neutralne. W ten sposób redukowany jest niepotrzebny szum w analizie. Dodatkowo model ten oparty jest na sieciach neuronowych potrafiących dokonywać predykcji dla 16 różnych języków (Mention, 2021). Dzięki temu klienci korzystający z tego systemu mogą w krótkim czasie wykryć cenne informacje w mediach społecznościowych na temat ich firmy oraz dowiedzieć się, czy nagły wzrost zainteresowania firmą jest efektem pozytywnego bądź negatywnego zjawiska (Mention, 2021). Ponadto firmy korzystające z tego rozwiązania mogą w ten sposób znaleźć osoby, które zechcą promować markę, a ich rekomendację użyć do budowania wizerunku firmy. Inną zaletą tego systemu bazującego na analizie sentymentu jest możliwość szybkiej reakcji na sytuacje kryzysowe dzięki analizie w czasie rzeczywistym. Analiza sentymentu w tym systemie umożliwia również monitorowanie działań firm konkurencyjnych (Mention, 2021). Przypadkami użycia monitoringu konkurencji są m.in. obserwacja niezadowolonych klientów firmach będących rywalami na rynku oraz przyczyna braku ich satysfakcji (analiza sentymentu bazująca na aspekcie) (Mention, 2021). Z drugiej strony, w przypadku udanych inwestycji, kampanii reklamowych bądź udanego wprowadzenia produktu na rynek analiza sentymentu pozwala na wykrycie w mediach społecznościowych pomysłów, które

przyniosły przewagę konkurencyjnej firmie. Organizacjami oraz korporacjami, które korzystają z usług Mention, są: Benq, Deliveroo, Microsoft czy też Prisma Media (Mention, 2021). Strona główna Mention jest dostępna pod adresem: www.mention.com.

Analiza sentymentu jest komercyjnie najczęściej wykorzystywana jako narzędzie do monitorowania reputacji organizacji, zadowolenia klientów z usług oraz produktów czy też do badania satysfakcji wśród zatrudnionych w firmie. Główną przyczyną, dlaczego analiza sentymentu jest tak wartościowa, jest wzrost popularności mediów społecznościowych oraz pojawienie się portali z recenzjami, stanowiące źródło obszernej i cennej wiedzy, która nie była w takim stopniu dostępna 20 czy 30 lat temu. Rozwój nauczania maszynowego oraz opracowanie modeli sieci LSTM czy sieci konwolucyjnych pozwoliło również na wykrywanie określonych emocji w tekście czy też na wykrycie mowy nienawiści we wpisach publikowanych w mediach społecznościowych. Analizę sentymentu w praktyce można implementować na wiele sposobów. W praktyce używa się analizy sentymentu szczegółowej, by lepiej i precyzyjniej zbadać odczucia użytkowników. Z kolei do określenia przyczyny wyniku sentymentu powszechnie używana jest analiza sentymentu bazująca na aspekcie. Inną implementacją analizy sentymentu, w której użytkownik może zbadać opinie autora tekstu na różne podmioty takie jak: filmy, książka, osoby, jest analiza sentymentu oparta na podmiocie. Precyzja analizy sentymentu w systemach jest też wysoka dzięki zastosowaniu takich struktur matematycznych jak sieci neuronowe, które poprzez przetwarzanie dużych ilości danych, są dopasowywane do rozwiązania konkretnego problemu biznesowego na poziomie 90%.

5. Aylien

Rozwiązanie to polega na określeniu opinii autora o wielu podmiotach wykrytych w tekście. Głównym zastosowaniem tego rozwiązania jest analiza danych informacyjnych z różnych portali oraz monitorowanie najnowszych danych finansowych (Aylien, 2022). W rezultacie firmy korzystają z tekstu analizy do oszacowania ryzyka oraz wykrycia potencjalnych inwestycji na rynku. Ten rodzaj analizy sentymentu jest wykorzystywany do określenia wydźwięku emocjonalnego dla każdego unikatowego podmiotu występującego w tekście, również wtedy gdy wartości sentymentu dla tych podmiotów mogą się znacząco różnić między sobą. Przykładem może być następujący fragment tekstu: „Książka A jest przyzwoita w porównaniu do książki B, która była beznadziejna (Aylien, 2022). Nie jest ona jednak aż tak dobra jak książka C”. Podstawowa, prosta analiza sentymentu oceni ten fragment najprawdopodobniej jako opinię negatywną, podczas gdy widoczne są różne odczucia emocjonalne autora w stosunku do każdego wymienionego podmiotu w tym tekście. Analiza sentymentu oparta na podmiocie dokona dla tego fragmentu tekstu predykcji w postaci tabeli składającej się z:

- nazwy podmiotu,

- wartości sentymentu oszacowanej przez model w postaci prawdopodobieństwa bycia przypisanym do danej klasy,
- rodzaju podmiotu,
- liczby wspomnień (Aylien, 2022).

Podmiot	Wartość sentymentu	Rodzaj podmiotu	Liczba wspomnień
A	0,57 (pozytywna)	Tytuł książki	1
B	0,78 (negatywna)	Tytuł książki	1
C	0,55 (pozytywna)	Tytuł książki	1

Drugim przykładem może być z kolei następujący fragment: „Książka A jest świetna, ale książka B jest jeszcze lepsza. Szkoda, że książka C nie jest nawet w połowie tak dobra jak one”. Analiza sentymentu oparta na podmiocie może wówczas dokonać predykcji w postaci:

Podmiot	Wartość sentymentu	Rodzaj podmiotu	Liczba wspomnień
A	0,65 (pozytywna)	Tytuł książki	1
B	0,78 (pozytywna)	Tytuł książki	1
C	0,63 (negatywna)	Tytuł książki	1

Analiza sentymentu oparta na podmiocie jest skuteczna szczególnie w przypadkach, gdy firma bądź organizacja planuje w swoim systemie monitorować informacje z kraju i ze świata oraz śledzić poszczególne podmioty, jak: konkretne osoby, firmy, państwa lub miasta w celu monitoringu inwestycji, sytuacji na rynku bądź sytuacji politycznej dla określonego podmiotu (Aylien, 2022).

Produktem rynkowym wykorzystującym ten rodzaj analizy sentymentu jest np. Aylien. Aylien udostępnia swoje usługi również w postaci REST API, co pozwala na integrację z systemami organizacji. Do klientów tej platformy zalicza się takie grupy i firmy, jak: IHS Markit, Wells Fargo oraz AON (Aylien, 2022). Strona tego produktu jest dostępna pod adresem: www.aylien.com.

3.4 Analiza sentymentu w produkcji leków

Zastosowanie analizy sentymentu w celu zbadania leków przynosi wiele korzyści. Jedną z zalet tego rozwiązania jest fakt, iż w przypadku wprowadzenia nowego produktu na rynek z modelu mogą skorzystać pacjenci, którzy dzięki temu mogą podjąć decyzję, jaki lek zakupić. Z kolei producenci leków oraz klinicyści są w stanie określić cenną informację nt. opinii pacjentów oraz

dowiedzieć się, czy są jakiekolwiek zdarzenia niepożądane (Na i Kyaing, 2015). Analiza sentymentu pozwala na podsumowanie komentarzy oraz wyróżnienie cech produktu, które mogą być przyczyną satysfakcji bądź niezadowolenia wśród pacjentów. Do badania komentarzy dotyczących leków powszechnie używaną implementacją analizy sentymentu jest analiza sentymentu bazująca na aspekcie (Na i Kyaing, 2015). Analiza bazująca na aspekcie pozwala na wyróżnienie takich istotnych szczegółów w opinii o leku, jak:

- ogólna opinia (określenie zadowolenia autora tekstu),
- efektywność (czy środek przyniósł poprawę stanu samopoczucia),
- skutki uboczne (czy wystąpiły efekty niepożądane),
- dolegliwość,
- koszt leku,
- dawka.

W celu uzyskania lepszego dopasowania modelu, w analizie sentymentu używanej do badania produktów farmaceutycznych używa się również rozbicia zdania na mniejsze części. Przykładem może być następujące zdanie: „Zażyłem ten lek i zadziałał on bardzo dobrze”, które zostanie podzielone na dwie części. Pierwszym zdaniem będzie: „Zażyłem ten lek”, a drugim zaś: „I zadziałał on bardzo dobrze”. Wśród takich fragmentów wyróżnia się dwie grupy: zdania podrzędne i zdania niezależne. Zdanie niezależne (inaczej nazywane zdaniem głównym) jest to takie zdanie, które zawiera podmiot i orzeczenie, dlatego też zachowuje ono samo w sobie logiczne znaczenie. Wiele zdań niezależnych może zostać połączonych za pomocą średnika lub przecinka z dodatkiem spójnika (są to takie słowa jak „gdyż”, „ale”, „więc”, „albo”, „i” itp.). Zdania podrzędne są to takie zdania, które wzbogacają zdania główne o dodatkową informację, lecz nie mogą występować same jako pojedyncze zdanie. Zdania podrzędne zaczynają się od takich spójników, jak: „pomimo”, „jeżeli”, „jakby”, „po” itp. Do rozbicia zdań na człony używa się struktur takich jak drzewa (konkretnie drzewa wyprowadzenia) (Na i Kyaing, 2015). Ważnym czynnikiem w poprawnej analizie sentymentu w badaniu leków jest korzystanie z obszernego leksykonu sentymentu. W odróżnieniu do analizy innych produktów, słowa bądź związki niektórych wyrazów mogą mieć bardzo spolaryzowany wynik sentymentu. Do przypadków, które w analizie sentymentu mogą mieć duże znaczenie, może należeć np. słowo „działać”. Wówczas takie słowo powinno mieć wysoki scoring sentymentu ze względu na to, że wskazuje, że dany lek działa. Natomiast takie związki wyrazów, jak: „wysoki poziom cukru”, „wzrost ciśnienia” czy też „utrata włosów”, wskazują na negatywną opinię i algorytm analizy sentymentu powinien być w stanie to określić (Na i Kyaing, 2015). W analizie sentymentu służącej do badania leków używane są również specjalne reguły umożliwiające uzyskanie lepszego dopasowania modelu.

1. Intensyfikacja, łagodzenia, minimalizacja, maksymalizacja

W podstawowych implementacjach analizy sentymentu nie uwzględniana jest reguła intensyfikacji, łagodzenia, minimalizacji czy maksymalizacji. Intensyfikacja jest to przypadek, w którym zostanie użyty przysłówek wzmacniający wydźwięk emocjonalny w tekście. Przykładem intensyfikacji może być użycie takich słów, jak: „bardzo”, „ogromnie”, „zdecydowanie”. Wówczas para wyrazów: „zdecydowanie lepszy” uzyska znacząco wyższy wynik sentymentu, niż w przypadku gdy we fragmencie tekstu znajdzie się tylko słowo „lepszy” (Na i Kyaing, 2015). Odwrotnym procesem do intensyfikacji jest działanie łagodzące. Do słów łagodzących będą należeć takie przysłówki, jak np. „trochę”, „nieco” itp. Wtedy para wyrazów „trochę lepszy” będzie mieć o połowę mniejszy wynik sentymentu niż w przypadku, gdy we fragmencie znajdzie się tylko słowo „lepszy”. Proces maksymalizacji jest natomiast działaniem wpływającym na polaryzację słowa, minimalizacja zaś powoduje zbliżenie wyniku sentymentu do zera (Na i Kyaing, 2015). Do słów maksymalizujących będą należeć takie wyrazy, jak: „kompletnie”, „totalnie”, „całkowicie”. Słowa minimalizującymi będą zaś: „nikła”, „minimalna”, „pomijalnie” itp.

2. Wartościowanie pozytywne i negatywne

Zasada wartościowania pozytywnego i negatywnego dotyczy słów, które określają z góry, czy wartość sentymentu jest pozytywna bądź negatywna. Do słów wartościujących negatywnie należą czasowniki takie jak „nienawidzić” czy „cierpieć”. Pozytywnie wartościującymi słowami będą takie czasowniki, jak: „pomóc”, „poprawić” czy „polepszyć”. W przypadku zdań o neutralnym wydźwięku emocjonalnym słowa wartościujące pozytywnie zmieniają wartość sentymentu z 0 na 0,5 – zaś negatywnie wartościujące na -0,5 (Na i Kyaing, 2015).

3. Redukcja zaburzenia

Jest to zasada stosowana w przypadkach, gdy występują następujące związki:

- czasownik o znaczeniu ograniczającym wraz z zaburzeniem bądź chorobą (w standardowej implementacji analizy sentymentu zdanie: „Ten lek ograniczył ból”, może zostać błędnie zaklasyfikowane jako zdanie o sentymencie negatywnym) (Na i Kyaing, 2015);
- choroba wraz z czasownikiem o znaczeniu ograniczającym (Przykład: „Ból głowy ustąpił”) (Na i Kyaing, 2015).

4. Połączenie przyimka z dolegliwością

Zasada ta określa zależność pomiędzy przyimkami a słowami oznaczającymi chorobę bądź dolegliwość. Przykładowo, bez zastosowania tej reguły wyrażenie: „dobry lek na gorączkę”, może zostać błędnie sklasyfikowane jako opinia negatywna (wartość bezwzględna wyniku sentymentu słowa: „dobry” jest niższa od bezwzględnej wartości sentymentu słowa: „gorączka”) (Na i Kyaing, 2015). Natomiast przy użyciu tej zasady wyrażenie to zostanie sklasyfikowane jako opinia

pozytywna. Inne przykłady klasyfikacji przy użyciu tej reguły:

- „nieefektywny lek na ból” (opinia negatywna),
- „pomógł przy katarze” (opinia pozytywna),
- „umierać z bólu” (opinia negatywna).

5. Łączniki zaprzeczające

Reguła ta polega na zignorowaniu podrzędnych zdań zawierających zaprzeczające łączniki, takie jak: „pomimo”, „mimo że”, „przeciwnie”. Wówczas algorytm skupia się tylko na analizie sentymentu zdania głównego (Na i Kyaing, 2015). Na przykład w zdaniu: „Pomimo że lek zadziałał dobrze, doznałem bólów głowy”. W tym przypadku zdanie podrzędne: „Pomimo że lek zadziałał dobrze”, zostanie zignorowane i nie będzie brane pod uwagę w klasyfikacji całości zdania (Na i Kyaing, 2015).

6. Czasowniki frazowe (ang. *phrasal verbs*)

W języku angielskim czasowniki frazowe powstają poprzez połączenie czasownika z przyimkiem. Połączenie to powoduje, że czasownik frazowy ma nowe znaczenie, często zupełnie inne niż pojedyncze człony, z których powstał (Na i Kyaing, 2015). W implementacji analizy sentymentu, która nie uwzględnia tej zasady, wyrażenie *back off* może zostać sklasyfikowane błędnie jako związek o negatywnym wydźwięku emocjonalnym, ponieważ słowa „back” i „off” będą analizowane osobno oraz dosłownie (Na i Kyaing, 2015). Z kolei przy zastosowaniu tej reguły wyrażenie to zostanie ocenione jako pozytywne (ang. back off – pol. cofnąć się, wycofać się, odsunąć się) (Na i Kyaing, 2015).

7. Pytania

Jeżeli fragment tekstu kończy się znakiem zapytania, wówczas neutralizowany jest jego wydźwięk sentymentu, a jego scoring wynosi 0 (Na i Kyaing, 2015).

8. Znaczenie biznesowe

Analizę sentymentu w badaniu leków oprócz zastosowania wymienionych reguł opiera się również na algorytmach takich jak: maszyna wektorów nośnych, klasteryzacja tekstu oraz deep learning. Oprócz znajomości tych narzędzi istotna jest znajomość dziedziny biznesu – w tym przypadku farmacji oraz medycyny. Znajomość pojęć związanych z lekami, różnych nazw objawów oraz schorzeń jest kluczowa w kontekście stworzenia odpowiednio dopasowanego modelu analizy sentymentu. Ponadto dobrze dopasowany model ma dużą wartość biznesową zarówno dla pacjentów, którzy mogą korzystać z systemu opartego na tych rozwiązaniach w celu znalezienia najlepszego leku na daną chorobę, jak i dla producentów leku, którzy w krótkim czasie mogą być w stanie zidentyfikować braki w produkcji, skutki uboczne i zareagować na te czynniki.

4. Zastosowanie NLP w badaniu ocen leków

4.1 Opis zbioru danych i przedstawienie schematu działania systemu

Zbiór danych zawiera komentarze pacjentów dotyczące użytych przez nich leków na określoną dolegliwość. Komentarze zostały zebrane z różnych stron poświęconych recenzjom leków poprzez współpracę uniwersytetu Kansas State i Uniwersytetu Technicznego w Dreźnie (Kallumadi i Gräßer, 2018). Wszystkie komentarze zostały napisane w języku angielskim, lecz nie podano informacji o krajach, z których pacjenci pochodzili. Zbiór danych liczy łącznie ponad 215 tysięcy obserwacji. Oceniono w nim 3436 leków na 885 różnych dolegliwości. Dane pochodzą z lat 2008–2017 i składają się z poszczególnych zmiennych:

- uniqueID: identyfikator komentarza,
- drugName: nazwa leku,
- condition: dolegliwość lub choroba,
- review: treść komentarza,
- rating: ocena, którą wystawił pacjent w skali 1–10,
- date: data opublikowania komentarza,
- usefulCount: ilość użytkowników, którzy uznali dany komentarz za przydatny,
- Year: rok, w którym komentarz został opublikowany.

Zbiór danych jest dostępny pod adresem:

<https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29>

(Kallumadi i Gräßer, 2018)

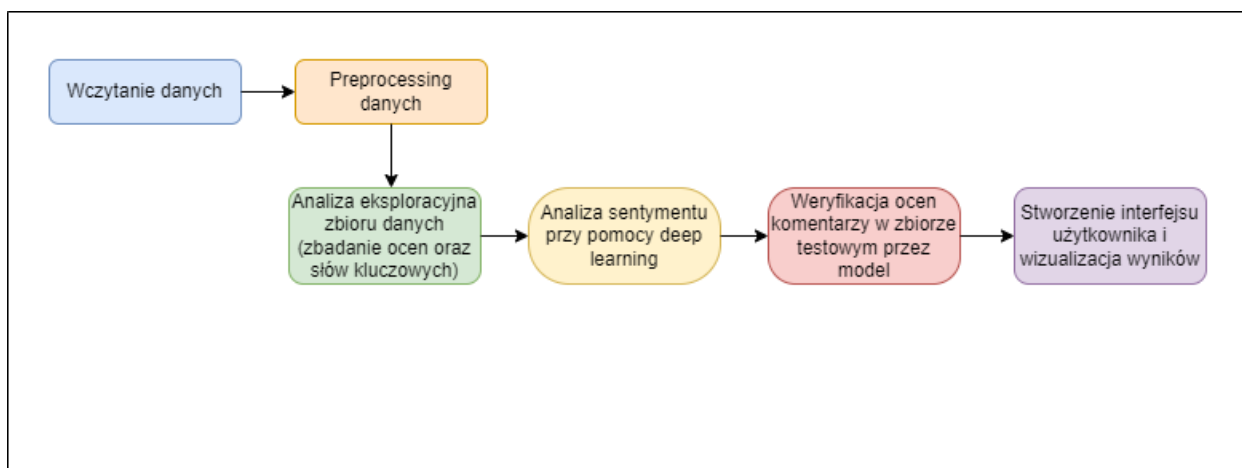


Diagram 1: Diagram przedstawiający przepływ procesów w systemie

Źródło: Opracowanie własne

Na diagramie 1 przedstawiono schemat przepływu procesów w systemie. W pierwszej kolejności dokonywane jest wgranie zbioru wejściowego, następnie przeprowadzany jest

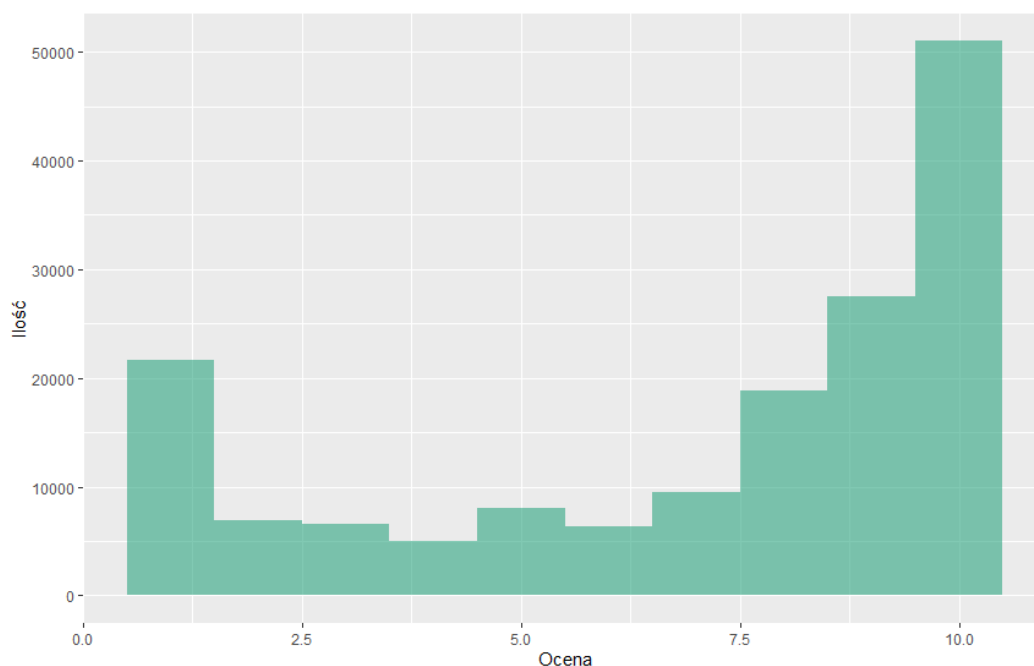
preprocessing w celu doprowadzenia danych do stanu, w którym komputer jest w stanie je analizować. Analiza w czasie odbywa się w następnej kolejności:

1. Analiza sentymentu poprzedzona analizą eksploracyjną (badanie rozkładu ocen, średnich itp.).
2. Ocena leków na podstawie komentarzy pacjentów za pomocą modelu analizy sentymentu opartej o sieć neuronową.

W kolejnych krokach dokonywana jest analiza eksploracyjna zbioru danych, badany jest m.in. rozkład ocen, słowa o najwyższej kontrybucji czy też leki wysoko oceniane przez pacjentów. W następnym etapie realizowana jest analiza sentymentu przy użyciu sieci neuronowej w celu klasyfikacji komentarzy pozytywnych i negatywnych. Po skonstruowaniu modelu i poddaniu go procesowi treningu oraz testowania, następuje użycie go do weryfikacji komentarzy niezawierających się w zbiorze treningowym.

4.2 Analiza eksploracyjna

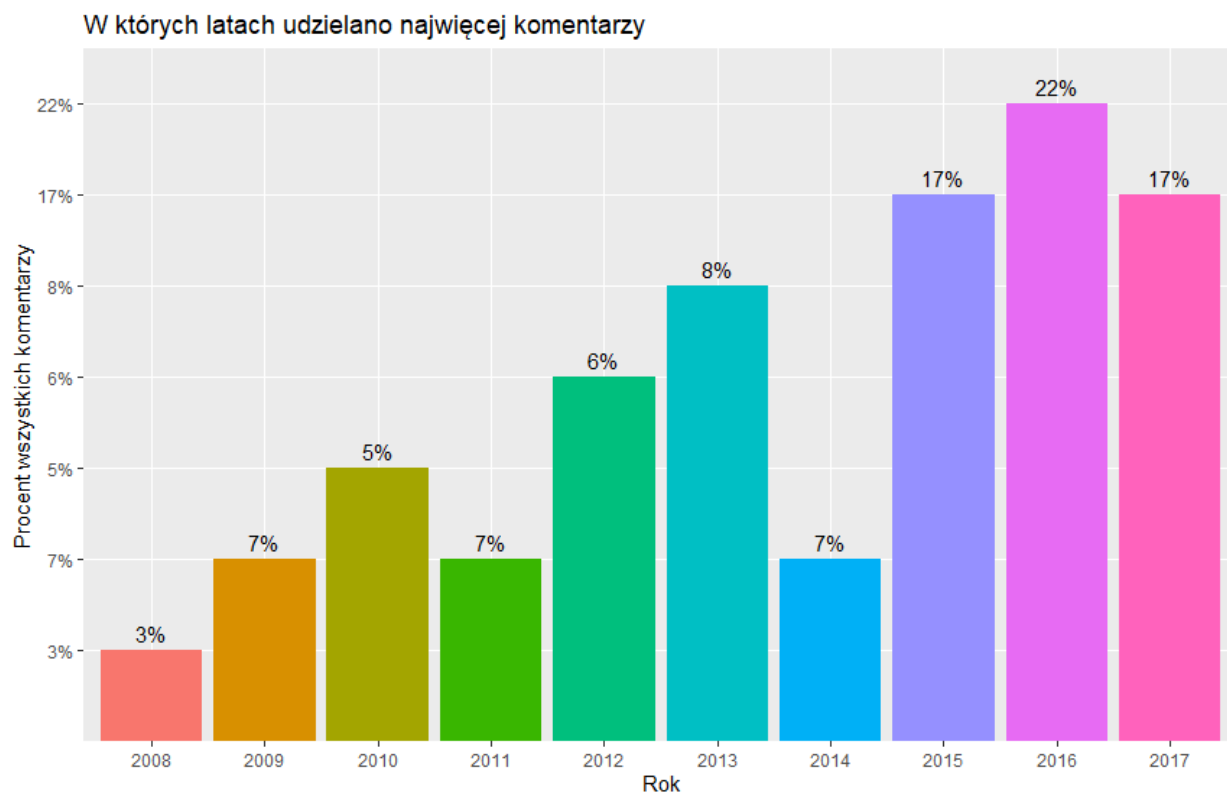
Do analizy eksploracyjnej użyto bibliotek graficznych takich jak: ggplot2 oraz plotly. Na początku zbadano histogram ocen leków. W następnym kroku dokonano analizy sentymentu opartej na sieci neuronowej stworzonej za pomocą biblioteki keras. Na wykresie 15 przedstawiono jego rozkład, który został stworzony przy pomocy biblioteki ggplot2 (Sievert, 2019) (Lander, 2017).



Wykres 6: Histogram ocen dla wszystkich leków

Źródło: Opracowanie własne

Według wykresu 6 najczęściej jest skrajnych ocen, czyli 1 oraz 9 i 10. Ta ostatnia ocena pojawia się najczęściej, oznacza to, że większość opinii w zbiorze jest bardzo pozytywna i rekomendująca dany lek. W następnym kroku zbadano jak często oceniano leki w latach 2008–2017.



Wykres 7: Procent wszystkich komentarzy w poszczególnych lat

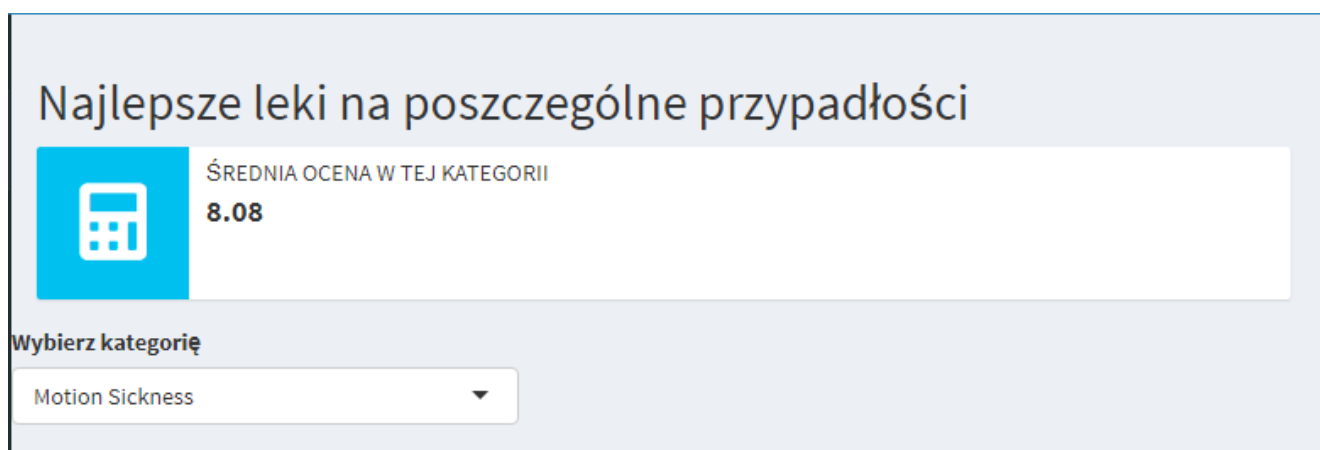
Źródło: Opracowanie własne

Według wykresu 7 najwięcej komentarzy zebrano w roku 2016 – ponad 20% z całego zbioru. Oprócz tego dużo komentarzy napisano też w latach 2015 i 2017 (ponad 15%), w pozostałych latach wyniki już były poniżej 10%.

W kolejnym kroku zbudowano interaktywny dashboard za pomocą biblioteki shiny. Na wykresach 8, 9, 10, 11 przedstawiono interaktywne wskaźniki KPI dla wybranej kategorii schorzeń czy dolegliwości. Ponadto w sekcji przedstawiono również wykres słupkowy dla najwyższej ocenianych leków w wybranej kategorii (Sievert, 2019).

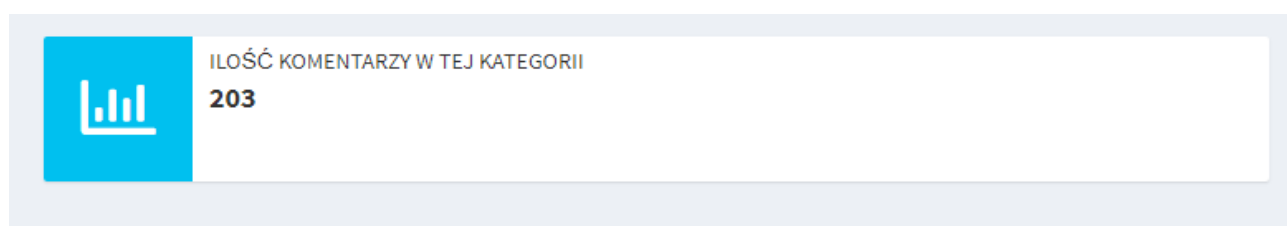
Wykresy te pokazują, że leki na chorobę lokomocyjną były oceniane bardzo wysoko, średnia ocena wynosi ponad 8 przy ponad 200 komentarzach. Ponadto komentarze te zebrały łącznie prawie 4000 poleceń, co wskazuje, że wiele osób poleca te środki. Do najwyższej ocenianych leków na tę dolegliwość należą Cyclizine, Marezine, Travel-Eze czy Dramamine: wszystkie z tych

4 leków zebrały średnią ocenę ponad 9,5.



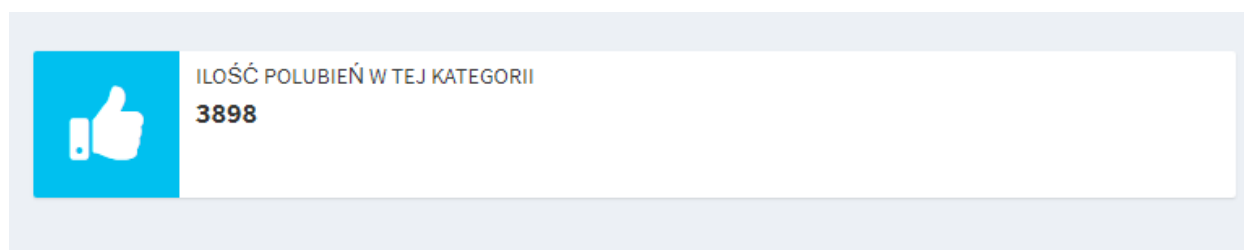
Wykres 8: Średnia ocena dla leków na chorobę lokomocyjną

Źródło: Opracowanie własne



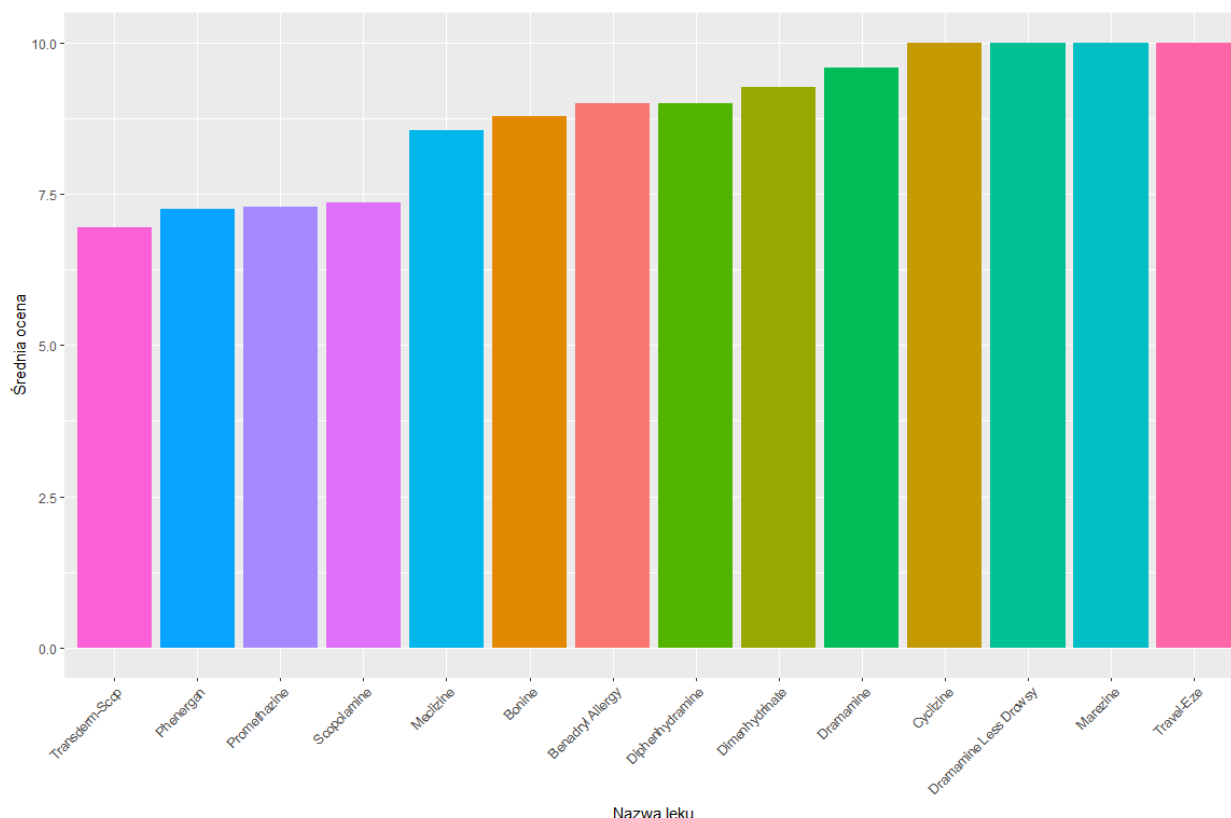
Wykres 9: Łączna liczba komentarzy dla leków na chorobę lokomocyjną

Źródło: Opracowanie własne



Wykres 10: Łączna liczba polubień dla leków na chorobę lokomocyjną

Źródło: Opracowanie własne



Wykres 11: Leki na przykładową dolegliwość – chorobę lokomocyjną, z najwyższą średnią ocen
 Źródło: Opracowanie własne

W następnej kolejności zaprezentowano wyniki klasyfikacji modelu, gdzie model sieci w zależności od podanego tekstu dokonywał weryfikacji, czy opinia jest pozytywna bądź negatywna. Do zbadania dopasowania modelu do danych wykorzystano bibliotekę pROC i caret.

4.3 Podział na n-gramy

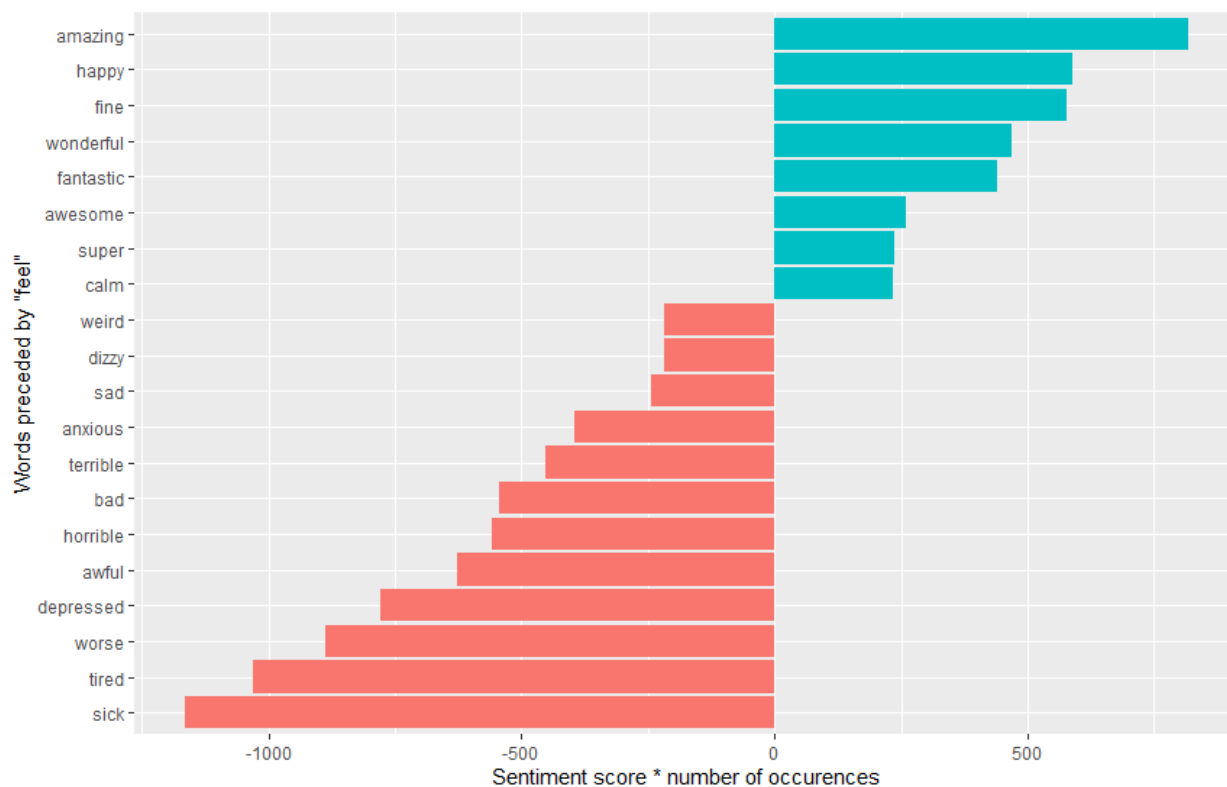
W pierwszym kroku usunięto ze zbioru komentarzy wyrazy znajdujące się w stop liście. Są to słowa budujące logikę zdania, mogą to być takie wyrazy jak np. spójniki („ponieważ”, „oraz”, „bo”) czy też słowa popularne („mp3”, „pc”). Słowa te nie wpływają na identyfikację tekstu oraz nie posiadają emocjonalnego wydźwięku, dlatego też usuwa się je w celu zredukowania wielkości zbiorów i oszczędzenia pamięci operacyjnej. Następnie przy pomocy biblioteki tidytext pobrany został leksykon AFINN (nazwa pochodzi od imienia i nazwiska autora leksykonu, Finna Årupa Nielsena) (dtu.dk, 2011). Jest to słownik, w którym każde słowo ma przypisaną wartość sentymentu. W tym leksykonie słowa o negatywnym wydźwięku przyjmują ujemne wartości, te o pozytywnym zaś zwracają wartości dodatnie (np. wyrazy *amazing* czy *brehtaking* zwracają 5, natomiast wulgaryzmy przyjmują wartość -5) (Silge i Robinson, 2017). W dalszej kolejności obliczane są liczności n-gramów, czyli występujących obok siebie n słów. Dla tego przypadku za n przyjęto 2, ponieważ można w ten sposób ukazać związek między bezpośrednio sąsiadującymi

ze sobą słowami. Dodatkowo pozwala to na obliczenie ich kontrybucji na podstawie częstości występowania tych słów oraz ich wartości sentymentu. Po obliczeniu wystąpienia słów obliczona została kontrybucja każdego słowa wzorem (Silge i Robinson, 2017):

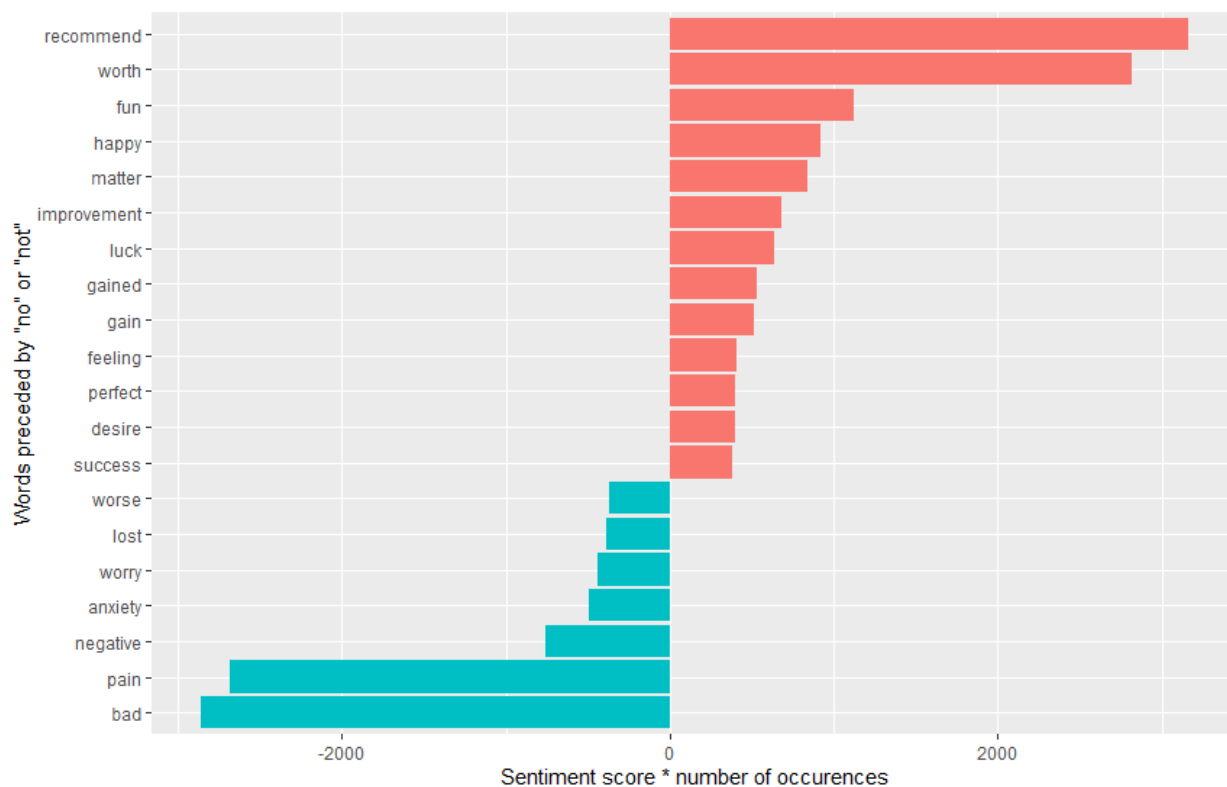
$$\text{Kontrybucja} = \text{Wartość sentymentu słowa} * \text{liczba wystąpień}$$

gdzie wartość sentymentu słowa oznacza wartość sentymentu dla danego słowa w leksykonie AFINN, a liczba wystąpień dotyczy liczby wystąpień w dokumencie (dla tego przypadku wzięto pod uwagę cały zbiór komentarzy).

Kontrybucja pozwala na obliczenie, jak bardzo dany term (czyli słowo) lub n-gram wpływa na wydźwięk tekstu (czy jest to zdanie nacechowane pozytywnie – przy wypadkowej kontrybucji o wartości większej niż 0 lub negatywnie w przypadku, gdy łączna kontrybucja danego zdania bądź tekstu jest ujemna). W następnym kroku zbadano odczucia pacjentów wynikające z komentarzy.



Wykres 12: Słowa, które były poprzedzone słowem „feel” w różnej odmianie gramatycznej
Źródło: Opracowanie własne



Wykres 13: Słowa które najczęściej były poprzedzone słowem „no” lub „not” o najwyższej kontrybucji

Źródło: Opracowanie własne

Wykres 12 pokazuje, że z takimi słowami jak *feel* (pol. czuć) w kontekście pozytywnym są związane takie słowa, jak: *amazing* (pol. niesamowicie), *happy* (pol. szczęśliwy) czy *fine* (pol. w porządku). Jeśli chodzi natomiast o słowa o negatywnym wydźwięku, które w dużym stopniu występowały ze słowem *feel*, dominują słowa typowe dla chorób, takie jak: *sick* (pol. chory), *tired* (pol. zmęczony) czy *depressed* (pol. przygnębiony). Ważnym słowem jest również *worse* (pol. gorzej), które sugeruje, że opinia pacjenta jest negatywna, a lek przyniósł odwrotny skutek do zamierzonego (Silge i Robinson, 2017).

Wykres 13 pokazuje zaś, że z takimi słowami jak *not* lub *no* powiązane są takie słowa, jak: *recommend* (pol. polecać), *worth* (pol. warto), *fun* (pol. zabawnie). W przypadku słów o negatywnym znaczeniu najczęściej występują takie słowa jak *bad* (pol. źle) oraz *pain* (pol. ból). Kolory na wykresie 8 zostały specjalnie odwrócone ze względu na zmieniony kontekst, tutaj słowa o pozytywnym znaczeniu będą oznaczać, że opinia o leku była najprawdopodobniej negatywna. Natomiast połączenie słów *no* i *pain* czy *not* i *worry* będzie wskazywać, że pacjent dany lek poleca.



Wykres 14: Chmura słów najczęściej przedstawiających się w komentarzach

Źródło: Opracowanie własne

Wykres 14 przedstawia chmurę słów w tekście. Słowa pokolorowane na zielono to wyrazy o pozytywnym znaczeniu, w kolorze czerwonym są zaś słowa o znaczeniu negatywnym. Wyrazy takie jak: *pain*, *anxiety*, *depression* czy *symptoms*, wskazują na problemy zdrowotne, jakie mieli pacjenci. Z kolei słowa: *helped*, *recommend*, *worth*, *effective*, dotyczą już oceny leku.

W następnym kroku zbadano współczynnik *tf-idf* do zbadania wagi słów na podstawie liczby ich wystąpień dla poszczególnych leków. Współczynnik *tf-idf* jest obliczany wzorem:

$$tf - idf_{i,j} = tf_{i,j} \times idf_i$$

gdzie:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = \log \frac{|D|}{|\{d: t_i \in d\}|}$$

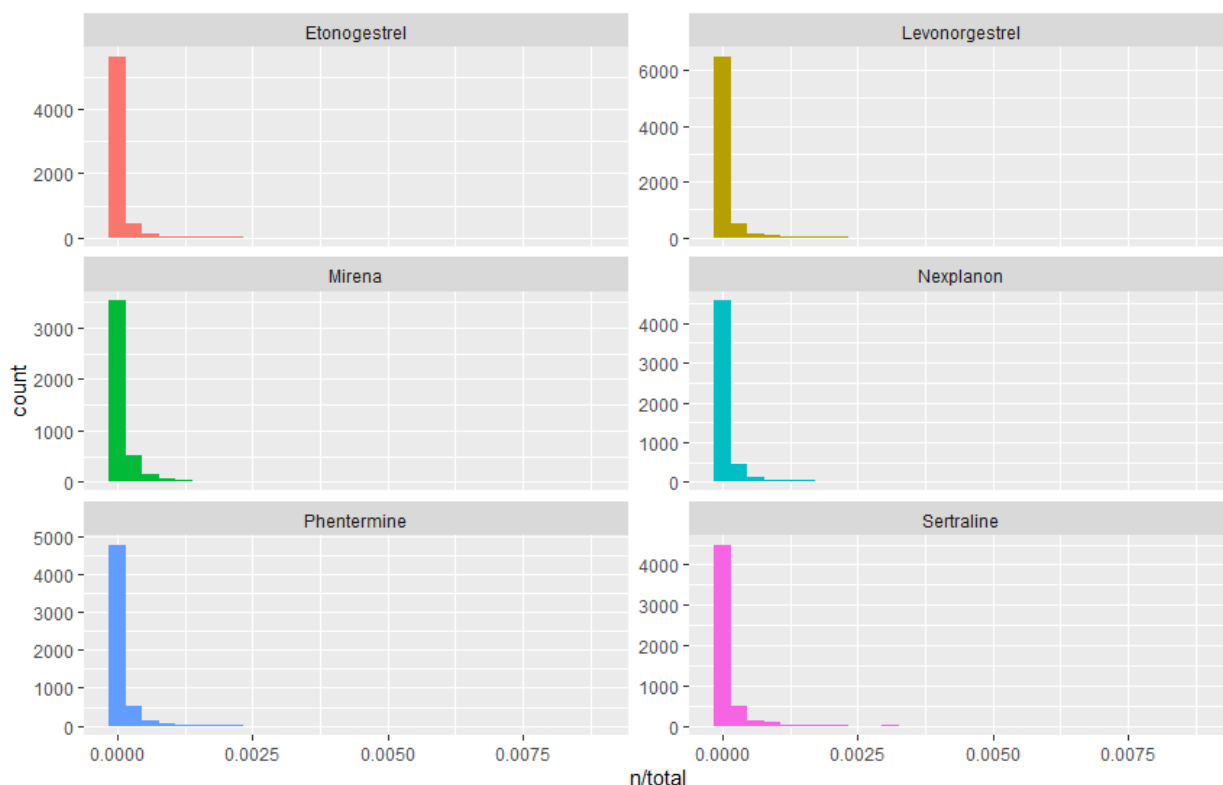
$|D|$ – oznacza liczbę dokumentów w korpusie (korpus to zbiór źródeł tekstowych używany do trenowania modelu),

$n_{i,j}$ – oznacza licznosc i -tego termu t_i w j -tym dokumencie d_j ,

$n_{k,j}$ – oznacza łączną ilość k termów w j -tym dokumencie d_j ,

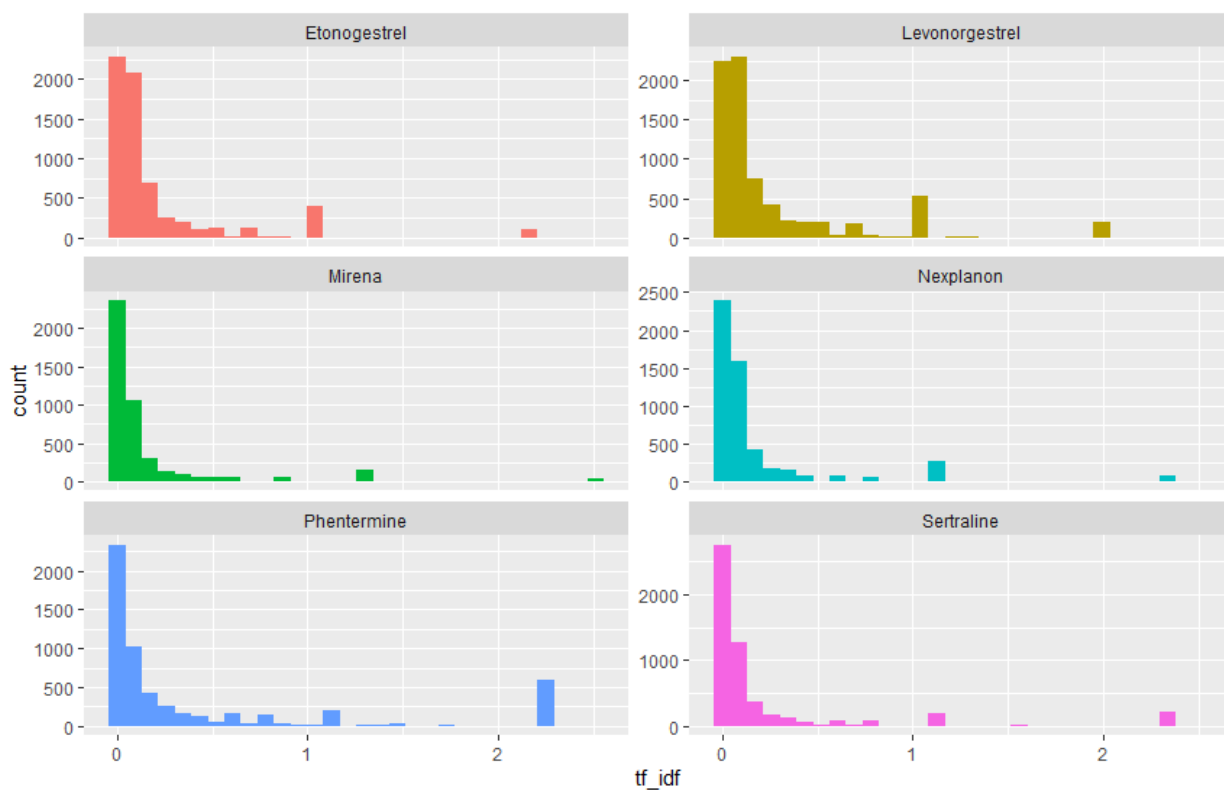
$|\{d: t_i \in d\}|$ – oznacza liczbę dokumentów zawierających przynajmniej jedno wystąpienie i -tego termu (Silge i Robinson, 2017).

Współczynnik tf-idf jest istotny, gdyż informuje, czy zbiór danych tekstowych zawiera istotne informacje oraz czy można na nim robić analizę NLP w celu osiągnięcia określonego celu biznesowego. Analiza ta została przeprowadzona dla całego ogółu zbioru komentarzy.



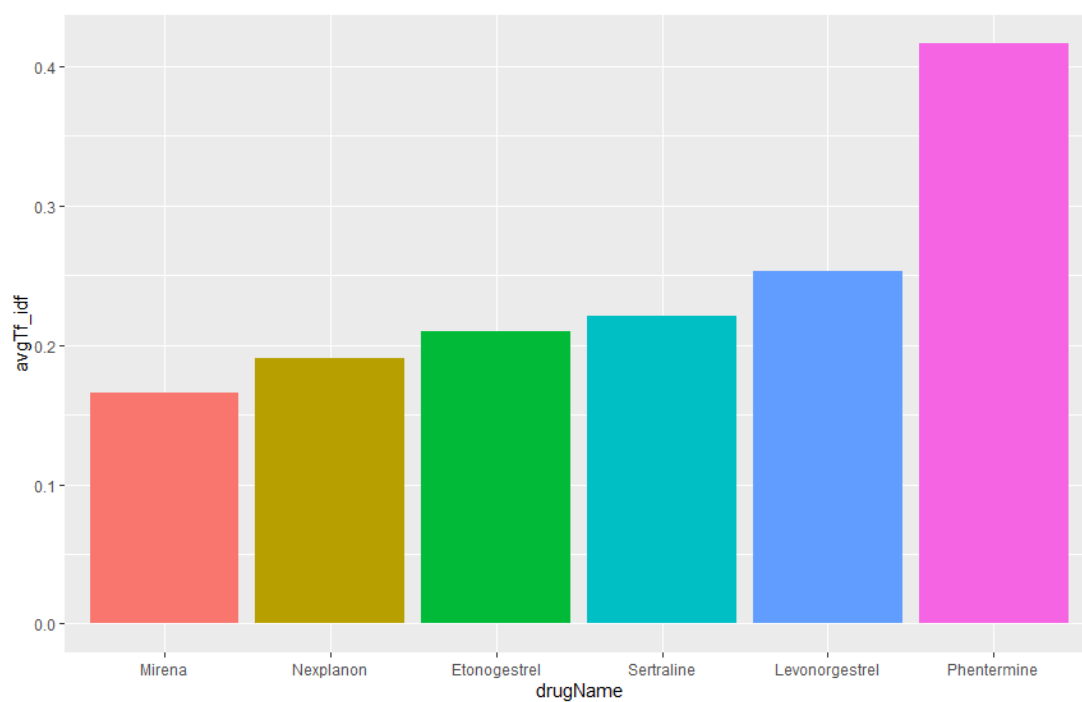
Wykres 15: Histogram współczynnika częstotliwości występowania słów(tf) w komentarzach dla 6 najczęściej komentowanych leków
Źródło: Opracowanie własne

Wykres 15 pokazuje, że dla każdego z 6 najbardziej ocenianych leków, tj. Etonogestrel, Levonorgestrel, Nexplanon i Mirena (wszystkie z nich są środkami antykoncepcyjnymi), Phentermine (lek wspomagający odchudzanie) i Sertraline (środek antydepresyjny), zdecydowanie najwięcej jest słów rzadko powtarzających się. Oznacza to, że wśród komentarzy dla tego leku nie brakuje słów istotnych dla znaczenia całego zdania zgodnie z prawem Zipfa (Silge i Robinson, 2017).



Wykres 16: Histogram tf_idf termów dla 6 najczęściej ocenianych leków

Źródło: Opracowanie własne



Wykres 17: Średni tf_idf dla 6 najczęściej ocenianych leków

Źródło: Opracowanie własne

Na histogramie tf_idf termów w komentarzach dla 6 najczęściej ocenianych leków

dominują słowa o małej wadze (ponad 4000 termów przeciętnie dla zbadanego leku o tf-idf bliskim 0). To nie oznacza jednak, że zbiory tych komentarzy są nie znaczące. Na wykresie 17 widać, że dla każdego ze zbadanych leków jest grupa słów o współczynniku tf-idf powyżej 1 czy nawet 2. Oznacza, że w zbiorze są słowa o dużej wadze, które pozwalają na zbadanie kontekstu wypowiedzi. Jest to kluczowe w możliwości stworzenia użytecznego modelu. Ma to odzwierciedlenie również w wykresie średnich współczynnika tf-idf (wykres 17), przy żadnym leku średnia tf_idf nie wynosi poniżej 0,1.

4.4 Budowa sieci neuronowej do analizy sentymentu

Po dokonaniu: analizy eksploracyjnej, zbadaniu ocen w zbiorze treningowym, zbadaniu słów o dużej kontrybucji i zbadaniu tf_idf, można dojść do wniosku, że w zbiorze komentarzy znajdują się wyrazy o wysokiej wadze istotności (czyli takie, które posiadają wysokie tf-idf) jak i bigramy (czyli n-gramy składające się z 2 słów) o dużej kontrybucji. Podział zbioru danych został dokonany przez publikatorów zbioru komentarzy i opublikowane je w postaci dwóch zbiorów danych (treningowego i testowego). Analizę wykonano na zbiorze treningowym składającym się ze 161 297 obserwacji oraz 8 zmiennych (zbiór testowy liczy 53 471 obserwacji). Na tej podstawie można dojść do wniosku, że ten zbiór danych nadaje się do skonstruowania modelu analizy sentymentu korzystającego z sieci neuronowej w celu zaklasyfikowania, czy dany komentarz był opinią pozytywną bądź negatywną. Najpierw dokonano konwersji zmiennej oceny na zmienną binarną. Za opinię negatywną przyjęto oceny od 0 do 6, za pozytywną recenzję uznano zaś oceny od 7 do 10. W tym przypadku uznano, że leków nie da się ocenić neutralnie, dlatego leki ocenione od 4 do 6 też zostały zaklasyfikowane jako ocenione negatywnie. W następnej kolejności zbudowano model sieci neuronowej. Jej architekturę przedstawiono na diagramie 2.

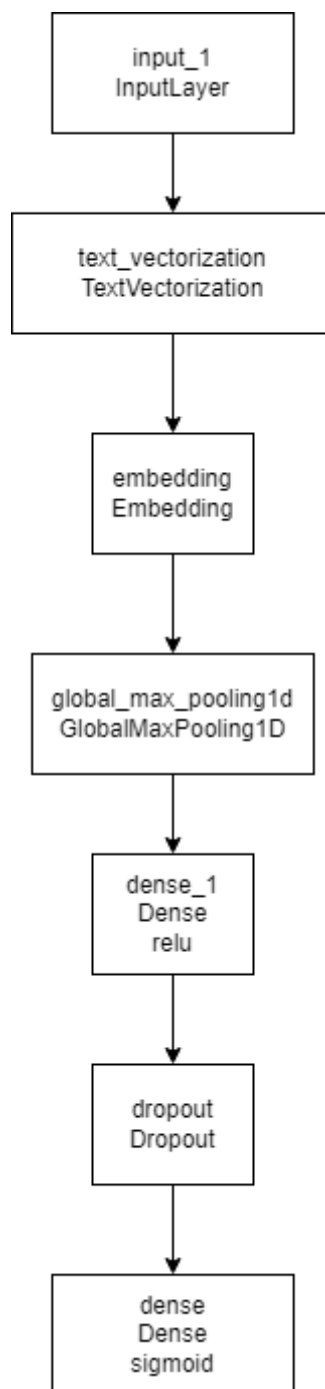


Diagram 2: Architektura modelu sieci neuronowej

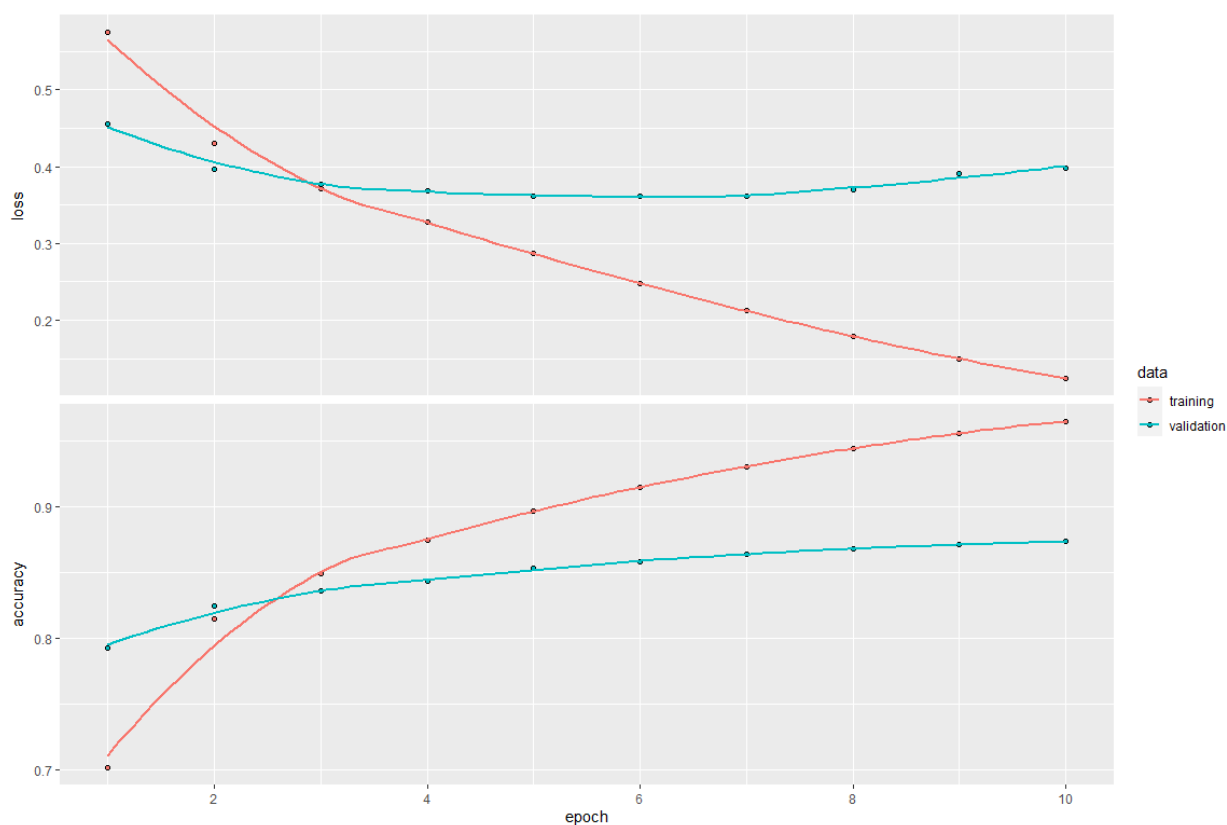
Źródło: Opracowanie własne

W pierwszej kolejności tworzona jest warstwa wejściowa jednowymiarowa, następnie warstwa odpowiadająca za wektoryzację tekstu (Chollet, 2018). Tekst jest wówczas konwertowany na wektor bitów. Jeżeli dane słowo wystąpi w danym tekście wejściowym, to wówczas wartość w indeksie tego słowa w wektorze będzie wynosiła 1 i analogicznie 0, gdy tego słowa nie będzie. Znacznie ułatwia to wówczas przetwarzanie tekstów przez komputer. Drugą warstwą jest embedding, warstwa ta przyjmuje na wejściu słowa w zwektoryzowanej wersji. Warstwa ta może być użyta do klasteryzacji tekstu lub jako część modelu uczenia głębokiego

(Brownlee, 2017). W tym przypadku zostanie użyta jako część modelu, która podobnie jak cały model, poddana będzie procesowi treningu. W kolejnej warstwie zainicjowana jest warstwa jednowymiarowego max pooling (pooling jest jednowymiarowy gdyż na wejściu znajdują się przekształcone w wektory słowa, a nie macierze). Max pooling umożliwia skupienie się na wyróżniających się wartościach w macierzach (bądź wektorach), w przeciwieństwie do pooling uśredniającego (Du i Shanker, 2015). W kolejnych fazach używana jest warstwa gęsta składająca się z 16 neuronów, w każdym z nich znajduje się funkcja aktywacji ReLU, po której dokonywany jest dropout na poziomie 0,5. Wartość tego parametru służy do ustawiania prawdopodobieństwa, z którym wartość na wyjściu neuronu zostanie odrzucona. W warstwach wejściowych rekomendowane jest ustawianie współczynnika odrzucenia na poziomie bliższym 1 jak np. 0,8 – z kolei w ukrytych warstwach najczęściej używaną wartością tego parametru jest 0,5 (Brownlee, 2018). Ostatnim elementem modelu jest warstwa wyjściowa gęsta, w której funkcją aktywacji jest funkcja sigmoidalna zwracająca prawdopodobieństwo, że komentarz jest pozytywny (Chollet, 2018).

Zagrożeniem wynikającym z budowy i publikacji takiego modelu jest złe dostrojenie modelu. W rezultacie użytkownicy mogą sugerować się złymi odpowiedziami modelu, przez co można narazić wielu ludzi na utratę zdrowia. Dlatego też bardzo istotną kwestią w konstrukcji sieci, jest przetestowanie jej na danych testowych i zbadaniu jej działania za pomocą różnych metryk.

Po zbudowaniu sieci poddano ją uczeniu na zbiorze treningowym. Dla znaczącego przyspieszenia procesu uczenia skorzystano ze wsparcia GPU w bibliotece tensorflow oraz keras. Za liczbę iteracji przejścia przez cały zbiór danych przyjęto 20 epok, za batch size zaś 512, oznacza to, że w jednej iteracji bierze udział 512 obserwacji. Przyjęto taką liczbę ze względu na dużą liczbę obserwacji przetworzoną w jednej iteracji. Przy dużej liczbie iteracji, w których jest niski batch, model jest bardziej dostosowany do problemów wymagających szczegółowego wglądu w obserwację. Z kolei decyzja o mniejszej liczbie iteracji, do której używa się batch o większym wolumenie obserwacji, jest rekomendowana w przypadku, gdy oczekiwana jest generalizacja problemu (Keskar, Mudigere, Nocedal, Smelyanskiy i Tang, 2017). Zbyt wysoka liczba epok mogłaby doprowadzić do nadmiernego dopasowania modelu do zbioru treningowego (ang. „*verfitting*”). Dlatego też, dla mniejszej liczby epok, wybrano batch size równy 512. Wyniki uczenia w postaci wykresów skuteczności modelu oraz wartości funkcji straty na zbiorze treningowym i walidacyjnym przedstawiona na wykresie 18.



Wykres 18: Skuteczność i wartość funkcji straty modelu sieci neuronowej na zbiorze treningowym i walidacyjnym

Źródło: Opracowanie własne

Wartość funkcji straty na zbiorze walidacyjnym wyniosła 0,4 w 10 epoce, dla zbioru treningowego zaś wyniosła 0,05. Z kolei precyzja modelu dla zbioru walidacyjnego wyniosła ok. 0,875, a dla treningowego ponad 0,95.

Tworzenie modelu

Wpisz hiperparametry sieci

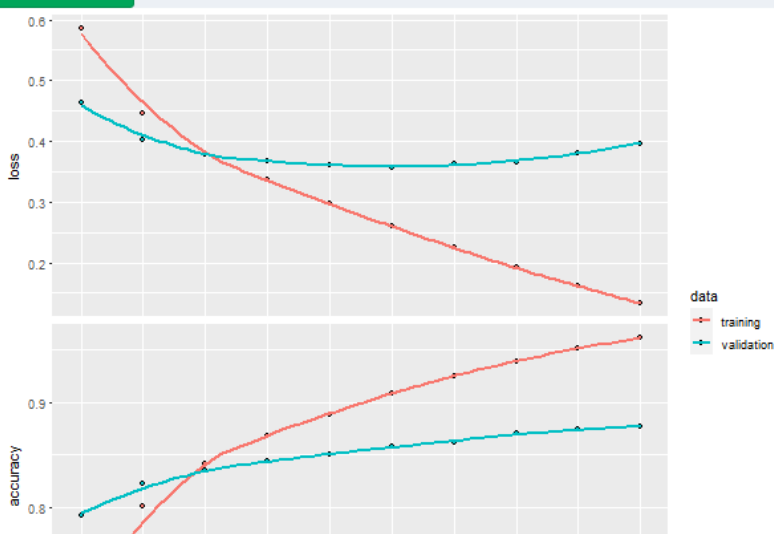
Wpisz batch size

Wpisz liczbę iteracji

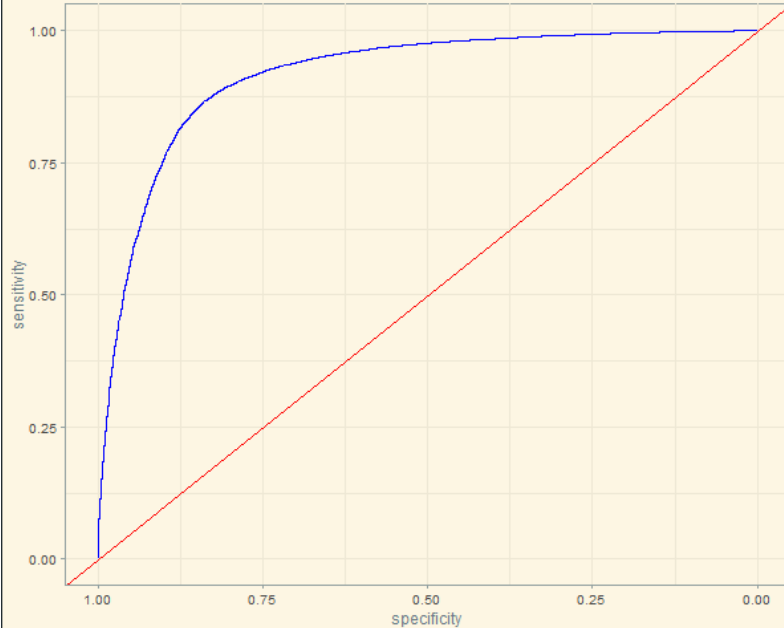
Wpisz liczbę neuronów w warstwie gęstej

Wpisz współczynnik dropoutu

Stwórz model



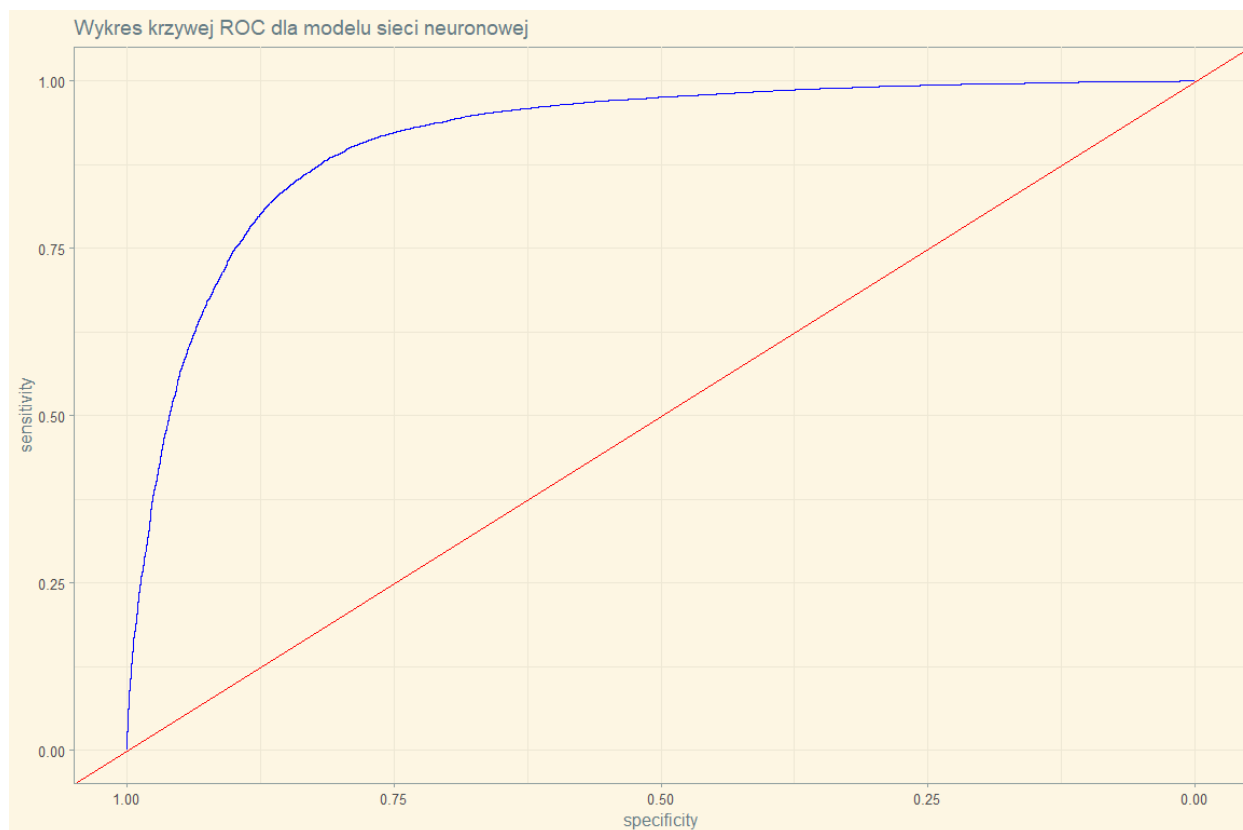
Wykres krzywej ROC dla modelu sieci neuronowej



Wykres 19: Dashboard w systemie służący do tworzenia modelu

Źródło: Opracowanie własne

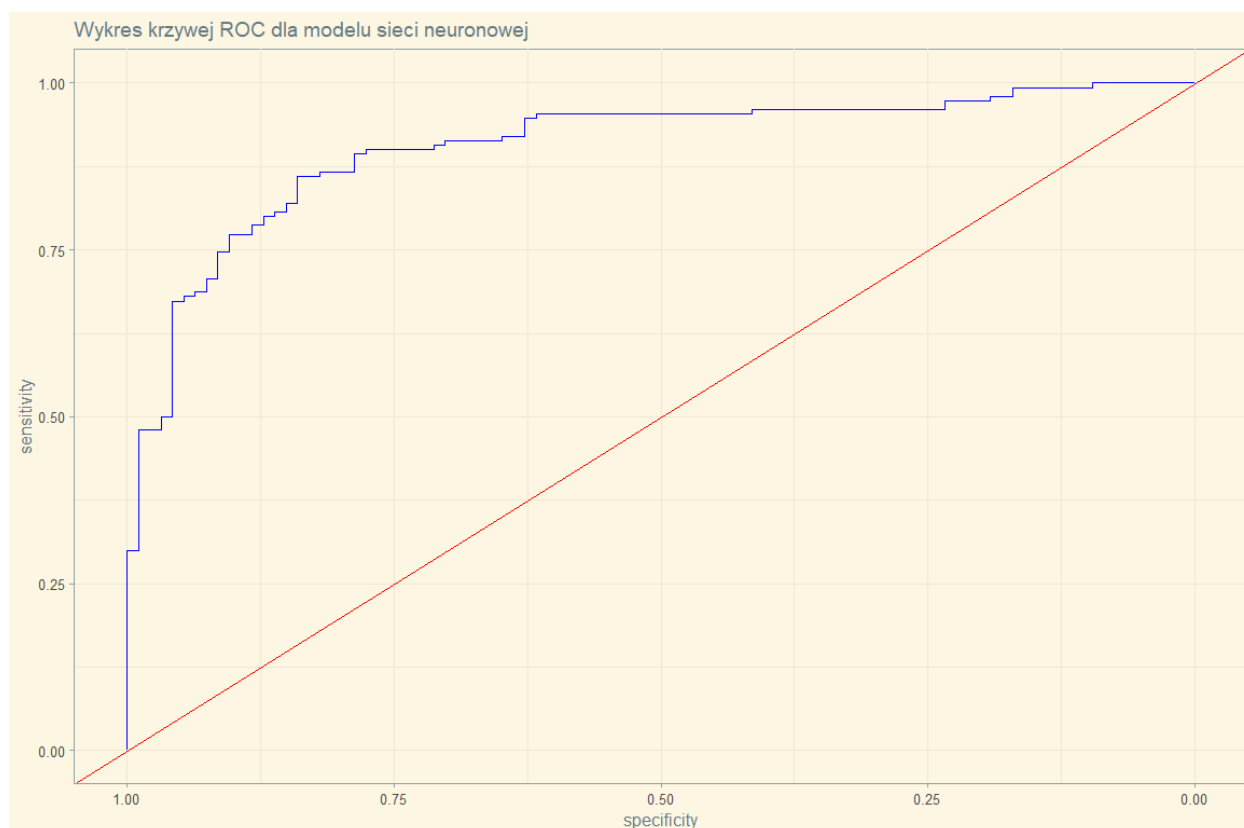
4.5 Ocena jakości modelu



Wykres 20: Krzywa ROC modelu sieci neuronowej klasyfikującej komentarze – ujęcie ogólne

Źródło: Opracowanie własne

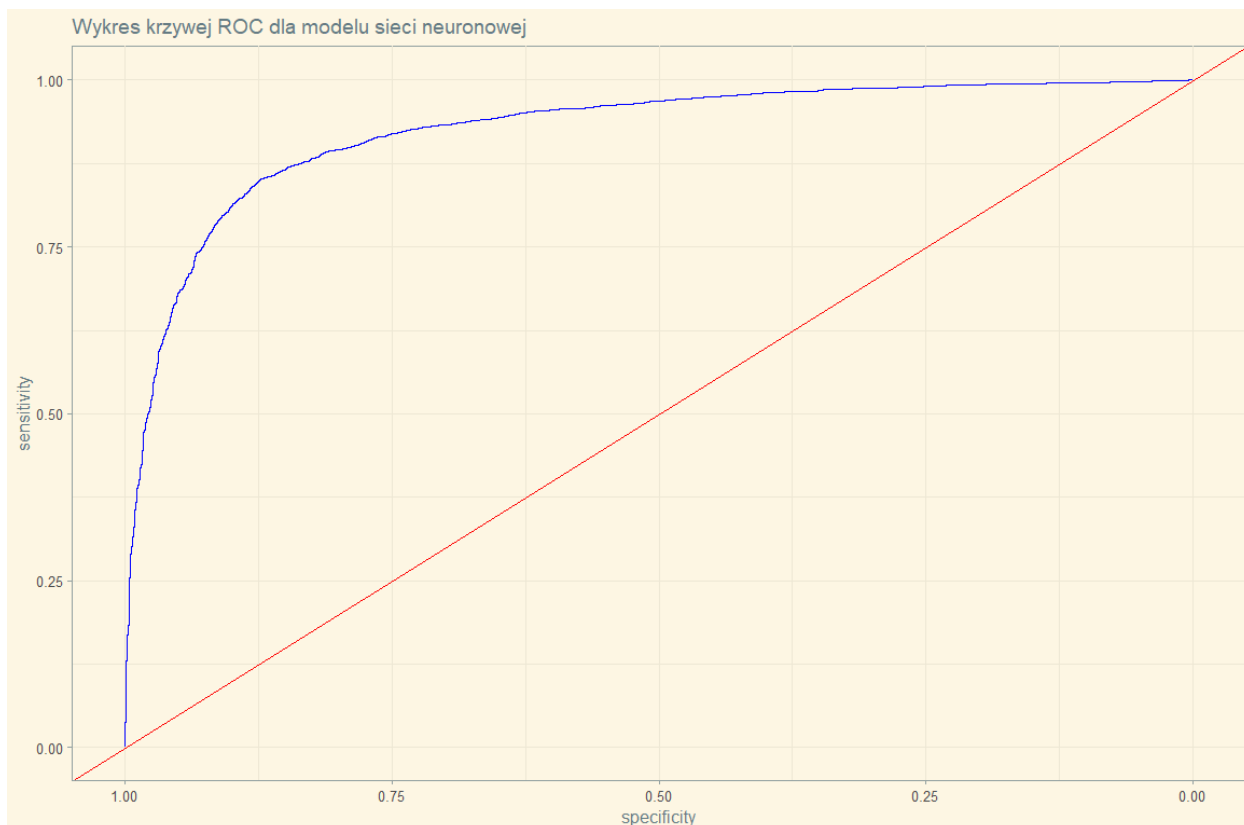
Pole pod krzywą ROC, która została pokazana na wykresie 20, wyniosło 0,91. Czułość wyniosła 0,78, swoistość 0,91, a precyzja wyniosła 0,89. Przełożyło się to na skuteczność modelu na poziomie 87% (Wickham, 2017) oraz na wynik wskaźnika F1 na poziomie 0,83. Sprawdzono również dopasowanie modelu na konkretnych grupach leków, jak np. leki przeciw różnym odmianom raka.



Wykres 21: Krzywa ROC modelu sieci neuronowej klasyfikującej komentarze - leki na raka

Źródło: Opracowanie własne

Pole pod przedstawioną na wykresie 21 krzywą ROC wyniosło 90,5. W porównaniu z całokształtem, swoistość dla leków na raka okazała się być na nieco niższym poziomie – 0.9. Przełożyło się to na minimalnie wyższą czułość modelu – 0.78, precyzja wyniosła 0.85, a wskaźnik F1 był na poziomie 0.82. Skuteczność modelu dla tej grupy leków wyniosła poziomie 85%. Oznacza to, że dla leków na raka, model w podobnym stopniu klasyfikuje negatywne i pozytywne opinie. Inną specyficzną grupą leków na której zbadano dopasowanie modelu były środki antykoncepcyjne (ang. „Birth control”).



Wykres 22: Krzywa ROC modelu sieci neuronowej klasyfikującej komentarze – środki antykoncepcyjne


Źródło: Opracowanie własne

Pole pod krzywą ROC wyniosło 0,925. Skuteczność sieci okazała się mieć wynik 0,87, w porównaniu do ujęcia ogólnego specyficzność okazała się niższa, bo wyniosła 0,83. Czulość za to wzrosła do 0,88, precyzja osiągnęła z kolei poziom 0,88, a wskaźnik F1, podobnie jak precyzja, wyniósł 0,88. Oznacza to, że model dla środków antykoncepcyjnych jest bardziej dopasowany do klasyfikowania recenzji jako pozytywne niż dla ogólnego ujęcia.

W ostatnim kroku zbadano odpowiedź modelu na wprowadzony przez użytkownika tekst.

Wpisz komentarz

Using that drug was a horrible experience, I do not recommend that



ODPOWIEDŹ MODELU
Opinia negatywna

Wykres 23: Odpowiedź modelu na prosty komentarz

Źródło: Opracowanie własne

Rozpoznanie opinii

Wpisz komentarz

Using that drug was at the beginning a horrible experience, then I started to feel fantastic as my headache disappeared. I highly recommend that



ODPOWIEŹ MODELU

Opinia pozytywna



Wykres 24: Odpowiedź modelu na bardziej rozbudowany komentarz

Źródło: Opracowanie własne

Na początku sprawdzono odpowiedź modelu na prosty komentarz, który w dość jasny sposób wskazuje na negatywną opinię. Na wykresie 24 zaś ten komentarz rozbudowano, początek wskazuje wstępnie na negatywną opinię, lecz dalsza część opinii pokazuje, że użytkownik był bardzo zadowolony z leku. Sieć udzieliła poprawnej odpowiedzi co wskazuje, że model ten poprawnie sobie radzi z rozbudowanymi zdaniami.

4.6 Użyte biblioteki oraz pakiety

Do stworzenia interfejsu użytkownika, opracowania modelu, dokonania analizy eksploracyjnej oraz analizy NLP użyto następujących bibliotek:

- dplyr (biblioteka z pakietu tidyverse do pracy na ramkach danych),
- stringr (biblioteka z pakietu tidyverse do pracy na zmiennych tekstowych oraz korzystania z wyrażeń regularnych),
- ggplot2 (biblioteka z pakietu tidyverse służąca do tworzenia wykresów),
- plotly (biblioteka służąca do tworzenia interaktywnych wykresów),
- tidytext (biblioteka do analizy NLP),
- wordcloud (biblioteka używana do obrazowania analizy NLP),
- keras (biblioteka do modeli deep learnig będąca rozszerzeniem biblioteki tensorflow),
- caret (biblioteka do tworzenia i badania modeli uczenia maszynowego),
- pROC (biblioteka służąca do przeliczenia parametrów swoistości i czułości, pozwalająca na tworzenie krzywej ROC),
- shiny (framework służący do tworzenia interaktywnych i webowych wykresów),
- shinydashboard (rozszerzenie biblioteki shiny o dodatkowe komponenty i funkcjonalności),
- ggthemes (rozszerzenie biblioteki ggplot2 o dodatkowe style wykresów, takie jak np. z tygodnika „The Economist” czy też z gazety „The Wall Street Journal”),
- DT (biblioteka do pracy z ramkami danych w bibliotece shiny).

Zakończenie

W pierwszej części pracy zostały wytłumaczone teoretyczne podstawy narzędzi oraz metod użytych w systemie analizy sentymentu. Osiągnięty został główny cel pracy, czyli stworzenie systemu opartego na analizie sentymentu, który klasyfikuje komentarze dotyczące leków. Dzięki temu użytkownik jest w stanie dowiedzieć się o działaniu konkretnego leku. Natomiast czytelnikowi przybliżone zostały kluczowe pojęcia oraz algorytmy dla takich dziedzin, jak NLP czy text mining. Zbiór danych zawierający komentarze pacjentów okazał się odpowiednim źródłem danych dla stworzenia modelu sieci neuronowej. Ponadto zaletą zbioru był fakt, iż zawierał on dużą liczbę obserwacji, bo 215 tysięcy komentarzy. Dlatego też model w podobny sposób dokonywał predykcji dla komentarzy dotyczących różnych leków na różne dolegliwości. Ogólna skuteczność dla modelu wyniosła 89%, przy wysokiej swoistości (ponad 90% – oznacza to, że model dobrze rozpoznaje opinie negatywne).

Na podstawie wyników dotyczących badań wyrażen kluczowych, tf-idf czy analizy eksploracyjnej, zbudowano system badający opinie o lekach, w którym zaimplementowana jest analiza sentymentu oparta na modelu sieci neuronowej. W systemie tym można zbadać komentarze na temat każdego leku, który został oceniony w zbiorze, odnaleźć najlepiej oceniane leki na wybraną chorobę lub dolegliwość, czy też zobaczyć statystyki dopasowania modelu sieci neuronowej. Dzięki temu użytkownik może szybko znaleźć lek na dokuczające mu objawy oraz sprawdzić opinie innych o danym leku. W sekcji poświęconej modelowi można też zbadać, jak poszczególne komentarze zostały rozpoznane przez model. Na zielono zostaną zaznaczone komentarze ocenione przez model jako opinie pozytywne, komentarze krytyczne zostały zaś wyświetlone w kolorze czerwonym. Ponadto użytkownik może sprawdzić, jaki procent pacjentów rekomenduje dany lek, w zależności od liczby zadowolonych pacjentów – ikona przy danych procentowych zmienia się. Dla progu od 0% do 49% ikoną będzie smutny wyraz twarzy, od 50% do 74% twarz o neutralnym wyrazie, od 75% zaś ikoną będzie uśmiechnięta twarz.



Wykres 25: Lek oceniony negatywnie przez pacjentów

Źródło: Opracowanie własne



PROCENT DOBRYCH ODPOWIEDZI DLA WYBRANEGO LEKU
91%



PROCENT OSÓB REKOMENDUJĄCYCH TEN LEK WEDŁUG MODELU
68%

Wykres 26: Lek oceniony niejednoznacznie przez pacjentów

Źródło: Opracowanie własne



PROCENT DOBRYCH ODPOWIEDZI DLA WYBRANEGO LEKU
97%



PROCENT OSÓB REKOMENDUJĄCYCH TEN LEK WEDŁUG MODELU
89%

Wykres 27: Lek oceniony pozytywnie przez pacjentów

Źródło: Opracowanie własne

Na wykresach 25, 26 i 27 przedstawiono, jak model ocenia komentarze na temat różnych leków. Interfejs systemu natomiast informuje użytkownika, czy dany lek jest polecany przez pacjentów: odpowiednio na wykresie 25 pokazane jest, że według modelu dany lek nie jest rekomendowany. Wykres 26 wskazuje natomiast na różne opinie wśród pacjentów, gdzie pomimo większej liczby opinii pozytywnych, znalazło się ponad 30% komentarzy krytycznie oceniających dany lek. Wykres 27 pokazuje lek, który został oceniony pozytywnie przez ponad 75% komentujących, dlatego też interfejs użytkownika pokazuje, że ten lek jest odpowiednim środkiem na daną dolegliwość.

Użytkownik ma też możliwość wpisać swój komentarz i sprawdzić odpowiedź sieci neuronowej. Model osiąga dokładność odpowiedzi na poziomie ok. 85%, współczynnik F1 wynosi 0,83, czułość jest równa 0,78, a precyzja osiąga poziom 0,89. Dlatego można to uznać za wynik satysfakcjonujący (model klasyfikuje na zadawalającym poziomie opinie pozytywne oraz negatywne). Model największe trudności ma z weryfikacją zdań, w których występują zaprzeczenia. Ponadto sprawdzono działanie modelu na komentarzach dotyczących działania na różnych grupach leków w celu weryfikacji, czy model jest tak samo dopasowany do konkretnych grup opinii. System adresuje takie potrzeby klientów, jak:

- sprawdzenie, jaki procent użytkowników jest zadowolonych z leku,
- podanie w analizie eksploracyjnej informacji o najlepiej ocenianych lekach,
- łatwy dostęp do informacji i podsumowania dotyczących danego leku, bez konieczności czytania każdego komentarza osobno.

System mógłby zostać rozszerzony o kolejne funkcjonalności:

- rozszerzenie analizy komentarzy w innych językach europejskich takich jak: język niemiecki, polski czy francuski;

- oparcie systemu na oprogramowaniu służącym do przetwarzania bardzo dużej ilości danych, jak np. Spark (biblioteki sparkr i sparklyr);
- udostępnienie otwartego API, które pozwoliłoby programistom na integrację swojego oprogramowania z systemem rekomendacji;
- udoskonalenie interfejsu użytkownika o dodatkowe widżety oraz funkcjonalności, w celu poprawy doświadczenia użytkownika (czyli tzw. UX – ang. *User Experience*).

Ryzykiem jest niewielki procent błędnych odpowiedzi modelu, ponieważ nie jest możliwe, aby modele predykcyjne osiągały poziom 100% skuteczności w swoich predykcjach. W ten sposób istnieje ryzyko, że któryś z pacjentów zdecyduje się na użycie leku pod wpływem błędnej rekomendacji modelu. Stanowi to wówczas zagrożenie dla zdrowia, a czasami i nawet życia pacjenta. Dlatego przed publikacją systemu jest niezbędne, aby taki model został poddany szczegółowej walidacji. Taki rozbudowany system rekomendacji mógłby być wykorzystany przez np. sklepy internetowe, apteki czy też używany przez firmy farmaceutyczne w celu zbadania satysfakcji klientów. Jednocześnie klientami systemu mogliby być również pacjenci szukający odpowiedniego leku, w tym przypadku mogliby na podstawie komentarzy sprawdzić, czy dany lek jest dobrze oceniany przez klientów na podstawie wyników podanych przez model sieci neuronowej. Ponadto pakiety takie jak shiny, ggplot2 bądź plotly dobrze realizują zobrazowanie wyników działania systemu poprzez generowanie responsywnych i interaktywnych wykresów. Do ograniczenia systemu należy fakt, iż implementacja analizy sentymentu nie uwzględnia aspektu, przez co algorytm nie wskazuje na konkretne przyczyny danej opinii. Dalsze badania nad tym systemem powinny dotyczyć:

- lepszego dostrojenia hiperparametrów modelu sieci neuronowej,
- zautomatyzowania procesu załadowywania danych oraz ich manipulacji przy użyciu takiego narzędzia jak Apache Airflow,

Jest to narzędzie służące do automatyzacji i harmonogramowania procesów oraz przepływu zadań. W ten sposób można ustawić przepływ zadań w postaci acyklicznego grafu skierowanego (ang. DAG, *Directed acyclic graph*). Dzięki temu możliwe jest budowanie procesów przetwarzania danych, gdzie sekwencyjnie są realizowane poszczególne procesy jak: wczytanie danych, ich transformacja, stworzenie modelu predykcyjnego i przedstawienie wyników predykcji. W ten sposób procesy w systemie mogłyby ulec automatyzacji poprzez ustawienie powtarzalnego wykonywania procesów.

- wykrywania w tekście czynników wskazujących na wystąpienie skutków ubocznych.

Rozbudowanie analizy sentymentu o aspekt pozwoli na zwiększenie wiedzy o przyczynach zadowolenia bądź niezadowolenie klienta. W ten sposób można będzie między innymi wykryć, czy zadowolenie bądź jego brak wynika z faktu wystąpienia efektów ubocznych.

W ten sposób można też stwierdzić, jak poważne okazały się efekty niepożądane. W przypadku gdy lek wywołał takie skutki jak senność czy obniżenie samopoczucia, efekty uboczne uznawane byłyby za lekkie. Natomiast gdy objawami byłyby takie dolegliwości, jak: biegunka, wymioty, wysypka czy też podniesienie ciśnienia, efekty zostałyby sklasyfikowane jako poważne. System wówczas stałby się jeszcze bardziej cenny dla klientów takich jak firmy farmaceutyczne, które prowadzą obszerne badania nad występowaniem niepożądanych efektów po zażyciu ich produktów.

Rozbudowany o te czynniki i funkcjonalności system mógłby być bardzo cennym produktem o dużej wartości na rynku, ponieważ istnieje niewiele platform adresujących problem badania wypisywanych przez pacjentów na temat leków, które zażyli.

System ten ma na celu przede wszystkim pomóc i ułatwić znalezienie informacji na temat danych leków. Nie zastępuje on pracy lekarzy, po których to stronie wciąż leży odpowiedzialność w kwestii przypisania odpowiednich środków dla swoich pacjentów. Użytkownicy dzięki temu systemowi, powinni głównie zaznajomić się z tym, co inni pacjenci uważają o danym leku, czy występują po jego zażyciu efekty uboczne oraz jak dotkliwe one są. Charakter tego systemu jest przede wszystkim informacyjny, a nie decyzyjny.

Spis tabel i wykresów

WYKRES 1: SCHEMAT DZIAŁANIA METOD BIBLIOTEKI TIDYTEXT PRZY PRACY Z MACIERZAMI DTM I TDM ŹRÓDŁO: (SILGE I ROBINSON, 2017)	16
WYKRES 2: SCHEMAT DZIAŁANIA METOD BIBLIOTEKI TIDYTEXT W JĘZYKU R PRZY DZIAŁANIU Z ALGORYTMEM LDA	18
WYKRES 3: SCHEMAT BUDOWY NEURONU	23
WYKRES 4: SCHEMAT PERCEPTRONU ROSENBLATTA	23
WYKRES 5: PRZYKŁAD BINARNEJ FUNKCJI AKTYWACJI	24
WYKRES 6: HISTOGRAM OCEN DLA WSZYSTKICH LEKÓW	49
WYKRES 7: PROCENT WSZYSTKICH KOMENTARZY W POSZCZEGÓLNYCH LAT	50
WYKRES 8: ŚREDNIA OCENA DLA LEKÓW NA CHOROBE LOKOMOCYJNĄ	51
WYKRES 9: ŁĄCZNA LICZBA KOMENTARZY DLA LEKÓW NA CHOROBE LOKOMOCYJNĄ	51
WYKRES 10: ŁĄCZNA LICZBA POLUBIEŃ DLA LEKÓW NA CHOROBE LOKOMOCYJNĄ	51
WYKRES 11: LEKI NA PRZYKŁADOWĄ DOLEGLIWOŚĆ – CHOROBE LOKOMOCYJNĄ, Z NAJWYŻSZĄ ŚREDNIĄ OCEN	52
WYKRES 12: SŁOWA, KTÓRE BYŁY POPRZEDZONE SŁOWEM „FEEL” W RÓŻNEJ ODMIANIE GRAMATYCZNEJ ŹRÓDŁO: OPRACOWANIE WŁASNE	53
WYKRES 13: SŁOWA KTÓRE NAJCZĘŚCIEJ BYŁY POPRZEDZONE SŁOWEM „NO” LUB „NOT” O NAJWYŻSZEJ KONTRYBUCJI ŹRÓDŁO: OPRACOWANIE WŁASNE	54
WYKRES 14: CHMURA SŁÓW NAJCZĘŚCIEJ PRZEDSTAWIAJĄCYCH SIĘ W KOMENTARZACH	55
WYKRES 15: HISTOGRAM WSPÓŁCZYNNIKA CZĘSTOTLIWOŚCI WYSTĘPOWANIA SŁÓW(TF) W KOMENTARZACH DLA 6 NAJCZĘŚCIEJ KOMENTOWANYCH LEKÓW ŹRÓDŁO: OPRACOWANIE WŁASNE	56
WYKRES 16: HISTOGRAM TF-IDF TERMÓW DLA 6 NAJCZĘŚCIEJ OCENIANYCH LEKÓW	57
WYKRES 17: ŚREDNI TF_IDF DLA 6 NAJCZĘŚCIEJ OCENIANYCH LEKÓW	57
WYKRES 18: SKUTECZNOŚĆ I WARTOŚĆ FUNKCJI STRATY MODELU SIECI NEURONOWEJ NA ZBIORZE TRENINGOWYM I WALIDACYJNYM ŹRÓDŁO: OPRACOWANIE WŁASNE	61
WYKRES 19: DASHBOARD W SYSTEMIE SŁUŻĄCY DO TWORZENIA MODELU	62
WYKRES 20: KRZYWA ROC MODELU SIECI NEURONOWEJ KLASYFIKUJĄCEJ KOMENTARZE – UJĘCIE OGÓLNE	63
WYKRES 21: KRZYWA ROC MODELU SIECI NEURONOWEJ KLASYFIKUJĄCEJ KOMENTARZE - LEKI NA RAKA	64
WYKRES 22: KRZYWA ROC MODELU SIECI NEURONOWEJ KLASYFIKUJĄCEJ KOMENTARZE – ŚRODKI ANTYKONCEPCYJNE	65
WYKRES 23: ODPOWIEDŹ MODELU NA PROSTY KOMENTARZ	65
WYKRES 24: ODPOWIEDŹ MODELU NA BARDZIEJ ROZBUDOWANY KOMENTARZ ŹRÓDŁO: OPRACOWANIE WŁASNE	66
WYKRES 25: LEK OCENIONY NEGATYWNIE PRZEZ PACJENTÓW	67
WYKRES 26: LEK OCENIONY NIEJEDNOZNACZNIE PRZEZ PACJENTÓW	68
WYKRES 27: LEK OCENIONY POZYTYWNIE PRZEZ PACJENTÓW	68
DIAGRAM 1: DIAGRAM PRZEDSTAWIAJĄCY PRZEPLÝW PROCESÓW W SYSTEMIE	48
DIAGRAM 2: ARCHITEKTURA MODELU SIECI NEURONOWEJ	59

Bibliografia

1. Aylien. (2022). *Using Entity-level Sentiment Analysis to understand News Content*.
2. Badjatiya, P., Gupta, S., Gupta, M. i Varma, V. (2017). *Deep Learning for Hate Speech Detection in Tweets*.
3. Baheti, P. (2022). *12 Types of Neural Network Activation Functions: How to Choose?*
4. Berisha, B., Meziu, E. i Shabani, I. (2022). *Big data analytics in Cloud computing: an overview*.
5. Brownlee, J. (2017). *How to Use Word Embedding Layers for Deep Learning with Keras*.
6. Brownlee, J. (2018). *A Gentle Introduction to Dropout for Regularizing Deep Neural Networks*.
7. Brownlee, J. (2019). *Loss and Loss Functions for Training Deep Learning Neural Networks*.
8. Cachola, I., Holgate, E., Preotiuc-Pietro, D. i Li, J. J. (2018). *Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media*.
9. Cherry, K. (2021). *The Role of Neurotransmitters*.
10. Chollet, F. (2018). *Deep Learning. Praca z językiem R i biblioteką Keras*.
11. Clarabridge. (2019). *The Pillars of Text Analytics: Sentiment, Categorization, Effort and Emotion*.
12. Dixon, S. (2022). *Number of social media users worldwide from 2018 to 2027*.
13. Dorash, M. (2017). *Machine Learning vs. Rule Based Systems in NLP*.
14. dtu.dk. (2011). *AFINN lexicon*.
15. Du, T. i Shanker, V. K. (2015). *Deep Learning for Natural Language Processing*.
16. Eliacik, E. (2022). *Follow the latest AI trends to survive tomorrow*.
17. Furbush, J. (2021). *Understand these 5 key deep learning classification metrics for better application success*.
18. Gavran, K. (2022). *Social Media Sentiment Analysis: A Guide*.
19. Gebel, Ł. (2020). *Why we need bias in neural networks*.
20. Kallumadi, S. i Gräßer, F. (2018).
<https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29>.
21. Kaneda, T. (2021). *How Many People Have Ever Lived on Earth?*
22. Kanna, C. (2021). *Word, Subword, and Character-Based Tokenization: Know the Difference*.
23. Keskar, N., Mudigere, D., Nocedal, J., Smelyanskiy, M. i Tang, P. (2017). *On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima*.
24. Kibble, R. (2013). *Introduction to natural language processing*. Goldsmith University of London.
25. Kohn, K. W. (2020). *Drugs Against Cancer: Stories of Discovery and the Quest for a Cure*.
26. Korbicz, J., Obuchowicz, A. i Uciński, D. (1994). *Sztuczne sieci neuronowe. Podstawy i zastosowania*.
27. Kunwar, S. (2013). *Text Documents Clustering using K-Means Algorithm*.
28. Lander, J. (2017). *R dla każdego*.
29. Loiseau, J.-C. B. (2019). *Rosenblatt's perceptron, the first modern neural network*.
30. Malik, F. (2020). *Sensitivity Vs Specificity In Data Science*.
31. Mamczur, M. (2021). *Jak działają konwolucyjne sieci neuronowe?*
32. Manning, C., Schütze, H. i Raghavan, P. (2009). *Introduction to information retrieval*.
33. McKinsey. (2016). *The CEO guide to customer experience*.
34. Mention. (2021). *Sentiment Analysis*.
35. MonkeyLearn. (2022). *What is Text Mining?*
36. Na, J.-C. i Kyaing, W. (2015). *Sentiment Analysis of User-Generated Content on Drug Review Websites*.
37. Neppali, V., Caragea, C. i Caragea, D. (2018). *Deep Neural Networks versus Naive Bayes Classifiers for Identifying Informative Tweets during disasters*.
38. Newberry, C. (2020). *How to Conduct a Social Media Sentiment Analysis*.

39. Ng, R. (2009). *Drugs: From Discovery to Approval, Second Edition*.
40. Opitz, L. (2017). *Sentiment Analysis Just Got Smarter*.
41. Pai, A. (2020). *What is Tokenization in NLP? Here's All You Need To Know*.
42. Pascual, F. (2019). *Analyze Sentiment in Product Reviews*.
43. Pascual, F. (2019). *Guide to Aspect-Based Sentiment Analysis*.
44. Patel, S. (2022). *A Guide to Customer Sentiment Analysis*.
45. Pinkowska, M. (2020). *How to use sentiment analysis for stock market?*
46. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E. i Mihalcea, R. (2019). *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations*.
47. Pozzi, F. A. (2018). *The role of emojis in sentiment analysis*.
48. Repustate. (2021). *10 Sentiment Analysis Data Sources For Strategic Data Analytics*.
49. Repustate. (2021). *Why Should We Use Sentiment Analysis In Social Media Mining?*
50. Robinson, S. (2021). *Sentiment analysis: Why it's necessary and how it improves CX*.
51. Roldós, I. (2020). *5 Sentiment Anlysis Examples in Business*.
52. Roldós, I. (2020). *What is Sentiment Analysis?*
53. Rosencrance, L. (2020). *Employee sentiment analysis*.
54. Schmelzer, R. (2022). *Top trends in big data for 2022 and beyond*.
55. sciencemuseum.org.uk. (2019). *Thalidomide*.
56. Seth, N. (2021). *Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim and Sklearn*.
57. Sievert, C. (2019). *Interactive Web-Based Data Visualization with R, plotly and shiny*.
58. Silge, J. i Robinson, D. (2017). *Text mining with R - A tidy approach*.
59. Sreemany, T. (2021). *Essential Text Pre-processing Techniques for NLP*.
60. Stedman, C. (2020). *Text mining (text analytics)*.
61. Verma, Y. (2021). *A Guide to Term-Document Matrix with Its Implementation in R and Python*.
62. Wickham, H. (2017). *R for Data Science*.
63. Wonderflow. (2018). *10 Sentiment Analysis Examples That Will Help Improve Your Products*.
64. Zalewska, A. (2018). *Customer experience, czyli czym jest doświadczenie klienta?*