SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE
WARSAW SCHOOL OF ECONOMICS

Studium ……

Kierunek/makrokierunek: ……..
Specjalność:** ……..……
Forma studiów: ……..

Maciej Sadkowski
MS107984

Tytuł pracy

Praca magisterska
napisana w instytucie informatyki i gospodarki cyfrowej
pod kierunkiem naukowym Mariusza Rafało

Warszawa 20…

*Zastosować właściwe

** W przypadku braku specjalności lub braku deklaracji o specjalności wiersz należy pominąć

1. Wstęp

Znaczenie leków dla ludzi

Odkrywanie i rozwój leków datuje się od pierwszych lat ludzkiej cywilizacji. Powstanie tradycyjnej medycyny chińskiej są szacowane jest na 3500 r. p.n.e. za panowania legendarnego cesarza Shennonga (Ng, 2009). Spis leków z tej dziedziny został później wykorzystany w zachodnich lekach np. rezerpina występująca w korzeniach rauwolfii żmijowej występującej w Azji jest używana w środkach hipotensyjnych (zapobiegających nadciśnieniu) oraz uspokajających. Innym przykładem może być efedryna występująca w przęśli chińskiej, środek ten używany jest m.in. do leczenia nieżytu nosa oraz jako środek zapobiegający niedociśnieniu (Ng, 2009). Przykładem innej dawnej cywilizacji, która również posiadała rozwiniętą medycynę był starożytny Egipt. Papirus Ebersa, który został stworzony co najmniej 3500 lat temu, zawierał 877 recept na choroby wewnętrzne, schorzeń oczu oraz skóry czy też dolegliwości ginekologiczne. Natomiast papirus Kahuna, który jest datowany na 1800 r. p.n.e. był kolekcją kuracji oraz metod leczenia na problemy ginekologiczne (Ng, 2009). Medycyna była również fundamentem cywilizacji starożytnej Grecji oraz Rzymu. W „De materia medica" autorstwa Pedaniosa Dioskurydesa opisane były środki lecznicze oparte w 80% na roślinach, 10% było pochodzenia zwierzęcego, a 10% bazowało na minerałach (Ng, 2009).

Początki medycyny nowoczesnej na przełomie XIX i XX wieku dokonały przełomu w historii ludzkości. Jeszcze na początku XX wieku dostępnymi lekami były tylko: Digitalis (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 penicyliny na dużą skalę, która stałą 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 wiele 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 dla mężczyzn była na poziomie 46 lat oraz 48 dla kobiet. Rozwój medycyny oraz powszechny dostęp do leków wpłynęły również na światową populację. W 1900 roku światowa populacja liczyła 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 tych okresach ilość urodzeń na 1000 osób wynosiła 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).

Znaczenie ocen leków i zaproponowanie modelu do ich określenia

Do jednych z zagrożeń zażywania leków są tzw. efekty uboczne. W latach pięćdziesiątych XX wieku zachodnioniemiecki koncern farmaceutyczny Chemie Grünenthal GmbH opracował lek o nazwie Talidomid. Środek ten był pierwotnie przeznaczony jako lek usypiający. W latach 1957-1961 lek ten był powszechnie stosowany jako środek przeciwbólowy dla kobiet w ciąży, używano go również przeciwko przeziębieniu czy grypie. Lek ten jednak okazał się wykazywać silne działania teratogenne (tzn. uszkadzające płód) w pierwszych fazach okresu prenatalnego co doprowadziło do narodzin ponad 10 000 tysięcy dzieci z poważnymi wadami jak deformacje stawów czy kończyn. Według szacowań, ok. 50% spośród tych narodzonych dzieci nie dożyło jednego roku (sciencemuseum.org.uk, 2019). W czasach współczesnych powszechne są już portale oraz strony w których pacjenci mogą opisać lek oraz opisać jego działanie oraz 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 do znalezienia odpowiedniego leku oraz 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 biznesowym, które mogłoby wpłynąć na poprawę doświadczenia użytkownika oraz na skrócenie czasu potrzebnego do osiągnięcie danego celu biznesowego jest opracowanie systemu rekomendacji leków. W systemie tym użytkownik uzyskałby dostęp do danych nt. leków przeciwko różnym schorzeniom oraz uzyskałby sugestię od systemu, który lek jest najwyżej oceniany przez użytkowników. Ponadto, model ten byłby również rozwinięty o model analizy sentymentów wykorzystujący sieć neuronową służący do weryfikacji komentarzy i rozróżniający opinię pozytywną od negatywnej. W ten sposób użytkownik uzyskiwałby informację o ilości pozytywnych opinii oraz o ilości komentarzy krytycznych. Do budowy tego systemu składającego się z interfejsu graficznego użytkownika, silnika przetwarzającego dane oraz modelu sieci neuronowej zdecydowano się na użycie języka programowania R wraz z takimi bibliotekami i pakietami jak: shiny, tidytext, keras,

dplyr oraz caret.

1. Opisanie NLP oraz użytych algorytmów do analizy 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 wersów i kolumn. W takiej formie łatwiej przechowuje się i procesuje 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 takie jak pliki audio czy wideo) (Stedman, 2020)
- Pół-ustrukturyzowany (Jest to pewne połącznie danych ustrukturyzowanych i nieustrukturyzowanych, dane są w pewny sposób uporządkowany, ale nie spełniają kryteriów relacyjnej bazy danych. Przykładem takiego formatu są pliki XML, JSON czy też HTML) (Stedman, 2020)

Szacuję się, że ok. 80% danych świecie jest w formie nieustrukturyzowanej co czyni text mining wyjątkowo cennym narzędziem do analizy danych tekstowych. Oprócz badania korelacji między słowami w tekście, text mining też skupia się na częstotliwości ich występowania oraz powtarzalnych wzorców 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. Text mining składa się z metod można wyróżnić następujące pojęcia:

- Klasyfikacja tekstu
- Ekstrakcja tekstu
- Klasteryzacja tekstu

Klasyfikacja tekstu jest procesem w którym nieustrukturyzowanym danom tekstowym przypisywana jest kategoria(tag). W ten sposób umożliwione jest analizowanie różnych źródeł danych i uzyskiwanie wartościowych wniosków w szybki i mało kosztowny sposób. 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 do 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 w ten sposób zgłoszenie może zostać szybko skierowane do odpowiedniego regionu.

Ekstrakcja tekstu to technika w której ze źródła danych tekstowych wyciągane są specyficzne elementy jak 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 klasyfikacją tekstu. 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. Wykrywanie cech jest często używane 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 ten sposób by w jednej grupie znajdowały się teksty podobne do siebie we 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
- Funkcja kryterium która pozwala obliczyć odpowiednie dopasowanie klastrów
- Algorytm optymalizujący funkcję kryterium

Klasteryzacja tekstu znajduje zastosowanie w :

- Identyfikacji fake news (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
- Analiza zgłoszeń w centrum wsparcia (Identyfikacja i przypisanie zgłoszeń do przyjętych wcześniej raportów) (Kunwar, 2013)

W ten sposób organizacje mogą znaleźć potencjalnie cenne spostrzeżenia w źródłach danych jak: firmowe dokumenty, wiadomości od klientów, rejestry rozmów z centrali rozmów, ankiety, recenzje, wpisy użytkowników na portalach mediów społecznościowych, zapisy z instytutów medycznych czy inne źródła danych tekstowych (Stedman, 2020).

1.2 Definicja NLP

Przetwarzanie języka naturalnego (ang. „Natural-language processing" – NLP) jest to zbiór technik pozwalających na zrozumienie komputerowi języka. W przetwarzaniu tym wyróżnia się dwa etapy

- Preprocessing danych
- Rozwój algorytmu

Preprocessing danych to zbiór metod służących do przygotowania oraz oczyszczenia danych dla komputera w celu przygotowania formatu danych w taki sposób by można było na nich pracować. Wśród metod składających się na preprocessing można wyróżnić takie etapy jak:

- Tokenizację
- Eliminowanie słów stop listą
- Stemming
- Dzielenie na części zdania

Metoda tokenizacji polega na dzieleniu tekstu na mniejsze części i zostanie bardziej 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 powstałym w wyniku tokenizacji zbioru treningowego. Do możliwych rozwiązań należy zebranie ze zbioru testowego tzw. 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 k 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 natomiast

jest częściowa utrata informacji podczas odrzucania rzadziej występujących słów, które mogą mieć wysoki stopień istotności zgodnie z prawem Zipfa. 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", „ą", „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 podsłowa np. wówczas słowo „smartest" zostałoby podzielone na dwa tokeny „smart" i „est" (Pai, 2020).

Eliminacja słów stop listą polega na usunięcia ze źródła danych wyrazów, które znajdują się w tzw. stop liście (ang. „stop words"). 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. W niektórych przypadkach takich jak analiza sentymentu, eliminacja słów stop listą może wpłynąć niekorzystnie na uzyskany rezultat. Wówczas następujące zdanie „This drug was not good at all" zostałby uproszczony do postaci „drug good" co wpłynęłoby na błędnie sklasyfikowane zdanie jako pozytywne (Kanna, 2021). Kolejnym ograniczeniem jest też fakt, że różne stop listy mogą zawierać inne słowa, zbiory takich słów są udostępniane przez takie biblioteki programistyczne jak tidytext, scikit-learn czy też nltk.

Stemming polega zaś na usuwaniu formantów ze słów w celu uproszczenia zbioru tekstowego, dzięki temu wyciągane jest samo zdanie słowa. Jako przykład można podać słowa „programming" i „programmer" (Kibble, 2013). Słowa ta nie mają takiego samego znaczenia, lecz

wskazują na tę samą dziedzinę, więc można je uprościć poprzez stemming. Wówczas dostajemy w rezultacie słowo „programm", które będzie częściej się powtarzać oraz z racji swojej krótszej nazwy, jest też bardziej oszczędne pamięciowo i obliczeniowo dla komputera. (Kibble, 2013) Do najpowszechniej stosowanego algorytmu jakim jest stemming w języku angielskim należy algorytm Portera (Manning, Schütze i Raghavan, 2009). Algorytm ten składa się z 5 kroków skracania upraszczania słowa następujących bezpośrednio po sobie. W każdym kroku należy wybrać zasadę uproszczenia końcówki słowa dla najdłuższego sufiksu. 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, która zostanie poddana transformacji według tych zasad:

$$caresses \rightarrow caress$$

$$ladies \rightarrow ladi$$

$$princess \rightarrow princess$$

$$dogs \rightarrow dog$$

W późniejszych krokach algorytmu często stosowanie pojęcie tzw. miary słowa. Koncept ten polega na sprawdzaniu ilości 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 następujący przykład:

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

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:

1. Systemy oparte o reguły

Ten rodzaj klasyfikacji opiera się o lingwistyczne reguły, według opracowanych przez człowieka reguły lingwistyczne 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. Reguły opierają się o wzorce składniowe, morfologiczne oraz leksykalne. Zasady te mogą też być związane z aspektami semantycznymi oraz fonologicznymi. Przykładową regułą może być następujący przypadek:

$$(Czarny \mid Biały \mid Niebieski \mid Zielony \mid Czerwony) \rightarrow Kolor$$

Na podstawie tej reguły, system przypisze kategorię „Kolor" w momencie, gdy jakikolwiek z podanych w regule wyrazów pojawi się w tekście. Systemy oparte o reguły 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. W rezultacie, wadą systemów opartych o reguły jest problem ze skalowalnością rozwiązania. Kolejnym ograniczeniem tych systemów jest również wymaganie posiadania konkretnej znajomości lingwistyki i danych, które należy analizować.

2. Systemy oparte o uczenie maszynowe

Te systemy opierają się modelach, które zostały nauczone poprzez poprzednie dane, które zostały sklasyfikowane prawidłowo. W tym celu zbiory treningowe muszą być spójne oraz reprezentatywne w celu osiągnięcia modelu, który będzie dokonywać poprawnej klasyfikacji. Na początku tego procesu, dane tekstowe są poddawane tzw. wektoryzacji. Ze zdefiniowanego zbioru wyrazów, dla każdego słowa obliczona jest ilość jego wystąpienia w tekście. Przetransformowane informacje wraz z oczekiwanymi predykcjami są przetwarzane przez algorytm uczenia maszynowego. 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 są oparte o teorię prawdopodobieństwa oraz teorię Bayesowską do zaklasyfikowania tekstu. Dane wówczas są zwektoryzowane do wektorów prawdopodobieństw tekstu należącego do określonej kategorii. Klasyfikatory te dokonują poprawnej oceny nawet w przypadku gdy zbiór treningowy zawiera niewiele danych)
- Maszyna wektorów nośnych (Algorytm ten dzieli zaklasyfikowane wektory na dwie grupy: grupę w której znajdują się wektory należące do określonej klasy i drugą

zawierającą te wektory, które do tej klasy nie należą. Algorytm ten osiąga lepsze wyniki od klasyfikatorów bayesowskich lecz są znacznie trudniejsze w zaimplementowaniu)

- Deep learning (Systemy oparte o głębokie sieci neuronowe – ich działanie zostanie opisane 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.

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

1. Skuteczność (ang. „accuracy") – jest 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* oznacza ilość poprawnych klasyfikacji pozytywnych (czyli odpowiedzi twierdzącej), *TN* to ilość poprawnych klasyfikacji negatywnych (czyli odpowiedzi przeczącej). *FN* jest za to ilością błędnych klasyfikacji negatywnych (czyli błędne zaklasyfikowanie obserwacji jako nie należącej do danej klasy), a *FP* natomiast jest ilością błędnych klasyfikacji pozytywnych (czyli będących zaklasyfikowanych jako przynależące do danej klasy niezgodnie z prawdą). 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 innych), może dojść do tzw. paradoksu skuteczności. Paradoks ten polega na sytuacji w której wysoka skuteczność modelu może nie oznaczać, że model jest dobrze dopasowany, gdyż 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 występujące jest o wiele istotniejsze. Metrykami które w sposób bardziej wiarygodny opisują model są czułość, swoistość i precyzja.

2. Czułość (ang. „sensitivity" lub „recall") – metryka ta 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 do wszystkich pozytywnych klasyfikacji. Im wyższa wartość współczynnika czułości, tym model w lepszym stopniu klasyfikuje pozytywne obserwacje (Malik, 2020).

3. Swoistość (ang. „Specificity") – statystyka ta jest 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 do wszystkich negatywnych klasyfikacji. Wysoka wartość współczynniku swoistości oznacza, że model rozpoznaje negatywne obserwacje na wysokim poziomie (Malik, 2020).

4. Precyzja (ang. „Precision") – precyzja 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 czy też wysoka czułość modelu, dzięki której prawie wszyscy pacjenci z chorobą zostaliby poprawnie zdiagnozowani.

5. Wskaźnik F1 (ang. „F1 score")

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

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

1.3 Analiza Sentymentu oraz jego rodzaje

Analiza sentymentu polega na wykorzystaniu przetworzenia języka naturalnego, analizy tekstu, lingwistyki komputerowej do systematycznego identyfikowania, wyodrębniania, określania emocjonalnego wydźwięku źródła tekstowego, czy był on pozytywny, negatywny bądź neutralny. Znaczenie analizy sentymentu jest tak duże ze względu na to że może ono być w pełni zautomatyzowane, dzięki czemu oszczędza to dużo pracy człowieka i jego czasu. 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) czy też badanie rynku. Ogromną skarbnicą takich danych są różne platformy społecznościowe jak:

twitter, facebook, reddit gdzie codziennie miliony użytkowników udziela wpisów na różne tematy (Silge i Robinson, 2017).

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

Do głównych wyzwań i trudności z jakimi analiza sentymentu się musi mierzyć są

- Zaprzeczenia
- Złożone zdania o kontrastującym wydźwięku
- Sarkazm
- Rozpoznanie nazw własnych (Named-entity recognition)

Przykładem zaprzeczenia są wszelakie zdania z takim wyrazami jak „nie", „niezbyt" czy „wcale". Przypadkiem tego można uznać za dwa zdania, „Biorąc pod uwagę wszelkie wydarzenia, należy uznać dzisiejszy dzień za udany" wskazuje na pozytywny wydźwięk wypowiedzi z którym analiza sentymentu nie powinna mieć błędu. Dla zdania „Biorąc pod uwagę wszelkie wydarzenia, nie należy wcale uznać dzisiejszy dzień za udany" z kolei już nie można być pewnym poprawności wyniku analizy ze względu na to, że przy zsumowaniu lub uśrednieniu wyniku sentymentu poszczególnych słów wciąż wynik może wskazywać na pozytywny ton wypowiedzi mimo, że w rzeczywistości tak nie jest.

W przypadku zdań o złożonych o kontrastującym wydźwięku problemem z jakim algorytm musi się zmierzyć jest występowanie słów o kontrastującym wydźwięku co może wpływać na błędny wynik analizy. Przykładem takiego zdania może być : „Dzisiaj rano czułem się okropnie, ale mimo to na przyjęciu bawiłem się przednio". Takie zdanie jest dla modelu ciężkie do poprawnego rozpoznania ze względu na obecność słów o bardzo rozbieżnym wydźwięku, występują tu słowa o bardzo pozytywnym wydźwięku jak „bawiłem się" i „przednio", ale też jest słowo „okropnie", które ma bardzo negatywny wynik sentymentu.

Modele analizy sentymentu nie są natomiast w ogóle w stanie poprawnie sklasyfikować opinii sarkastycznych. Dzieje się tak ponieważ, wydźwięk emocjonalny nie jest przenoszony wprost w słowach dlatego też takie zdanie „Bardzo się cieszę, że mój pociąg przyjedzie z

dwugodzinnym opóźnieniem" nie jest w stanie być prawidłowo sklasyfikowanym przez jakąkolwiek analizę sentymentu.

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

- Standardowa analiza sentymentu
- Szczegółowa analiza sentymentu
- Wykrywanie emocji
- Analiza sentymentu oparta o aspekt
- 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-graned sentyment 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, 5 Sentiment Anlysis Examples in Business, 2020).

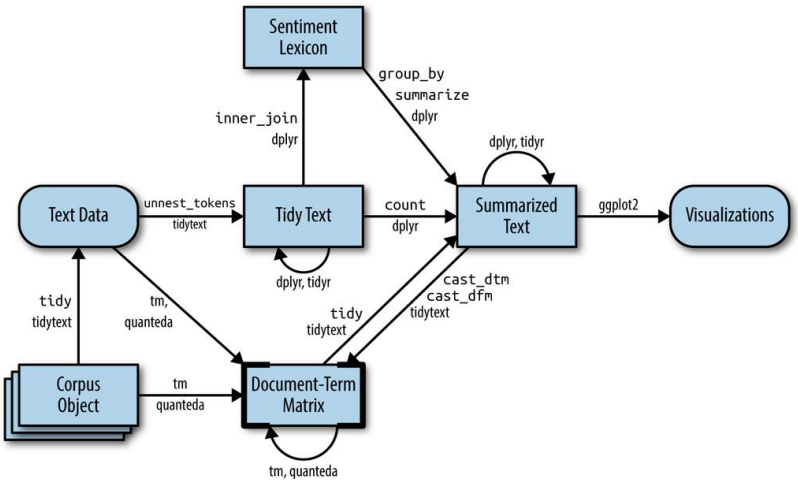
Analizę sentymentu, która opiera się o wykrywanie emocji używa się do rozpoznania emocji autora tekstu. W przeciwieństwie do standardowej i szczegółowej analizy sentymentu wykrywanie emocji jako rezultaty wyjściowe przyjmie wartości takie jak: smutek, radość, złość, strach czy zmartwienie. Systemy oparte o wykrywanie 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, 5 Sentiment Anlysis Examples in Business, 2020). Uczenie maszynowe w wykryciu emocji może być skuteczniejsze od korzystania z leksykonów ze względu na fakt, że 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, 5 Sentiment Anlysis Examples in Business, 2020).

Analiza sentymentu oparta o aspekt polega na rozszerzeniu analizy sentymentu o

dokonania oceny przedmiotu opinii autora. Jest to szczególnie użyteczne w przypadku badania opinii klienta o produkcie, gdyż można w ten sposób określić co jest przyczyną zadowolenia bądź niezadowolenia klienta. Zaletą tego rozwiązania jest również jego skalowalność: analiza sentymentu oparta o aspekt pozwala na zautomatyzowaną analizę dużej ilości danych tekstowych, pozwalającą na oszczędność zasobów finansowych oraz czasowych. Inną korzyścią analizy sentymentu opartej o aspekt jest możliwość dokonywania analizy w czasie rzeczywistym, dzięki czemu można w znacznie szybszy sposób wykryć niepożądane efekty i podjąć kroki w celu ich eliminacji. Oprócz możliwości szybszej reakcji na zdanie klienta, analiza sentymentu oparta o aspekt pozwala na lepsze zrozumienie produktu oraz 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 te jest stosowane w chatbotach oraz systemach inteligentnego dialogu, gdzie bardzo istotne jest poprawna interpretacja potrzeb i zamiarów użytkownika. Do analizy zamiaru używa się transformerów czyli modeli deep learning pozwalających na wyliczenie istotności poszczególnych elementów wejściowego zbioru danych. Inną strukturą wykorzystywaną do analizy zamiaru jest kapsułowa sieć neuronowa. Ten model jest używany do modelowania hierarchicznych związków, neurony w tej sieci są kapsułami będącymi wektorami prawdopodobieństw wystąpienia określonej cechy (Pascual, 2019).

1.4 DocumentTermMatrix i TermDocumentMatrix



Wykres 1: Schemat działania metod biblioteki tidytext przy pracy z macierzami DTM i TDM Źródło: „Text mining with R" – Julia Silge i David Robinson

W eksploracji tekstu do jednych z najczęściej używanych struktur analitycznych należą macierze DocumentTermMatrix i TermDocumentMatrix. Macierz DocumentTermMatrix jest to macierz w której:

• Każdy wiersz reprezentuje dokument (czyli źródło tekstowe takie jak książka czy

artykuł)

- Każda kolumna reprezentuje term (czyli słowo)
- Każda wartość reprezentuje ilość danego termu 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 skompresowania i co za tym idzie, mniejszym obciążeniem pamięciowym. Mierzenie rzadkości (ang. „sparsity”) takiej macierzy jest dane wzorem:

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

Transponowana macierz DTM jest nazywaną macierzą TermDocumentMatrix. Wtedy w to w danej kolumnie znajdują się liczności 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-df.

Zastosowanie macierzy DTM czy TDM do różnych zadań w przetwarzaniu języka naturalnego, do takich czynności należą:

- Poprzez rozbicie tekstu i dokonania atomizacji, można dzięki macierzom DTM i TDM poprawić działanie silników wyszukiwania poprzez ujednoznacznienie słów o różnym znaczeniu i wyszukiwanie synonimów (Verma, 2021)
- Większość procesów NLP dotyczy eksploracji więcej niż jeden rodzaj danych behawioralnych w tekście. Macierze TDM są bardzo użyteczne w ekstrakcji danych behawioralnych, a poprzez przeprowadzanie wielowymiarowej analizy macierzy DTM można określić wiele tematów poruszonych w danych (Verma, 2021)

Prawo Zipfa głosi, że częstotliwość występowania danego termu w dokumencie jest odwrotnie proporcjonalna do jego rangi istotności, co oznacza, że iloczyn rangi istotności termu i częstotliwości występowania jest stała (Silge i Robinson, 2017).

1.5 N-gramy

Jedną z technik działania na ogromnych i skomplikowanych zbiorów tekstowych jest dzielenie tekstu na tzw. n-gramy. N-gramy są to sekwencje składające się z n słów. Pozwala to na łatwiejsze operowanie na tekście i ograniczenie czasowe działania algorytmów. Do uproszczenia tekstu używa się także stoplisty. Stoplista 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 logicznego sensu wypowiedzi, lecz nie nadają

jej żadnego znaczenia, tonu czy wydźwięku. Ilość n-gramów w jednym zdaniu lub fragmencie tekstu K można określić wzorem:

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

Gdzie X oznacza ilość 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 to zdanie: „Dzisiaj przewidywane są przelotne opady". Gdy rozbije się je na bigramy otrzyma się wówczas:

- „Dzisiaj przewidywane"
- „przewidywane są"
- „są przelotne"
- „przelotne opady

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

- „Dzisiaj przewidywane są"
- „przewidywane są przelotne"
- „są przelotne opady"

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

1.6 Algorytm LDA

Jednym z zagadnień dziedziny jaką jest text mining, jest modelowanie tematyki (ang. „topic modeling"), które polega na znajdowaniu wśród dokumentów grup o zbliżonej tematyce. Jedną z najpowszechniej stosowanych metod w modelowaniu tematyki jest algorytm LDA („Latent Dirchlet Allocation"). W algorytmie tym występują dwie zasady:

1. Każdy dokument jest zbiorem tematów

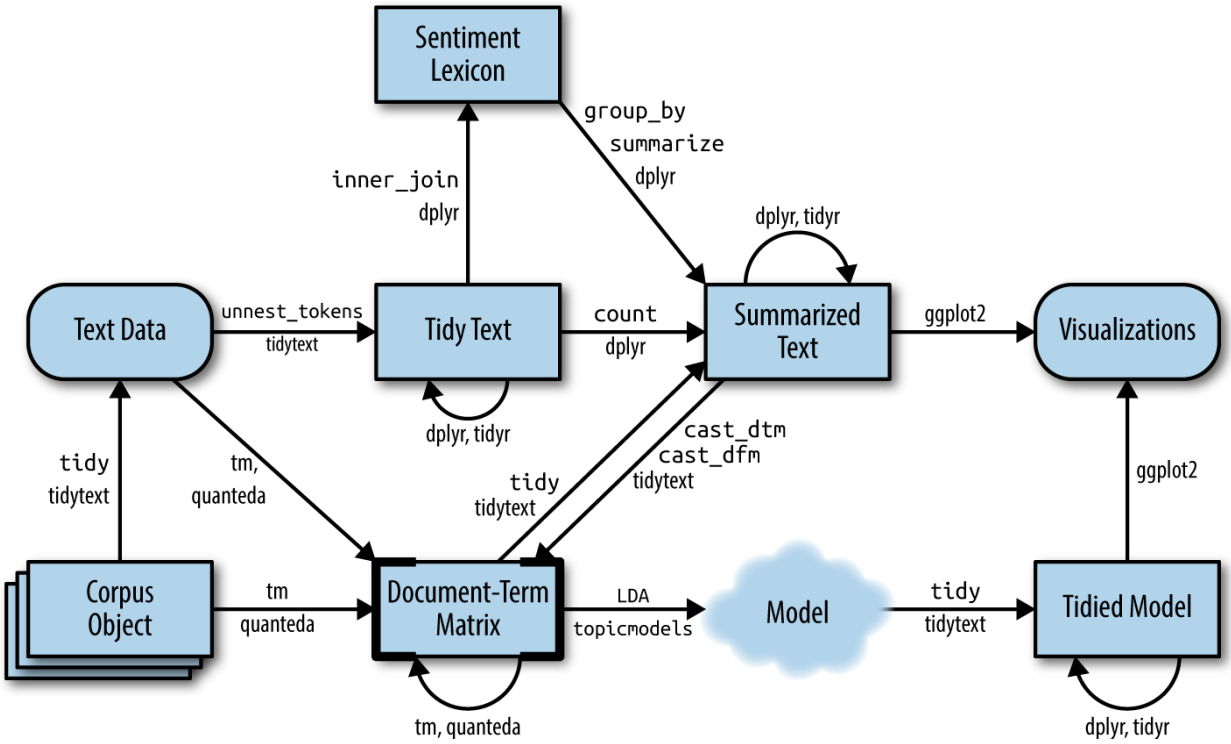
Należy przyjąć, że każdy zbiór tekstowy może zawierać słowa o różnej tematyce w szczególnych proporcjach. Przykładem może być dwutematyczny model, wówczas zbiory tekstowe mogą zostać sklasyfikowane w następujący sposób (Silge i Robinson, 2017):

- 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 to badania wiadomości w portalu medialnym, który dzieli wiadomości na dwie kategorie: „polityka" i „rozrywkowa". Do najczęściej pojawiających

się słów w informacjach poświęconych polityce będą takie wyrazy jak: „prezydent”, „parlament”, „rząd”, „ministrowie”. Z kolei wśród wiadomości ze świata 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 przy działaniu z algorytmem LDA
Źródło: „Text mining with R” – Julia Silge i David Robinson

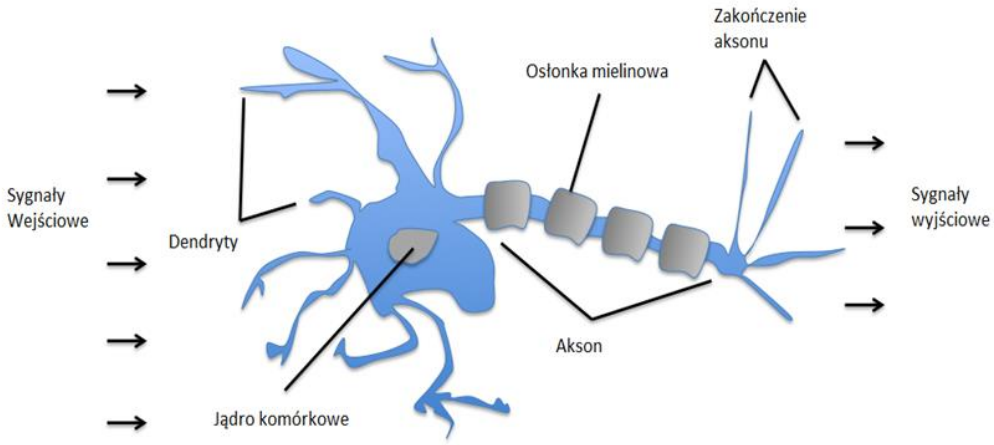
LDA jest matematyczną metodą, która jednocześnie określa grupę słów powiązaną z danym tematem oraz określa zbiór tematów, które opisują dany dokument. Do poddania dokumentów analizie, należy najpierw je poddać preprocessingowi, tokenizacji a następnie przekształcić w macierz DTM składającą się z *m* 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 należności tematu do dokumentu
- β (beta), prawdopodobieństwo należności 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).

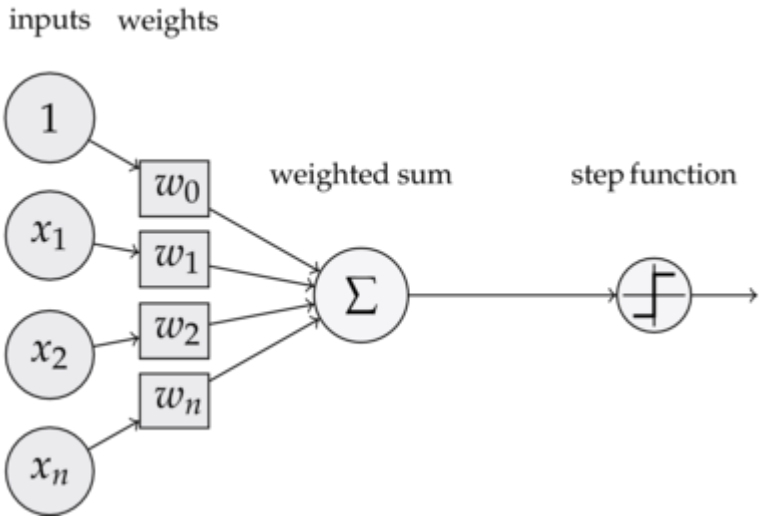
1.7 Czym są sieci neuronowe i jakie są ich zastosowania

Do klasyfikacji komentarzy użyto metody uczenia głębokiego (ang. „deep learning"). Opiera się ona o struktury matematyczne zwane sieciami neuronowymi. Nazwa tych struktur, nie jest przypadkowa, ich działanie jest inspirowane naturą, konkretnie mózgami zwierząt. Sam neuron jest komórką składająca się z ciała komórkowego (perikarionu), czyli wypustek określanych mianem dendrytów i aksonów. Neurony potrafią odbierać i przesyłać sygnały elektryczne oraz wszelakie informacje (Cherry, 2021). Dendryt z kolei to wypustka stanowiąca przedłużenie komórki nerwowej i odpowiada za odbieranie impulsów i 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 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).

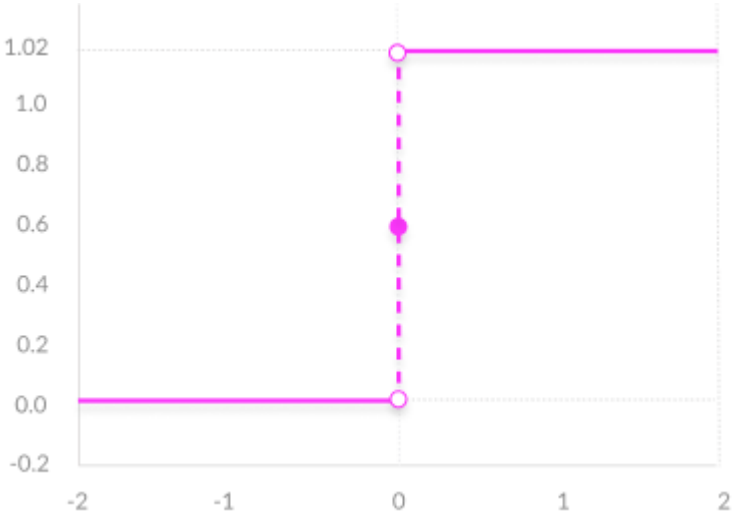


Wykres 4: Schemat perceptronu Rosenblatta
Źródło: www.statworx.com

Na wykresie 3 przedstawiono perceptron z tzw. uprzedzeniem (ang. „bias"), czyli do parametrów wejściowych dodano również wartość równa 1 razem ze swoją wagą w celu łatwiejszego przypisania. 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 to funkcji aktywacji, gdzie wartość sumy jest przekształcona na wyjście. Pierwotnie były to binarne funkcje dyskretne, które w zależności od tego czy suma jest większa od wartości progowej zwracały 1 lub 0 , bądź 1 lub – 1 (Loiseau, 2019).



Wykres 5: Przykład binarnej funkcji aktywacji
Źródło: www.mygreatlearning.com

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

Rozwiązaniem tego problemu stanowią nieliniowe funkcje aktywacji:

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

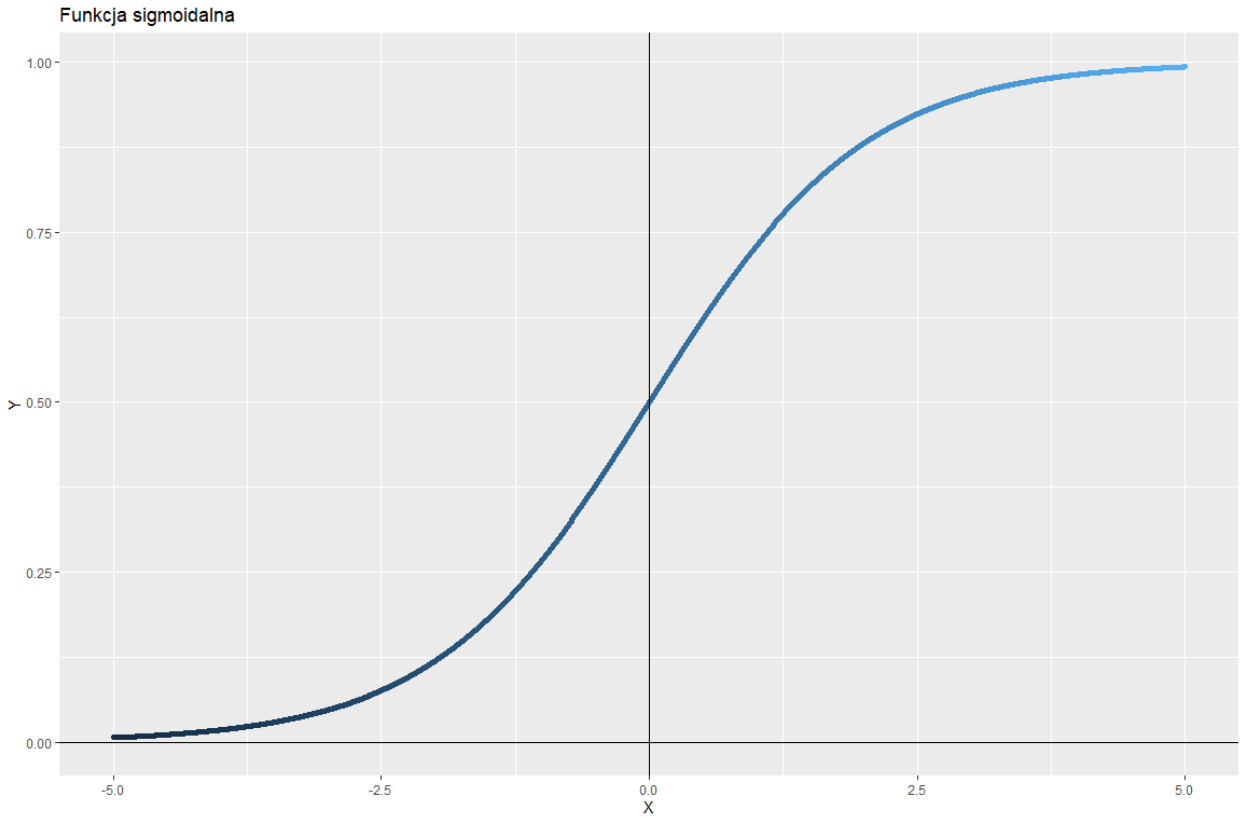
Do najczęściej używanych funkcji aktywacji należą:

1. Sigmoidalna funkcja aktywacji

Funkcja ta jest dana wzorem:

$$f(x) = \frac{1}{1 + e^{-x}}$$

Jest to funkcja bardzo często używana do modelowania prawdopodobieństwa ze względu na zakres jej wartości równy <0, 1>. Do kolejnej zalety tej funkcji należy jej różniczkowalność, która zapewnia łagodny gradient przez co nie ma skoków w wartościach wyjściowych.



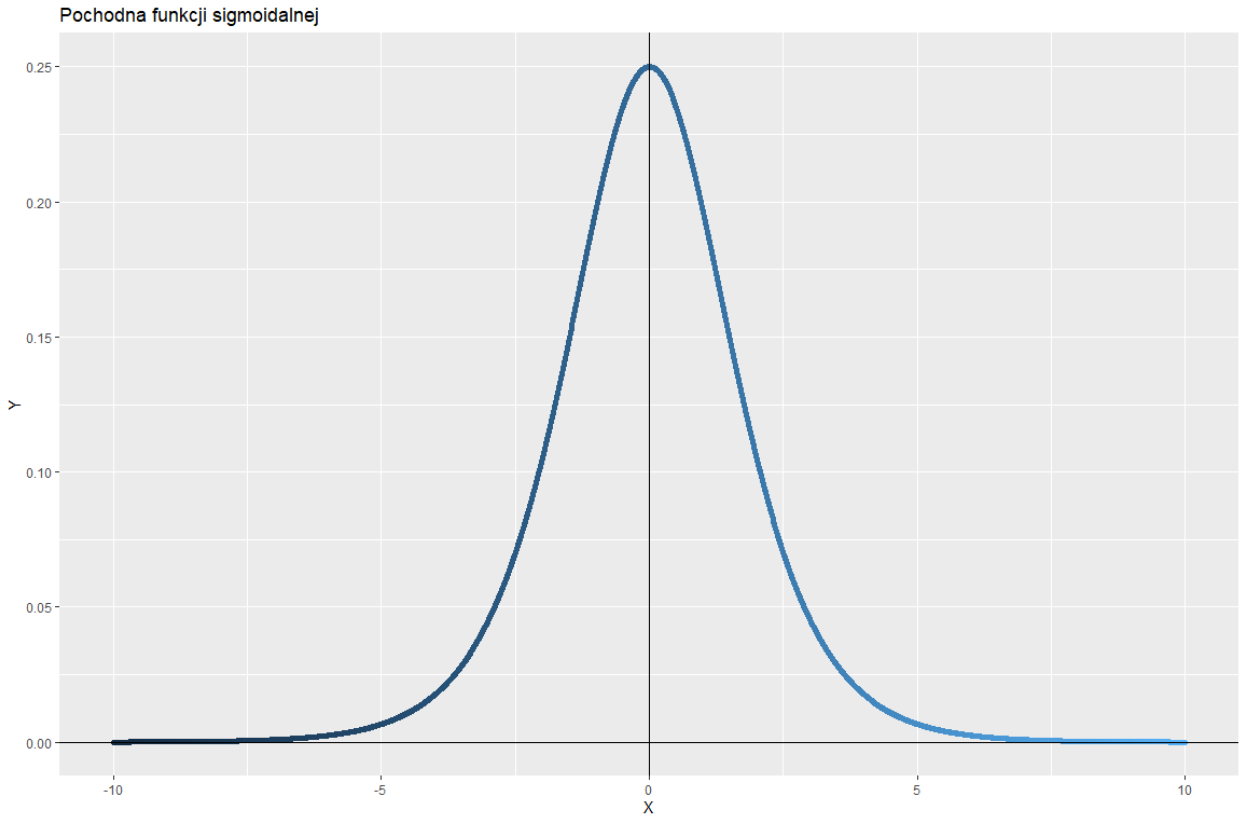
Wykres 6: Przedstawienie funkcji sigmoidalnej

Źródło: Opracowanie własne

Wadą funkcji sigmoidalnej jest natomiast mały zakres wartości jej pochodnej. Pochodna jest ta dana wzorem:

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

Gdzie $\sigma(x)$ oznacza funkcję sigmoidalną.



Wykres 7: Przedstawienie pochodnej funkcji sigmoidalnej

Źródło: Opracowanie własne

Wykres 6 pokazuje, że wartości są znaczące tylko na przedziale argumentów <-5,5>. Dla pozostałych argumentów, krzywa pochodnej funkcji sigmoidalnej zbiega do 0. To oznacza, że wartość gradientu w tych argumentach będzie bardzo niska. Jest to negatywne zjawisko, gdyż w przypadku gdy wartość gradientu osiąga 0, sięc przestaje się uczyć i dotyka ją tzw. problem zanikającego gradientu (ang. „vanishing gradient problem"). Kolejnym problemem jest kwestia wartości wyjściowych (Baheti, 2022). Na wykresie 5 można zauważyć, że krzywa nie jest symetryczna względem zera co oznacza, że wyjścia u wszystkich neuronów będzie tego samego znaku. Fakt ten utrudnia proces trenowania sieci.

2. Funkcja ReLU

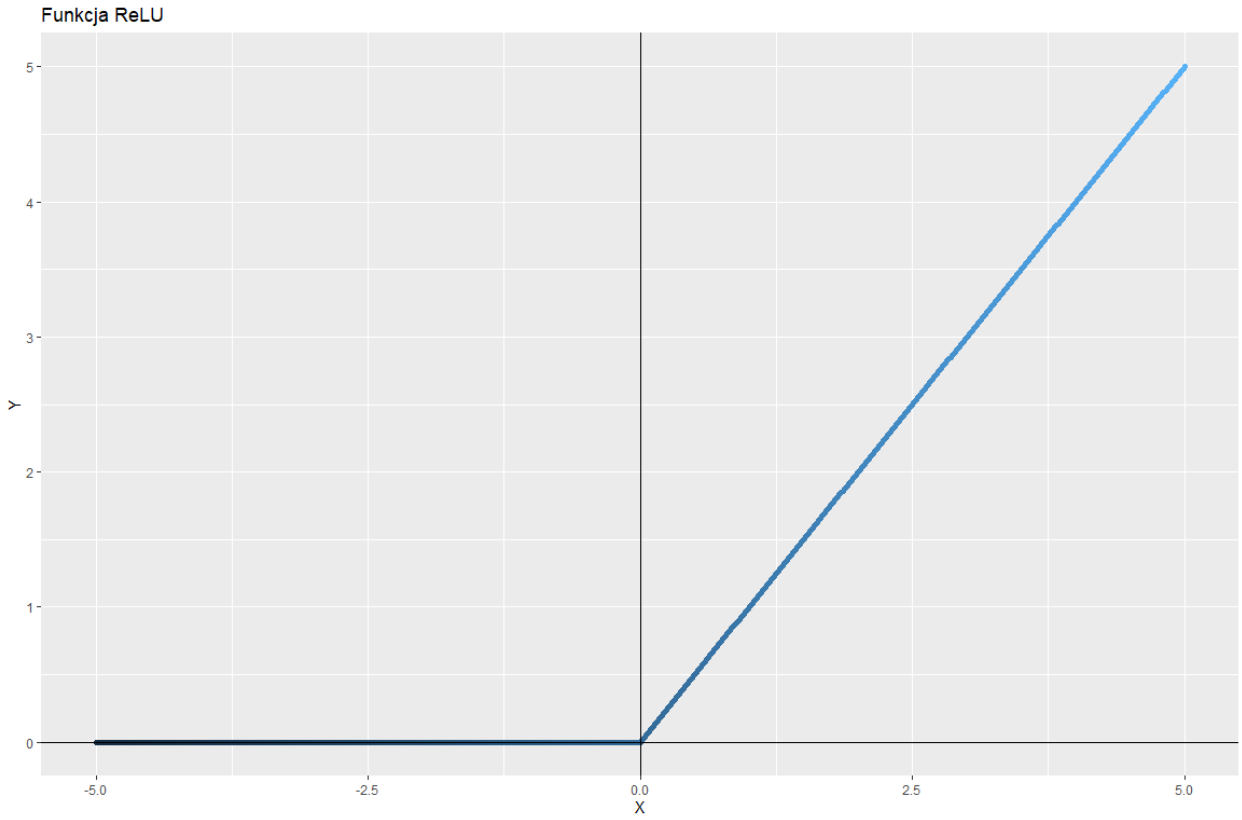
Funkcja ta jest dana wzorem:

$$f(x) = \max(x, 0)$$

Akronim ReLU oznacza „Rectified Linear Unit". Skorygowana funkcja aktywacji ReLU jest użyteczna ze względu jej różniczkowalność oraz możliwość korygowania wag algorytmem propagacji wstecznej przy jednoczesnym zapewnieniu wydajności obliczeniowej. Cechą

charakterystyczną funkcji ReLU jest to, że nie wszystkie neurony są aktywowane w tym samym czasie. W przypadku gdy wynik transformacji ReLU będzie równy 0, neuron zostanie deaktywowany. Do zalet funkcji ReLU należą:

- Większa wydajność obliczeniowa w stosunku do funkcji sigmoidalnej czy tanh ze względu na fakt, że tylko niektóre neurony są aktywowane
- Funkcja ReLU przyspiesza zbieganie spadku gradientu do globalnego minimum funkcji straty

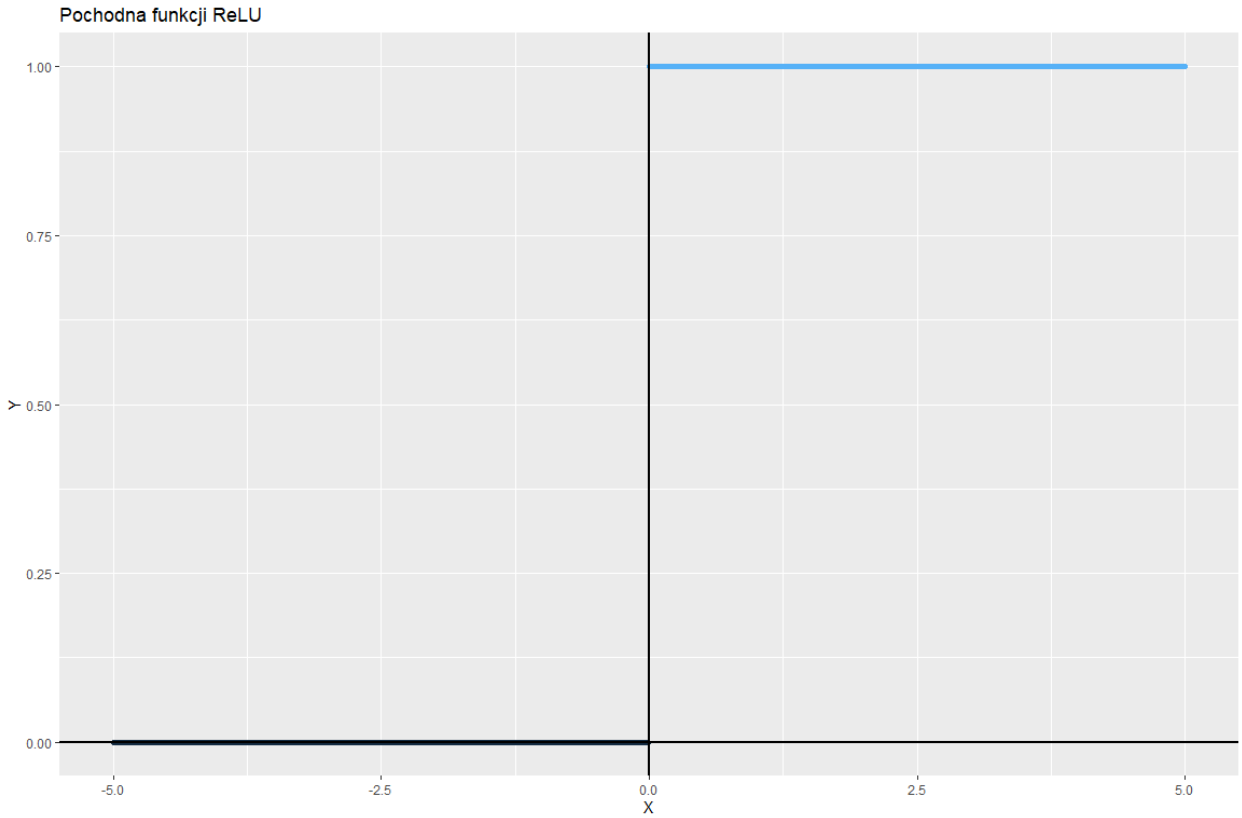


Wykres 8: Przedstawienie funkcji reLU
Źródło: Opracowanie własne

Do wad funkcji ReLU należy natomiast tzw. problem umierającego ReLU (ang. „dying ReLU problem). Pochodną tą można opisać wzorem:

$$f'(x) = g(x) = 1 \; jeżeli \; x \geq 0$$

$$g(x) = 0 \; jeżeli \; x < 0$$



Wykres 9: Przedstawienie pochodnej funkcji ReLU

Źródło: Opracowanie własne

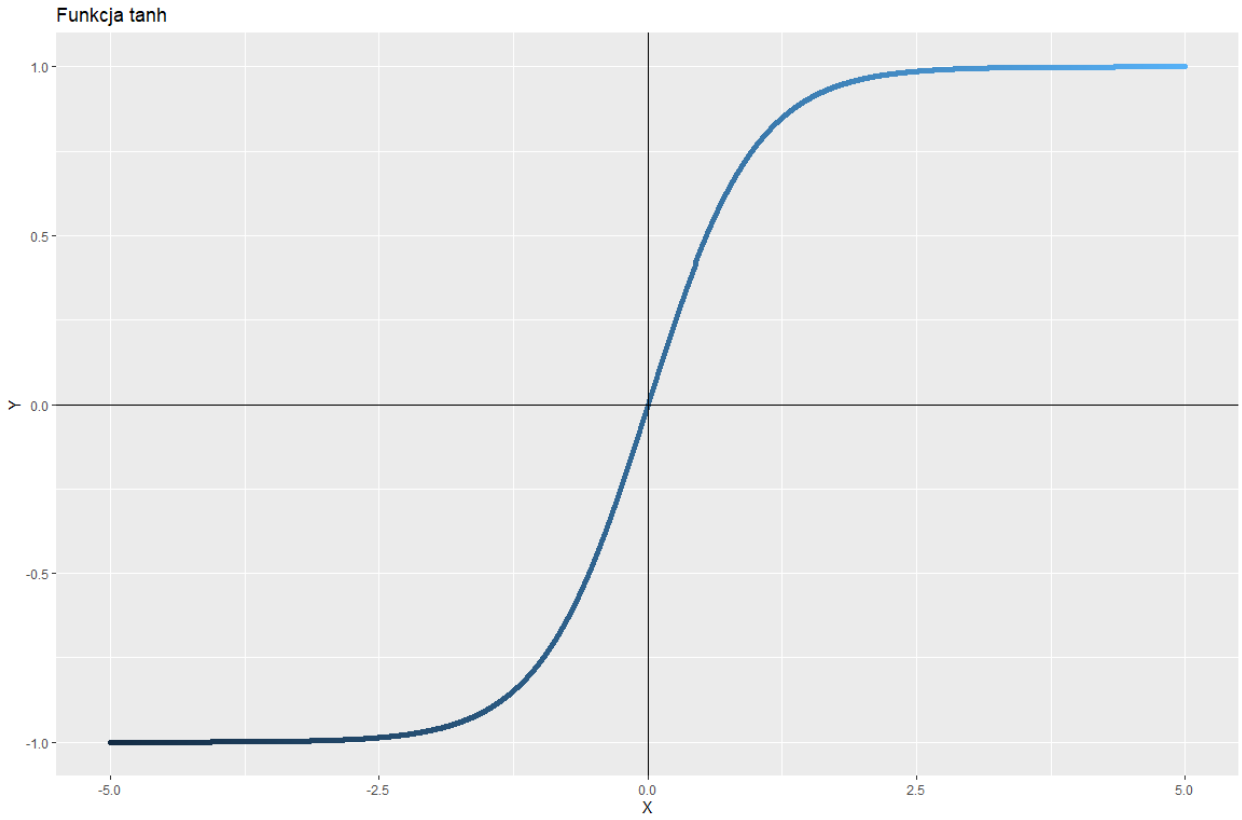
Na wykresie 8 widać, że dla ujemnych argumentów wartość gradientu będzie wynosić 0. Z tego powodu, dla niektórych, podczas kroku propagacji wstecznej, wagi w niektórych neuronach nie zostaną poprawione. To z kolei może skutkować w utworzeniu martwych neuronów, które nigdy nie zostaną aktywowane.

3. Funkcja tanh

Funkcja ta jest dana wzorem:

$$f(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}$$

Funkcja tanh jest bardzo podobna do funkcji sigmoidalnej. Jej krzywa podobnie jak krzywa sigmoidalna układa się w kształt przypominający literę „S". Różni się ona natomiast zakresem wartości, funkcja tanh przyjmuje wartości także ujemne czyli od -1 do 1. Funkcja ta zbiega dla coraz mniejszych argumentów do -1, z kolei dla coraz większych argumentów wartość funkcji zbiega do 1 (Baheti, 2022).

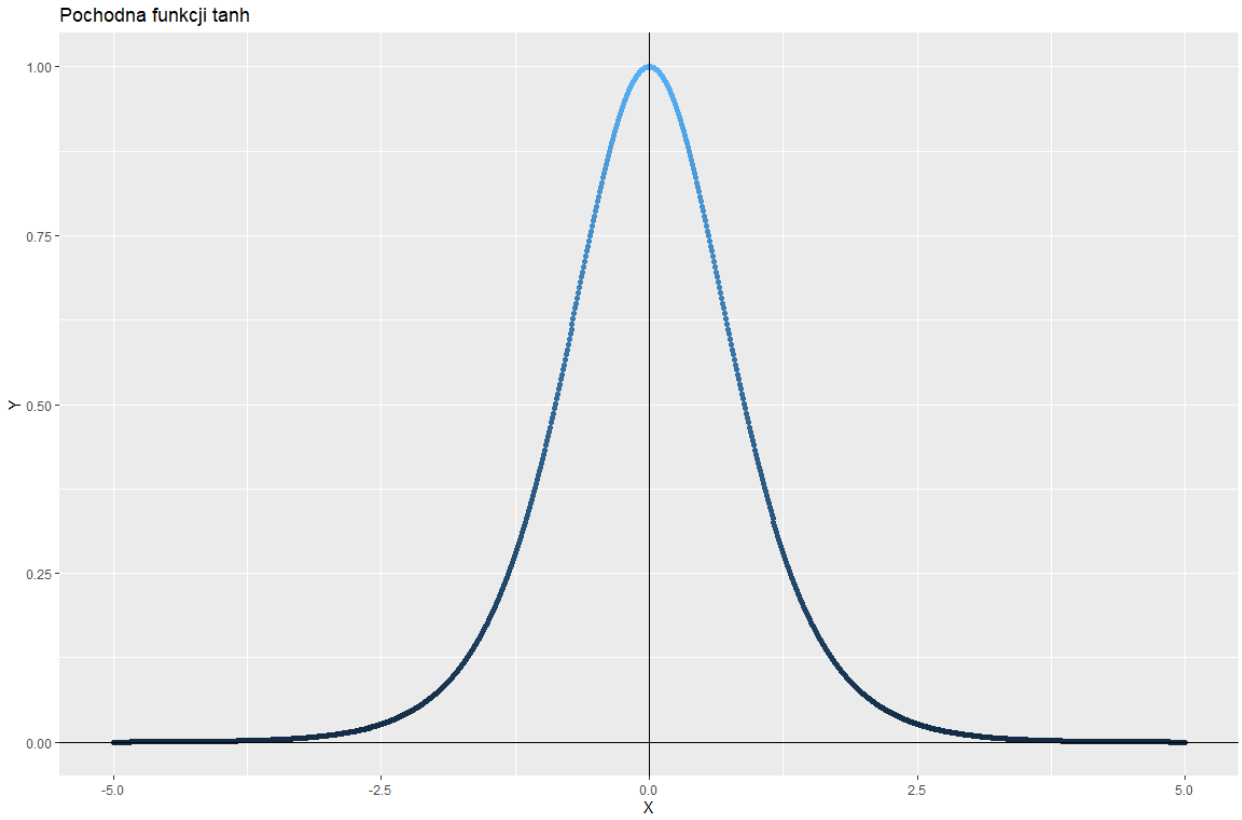


Wykres 10: Przedstawienie funkcji tanh

Źródło: Opracowanie własne

Funkcja tanh ma jednak bardzo podobne ograniczenia co funkcja sigmoidalna. Podobnie tylko dla wąskiego zakresu argumentów gradient ma wartości znaczniej różniące się od 0. Pochodna funkcji tanh jest dana wzorem:

$$tanh'(x) = 1 - tanh^2(x)$$



Wykres 11: Przedstawienie pochodnej funkcji tanh

Źródło: Opracowanie własne

Na wykresie 10 można zauważyć, że również problem zanikającego gradientu dotyczy również funkcję tanh. Różnicą natomiast między krzywą funkcji sigmoidalnej a tanh jest większa stromość krzywej funkcji tanh.

4. Funkcja softmax

Jest to funkcja, która za argument przyjmuje wektor prawdopodobieństw (mogą to też być wyliczone funkcją sigmoidalną predykcje). Aby ta funkcja zadziałała, musi zostać spełniony warunek: suma wartości w wektorze musi być równa 1. Następnie każdemu elementowi wektoru jest przypisywana wartość według następującego wzoru:

$$softmax(x_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

Dlatego dla wektora wyjść neuronów [0,34. 1.34. 0.85] funkcja softmax wyliczy wektor prawdopodobieństw [0,19; 0,5 i 0.31]. W następnej kolejności funkcja ta zwróci wektor zerojedynkowy, w którym najwyższej wartości wyjściowej zostanie przypisana jedynka, a reszcie zera. W rezultacie otrzymany zostanie końcowy wektor [0, 1, 0]. Funkcja softmax jest najczęściej używana w ostatniej warstwie sieci neuronowych w przypadkach klasyfikacji wielopoziomowych (Baheti, 2022).

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:

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

Gdzie

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

Przy czym n oznacz liczbę elementów w warstwie wyjściowej.

Problemem uczeni sieci neuronowych jest znalezienie globalnego minimum funkcji błędu ξ. Bardzo często używaną do tego metodą jest gradientowa metoda największego spadku. Polega to 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:

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

Gdzie η oznacza współczynnik proporcjonalności.

W przypadku liniowych elementów przetwarzających zachodzi równości:

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

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

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. „regułą 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 na przykład w modelach generatywnych które opierają się o sieci rekurencyjne czyli 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 sieci składające się od kilkunastu do kilkudziesięciu neuronów (może być w jednej warstwie 16, 32 czy nawet 64 neuronów). Najczęściej używaną funkcją aktywacji w tego rodzaju warstwie jest funkcja ReLU. Warstwę tę też można użyć jako warstwę końcową (np. w postaci neuronów z funkcją sigmoidalną). Innymi, szeroko stosowanymi warstwami sią 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 mniejszego zmniejszenia dalszych warstw co w rezultacie daje głębszą architekturę (tj. więcej warstw w sieci neuronowej). Większy rozmiar z kolei kosztem odbija się na mniejszej dokładności lecz lepszej generalizacji problemu, gdy np. w danych wejściowych nie trzeba przykładać uwagi do szczegółów. 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 szeroko używane w rozpoznawaniu i klasyfikacji obrazów (filtry 2D) czy też 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 ilość 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 (Wówczas wynik się uśrednia)

Jako przykład można podać macierz pikseli :

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

Stosując 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 nadmiernego przeuczania się sieci używa się warstw w których dokonywany jest tzw. dropout (ang. „odrzucenie"). Polega to na losowym wyłączaniu niektórych połączeń między neuronami. Powoduje to, że sieć nie łapie zbyt szybko wyuczonego wzorcu przez co problem przeuczenia jest znacznie zredukowany.

Oprócz rozpoznawania i klasyfikacji obrazów czy tekstów, sieci neuronowe można wykorzystywać również do analizy video 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, a nawet takie obiekty jak zdjęcia, dźwięk czy tekst. Sieci generatywne można np. zastosować w odtworzeniu bardzo starych nagrań czy przy odnowieniu jakichś starych zdjęć z brakującymi elementami (Mamczur, 2021).

2. Analiza sentymentu w badaniach rynkowych

2.1 Źródła danych

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

1. Media społecznościowe

Według szacowań, 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. Ponadto, wśród serwisów jak TikTok czy Tumblr dominującą grupą są ludzie w przedziale wiekowym od 16 do 24 lat, wśród użytkowników serwisu Facebook są to ludzie w grupie wiekowej 25-34 (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. 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, Why Should We Use Sentiment Analysis In Social Media Mining?, 2021). Analiza sentymentu dla tych portali jest bardzo wartościowa, gdyż informacje o ilości polubień, udostępnień czy komentarzy pod wpisem bądź filmem nie jest zawsze w stanie przedstawić

wiarygodną informację o zadowoleniu użytkowników oraz ich opinii na dany temat. Analizy sentymentu w mediach społecznościowych używa się do badania:

- Odkrywania nowych trendów na rynku
- Śledzenia świadomości marki
- Potencjalnych nowych produktów
- Działalności konkurencji
- 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 takie jak np. wady w produkcie czy też efekty uboczne w przypadku leków (Repustate, 10 Sentiment Analysis Data Sources For Strategic Data Analytics, 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 i wiedzę o potencjalnych możliwościach inwestycyjnych (Repustate, 10 Sentiment Analysis Data Sources For Strategic Data Analytics, 2021).

4. Interakcje z pracownikiem oraz jego feedback

Opinia zwrotna od pracownika stanowi kluczową rolę w badaniu doświadczenia pracownika (z ang.”employee experience”, EX) oraz poprawy kultury organizacyjnej przedsiębiorstwa czy też zadowolenia pracowników. Dane mogą zostać zebrane w postaci rozmów audio, chatbotów, nagrań video, komunikacji z działem HR oraz ankiet (Repustate, 10 Sentiment Analysis Data Sources For Strategic Data Analytics, 2021).

5. 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 Patient Voice w której pacjenci mogą udzielić opinię zarówno jak o wizycie u lekarza jak również o całej kuracji czy o obsłudze placówki medycznej. 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, 10 Sentiment Analysis Data Sources For Strategic Data Analytics, 2021).

2.2 Zastosowanie biznesowe

Do najczęściej realizowanych celów biznesowych analizą sentymentu należą takie pojęcia jak:

1. Monitorowanie marki

Monitoring marki oraz zarządzanie reputacją firmy to jedna 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 w 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 to wykrycia trendów oraz dokonanie estymacji potencjalnej reakcji klientów na określoną decyzję ze strony firmy (Roldós, 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, forów powoduje, że jedno złe doświadczenie może wywoływać straty w firmie wiele razy. Analiza sentymentu na danych tekstowych 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).

3. Badanie satysfakcji pracowników

Analiza sentymentu opierająca się o informacje zwrotne od pracowników w ankiecie czy też w opiniach użytkowników na portalach takich jak Glassdoor czy GoWork albo 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).

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 jakie zmiany wpłynęłyby pozytywnie na jego satysfakcję. Ponadto, wykorzystanie analizy sentymentu opartej na aspekt umożliwia ocenić potencjalne obszary do zmian w konkretnej cesze produktu takie jak np. funkcjonalność, interfejs czy też doświadczenie użytkownika (ang. „user experience", „UX") (Roldós, 2020).

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 oparta o aspekt pozwala na wskazanie konkretnych elementów, które należy uznać za szczególnie istotne (Wonderflow, 2018).

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 oraz można w ten sposób określić swoją pozycję na rynku w tej chwili. Ponadto, dzięki analizie sentymentu opartej na aspekcie można porównać sytuację firmy z konkurencją na różnych szczeblach takich jak: obsługa klienta, jakość produktów, doświadczenie użytkownika. Kolejną zaletą tego rozwiązania jest możliwość śledzenia postępów oraz zmian pomiędzy przedsiębiorstwem a jego konkurencją (Wonderflow, 2018).

7. Obserwacja mediów społecznościowych

Codziennie na portalach mediów społecznościowych publikowane jest 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, że serwisy takie jak Twitter czy YouTube udostępniają otwarte API, co pozwala na analizę danych w czasie rzeczywistym (Wonderflow, 2018).

8. Lepsze zarządzanie w kryzysowych sytuacjach

Ze względu na to, że analiza sentymentu odbywa się w czasie rzeczywistym, w systemie opartym o taką analizę, nagły wzrost negatywnych opinii nie może pozostać niezauważalnym. W ten sposób można natychmiast zapobiec wizerunkowemu upadkowi poprzez szybie podjęcie działań w przypadku gdy w portalach mediów społecznościowych, serwisach informacyjnych bądź stronach z opiniami i recenzjami użytkowników dojdzie do wzrostu krytycznych opinii niezadowolonych użytkowników (Wonderflow, 2018).

2.3 Przykłady zastosowań analizy sentymentu

Wśród komercyjnych zastosowań analizy sentymentu można wyróżnić takie produkty jak:

1. Talkwalker

Jest to narzędzie wykorzystujące NLP, w tym analizę sentymentu do badania opinii klientów czy ich odczuć. 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 oraz reputację firmy z konkurencją. Inną zaletą tego produktu jest jego wrażliwość na emocjonalny wydźwięk tekstu: model analizy sentymentu jest w stanie na przykład wykryć sarkazm w zdaniu co jest jednym z największych wyzwań przy analizie sentymentu. Z usług Talkwalker korzystają takie firmy oraz instytucje jak: Goodwill, Spotify, Orange, Dentsu International czy Europejski Bank Inwestycyjny. Strona produktu jest dostępna pod następującym linkiem: www.talkwalker.com

2. Clarabridge

Produkt ten wykorzystuje różne źródła danych takie 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. Narzędzie to jest w stanie rozpoznać temat wypowiedzi, wyróżnić kluczowe informacje. Główną funkcjonalnością tego produktu jest badanie zachowania klientów i ich predykcja. Do użytkowników tej platformy należą takie korporacje jak: BMW, Under Armour, Accenture, Deloitte czy też Ernst & Young. Strona produktu znajduje się pod linkiem: 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)
- Analizy VoC (ang. „Voice of the Customer") – czyli analizowanie informacji zwrotnej od klienta
- Analiza pracowników – wykrycie mocny i słabych stron pracowników w organizacji w celu poprawy ich satysfakcji oraz osiągnięcia lepszej produktywności
- Analiza mediów społecznościowych
- Serwisy kontaktowe 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 linkiem: www.meaningcloud.com

4. Aylien

Portal ten w odróżnieniu do poprzednio wymienionych komercyjnych produktów nie dokonuje analizy sentymentu danych tekstowych związanych z klientami bądź potencjalnymi konsumentami produktów. Głównym zastosowaniem tego produktu jest analiza danych informacyjnych z różnych portali oraz monitorowanie najnowszych danych finansowych. W rezultacie, firmy korzystają z tego rozwiązania w celu szacowania ryzyka oraz wykrycia potencjalnych inwestycji na rynku. 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. Strona tego produktu jest dostępna pod linkiem: www.aylien.com

5. Mention

Do rozwiązań tego systemu należy monitorowania sytuacji, w których zostanie wspomniana nazwa organizacji (ang. „mention" – wzmianka) w mediach społecznościowych, portalach informacyjnych czy też wyszukiwarkach internetowych. Głównym zastosowaniem analizy sentymentu w tej platformie jest monitoring reputacji organizacji na rynku oraz porównanie jej z firmami konkurencyjnymi. Organizacjami oraz korporacjami, które korzystają z usług Mention są: Benq, Deliveroo, Microsoft czy też Prisma Media. Strona główna Mention jest dana pod adresem: www.mention.com

Analiza sentymentów 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, które stanowią źródło obszernej i cennej wiedzy, która nie była w takim stopniu dostępna 20 czy 30 lat temu.

3. Analiza tekstu oraz weryfikacja komentarzy w języku R

3.1 Opis zbioru danych

Zbiór danych dotyczy komentarzy pacjentów na użyty przez nich w określonej dolegliwości. Komentarze były zbierane z różnych stron poświęconych recenzjom leków poprzez współpracę uniwersytetu Kansas State i Uniwersytetu Technicznego w Dreźnie. 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:

- 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

Zbiór danych jest dostępny pod linkiem
https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29
(Kallumadi i Gräßer, 2018):

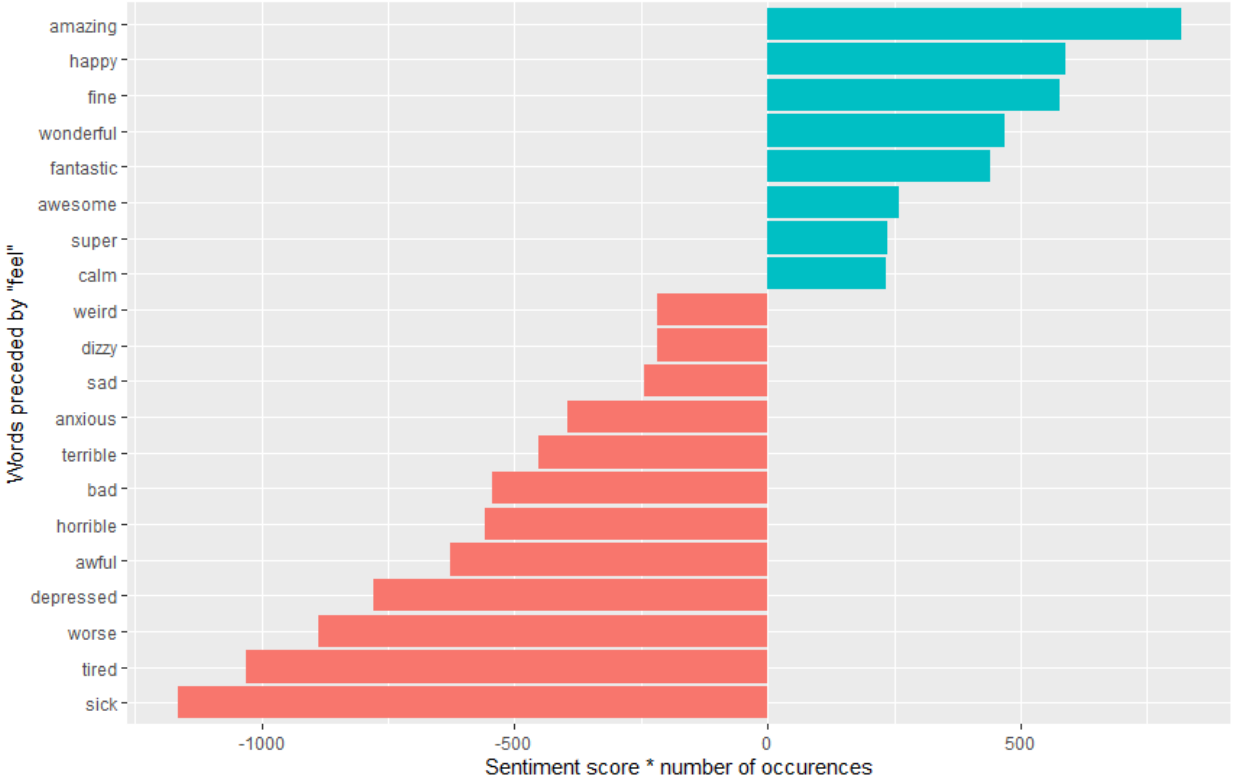
3.2 Wyszukanie związków o najwyższej kontrybucji i podział na n-gramy

Do wyszukania słów kluczowych oraz wykrycia tych wyrażeń o najwyższym współczynniku kontrybucji użyto biblioteki tidytext. W pierwszym kroku usunięto ze zbioru komentarzy wyrazy znajdujące się w stop liście. Są to słowa budujące logikę zdania takie 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 oszczędności pamięci operacyjnej. Następnie przy pomocy biblioteki tidytext pobrany został leksykon AFINN (Technical University of Denmark, 2011), który jest słownikiem w którym każde słowo ma przypisaną wartość sentymentu. W tym leksykonie słowa o negatywnym wydźwięku przyjmują ujemne wartości, zaś te o pozytywnym zwracają wartości dodatnie (np. wyrazy „amazing" czy „breathtaking" 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 gdyż można w ten sposób ukazać związek między bezpośrednio sąsiadującymi ze sobą słowami oraz obliczyć ich kontrybucję na podstawie częstości ich występowania oraz wartości sentymentu. Po obliczeniu wystąpienia słów obliczona została kontrybucja każdego słowa wzorem (Silge i Robinson, 2017):

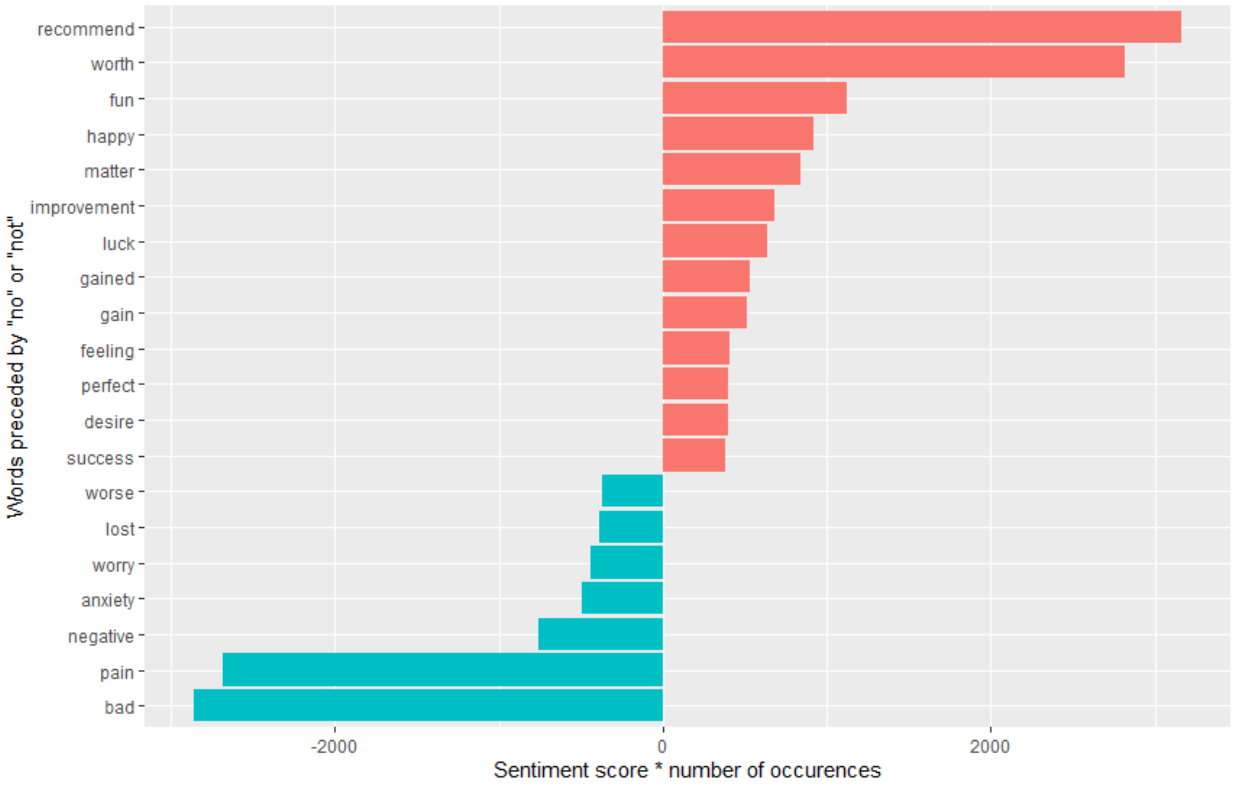
$$Kontrybucja = Wartość\ sentymentu\ słowa * ilość\ wystąpień$$

Gdzie wartość sentymentu słowa oznacza wartość sentymentu dla danego słowa w leksykonie AFINN, a ilość wystąpień dotyczy ilości 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 najczęściej były poprzedzone słowem „feel" w różnej odmianie
Źródło: Opracowanie własne



Wykres 13: Słowa które najczęściej były poprzedzone słowem "no" lub "not" o najwyższym scoringu bezwzględnym
Źródło: Opracowanie własne

Wykres 1 pokazuje, że z takimi słowami jak „feel" czyli „czuć" w kontekście pozytywnym są związane takie słowa jak „amazing"(niesamowicie), „happy"(szczęśliwy) czy „fine"(dobrze). Jeśli chodzi natomiast o słowa o negatywnym wydźwięku, które w dużym stopniu występowały ze słowem „feel" dominują słowa typowe dla chorób takie jak „sick"(chory), „tired" (zmęczony) czy „depressed". Ważnym słowem jest również „worse"(„gorzej", które sugeruje, że opinia pacjenta jest negatywna (Silge i Robinson, 2017)).

Wykres 2 pokazuje zaś, że z takimi słowami jak „not" lub „no" powiązane są takie słowa jak „recommend"(polecać), „worth"(warto), „fun"(zabawnie). W przypadku słów o negatywnym znaczeniu najczęściej występuje słowo „bad"(źle) oraz „pain"(ból). Kolory na wykresie 2 zostały specjalnie odwrócone ze względu na zmieniony kontekst, tutaj słowa o pozytywnym znaczeniu będą oznaczać, że opinia o leku była najprawdopodobniej negatywna, zaś połączenie słów „no" i „pain" czy „not" i „worry" będą wskazywać, że pacjent takowy lek poleca.

Wykres 14: Chmura słów najczęściej przedstawiających się w komentarzach
Źródło: Opracowanie własne

Wykres 3 przedstawia chmurę słów w tekście. Na zielono są słowa o pozytywnym znaczeniu, na czerwono słowa o znaczeniu negatywnym. Słowa takie jak „pain", „anxiety", „depression" czy „symptoms" wskazują na problemy zdrowotne jakie mieli pacjenci. Z kolei słowa „helped", „recommend", „worth", „effective" skupiają się już bardziej na ocenie leku.

.W następnym kroku zbadano współczynnik tf-idf do zbadania wagi słów w oparciu o liczbę ich wystąpień dla poszczególnych leków. Współczynnik tf-idf jest obliczany wzorem:

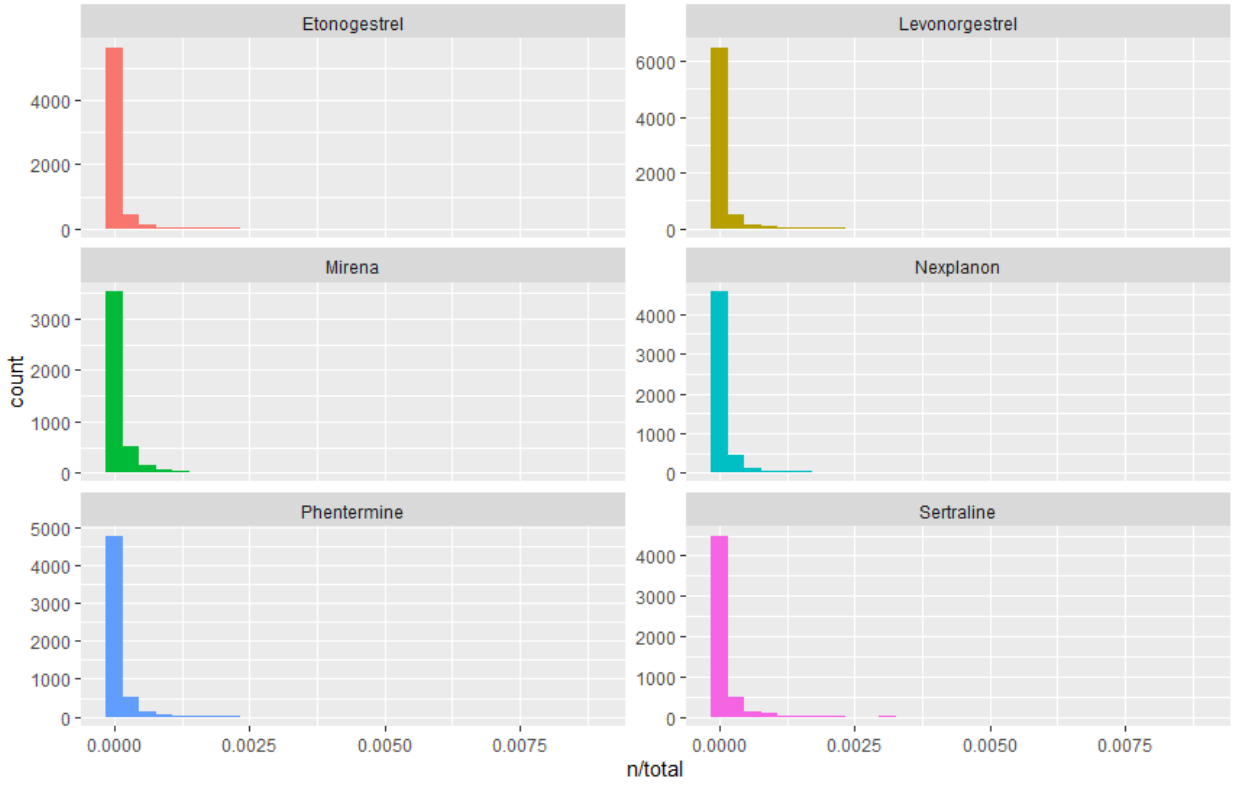
$$tf - idf_{i,j} = tf_{i,j} \times idf_i$$

Gdzie:

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

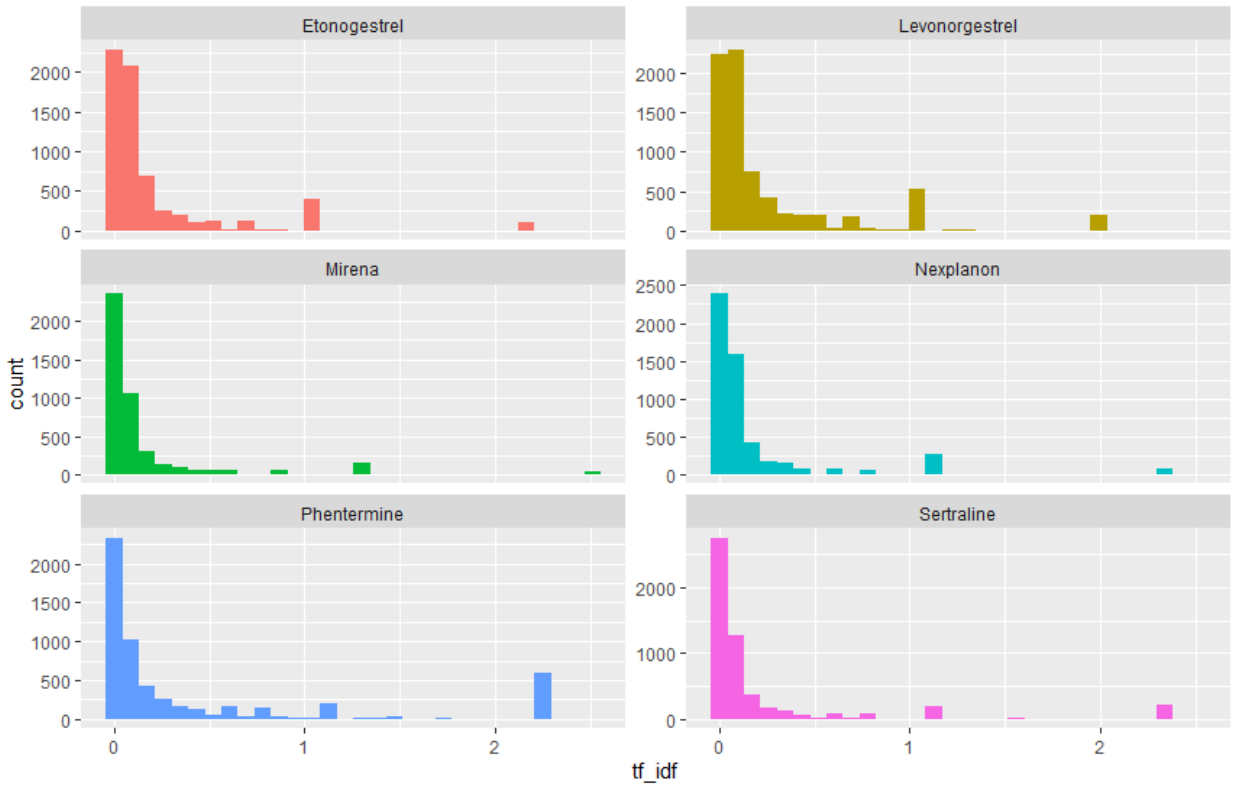
$$idf_i = \log \frac{|D|}{|\{d: t_i \in d\}|}$$

$|D|$ - oznacza liczbę dokumentów w korpusie, a $|\{d: t_i \in d\}$ 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 i można na nim robić analizę NLP w celu osiągnięcia określonego celu biznesowego. Analiza ta została przeprowadzona dla całego ogółu zbioru komentarzy.



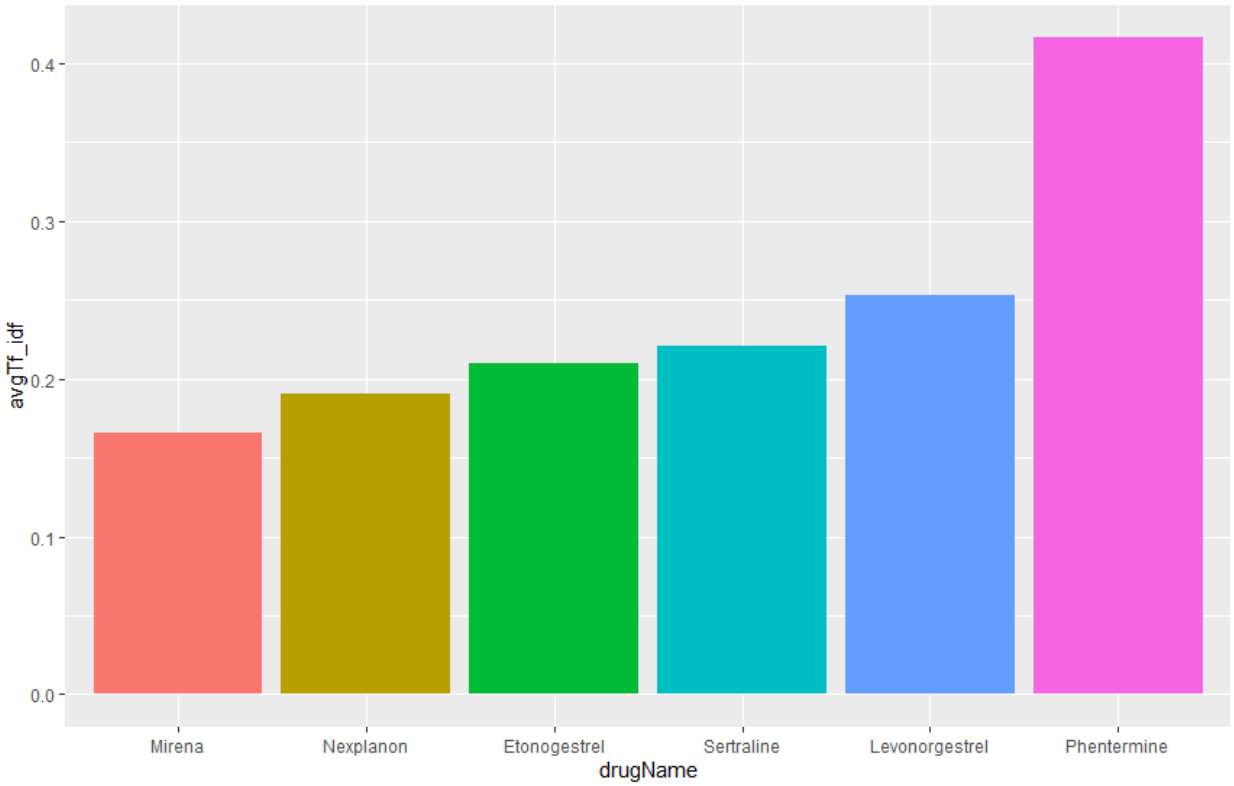
Wykres 15 Histogram współczynnika częstotliwości występowania słów(tf) w komentarzach dla 6 najczęściej komentowanych leków
Źródło: Opracowanie własne

Wykres 4 pokazuje, że dla każdego z 6 najbardziej ocenianych leków tj. Etonogestrel, Levonorgestrel, Nexplanon i Mirena (środki antykoncepcyjne), Phentermine (lek wspomagający odchudzanie) i Sertraline (środek antydepresyjny), zdecydowanie najwięcej jest słów mało 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



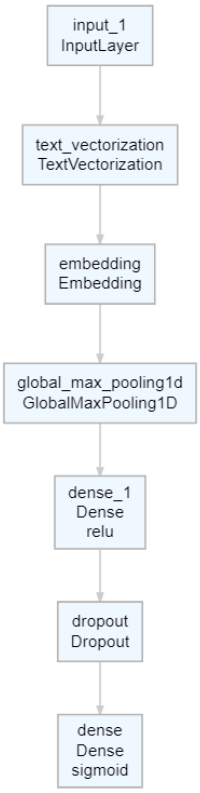
Wykres 17: Średni tf_idf dla 6 najczęściej ocenianych leków

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 nic nie znaczące. Na wykresie 5 widać, że dla każdego ze zbadanych leków jest grupa słów o współczynniku tf-df powyżej 1 czy nawet 2. Oznacza, że w zbiorze są słowa o dużej wadze, które pozwalają na zbadanie kontekstu wypowiedzi co jest kluczowe w możliwości stworzenia użytecznego modelu. Ma to odzwierciedlenie również w wykresie średnich współczynnika tf-idf (wykres 6), przy żadnym leku średnia tf_idf nie wynosi poniżej 0,1.

3.3 Analiza sentymentu przy wykorzystaniu sieci neuronowej do oszacowania oceny leku na podstawie komentarzy przy użyciu biblioteki keras

Po dokonaniu analizy eksploracyjnej, 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 o wysokim tf-idf) jak i bigramy (czyli n-gramy składające się z 2 słów) o dużej kontrybucji. Analizę wykonano na zbiorze składającym się ze z 161 297 obserwacji oraz 7 zmiennych. Na tej podstawie można dojść do wniosku, że ten zbiór komentarzy nadaje się do skonstruowania modelu sieci neuronowej w celu zaklasyfikowania czy dany komentarz był opinią pozytywną bądź negatywną. Najpierw dokonano konwersji zmiennej oceny na zmienną binarną. Za ocenę negatywną przyjęto oceny od 0 do 6, zaś za pozytywne oceny od 7 do 10. W następnej kolejności zbudowano model sieci neuronowej. Jej architekturę przedstawiono na wykresie.

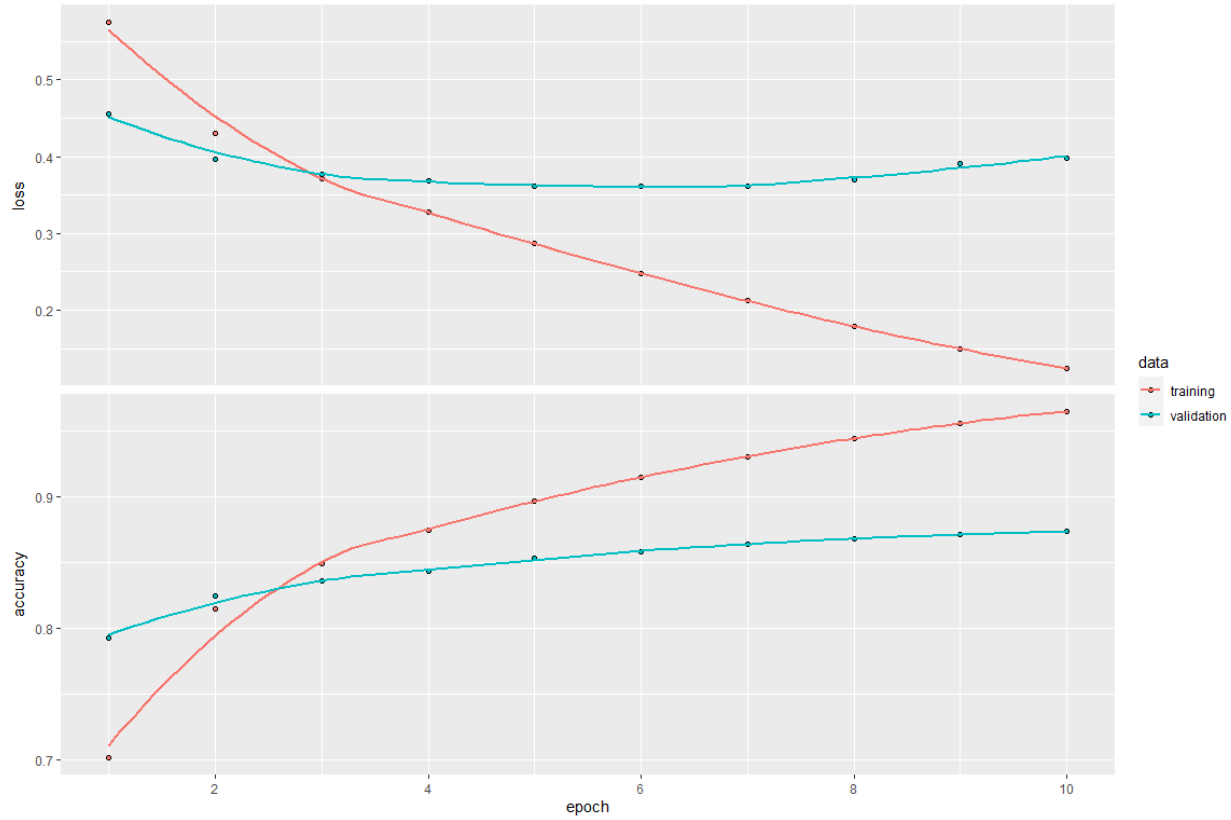


Wykres 18: Architektura modelu sieci neuronowej
Źródło: Opracowanie własne

W pierwszej kolejności stworzona jest warstwa wejściowa jednowymiarowa, następnie tworzona jest warstwa odpowiadająca za wektoryzację tekstu (Chollet, 2018). Tekst jest wówczas konwertowany na wektor bitów, gdzie jeżeli dane słowo wystąpi w danym tekście wejściowym, to wówczas wartość w indeksie tego słowa 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 warstwa wbudowana po której jest warstwa jednowymiarowego max poolingu (pooling jest jednowymiarowy gdyż na wejściu znajdują się przekształcone w wektory słowa). Max pooling umożliwia skupieni się na wyróżniających wartościach w macierzach(bądź wektorach) w przeciwieństwie do poolingu uśredniającego (Du i Shanker). 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 głębszych warstwach najczęściej używaną wartością tego parametru jest 0.5 (Brownlee, A Gentle Introduction to Dropout for

Regularizing Deep Neural Networks, 2018). Ostatnim elementem modelu jest warstwa wyjściowa gęsta w której funkcją aktywacji jest funkcja sigmoidalną zwracająca prawdopodobieństwo, że komentarz jest pozytywny (Chollet, 2018).

Po zbudowaniu sieci poddano ją uczeniu na zbiorze treningowym. Dla znaczącego przyspieszenia procesu uczenia skorzystano z wsparcia GPU w bibliotece tensorflow oraz keras. Za ilość iteracji przejścia przez cały zbiór danych przyjęto 10 zaś za batch 512, oznacza to, że w jednej iteracji bierze udział 512 obserwacji. Przyjęto taką ilość ze względu na dużą ilość obserwacji przetworzoną w jednej iteracji. Przy dużej ilości iteracji, w których jest niski batch model bardziej jest dostosowany do problemów wymagających szczegółowego wglądu w obserwację. Z kolei decyzja o mniejszej ilości iteracji, do której używa się batch o większym wolumenie obserwacji jest rekomendowane w przypadku gdy oczekiwana jest generalizacja problemu (Keskar, Mudigere, Nocedal, Smelyanskiy i Tang, 2017). Wyniki uczenia w postaci wykresów skuteczności modelu oraz wartości funkcji straty na zbiorze treningowym i walidacyjnym przedstawiona na wykresie 19.



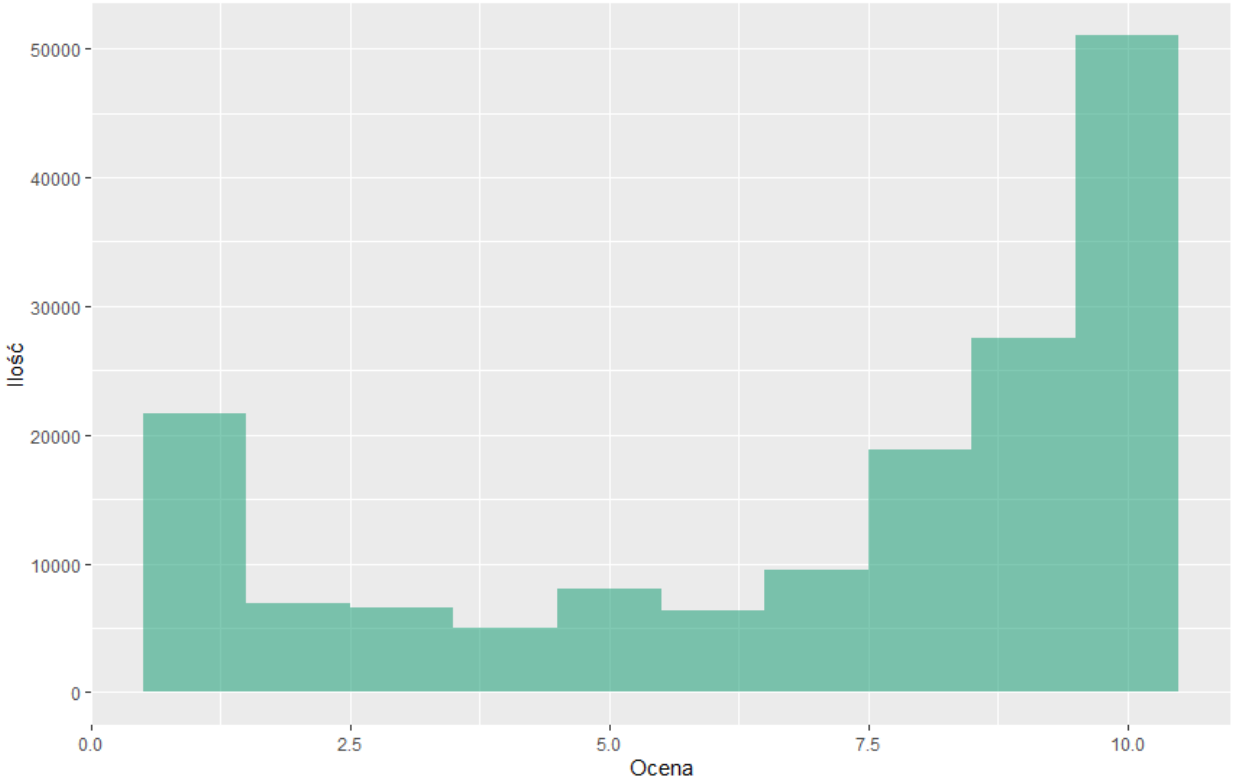
Wykres 19: Skuteczność i wartość funkcji straty modelu sieci neuronowej na zbiorze treningowym i walidacyjnym
Źródło: Opracowanie własne

Wartość funkcji straty na zbiorze walidacyjnym wyniosła 0,4 w 10 epoce, zaś dla zbioru

treningowego wyniosła 0,05. Z kolei precyzja modelu dla zbioru walidacyjnego wyniosła ok. 0,85 a dla treningowego ponad 0,95.

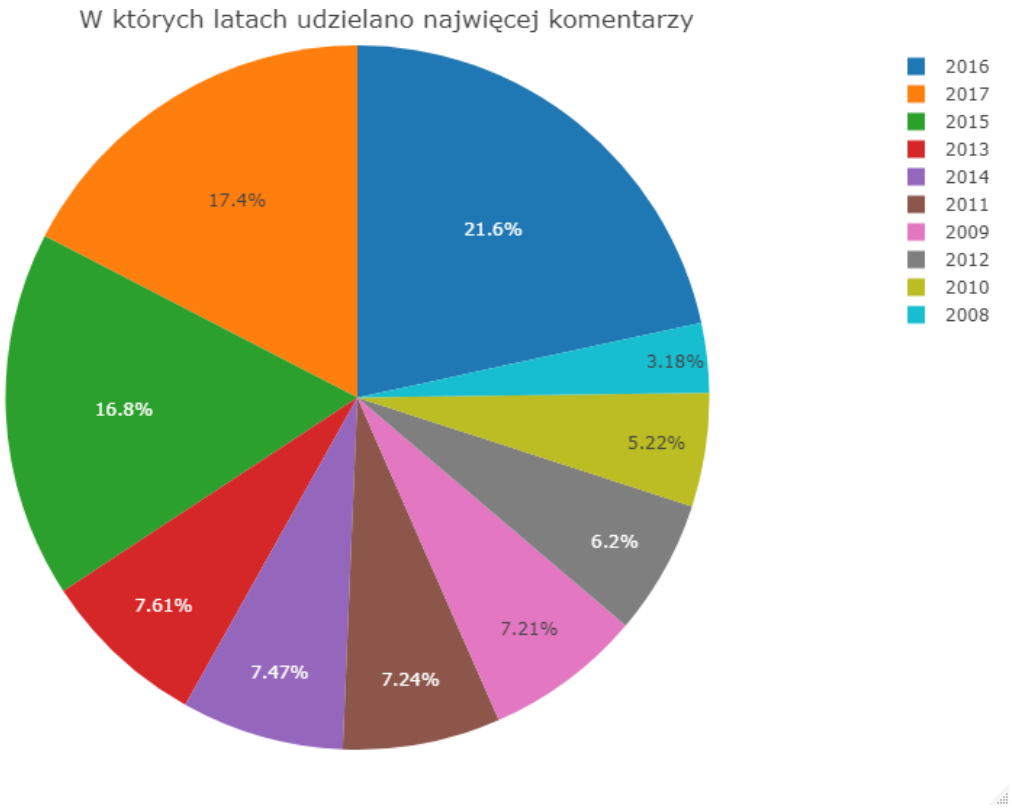
3.4 Przedstawienie wyników analizy eksploracyjnej oraz estymacji wyników dokonanych przez model

Do analizy eksploracyjnej użyto bibliotek graficznych takie jak: ggplot2 oraz plotly. Na początku zbadano histogram ocen leków. W następnym kroku dokonano analizy sentymentu opartej o sieć neuronową stworzoną za pomocą biblioteki keras. Na wykresie 20 przedstawiono jego wykres, który został stworzony przy pomocy biblioteki ggplot2 (Sievert, 2019) (Lander, 2017).



Wykres 20: Histogram ocen dla wszystkich leków
Źródło: Opracowanie własne

Według wykresu 20, najwięcej jest skrajnych ocen, czyli 1 oraz 9 i 10. Ta ostatnia ocena pojawia się najczęściej co oznacza, że większość opinii w zbiorze jest bardzo pozytywna i rekomendująca dany lek. W następnym kroku zbadano jak często oceniano leki w poszczególnych latach (2008-2017).



Wykres 21: Ilość komentarzy dla poszczególnych lat

Źródło: Opracowanie własne

Według wykresu 21 najwięcej komentarzy zebrano w roku 2016 – ponad 20 procent z całego zbioru. Oprócz tego dużo komentarzy napisano też w latach 2015 i 2017 (ponad 15%), w pozostałych latach wyniki już były poniżej 10%.

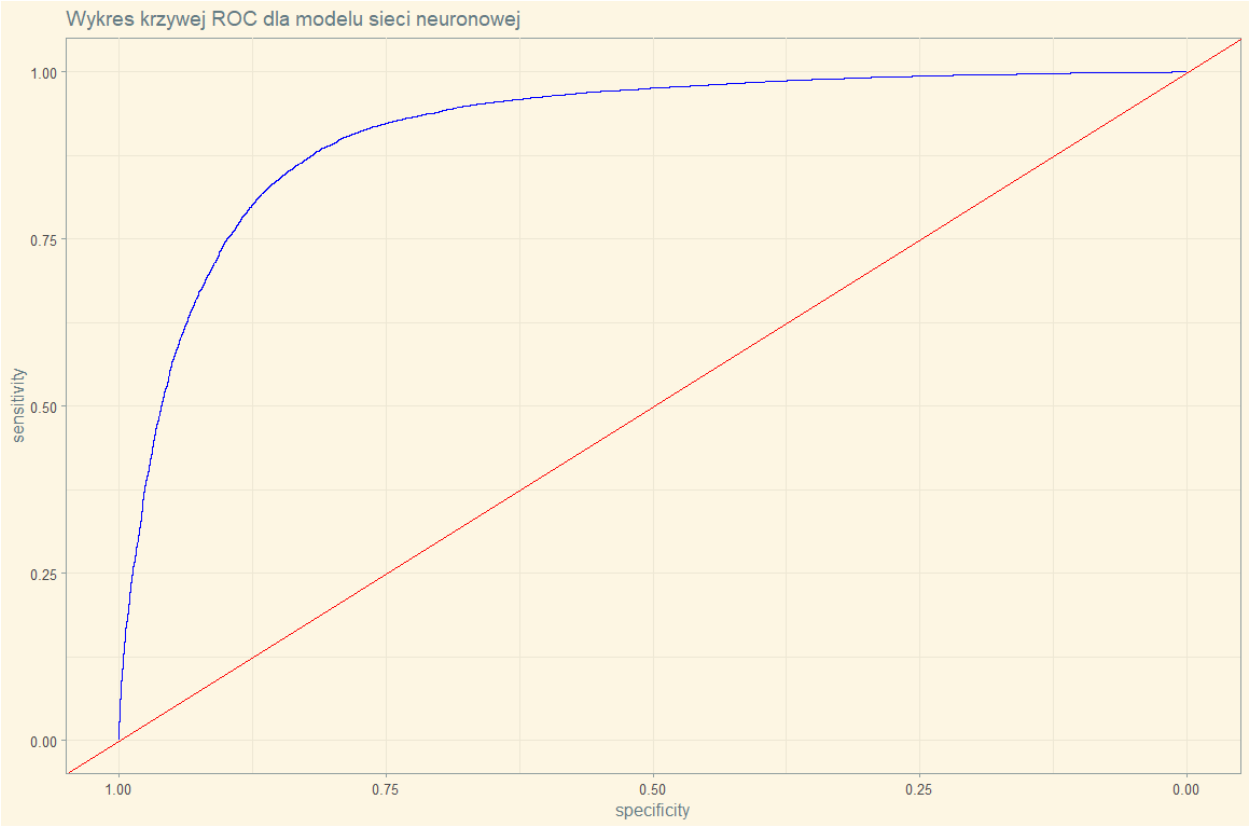
W kolejnym kroku zbudowano interaktywny dashboard za pomocą biblioteki shiny. Na wykresie 22 przedstawiono interaktywne wskaźniki KPI dla wybranej kategorii schorzeń czy dolegliwości. Ponadto w sekcji przedstawiono również wykres słupkowy dla najwyżej ocenianych leków w wybranej kategorii (Sievert, 2019).

Wykres 22 pokazuje, że leki na chorobę lokomocyjną były oceniane bardzo wysoko, średnia ocena wynosi aż ponad 8 przy ponad 200 komentarzach. Ponadto, komentarze te zebrały łącznie prawie 4000 poleceń co wskazuje, że wiele osób poleca te środki. Do najwyżej ocenianych leków na tą dolegliwość należą Cyclizine, Marezine, Travel-Eze czy Dramamine: wszystkie z tych 4 leków

Wykres 22: Kluczowe informacje nt. komentarzy o lekach na chorobę lokomocyjną
Źródło: Opracowanie własne

zbierały średnią ocenę ponad 9,5.

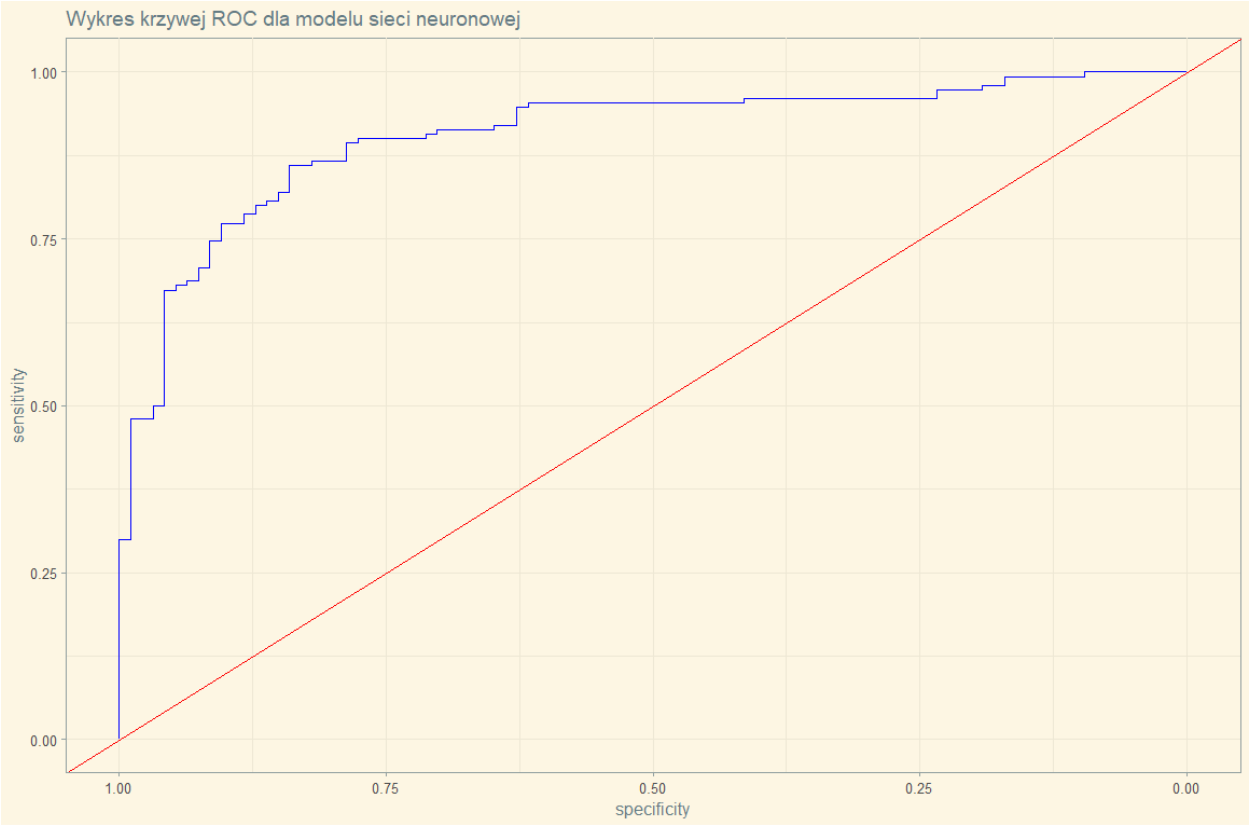
W następnej kolejności zaprezentowano wyniki klasyfikacji modelu, gdzie model sieci w zależności od podanego tekstu dokonywał weryfikacji czy opinia jest pozytywna bądź negatywna. Do zbadania dopasowania modelu do danych, wykorzystano bibliotekę pROC i caret.



Wykres 23: Krzywa ROC modelu sieci neuronowej klasyfikującej komentarze – ujęcie ogólne
Źródło: Opracowanie własne

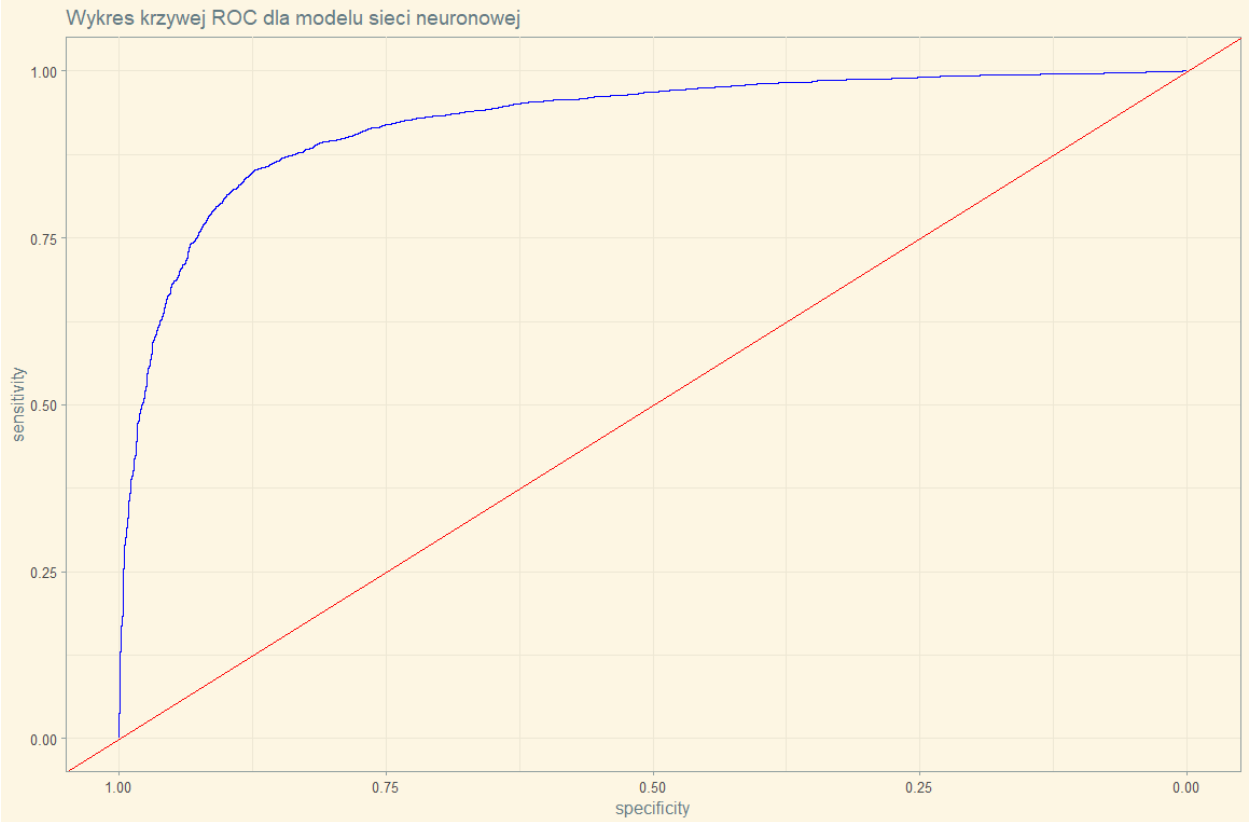
Pole pod krzywą ROC, która została pokazana na wykresie 10, wyniosło 0,91. Czułość wyniosła 0,78, swoistość 0,91, a precyzja wyniosła 0,89. Przełożyło się to na skuteczność modelu na poziomie 87% (Wickham, 2017) oraz na wynik wskaźnika F1 na poziomie 0,83. Sprawdzono również dopasowanie modelu na konkretnych grupach leków jak np. leki przeciw różnym

odmianom raka.



Wykres 24: Krzywa ROC modelu sieci neuronowej klasyfikującej komentarze - leki na raka
Źródło: Opracowanie własne

Pole pod przedstawioną na wykresie 24 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. Inna specyficzną grupą leków na której zbadano dopasowanie modelu były środki antykoncepcyjne (ang. „Birth control").

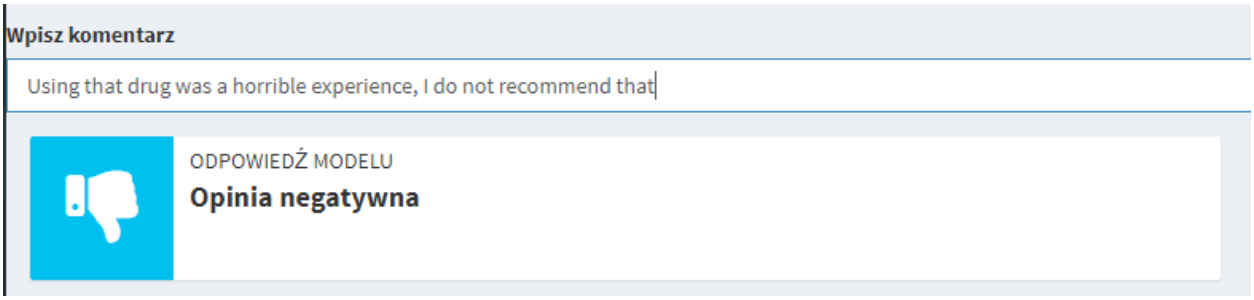


Wykres 25:Krzywa ROC modelu sieci neuronowej klasyfikującej komentarze

Źródło: Opracowanie własne

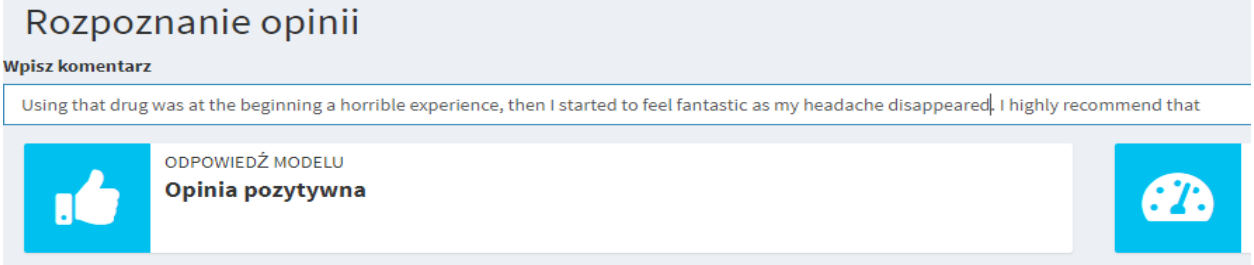
Pole pod krzywą ROC wyniosło 0,925. Skuteczność sieci okazała się mieć wynik 0,87, w porównaniu do ujęcia ogólnego specyficzność okazała się niższa, bo wyniosła 0,83. Czułość za to wzrosła do 0,88, precyzja osiągnęła z kolei poziom 0,88, a wskaźnik F1 podobnie jak precyzja wyniósł 0,88. Oznacza to, że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

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



Wykres 26: Odpowiedź modelu na prosty komentarz

Źródło: Opracowanie własne



Wykres 27: 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 12 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.

3.5 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 ggplot2 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 wykesów)
- 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)

3.6 Kod źródłowy

Pełny kod źródłowy jest dostępny pod linkiem:

https://github.com/HomeSeeker88/MastersThesis

Cały projekt można sklonować przy pomocy programu git lub ściągnąć bezpośrednio z linku.

4. Podsumowanie wyników pracy

W pierwszej części pracy zostały wytłumaczone teoretyczne podstawy narzędzi oraz metod użytych w systemie rekomendacyjnym leków.

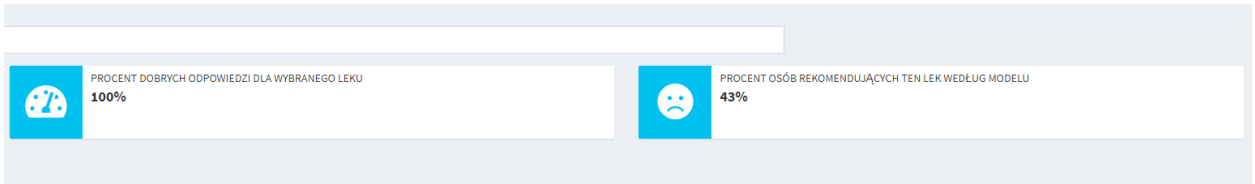
Na podstawie wyników dotyczących badań wyrażeń kluczowych, tf-idf czy analizy eksploracyjnej zbudowano system rekomendacyjny analizy sentymentu oparty o model sieci neuronowej. W systemie tym można zbadać komentarze na temat każdego leku, który został oceniony w zbiorze, odnaleźć najlepiej oceniane leki 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 na dany lek. W sekcji poświęconej modelowi można też zbadać jak poszczególne komentarze zostały rozpoznane przez model.

	uniqueID	review	rating	modelResponse	usefulCount	Correct
1	101140	"Have been on Actos for almost a year, gained 24 pounds and have swelling in hands and feet and am retaining a lot of water in my thighs. My sugar levels are good, my doctor lowered my dosage from 30 mg to 15 mg but refused to take me off. Will get a second opinion because the side effects are too much."	4	Opinia negatywna	42	true
2	101142	"Actos has worked well for me for 10 years and now comes FDA with another cancer scare. Anything can induce cancer and the scientific data on Actos causing cancer does not have that much credibility with me."	10	Opinia pozytywna	50	true
3	101139	"I have been taking Actos for 5 years. Recently I'm diagnosed me with heart failure. My EF is 30%. 2 weeks ago I stopped taking Actos and my fasting blood test is 223! The same reading as when I was taking Actos."	2	Opinia negatywna	23	true
4	101133	"Worked well at first, from 1 get 200 plus readings when I wake up."	3	Opinia negatywna	6	true
5	101105	"My blood sugar went with Actos is higher than without taking it."	1	Opinia negatywna	28	true
6	101141	"I have been taking Actos 15mg for about a month now, no ill effects like I feared. I keep watch for edema and so far nothing. My insurance would not cover the Januvia I used to be on- and for me Actos works better. My numbers are normal now (80-140) and they were high normal on the Januvia. I would recommend this medicine to people who cannot afford the DDP-4 (Januvia class) drugs."	9	Opinia pozytywna	41	true
7	101137	"Actos 15mg Metformin 850 mg works for me for almost 3 months now. Before I have fbs of 185 and 225 after meal. Now, almost everyday I prick, I got around 90 and 100 before and after meal. Better take these meds before breakfast."	10	Opinia pozytywna	13	true

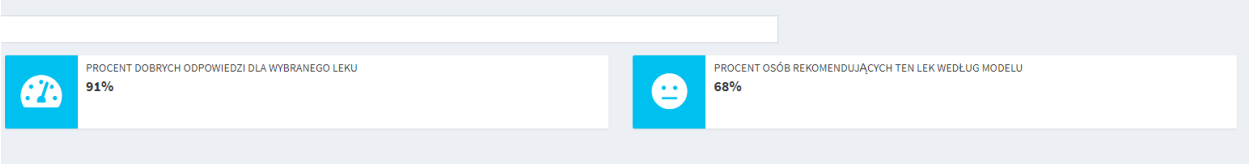
Showing 1 to 7 of 7 entries Previous 1 Next

Wykres 28: Tabela przedstawiające komentarze, ocenę oraz wynik modelu
Źródło: Opracowanie własne

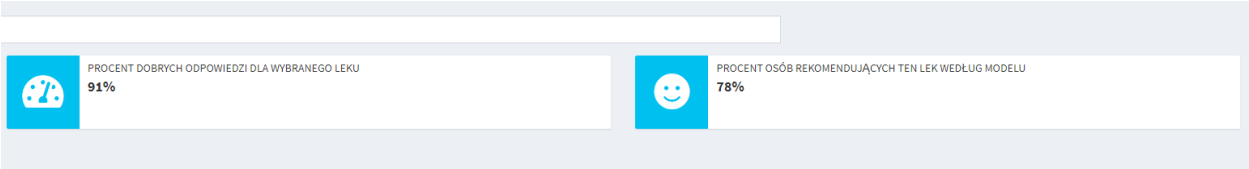
Na wykresie 28 na zielono zostały zaznaczone komentarze ocenione przez model jako opinie pozytywne, komentarze krytyczne zostały zaś wyświetlone w kolorze czerwonym. Ponadto, użytkownik może sprawdzić ile procent pacjentów rekomenduje dany lek, w zależności od ilości zadowolonych pacjentów, ikona przy danych procentowych zmienia się. Dla progu od 0 do 49% ikoną będzie smutny wyraz twarzy, od 50 do 74% twarz o neutralnym wyrazie, zaś od 75% ikoną będzie uśmiechnięta twarz.



Wykres 29: Lek oceniony negatywnie przez pacjentów
Źródło: Opracowanie własne



Wykres 30: Lek oceniony niejednoznacznie przez pacjentów
Źródło: Opracowanie własne



Wykres 31: Lek oceniony pozytywnie przez pacjentów
Źródło opracowanie własne

Na wykresach 29, 30 i 31 przedstawiono jak model ocenia komentarze na temat różnych leków, interfejs systemu natomiast informuje użytkownika czy dany lek jest polecany prez pacjentów: odpowiednio na wykresie 29 pokazane jest, że według modelu dany lek nie jest rekomendowany. Wykresie 30 wskazuje natomiast na różne opinie wśród pacjentów, gdzie większej ilości opinii pozytywnych, znalazło się ponad 30% komentarzy krytycznie oceniających dany lek. Wykres 31 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% co można uznać za wynik satysfakcjonujący, model największe trudności ma z weryfikacją zdań w których występują zaprzeczenia. Ponadto sprawdzono działanie modelu na komentarzach dotyczących działaniach różnych grupach leków w celu sprawdzenia czy model jest tak samo dopasowany to konkretnych grup opinii. System mógłby zostać rozszerzony o kolejne funkcjonalności:

- Rozszerzenie analizy komentarzy w innych językach europejskich takich jak 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 widgety oraz funkcjonalności, w celu poprawy doświadczenia użytkownika (czyli tzw. UX – ang. „User Experience”)

Taki bardziej rozbudowany system rekomendacji mógłby być wykorzystany przez np. sklepy

internetowe, apteki czy też mogłyby być używane przez firmy farmaceutyczne w celu zbadania satysfakcji klientów.

Bibliografia

1. Baheti, P. (2022). *12 Types of Neural Network Activation Functions: How to Choose?*
2. Brownlee, J. (2017). *How to Use Word Embedding Layers for Deep Learning with Keras.*
3. Brownlee, J. (2018). *A Gentle Introduction to Dropout for Regularizing Deep Neural Networks.*
4. Cherry, K. (2021). *The Role of Neurotransmitters.*
5. Chollet, F. (2018). *Deep Learing. Praca z językiem R i biblioteką Keras.*
6. Du, T. i Shanker, V. K. (brak daty). *Deep Learning for Natural Language Processing.*
7. Kallumadi, S. i Gräßer, F. (2018). *https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29.*
8. Kaneda, T. (2021). *How Many People Have Ever Lived on Earth?*
9. Kanna, C. (2021). *Word, Subword, and Character-Based Tokenization: Know the Difference.*
10. Keskar, N., Mudigere, D., Nocedal, J., Smelyanskiy, M. i Tang, P. (2017). *On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima.*
11. Kibble, R. (2013). *Introduction to natural language processing.* Goldsmith University of London.
12. Kohn, K. W. (2020). *Drugs Against Cancer: Stories of Discovery and the Quest for a Cure.*
13. Korbicz, J., Obuchowicz, A. i Uciński, D. (1994). *Sztuczne sieci neuronowe. Podstawy i zastosowania.*
14. Kunwar, S. (2013). *Text Documents Clustering using K-Means Algorithm.*
15. Lander, J. (2017). *R dla każdego.*
16. Loiseau, J.-C. B. (2019). *Rosenblatt's perceptron, the first modern neural network.*
17. Malik, F. (2020). *Sensitivity Vs Specificity In Data Science.*
18. Mamczur, M. (2021). *Jak działają konwolucyjne sieci neuronowe?*
19. Manning, C., Schütze, H. i Raghavan, P. (2009). *Introduction to information retrieval.*
20. McKinsey. (2016). *The CEO guide to customer experience.*
21. MonkeyLearn. (2022). *What is Text Mining?*
22. Ng, R. (2009). *Drugs: From Discovery to Approval, Second Edition.*
23. Pai, A. (2020). *What is Tokenization in NLP? Here's All You Need To Know.*
24. Pascual, F. (2019). *Guide to Aspect-Based Sentiment Analysis.*
25. Repustate. (2021). *10 Sentiment Analysis Data Sources For Strategic Data Analytics.*
26. Repustate. (2021). *Why Should We Use Sentiment Analysis In Social Media Mining?*
27. Roldós, I. (2020). *5 Sentiment Anlysis Examples in Business.*
28. Roldós, I. (2020). *What is Sentiment Analysis?*
29. sciencemuseum.org.uk. (2019). *Thalidomide.*
30. Seth, N. (2021). *Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim and Sklearn.*
31. Sievert, C. (2019). *Interactive Web-Based Data Visualization with R, plotly and shiny.*
32. Silge, J. i Robinson, D. (2017). *Text mining with R - A tidy approach.*
33. Stedman, C. (2020). *Text mining (text analytics).*
34. Technical University of Denmark. (2011). *http://www2.imm.dtu.dk/pubdb/pubs/6010-full.html.*
35. Verma, Y. (2021). *A Guide to Term-Document Matrix with Its Implementation in R and Python.*
36. Wickham, H. (2017). *R for Data Science.*

37. Wonderflow. (2018). *10 Sentiment Analysis Examples That Will Help Improve Your Products.*

OŚWIADCZENIE AUTORA PRACY DYPLOMOWEJ 1 LICENCJACKIEJ/MAGISTERSKIEJ

pod tytułem .....
.....
napisanej przez: .....**nr albumu** .....
pod kierunkiem .....

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca dyplomowa nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy dyplomowej jest identyczna z załączoną wersją elektroniczną.

Wyrażam zgodę na poddanie pracy dyplomowej kontroli, w tym za pomocą programu wychwytującego znamiona pracy niesamodzielnej, zwanego dalej programem, oraz na umieszczenie tekstu pracy dyplomowej w bazie porównawczej programu, w celu chronienia go przed nieuprawnionym wykorzystaniem, a także przekazanie pracy do Ogólnopolskiego Repozytorium Prac Dyplomowych.

Wyrażam także zgodę na przetwarzanie przez Szkołę Główną Handlową w Warszawie moich danych osobowych umieszczonych w pracy dyplomowej w zakresie niezbędnym do jej kontroli za pomocą programu oraz w zakresie niezbędnym do jej archiwizacji i nieodpłatnego udostępniania na zasadach określonych w zarządzeniu.

……. ……..
(data) (podpis autora)

1
Zastosować właściwe.