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

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 (Kallumadi i Gräßer, 2018):

3.2 Analiza sentymentu oraz podziału na n-gramy przy pomocy pakietu tidytext

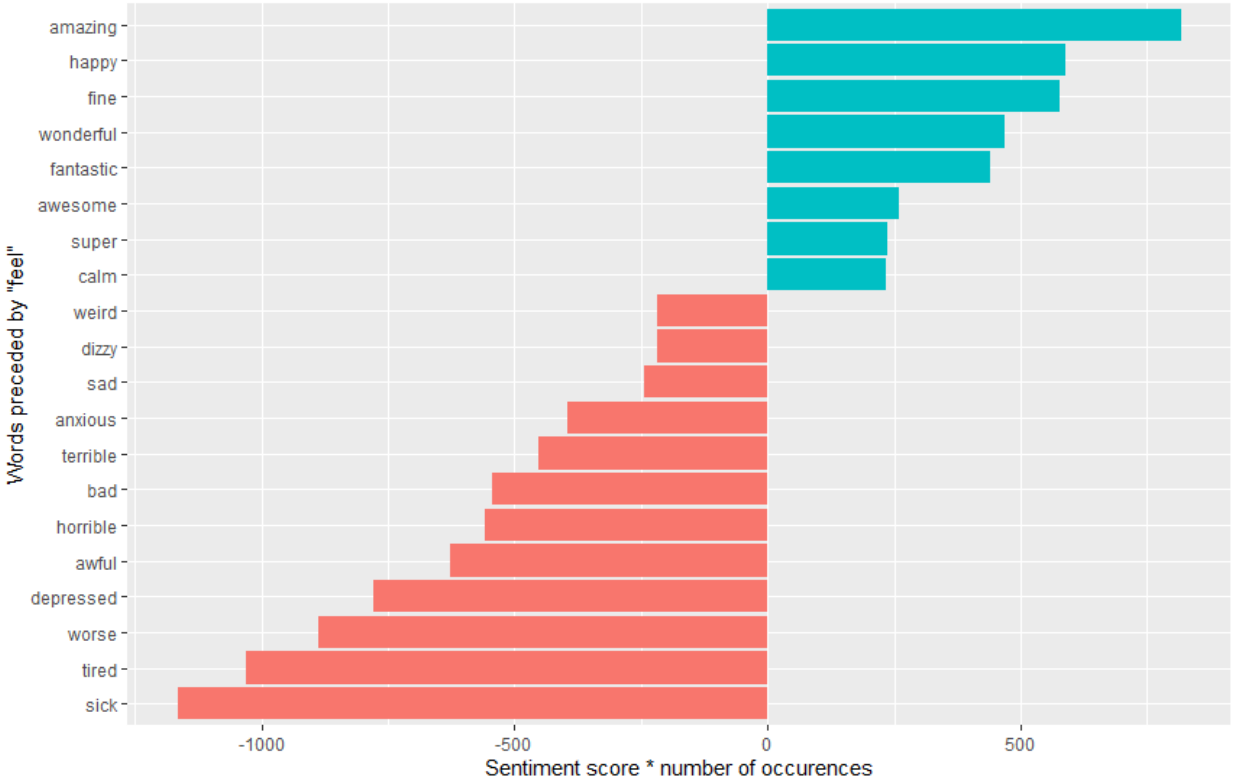
Do analizy sentymentu użyto biblioteki tidytext, aby pozbyć się niepotrzebnych słów tzw. „stop words". Są to słowa o małym znaczeniu takie jak np. spójniki („ponieważ", „oraz", „bo") czy też słowa popularne („mp3", „pc"). Słowa te nie wpływają na identyfikację tekstu dlatego też usuwa się je w celu zredukowania 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):

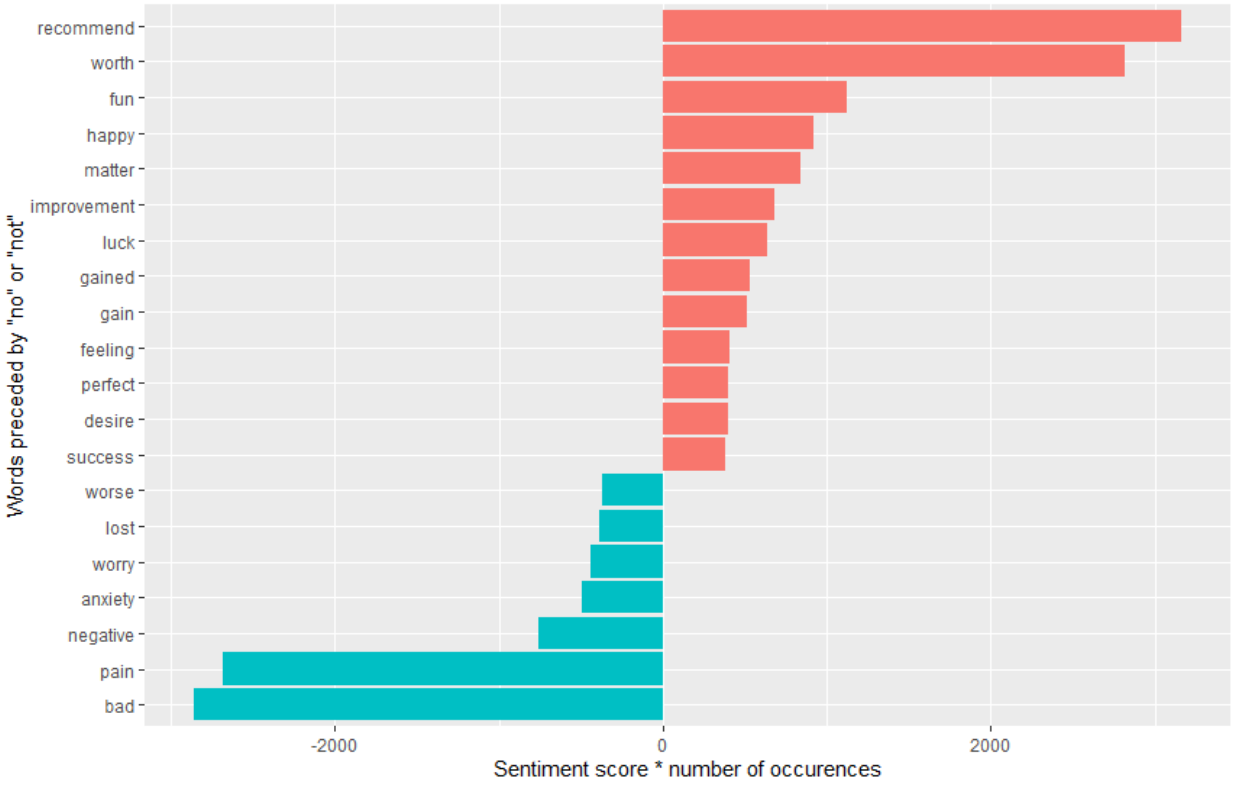
$$Kontrybucja = Wartość\ sentymentu\ słowa * ilość\ wystąpień$$

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 1: Słowa które najczęściej były poprzedzone słowem „feel" w różnej odmianie
Źródło: Opracowanie własne



Wykres 2: 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 3: 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:

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

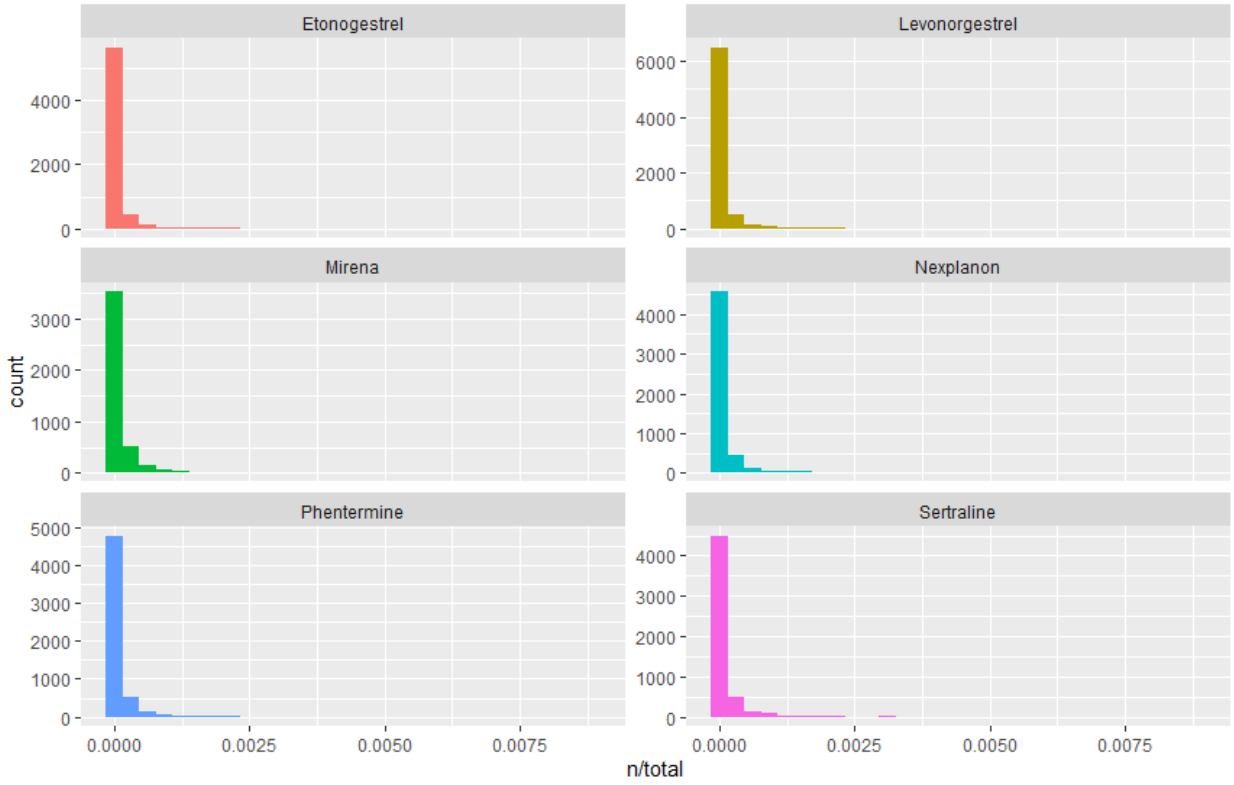
Gdzie:

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

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

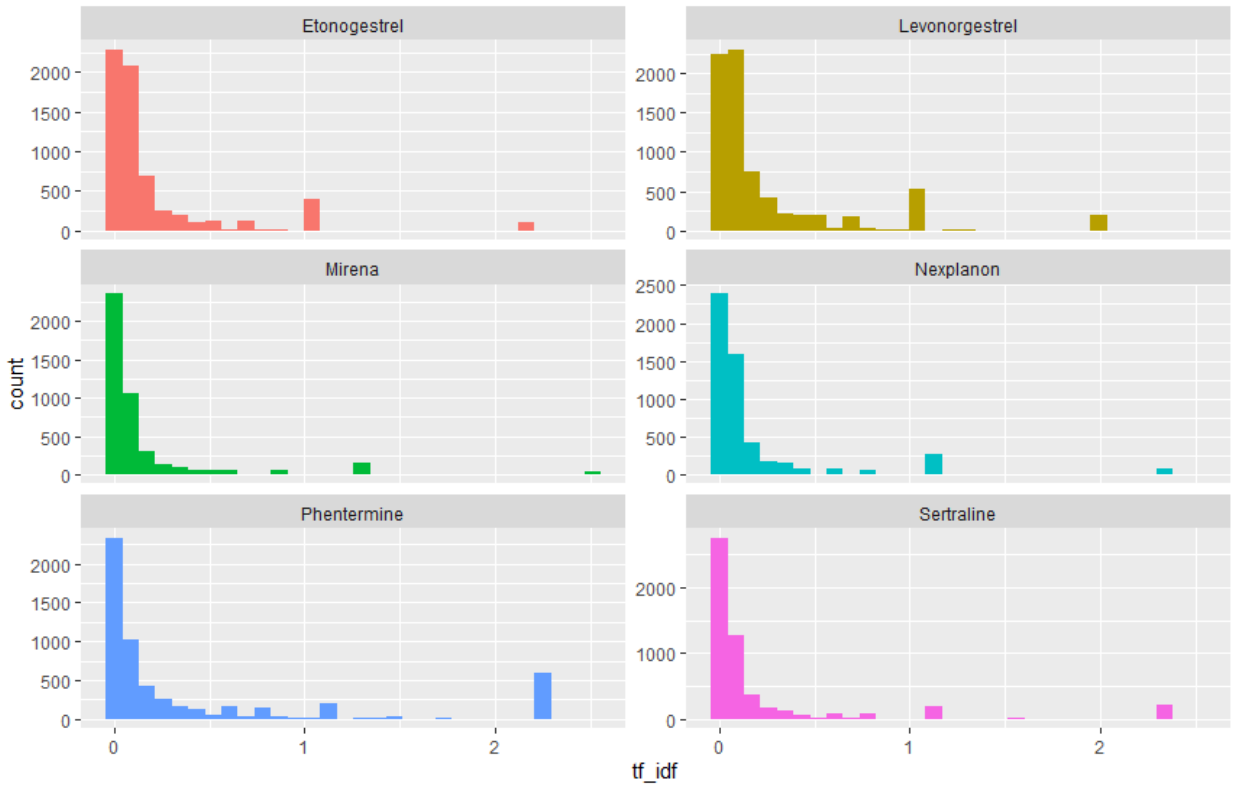
$|D|$ - oznacza liczbę dokumentów w korpusie, a $|\{d: t_i \in d\}$ oznacza liczbę dokumentów zawierających przynajmniej jedno wystąpienie i-tego termu (Silge i Robinson, 2017).

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.



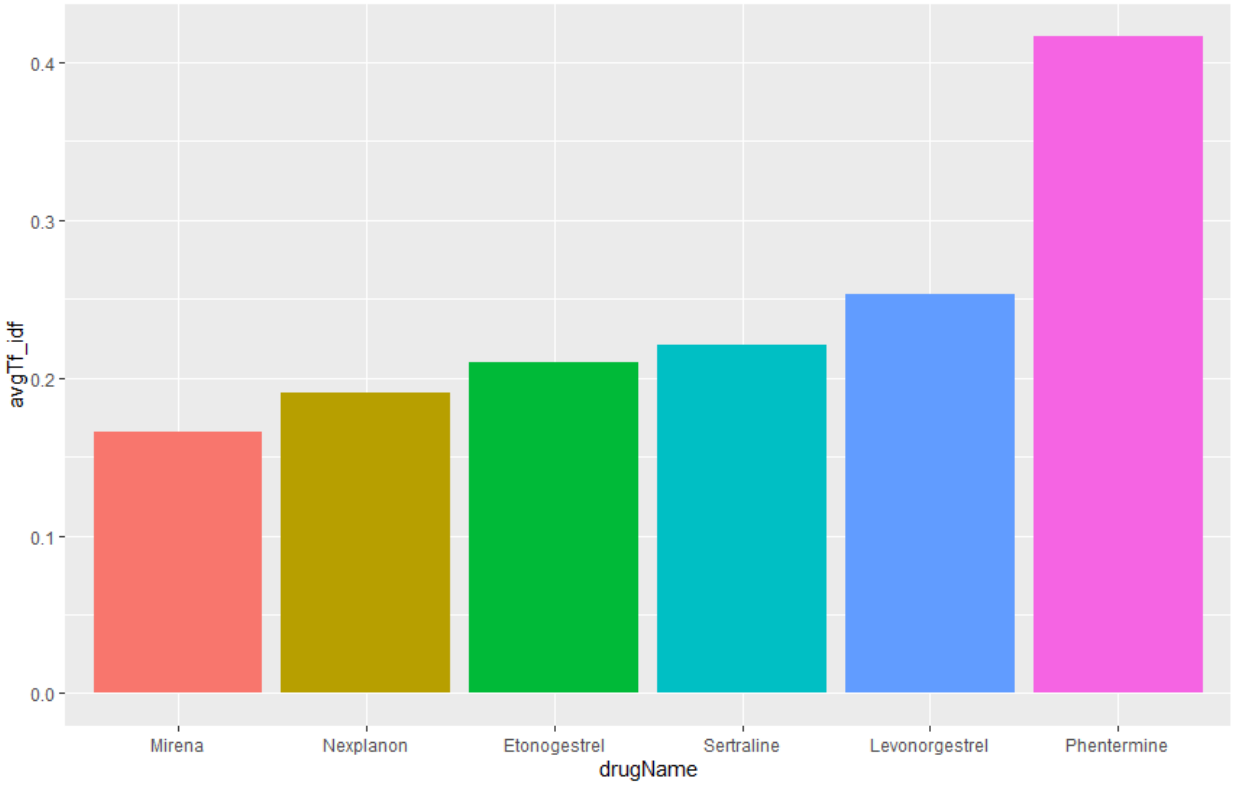
Wykres 4 Histogram współczynnika częstotliwości występowania słów(tf) w komentarzach dla 6 najczęściej komentowanych leków
Źródło: Opracowanie własne

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 5: Histogram tf-idf termów dla 6 najczęściej ocenianych leków

Źródło: Opracowanie własne



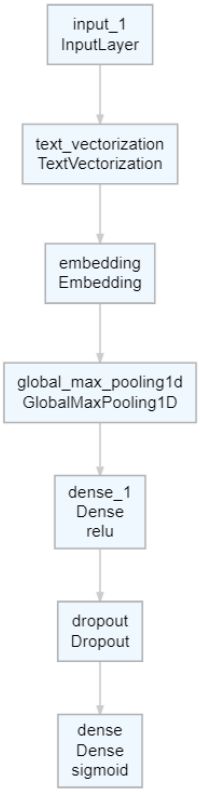
Wykres 6: Średni tf_idf dla 6 najczęściej ocenianych leków

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 Wykorzystanie sieci neuronowej do oszacowania oceny leku na podstawie komentarzy przy użyciu biblioteki keras

Po dokonaniu analizy sentymentu, 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. 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.

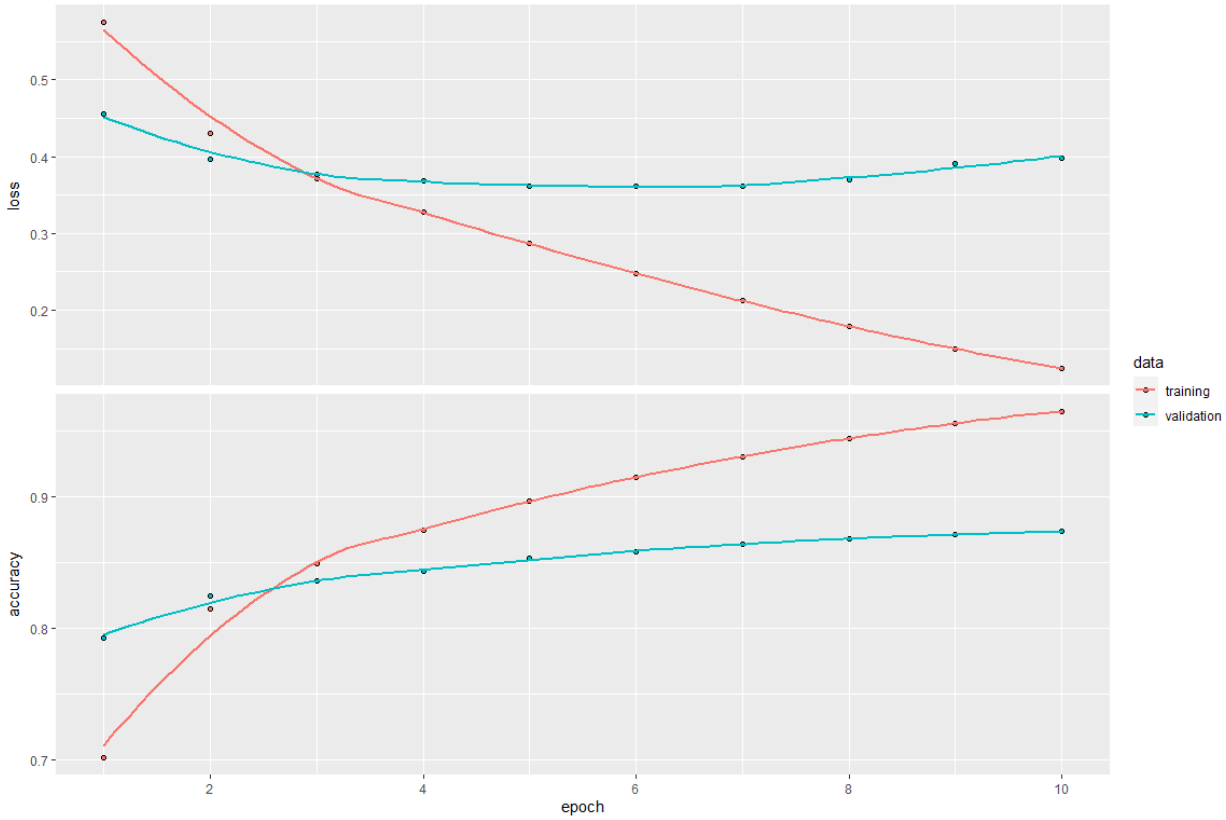


Wykres 7: Architektura modelu sieci neuronowej
Źródło: Opracowanie własne

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, metoda powszechnie stosowana w NLP. Każde słowo jest konwertowane na wektor określonej długości. 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 jest tutaj wskazany, gdyż chcemy się skupić na wyróżniających się wartościach w macierzach 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, a na samym końcu znajduje się 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 rozmiar batchowania 512.

Wyniki uczenia w postaci wykresów skuteczności modelu oraz wartości funkcji straty na zbiorze treningowym i walidacyjnym przedstawiona na wykresie 6.

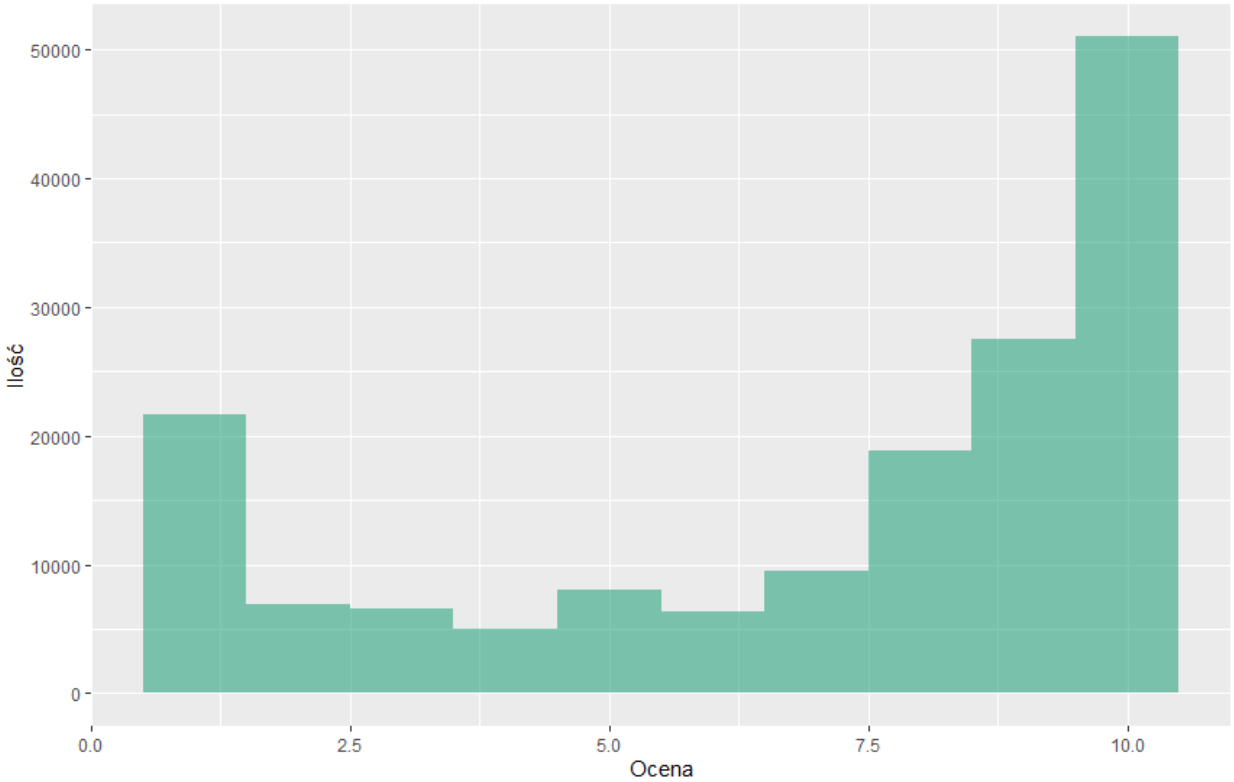


Wykres 8: Skuteczność i wartość funkcji straty modelu sieci neuronowej na zbiorze treningowym i walidacyjnym
Źródło: Opracowanie własne

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

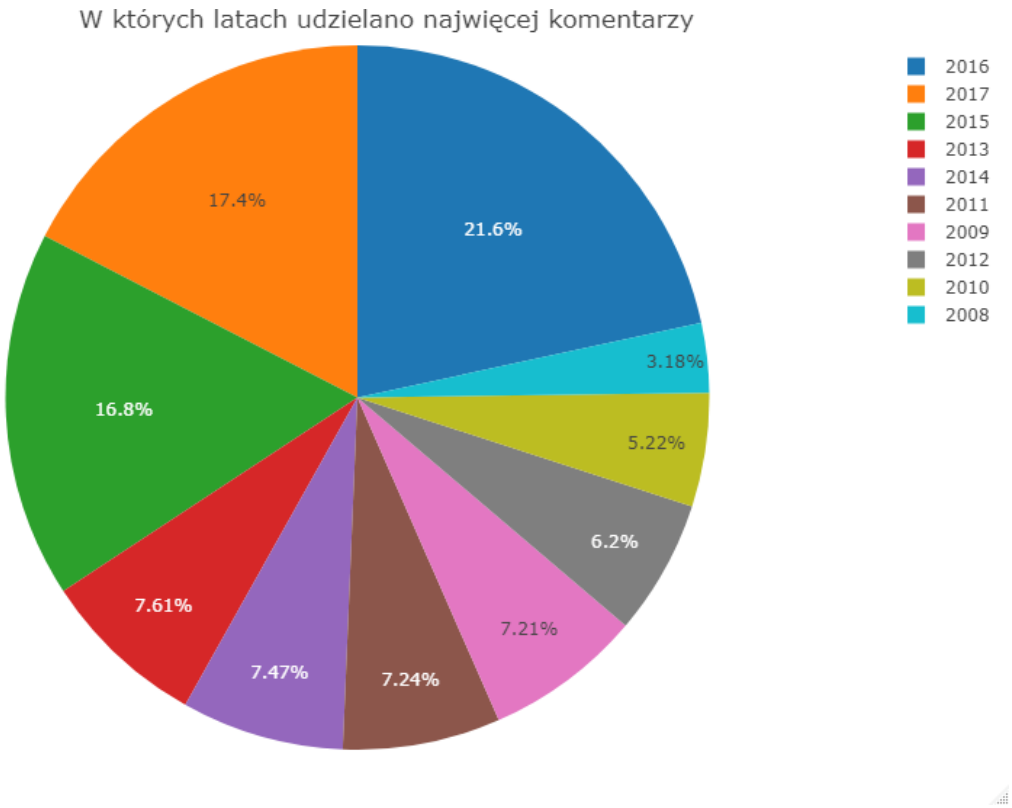
Do analizy eksploracyjnej użyto bibliotek graficznych takie jak: ggplot2 oraz plotly. Na początku zbadano histogram ocen leków. Na wykresie 5 przedstawiono jego wykres, który został stworzony przy pomocy biblioteki ggplot2 (Sievert, 2019) (Lander, 2017).



Wykres 9: Histogram ocen dla wszystkich leków

Źródło: Opracowanie własne

Według wykresu 7, 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 10: Ilość komentarzy dla poszczególnych lat

Źródło: Opracowanie własne

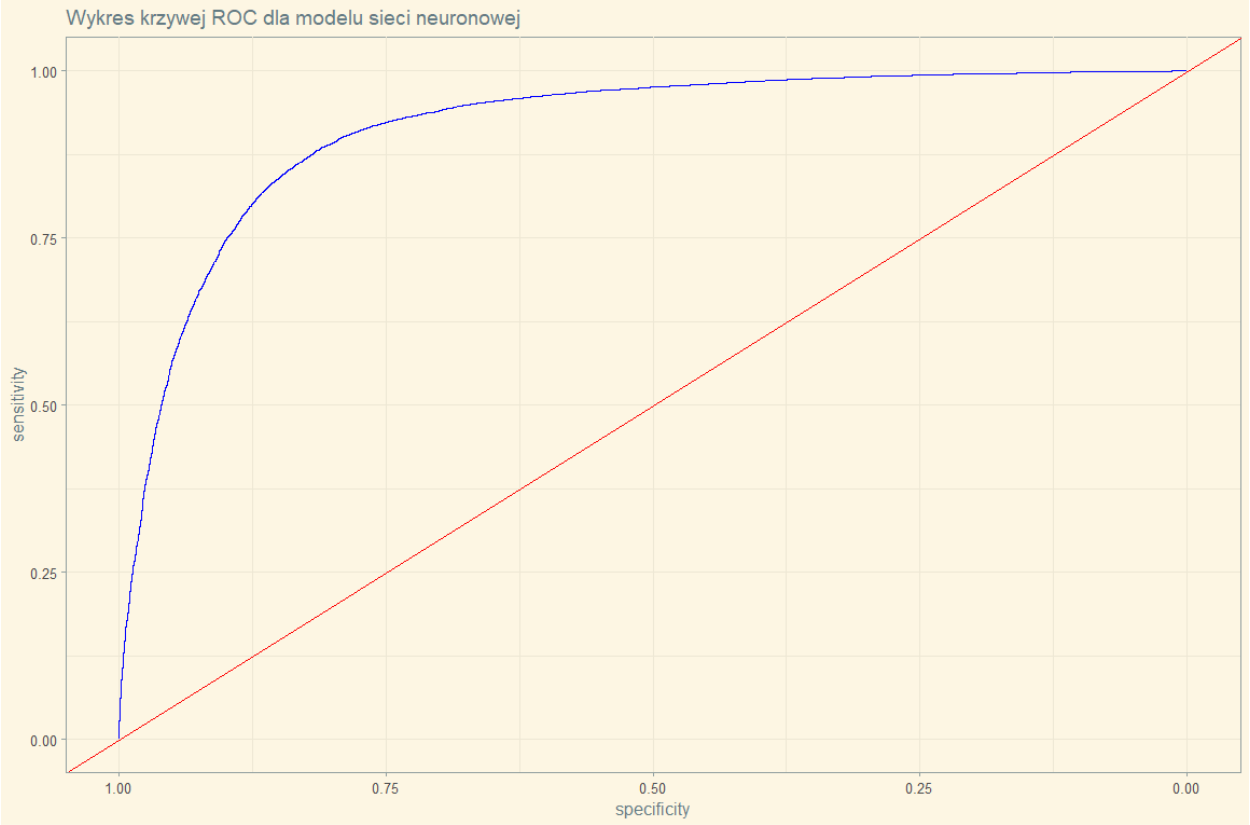
Według wykresu 8 najwięcej komentarzy zebrano w roku 2016 – ponad 20 procent z całego zbioru. Oprócz tego dużo komentarzy napisano też w latach 2015 i 2017 (ponad 15%), w pozostałych latach wyniki już były poniżej 10%.

W kolejnym kroku zbudowano interaktywny dashboard za pomocą biblioteki shiny. Na wykresie 9 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 11: Kluczowe informacje nt. komentarzy o lekach na chorobę lokomocyjną
Źródło: Opracowanie własne

Wykres 9 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 polubień 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 zbierały średnią ocenę ponad 9,5.

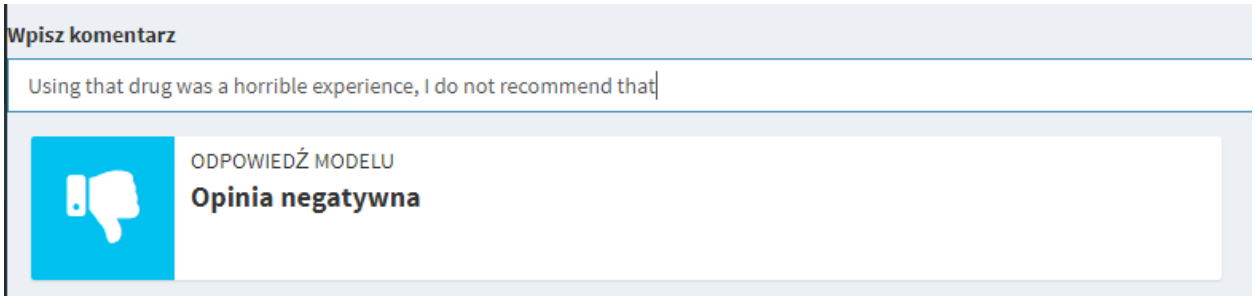
W następnej kolejności zaprezentowano wyniki klasyfikacji modelu, gdzie model sieci w zależności od podanego tekstu dokonywał weryfikacji czy opinia jest pozytywna bądź negatywna. Do zbadania dopasowania modelu do danych, wykorzystano bibliotekę pROC i caret.



Wykres 12: Krzywa ROC modelu sieci neuronowej klasyfikującej komentarze

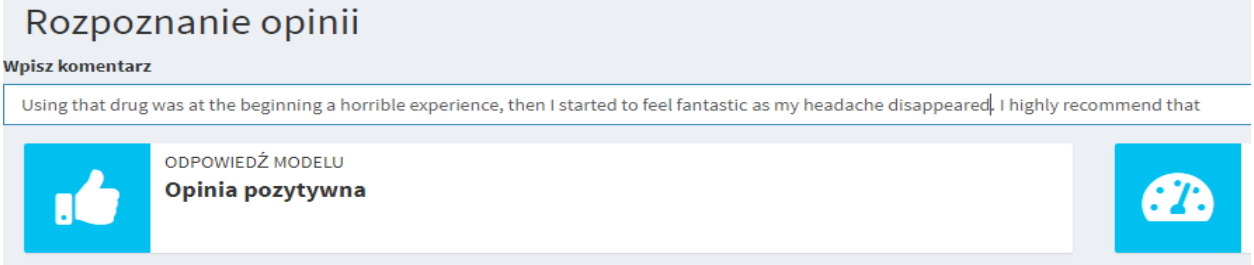
Źródło: Opracowanie własne

Pole pod krzywą ROC, która została pokazana na wykresie 10, wyniosło 0,91. Czułość wyniosła 0,88 a swoistość 0,81. Przełożyło się to na skuteczność modelu na poziomie 86% (Wickham, 2017). W ostatnim kroku spadano odpowiedź modelu na wprowadzony przez użytkownika tekst.



Wykres 13: Odpowiedź modelu na prosty komentarz

Źródło: Opracowanie własne



Wykres 14: 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 frameworki

Do stworzenia dashboardów, 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 deep learningu 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 dashboardów)
- shinydashboard (rozszerzenie frameworka 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 we frameworku 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.

5. Bibliografia

1. Chollet, F. (2018). *Deep Learing. Praca z językiem R i biblioteką Keras.*
2. Du, T. i Shanker, V. K. (brak daty). *Deep Learning for Natural Language Processing.*
3. Kallumadi, S. i Gräßer, F. (2018). *https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29.*
4. Lander, J. (2017). *R dla każdego.*
5. Sievert, C. (2019). *Interactive Web-Based Data Visualization with R, plotly and shiny.*
6. Silge, J. i Robinson, D. (2017). *Text mining with R - A tidy approach.*
7. Technical University of Denmark. (2011). *http://www2.imm.dtu.dk/pubdb/pubs/6010-full.html.*
8. Wickham, H. (2017). *R for Data Science.*

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.