

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…

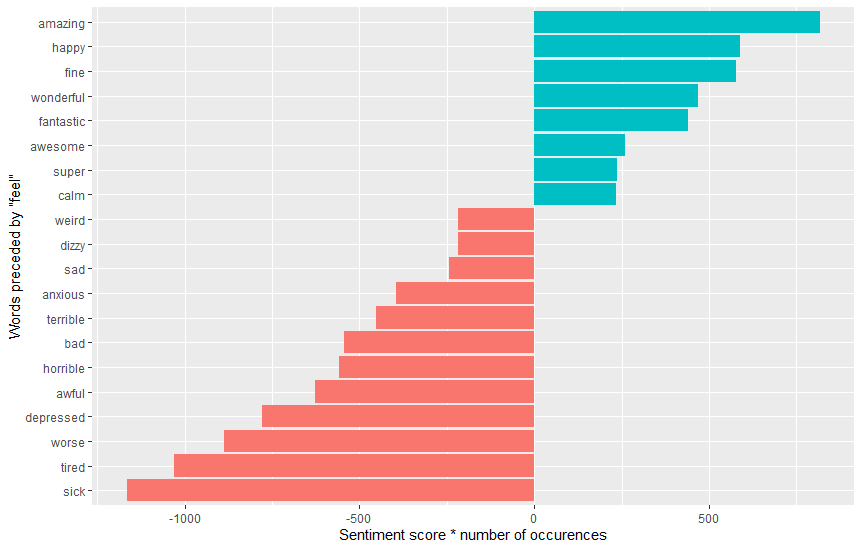
\*Zastosować właściwe

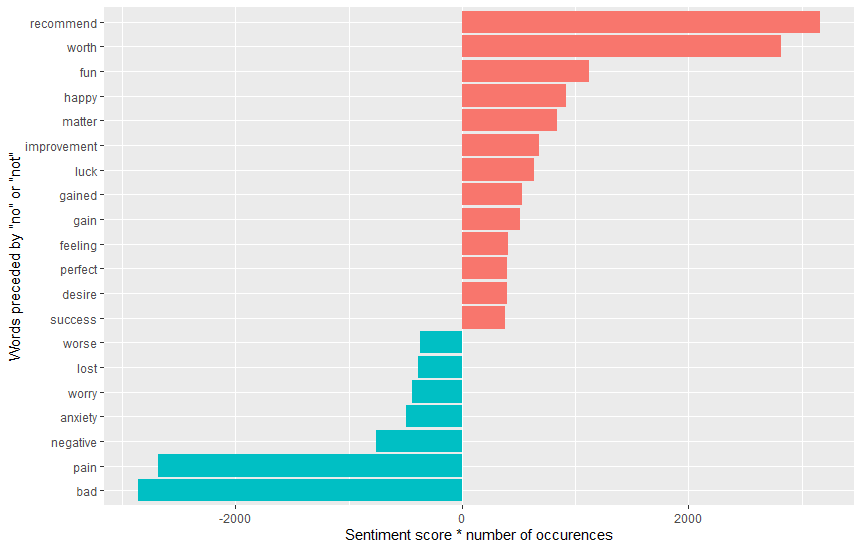
\*\* W przypadku braku specjalności lub braku deklaracji o specjalności wiersz należy pominąć

**3. Część implementacyjna**

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

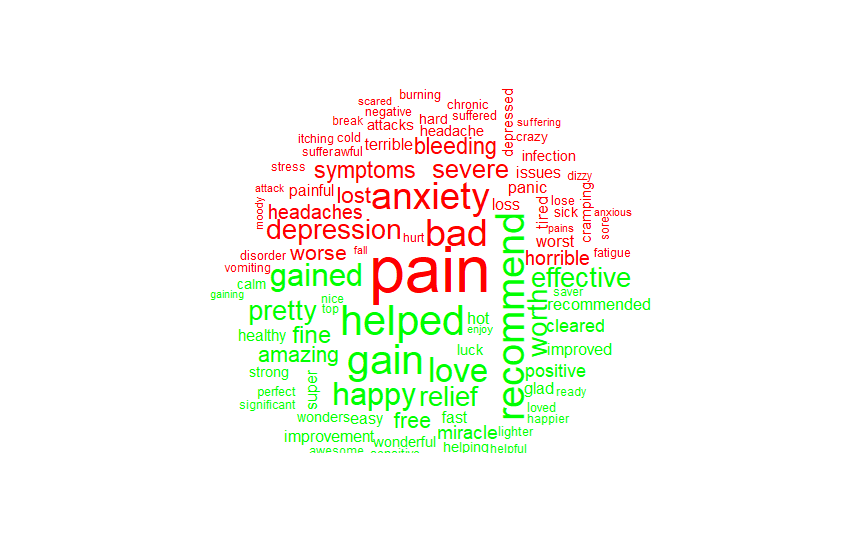
Do analizy sentymentu użyto biblioteki tidytext, aby pozbyć się niepotrzebnych słów. Następnie z leksykonu AFINN, pobrany został zbiór scoringu poszczególnych słów. Słowa w nim o negatywnym wydźwięku przyjmują ujemne wartości, zaś te o pozytywnym zwracają wartości dodatnie (np. wyrazy „amazing” czy „breathtaking” zwracają 5, natomiast wulgaryzmy przyjmują wartość -5). W dalszej kolejności obliczane są liczności n-gramów czyli występujących obok siebie *n* słów. Dla tego przypadku za *n* przyjęto 2. Po obliczeniu wystąpienia słów obliczona została kontrybucja każego słowa wzorem:

  
*Wykres 1: Słowa której najczęściej były poprzedzone słowem feel w różnej odmianie*

  
*Wykres 2: Słowa które najczęściej były poprzedzone słowem "no" lub "not" o najwyższym scoringu bezwzględnym*

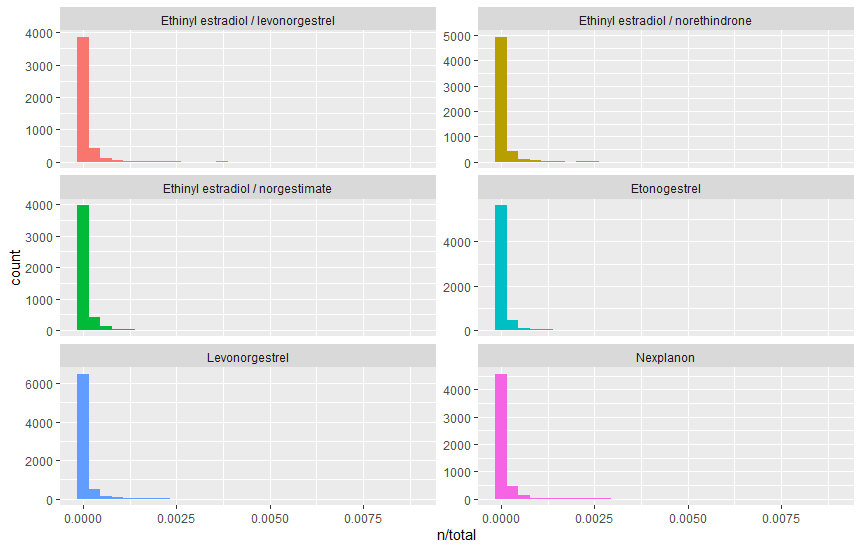
Wykres 1 pokazuje, że z takimi słowami jak „feel” czyli czuciem w kontekście pozytywnym są związane takie słowa jak „amazing”(niesamowicie), „happy”(szczęśliwy) czy „fine”(dobrze). Jeśli chodzi natomiast o słowa o negatywnym wydźwięku, które w dużym stopniu występowały ze słowem „feel” dominują typowe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

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*

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

Następnie zbadano tf\_idf

  
*Wykres 4: Tf\_idf w komentarzach dla 6 wybranych leków*

Wykres 4 pokazuje, że dla każdego z 6 wybranych leków 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.

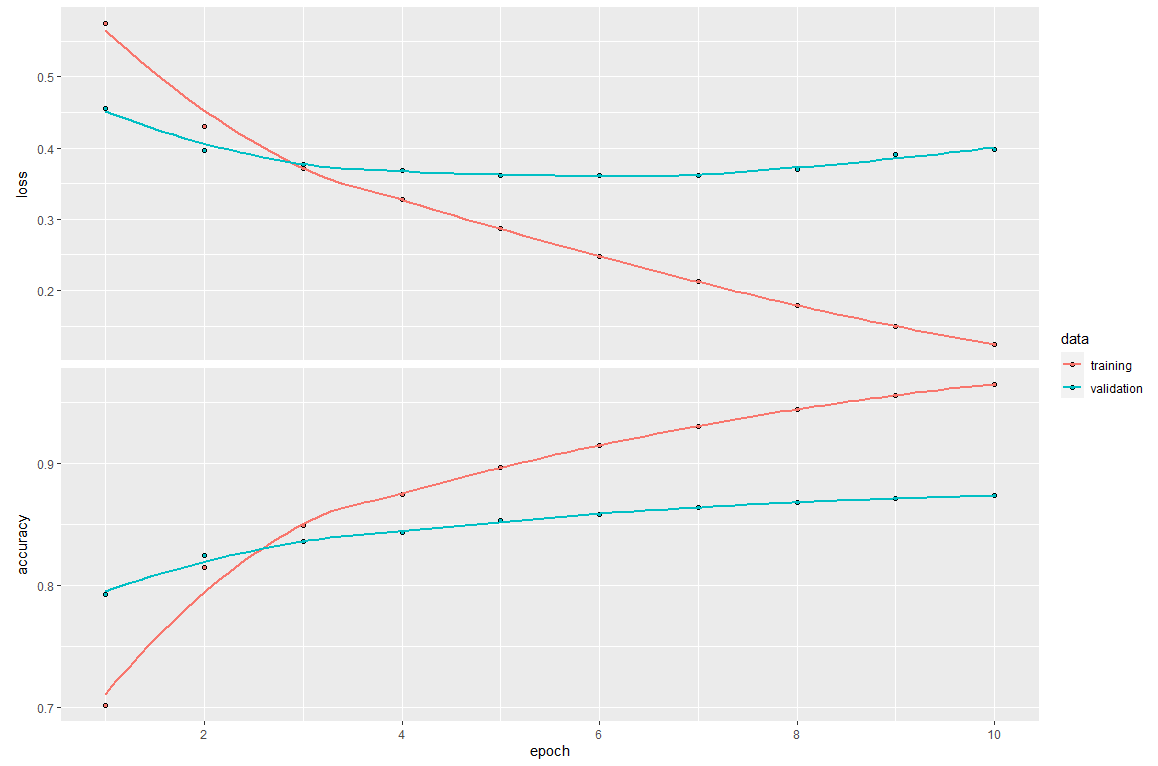
3.2 Wykorzystanie sieci neuronowej do oszacowania oceny leku na podstawie komentarzy przy użyciu biblioteki keras

Po dokonaniu analizy sentymentu przystąpiono do budowy modelu deep learningowego w celu zaklasyfikowania czy dany komentarz był opinią poztywną i 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 5: Architektura modelu sieci neuronowej*

W pierwszej kolejności stworzona jest warstwa wejściowa jednowymiarowa, następnie tworzona jest warstwa odpowiadająca za wektoryzację tekstu. 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. W kolejnej warstwa wbudowana po której jest warstwa jednowymiarowego max poolingu. Max pooling jest tutaj wskazany, gdyż chcemy się skupić na wyróżniających się wartościach w macierzach w przeciwieństwie do poolingu uśredniającego. 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 wyjściowa w której funkcją aktywacji jest funkcja sigmoidalna zwracająca prawdopodobieństwo, że komentarz jest pozytywny.

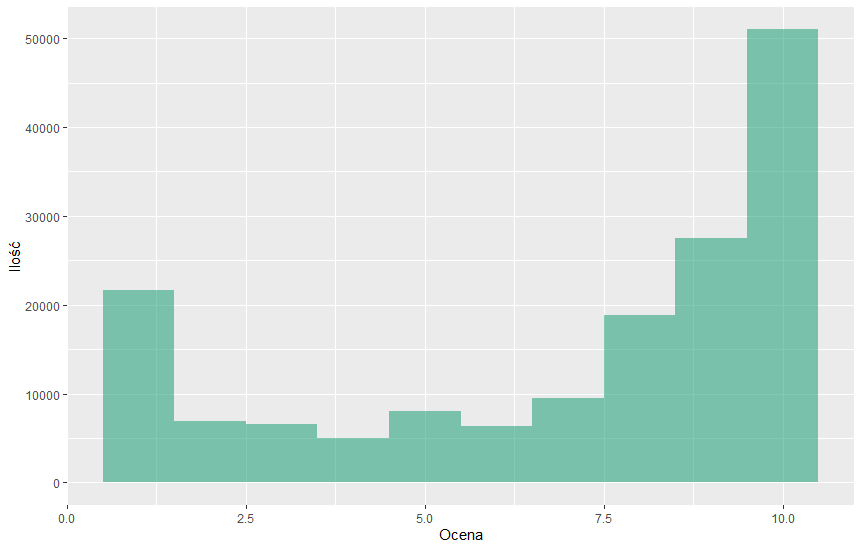
Po zbudowaniu sieci poddano ją uczeniu na zbiorze treningowym. Dla znaczącego przyspieszenia procesu uczenia skorzystano z wsparcia GPU w bibliotece tensorflow oraz biblioteki. Za ilość iteracji przejścia przez cały zbiór danych przyjęto 10, zaś za rozmiar batchowania 512. Wyniki uczenia w postaci wykresów skuteczności modelu oraz wartości funkcji straty na zbiorze treningowym i walidacyjnym przedstawiona na wykresie 6.

  
*Wykres 6: Skuteczność i wartość funkcji straty modelu sieci neuronowej na zbiorze treningowym i walidacyjnym*

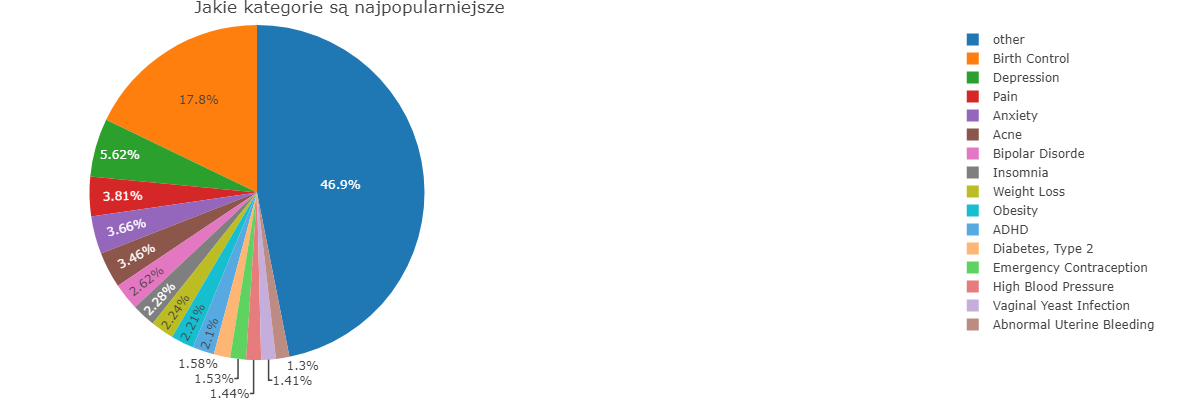
Wartość funkcji straty na zbiorze walidacyjnym wyniosła 0,4 w 10 epoce, zaś dla zbioru treningowego wyniosła 0,05. Z kolei precyzja modelu dla zbioru walidacyjnego wyniosła ok. 0,85 a dla treningowego ponad 0,95.

3.3 Przedstawienie wyników analizy eksploracyjnej oraz estymacji wyników dokonanych przez model

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

  
*Wykres 7: Histogram ocen dla wszystkich leków*

Według wykresu 5, 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.

  
*Wykres 8: Które kategorie był najczęściej oceniane*

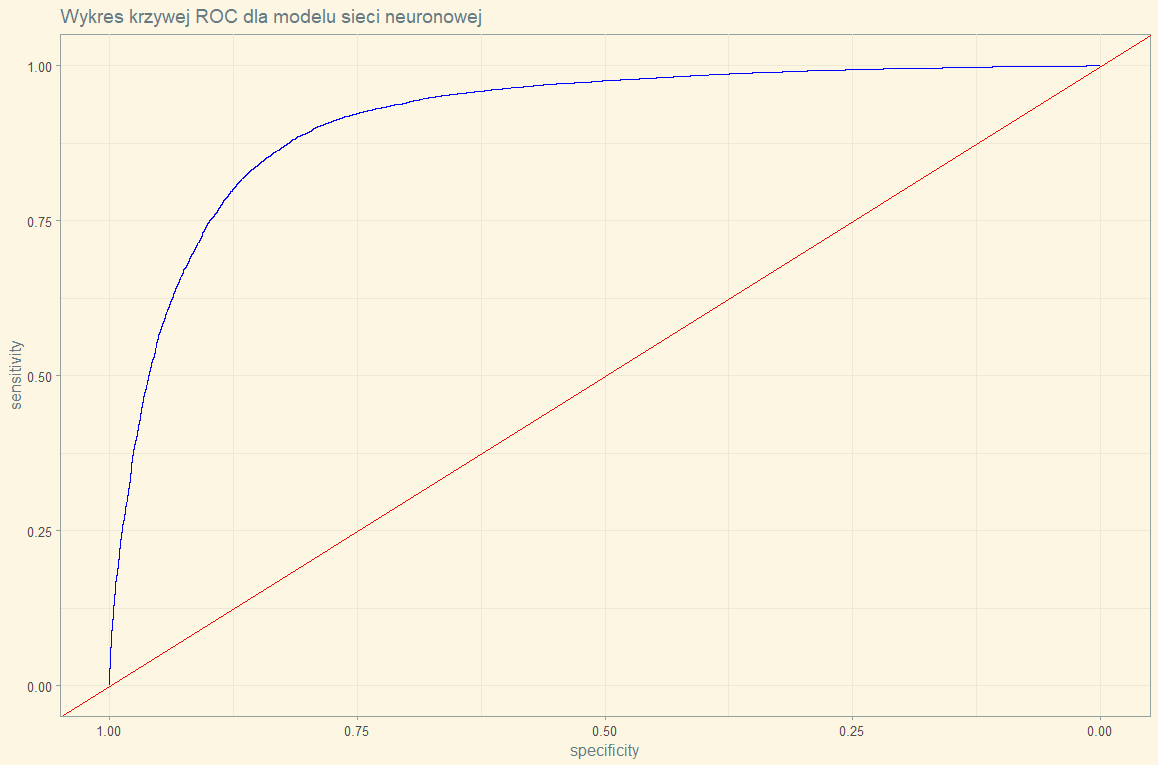
Według wykresu 8 najczęściej ocenianymi lekami są różnorakie środki antykoncepcyjne (17,8%), leki na depresję or środki przeciwbólowe i przeciwlękowe (odpowiednio 5,62%, 3,81% i 3,66%).

W kolejnym kroku zbudowano interaktywny dashboard we frameworku Shiny. Na wykresie 9 przedstawiono interaktywne wskaźniki KPI dla wybranej kategorii schorzeń czy dolegliwości. Ponadto, w tabie przedstawiono również wykres słupkowy dla najwyżej ocenianych leków w wybranej kategorii.

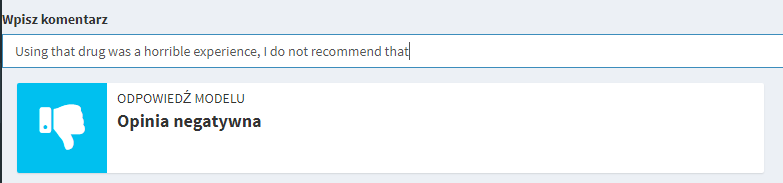
Wykres 9 pokazuje, że leki na chorobę lokomocyjną były oceniane bardzo wysoko, średnia ocena wynosi aż ponad 8 przy ponad 200 komentarzach. Ponadto, komentarze te zebrały łącznie prawie 4000 polubień co wskazuje, że wiele osób poleca te środki. Do najwyżej ocenianych leków na tą dolegliwość należą Cyclizine, Marezine, Travel-Eze czy Dramamine: wszystkie z tych 4 leków zbierały ocenę średnią ponad 9,5.

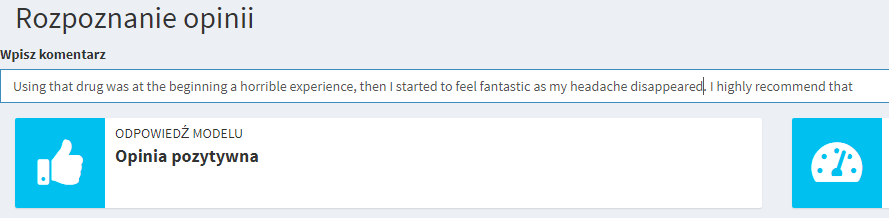
  
*Wykres 9: Kluczowe informacje nt. komentarzy o lekach na chorobę lokomocyjną*

W następnej kolejności zaprezentowano wyniki klasyfikacji modelu, gdzie sieć model sieci w zależności na podany tekst dokonywał weryfikacji czy opinia jest pozytywna bądź negatywna.

 *Wykres 10: Krzywa ROC modelu sieci neuronowej klasyfikującej komentarze*

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

  
*Wykres 11: Odpowiedź modelu na prosty komentarz*

  
*Wykres 12: Odpowiedź modelu na bardziej rozbudowany komentarz*

Na początku sprawdzono odpowiedź modelu na prosty komentarz, który w dość prosty 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.

3.4 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.

**Oświadczenie autora pracy dyplomowej**

**licencjackiej/magisterskiej1**

**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.