



POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI
Instytut Informatyki

Praca dyplomowa magisterska

ROZPOZNAWANIE EMOCJI W MEDIACH SPOŁECZNOŚCIOWYCH Z WYKORZYSTANIEM GŁĘBOKICH SIECI NEURONOWYCH

inż. Jakub Zdanowski, 127239

Promotor

dr hab. inż. Agnieszka Ławrynowicz

Opiekun

mgr inż. Maksymilian Marcinowski

POZNAŃ 2020

Tutaj będzie karta pracy dyplomowej;
oryginał wstawiamy do wersji dla archiwum PP, w pozostałych kopiach wstawiamy ksero.

Spis treści

1	Wstęp	1
1.1	Cel i zakres pracy	1
1.2	Struktura pracy	2
2	Podstawy teoretyczne	3
2.1	Rozpoznawanie emocji	3
2.2	Modele emocji	3
2.3	Głębokie uczenie	5
3	Analiza eksploracyjna	7
3.1	Zbiór danych - EmoContext	7
4	Przetwarzanie danych	10
4.1	Wczytanie i oczyszczenie danych	10
4.2	Tokenizacja i wyrównanie	10
4.3	Zagłębienia słów	11
5	Budowa modeli	12
5.1	Jednowarstwowa architektura LSTM	12
5.2	Głęboka architektura LSTM	13
6	Ewaluacja modeli	16
6.1	Nauka modeli	16
6.1.1	Dobór parametrów	16
6.1.2	Przebieg nauki	17
6.2	Metryki	18
6.2.1	Macierz pomyłek	18
6.2.2	Miara jakości nauki	18
6.2.3	Ostateczna ocena modelu	18
6.3	Wyniki	19
7	Podsumowanie	20
	Literatura	21
A	Płyta CD	22

Rozdział 1

Wstęp

Emocje międzyludzkie są podstawą codziennych interakcji z innymi osobami, a badania naukowców pokazują, że emocje są zjawiskiem uniwersalnym dla ludzi bez względu na pochodzenie. Jednak wpływy kulturowe jak i interpersonalne odgrywają kluczową rolę w identyfikacji konkretnych nawet podstawowych emocji, takich jak radość, miłość, gniew, strach i złość. Im bardziej wyszczególnione są etykiety emocji, tym trudniej jest wykryć tę właściwą. System rozpoznawania emocji może okazać się przydatny do wzajemnego zrozumienia między osobami, poprzez dostarczenie niewykrytego sygnału emocji.

Doskonałym przykładem są emocje występujące w mediach społecznościowych, które w niektórych sytuacjach są bardzo wyraziste, lecz bywają też niejednoznaczne i przy tym mogą być wyrażane w wulgarny sposób. System wykrywania emocji w obszarze dialogów między ludźmi w postaci rozmowy na forum internetowym lub komentarzy pod postem może okazać się bardzo pomocny w poprawieniu bezpieczeństwa w Internecie dla ludzi młodych i dzieci. Przykładów zastosowań jest mnóstwo, kilka z nich to filtrowanie treści, blokada słów wulgarnych i obraźliwych oraz zdań z ukrytym podtekstem. Jest to także rekomendacja treści, grupowanie tekstu o podobnym znaczeniu, czy nawet badanie rynku. Można by wykorzystać taki system do zbadania większej liczby komentarzy, np. w przypadku koncernów samochodowych, czy firm produkujących elektronikę jakie emocje przewyższają w komentarzach pod wyświetlanymi reklamami w mediach społecznościowych. Czy jest to zachwyt, zadowolenie, czy rozczarowanie. Działania te mogły by odpowiedzieć na pytanie czy wydany właśnie przez nich produkt przyjmie się na rynku oraz co warto by poprawić. W takich przypadkach etykietowanie komentarzy przez człowieka mogłoby okazać się trudne do wykonania i niejednoznaczne oraz bardzo kosztowne i czasochłonne.

1.1 Cel i zakres pracy

Głównym celem pracy jest opracowanie modelu uczenia maszynowego opartego na głębokich sieciach neuronowych (klasyfikatora wieloklasowego) w celu detekcji emocji w tekście postów z dialogów z mediów społecznościowych. W ramach tego celu niezbędne będzie zapoznanie się z literaturą dotyczącą przetwarzania języka naturalnego i dostępnymi frameworkami do uczenia głębokiego, przeprowadzenie analizy eksploracyjnej wybranych

zbiorów danych, wstępne przetworzenie tych danych po czym nastąpi budowa i ewaluacja kilku modeli o różnych architekturach w celu przeprowadzenia analizy porównawczej.

1.2 Struktura pracy

Struktura pracy jest następująca. Rozdział 2 przedstawia podstawy teoretyczne wymagane do zrozumienia dalszych etapów pracy wraz z przeglądem literatury. Rozdział 3 jest poświęcony analizie eksploracyjnej wybranych zbiorów danych. Rozdział 4 zawiera techniki przetwarzania wstępnego zbiorów danych. Rozdział 5 ukazuje budowę modeli a rozdział 6 ich ewaluację. Rozdział 7 zawiera podsumowanie pracy.

Rozdział 2

Podstawy teoretyczne

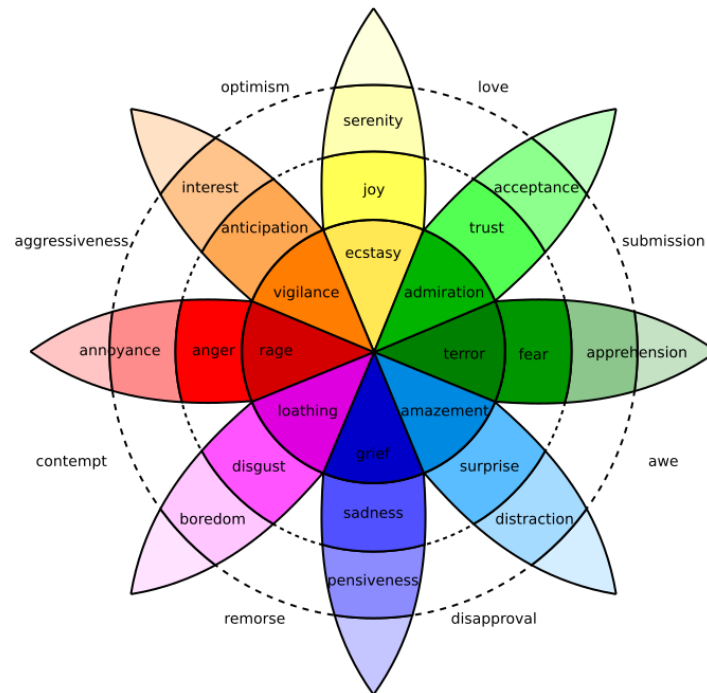
2.1 Rozpoznawanie emocji

Rozpoznawanie emocji w dialogach koncentruje się na wydobyciu emocji przekazanej w rozmowie pomiędzy co najmniej dwoma rozmówcami. Problem ten stawia bardzo dużo wyzwań, takich jak obecność sarkazmu w rozmowie, przesunięcie emocji do kolejnych wypowiedzi tego samego rozmówcy oraz uchwycenie szerszego kontekstu pomiędzy wypowiedziami różnych osób. Dużym plusem w tej dziedzinie jest bardzo dobra dostępność do danych, które pochodzą z platform społecznościowych takich jak Facebook, Youtube, Reddit, Twitter [8]. Poprzez łatwy dostęp do danych rozpoznawanie emocji w rozmowie staje się coraz bardziej popularne, a trudność tego problemu stwarza coraz to bardziej odległe granice co sprowadza się do wysokiego zainteresowania tą dziedziną przetwarzania języka naturalnego (ang. *natural language processing* - *NLP*).

Bardzo ważnym elementem w rozpoznawaniu emocji jest możliwość zrozumienia danego przekazu w kontekście, od którego może zależeć rodzaj emocji. Szczególnie trudnym przypadkiem jest zrozumienie i zapamiętanie kontekstu w konwersacji, co jak pokazują autorzy artykułu na temat architektury głębokiego uczenia do rozpoznawania emocji w rozmowach tekstowych [13], może okazać się kluczowym czynnikiem skuteczności rozpoznawania emocji. Do uzyskania satysfakcjonujących wyników nie wystarczają tradycyjne metody uczenia maszynowego lub najbardziej podstawowe architektury sieci neuronowych. Modele te wykorzystują zaawansowane techniki architektury transformera (ang. *the Transformer*) [11], które korzystają z podejścia mającego na celu poprawę modelowania sekwencji do sekwencji (ang. *Seq2Seq*) poprzez samoobserwację (ang. *self-attention*) i kodowanie pozycji (ang. *positional encoding*).

2.2 Modele emocji

Aby dobrze zrozumieć postawiony problem niezbędne będzie określenie czym są emocje. Wszyscy ludzie posiadają wrodzony zestaw podstawowych emocji, które można rozpoznać za pomocą gestów, czynów lub wypowiedzianych słów. Możemy wyróżnić dyskretne emocje, aby móc odróżnić je od siebie. Istnieje kilka definicji różnych modeli emocji, jednym z nich jest model zaproponowany przez Paula Ekmana [3]. Paul wraz ze współpra-



Rysunek 2.1. Koło emocji Plutchika [7].

cownikami stwierdzili, że istnieje sześć podstawowych emocji: gniew, obrzydzenie, strach, szczęście, smutek i zaskoczenie, a z każdą z tych emocji związane są jakieś cechy. Dzięki temu można wyrazić emocje w różnym stopniu a każda z nich jest zdefiniowana jako dyskretna kategoria, co pozwala na dość łatwą klasyfikację konkretnej emocji.

Kolejną definicję modelu emocji przedstawił Robert Plutchik, który podzielił emocje na osiem podstawowych typów, z których każdy ma drobniejsze podtypy pokrewne [7], zaprezentowane na rysunku 2.1 za pomocą koła emocji. Prezentuje on emocje jako koncentryczne kręgi, gdzie wewnętrzne części odpowiadają za podstawowe emocje a te zewnętrzne za bardziej złożone. Model ten jest dyskretny, lecz widać w nim pewne zależności i podobieństwa pomiędzy sąsiadującymi częściami koła emocji. Budowa ta wynika ze złożoności emocji i możliwości wyrażania ich intensywności.

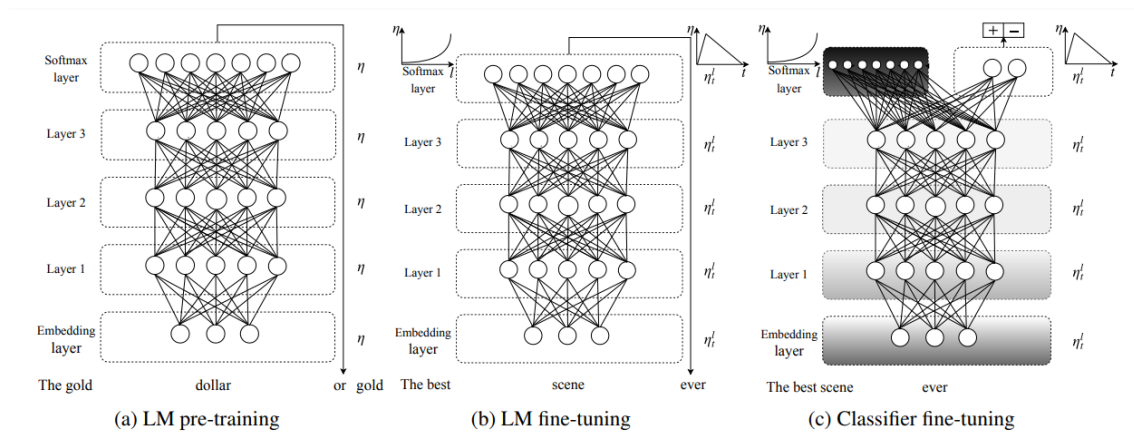
Podsumowując wymienione modele emocji możemy wydzielić dwa główne typy: kategoryczne oraz wymiarowe. Modele wymiarowe mapują emocję w sposób ciągły na wektory. Modele kategoryczne klasyfikują emocję do konkretnej emocji dyskretnnej, np. jednej z wybranego modelu emocji Ekmana lub Plutchika. Modele kategoryczne mają pewne wady. Jedną z nich jest brak możliwości opisanie innych emocji oraz utrudnione opisywanie emocji złożonej z kilku różnych podtypów zdefiniowanych w dyskretnym modelu. Drugą wadą jest brak możliwości porównywania emocji, co umożliwiłby model wymiarowy, za pomocą porównywania dwóch wektorów. Wybór odpowiedniego modelu emocji nie jest łatwy, a jednocześnie jest bardzo ważnym elementem do późniejszej klasyfikacji emocji. Decydując się na kategoryczny typ emocji, z jednej strony mamy prosty model Ekmana który nie jest w stanie zamodelować złożonych emocji. Z drugiej strony w modelu Plutchika może

być bardzo trudno rozróżnić drobnoziarniste emocje od siebie. Wybór ten należy zatem dokonać mając na uwadze wielkość oraz jakość zbioru danych.

2.3 Głębokie uczenie

W przetwarzaniu języka naturalnego z użyciem głębokich sieci neuronowych coraz częściej używane są techniki transferu wiedzy (ang. *transfer learning*) oraz adaptacji domenowej. Model języka jest kluczowym elementem do zastosowania powyższych technik. Umożliwia on przewidzenie kontekstu w jakim dane słowo znajduje się w zdaniu i na tej podstawie umożliwia odkryć jego prawdziwy sens. Jest uważany za bardzo istotny element w dziedzinie *NLP*, który stanowi podstawę do wszelkich zastosowań przetwarzania języka naturalnego. Najważniejsze jego cechy to zrozumienie długofalowych zależności i hierarchicznej struktury tekstu, a największe zalety to otwarte i wolne zasoby do jego stworzenia. Jest tworzony za pomocą nienadzorowanego procesu uczenia, który potrzebuje tylko korpusu nieoznakowanego tekstu.

Znakomitym przykładem użycia transferu wiedzy za pomocą wielokrotnego uczenia modelu języka jest metoda *ULMFIT* [4] (ang. *Universal Language Model Fine-tuning for Text Classification*), która może być zastosowana do każdego zadania w *NLP*. Zastosowane są w niej techniki, które są kluczowe dla dostrojenia modelu językowego. Używane są w niej 3 warstwy sieci neuronowej wykorzystującej komórki *LSTM* (ang. *Long Short-Term Memory*), które w przeciwieństwie do standardowych komórek sieci neuronowych, posiadają połączenia zwrotne, umożliwiające zapamiętanie sąsiednich stanów w sieci. Dodatkowo zastosowana jest technika przerywania (ang. *dropout*) niwelująca problem przeuczenia. Cały etap nauki w metodzie *ULMFIT* składa się z 3 etapów zaprezentowanych na rysunku 2.2. Na początku następuje szkolenie wstępne modelu językowego na dowolnym korpusie, następnie dostrojenie modelu językowego na zadaniu docelowym i na końcu dostrojenie klasyfikatora na zadaniu docelowym. Dzięki zastosowaniu tych technik możliwe jest wyuczenie wstępne modelu języka na dowolnych danych, (np. korpus Wikipedii) a następnie wykorzystanie tego wstępnie wyuczonego modelu na zadaniu docelowym. Na podobnej zasadzie działają dzisiejsze najbardziej wyrafinowane architektury głębokich sieci neuronowych w zastosowaniu przetwarzania języka naturalnego nazywane aktualnym stanem techniki (ang. *state of the art*).



Rysunek 2.2. 3 etapy nauki modelu języka (ang. *LM*) w metodzie ULMFIT [4].

Rozdział 3

Analiza eksploracyjna

3.1 Zbiór danych - EmoContext

Zbiór danych został udostępniony przez zespół z Międzynarodowych Warsztatów Ewaluacji Semantycznej (ang. *SemEval-2019 International Workshop on Semantic Evaluation*) jako jedno z zadań konkursowych o nazwie EmoContext¹ (ang. *Contextual Emotion Detection in Text*). Celem jest odkrycie prawidłowej etykiety emocji dla danego dialogu, składającego się z trzech wypowiedzi. W zadaniu tym użyty jest uproszczony model emocji zaproponowany przez Paula Ekmana [3], składający się z czterech klas: *Happy*, *Sad*, *Angry* oraz *Others*. Przykładowe dialogi oraz odpowiadające im etykiety emocji pochodzące ze zbioru treningowego zaprezentowane są na rysunku 3.1.

	id	turn1	turn2	turn3	label
0	0	Don't worry I'm girl	hmm how do I know if you are	What's ur name?	others
1	1	When did I?	saw many times i think -_-	No. I never saw you	angry
2	2	By	by Google Chrome	Where you live	others
3	3	U r ridiculous	I might be ridiculous but I am telling the truth.	U little disgusting whore	angry
4	4	Just for time pass	wt do u do 4 a living then	Maybe	others
5	5	I'm a dog person	youre so rude	Whaaaaat why	others
6	6	So whatsup	Nothing much. Sitting sipping and watching TV....	What are you watching on tv?	others
7	7	Ok	ok im back!!	So, how are u	others
8	8	Really?	really really really really really	Y saying so many times...i can hear you	others
9	9	Bay	in the bay	👉 love you	others

Rysunek 3.1. Przykładowe dialogi w danych z EmoContext wraz z etykietą emocji.

Przez organizatorów konkursu udostępnione zostały następujące zbiory do ewaluacji własnych modeli: zbiór treningowy (*train*), zbiór walidacyjny (*dev*) oraz zbiór testowy (*test*). Każdy z tych zbiorów zawiera tę samą strukturę danych, każdy przykład składa się z identyfikatora (*Id*), trzech wypowiedzi w dialogu oraz etykiety emocji. Szczegółowe informacje na temat tych zbiorów ukazuje tabela 3.1.

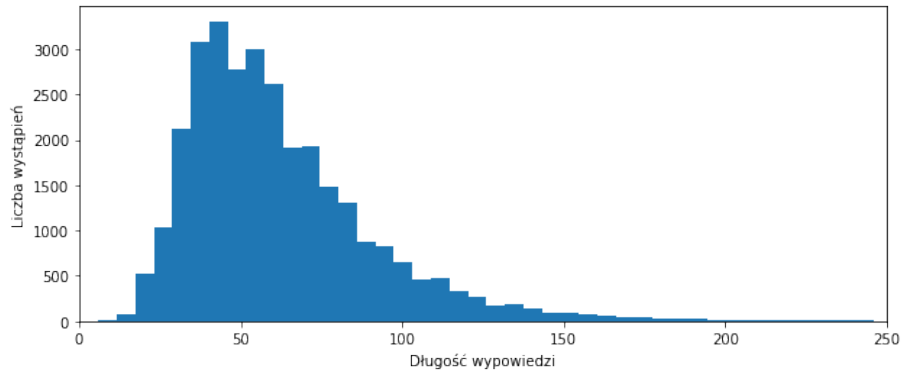
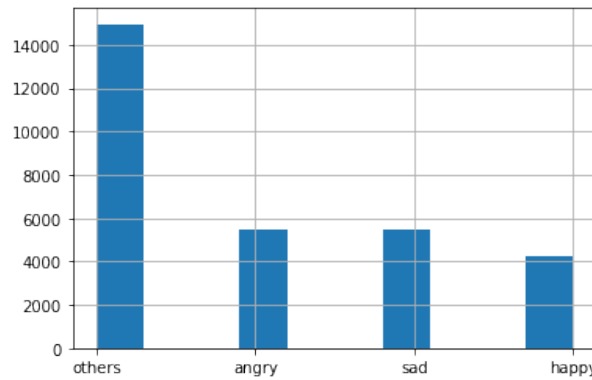
Aby lepiej zrozumieć dane treningowe przeprowadzone zostały podstawowe analizy eksploracyjne tego zbioru. Jedną z nich jest rozkład długości całego dialogu zaprezen-

¹<https://www.humanizing-ai.com/emocontext.html>

Tablica 3.1

Tabela prezentująca informacje o zbiorach w danych z EmoContext.

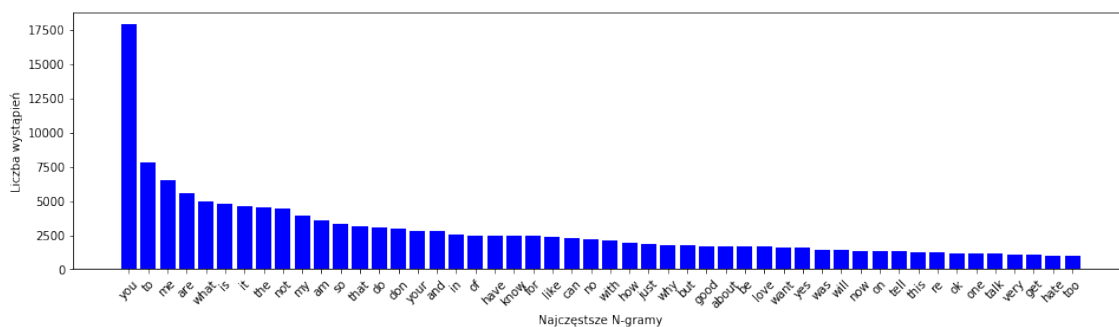
zbiór	liczba przykładów	najczęstsza klasa
train	30160	<i>Others</i> 50%
dev	2755	<i>Others</i> 84%
test	5509	<i>Others</i> 84%

**Rysunek 3.2.** Rozkład długości dialogu w danych z EmoContext.**Rysunek 3.3.** Niezbalansowany rozkład etykiet emocji w danych z EmoContext.

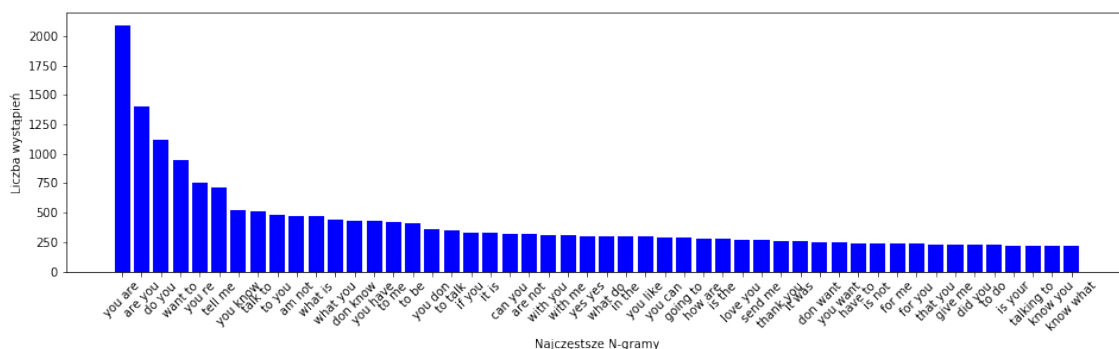
wany na rysunku 3.2. Widzimy na nim że przeważające dialogi są dosyć krótkie, średnio 62 znaki, co może utrudnić zadanie rozpoznania właściwej emocji. Zauważyć można także występujący długi ogon w kierunku coraz dłuższych dialogów, najdłuższy z nich ma 692 znaki.

Kolejna analiza to histogram częstości występowania danej etykiety w zbiorze treningowym, dzięki temu można sprawdzić czy występuje niezbalansowanie klas. Rysunek 3.3 prezentuje dominację klasy *Others* nad pozostałymi klasami, jest średnio ponad dwukrotnie liczniejszy od każdej z pozostałych etykiet.

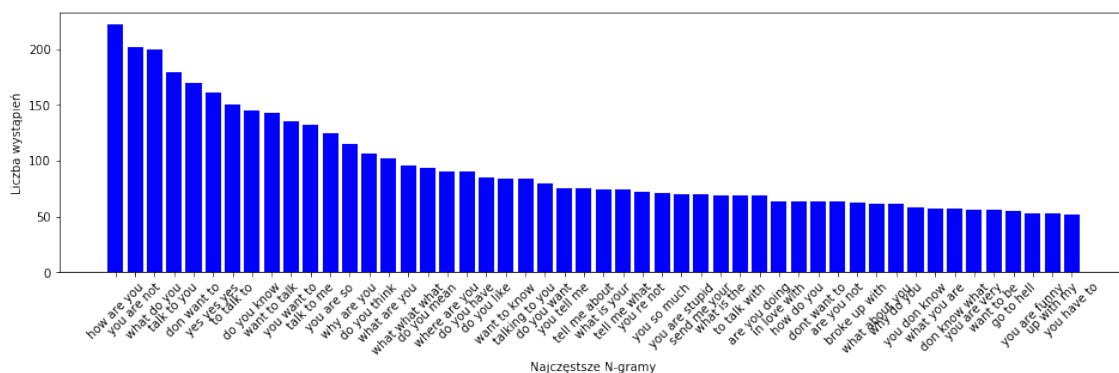
Ostatnim elementem analizy eksploracyjnej tego zbioru jest identyfikacja najczęściej występujących *N-gramów*, gdzie *N* to liczba zlepek słów występujących obok siebie w dialogach. Rysunki 3.4, 3.5 oraz 3.6 ukazują histogramy najczęstszych *N-gramów*. Można zauważyć że najczęściej występującymi słowami są słowa należące do grupy słów nie wnoszących znaczenia (ang. *stop words*), np.: słowa popularne, spójniki, przedimki, jednak



Rysunek 3.4. Rozkład wystąpień najczęstszych 1-gramów (wyrazów) w danych z EmoContext.



Rysunek 3.5. Rozkład wystąpień najczęstszych 2-gramów w danych z EmoContext.



Rysunek 3.6. Rozkład wystąpień najczęstszych 3-gramów w danych z EmoContext.

często występujące słowa to także wnoszące dużo znaczenia do wypowiedzi takie jak *love*, *like*, *good*, *hate*.

Rozdział 4

Przetwarzanie danych

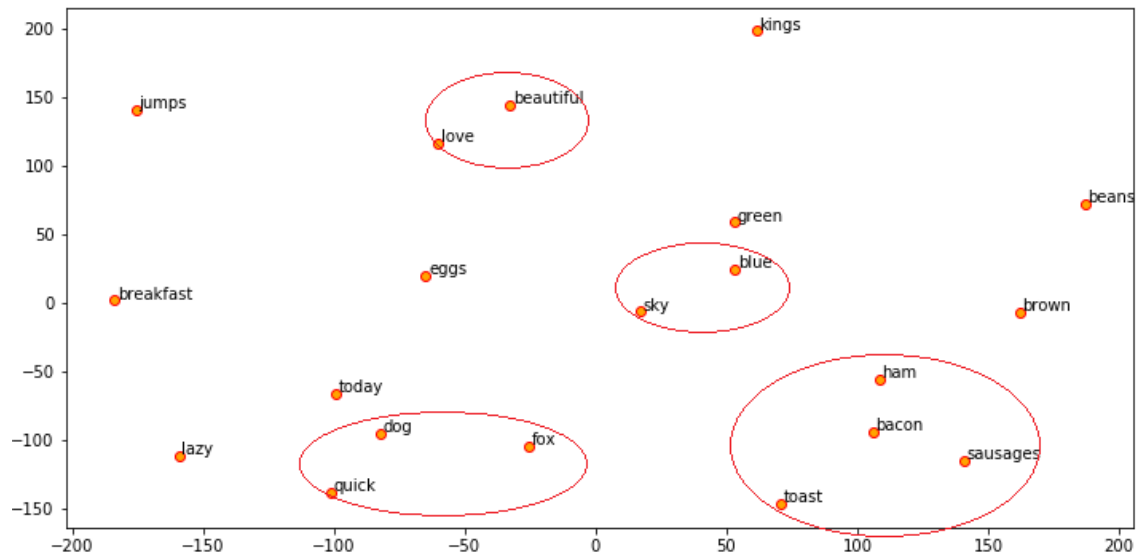
4.1 Wczytanie i oczyszczenie danych

Pierwszym etapem przygotowania danych do użycia w modelu jest ich wczytanie oraz przetworzenie. Zbiór danych z EmoContext zawiera pięć kolumn (rys. 3.1). W pierwszej kolumnie znajduje się identyfikator, w kolumnach od 2 do 4 zawarty jest docelowy tekst który będzie użyty jako wejście do modelu, a w ostatniej kolumnie znajduje się etykieta emocji. Po wczytaniu danych następuje połączenie trzech kolumn z tekstem w jeden ciąg znaków, oddzielone znakiem specjalnym *EOS* (ang. *end of sentence*). Jest to niezbędna operacja która umożliwi docelowemu modelowi oddzielić trzy wypowiedzi od siebie. Następnie na tak otrzymanym ciągu znaków wykonywane są operacje usuwania powtarzających się znaków interpunkcyjnych (np. *"!!!!"* zostanie zamienione na pojedynczy znak *"!"*). Na końcu wykonywane jest oddzielenie znaków interpunkcyjnych od wyrazów, usunięcie powtarzających się spacji oraz zamiana wszystkich dużych liter na małe.

4.2 Tokenizacja i wyrównanie

Modele głębokie przetwarzania języka naturalnego nie operują na jawnym tekście w postaci ciągu znaków, tylko na reprezentacji w postaci liczbowej. Do reprezentacji wyrazów używane są słowniki które przechowują wszystkie znane słowa z korpusu uczącego. Do tej zamiany użyta została metoda *Tokenizer* z pakietu TensorFlow. Klasa ta pozwala na wektoryzację korpusu tekstu, poprzez przekształcenie każdego wyrazu w liczbę całkowitą. Każda z tych liczb odpowiada indeksowi tego słowa w słowniku.

Po przekształceniu każdego słowa w token następuje wyrównanie długości wszystkich przykładów w zbiorze. Przykładowa liczba użyta do wyrównania (ang. *padding*) to 20. Wszystkie przykłady, które mają mniejszą liczbę tokenów, wypełniane są tokenem zerowym aż do osiągnięcia długości 20, a wszystkie przykłady które mają większą liczbę tokenów są skracane od końca. W ten sposób została utworzona macierz danych która może być wykorzystana do użycia w nauce i ewaluacji modeli.



Rysunek 4.1. Przykładowe słowa w przestrzeni wektorowej GloVe [1].

4.3 Zagłębienia słów

W przypadku głębokich sieci neuronowych z rekurencyjnymi komórkami LSTM wysoce efektywne jest użycie zagłębień słów (ang. *words embeddings*) jako wejścia do pierwszej warstwy sieci neuronowej. Osadzanie słów jest przeprowadzone za pomocą mapowania tych słów na wektory liczb rzeczywistych. Reprezentacja słów jako wektory liczb jest rozszerzeniem reprezentacji *jeden z N* (ang. *one hot encodings*), która jest używana w celu zwiększenia wydajności modeli NLP. Jest to także możliwość użycia transferu wiedzy w postaci wstępnie wytrenowanych zagłębień słów jako reprezentacji danych wejściowych. Jednym z algorytmów, który umożliwia stworzenie takiej reprezentacji jest metoda GloVe [1] (ang. *Global Vectors for Word Representation*), stworzona przez zespół z Uniwersytetu Stanford. Udostępniona przez nich macierz umożliwia użycie tej reprezentacji, jako odwzorowanie słów na wektory liczb rzeczywistych. Słowa zmapowane w tej przestrzeni mają zachowane pewne właściwości, odległość między nimi jest powiązana z podobieństwem semantycznym. Na rysunku 4.1 ukazane są przykładowe słowa i zachowane podobieństwo np. między słowami *love* i *beautiful*. Sposób uzyskania tej reprezentacji bazuje na współwystępowaniu danych słów w korpusie w otoczeniu które je definiuje. Główną intuicją tego modelu jest założenie, że proporcje prawdopodobieństwa współwystępowania słów mają potencjał do kodowania jakiejś formy znaczenia. Wynikiem tej metody są wektory słów które bardzo dobrze radzą sobie z zadaniami opartymi o podobieństwo, analogię, a także odkrywaniu semantyki emocjonalnej słów i wiele innych wymienionych w pakiecie *word2vec* [6].

Rozdział 5

Budowa modeli

5.1 Jednowarstwowa architektura LSTM

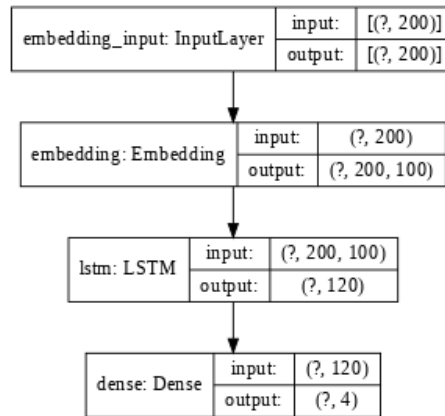
Pierwszy model bazuje na zaproponowanej przez organizatorów konkursu jednowarstwowej architekturze sieci rekurencyjnej z komórkami LSTM. Zastosowanie tej budowy modelu zapewnia zdolność do uczenia się długoterminowych zależności przy jednoczesnym unikaniu długotrwałego problemu uzależnienia.

Przygotowawczym etapem modelu jest przedstawienie wymagań dla danych, które mogą być przetwarzane w warstwie wejściowej (ang. *Input Layer*). Każdy przykład jest przetworzony zgodnie z opisem w rozdziale 4. Jednym z ostatnich etapów przetwarzania jest wyrównanie wszystkich przykładów do równej długości w tym przypadku 200, zgodnie z rysunkiem 5.1 przedstawiającym budowę jednowarstwowej architektury LSTM.

Kolejną warstwą jest zamiana reprezentacji słowa na gęstą reprezentację wektorową opisaną w sekcji 4.3. Do uzyskania tej reprezentacji użyto gotowe macierze udostępnione przez autorów metody na stronie internetowej¹. Do wyboru były macierze przygotowane na różnych zbiorach danych oraz o różnych szerokościach wektorów. Testy przeprowadzono z użyciem zagłębień słów wytrenowanych na zbiorze danych pochodzących z Wikipedii oraz z archiwum danych tekstowych Gigaword 5 (ang. *English Gigaword Fifth Edition*). Wybór szerokości wektorów podyktowany był złożonością modelu i czasem nauki, z pośród zbioru 50, 100, 200, 300 zdecydowano się na szerokość 100. Podsumowując złożoność tego etapu zawiera on prawie 2 miliony wewnętrznych parametrów (wag), które są zamrożone na czas uczenia. Rysunek 5.2 przedstawia szczegóły budowy jednowarstwowej architektury LSTM, wraz z tą warstwą o nazwie *Embedding*.

Do stworzenia kolejnej warstwy użyto gotową reprezentację LSTM z pakietu tensorflow o nazwie `keras.layers.LSTM`, która udostępnia możliwość definiowania konkretnych szczegółów implementacyjnych. W warstwie tej użyto domyślną funkcję aktywacji jakim jest tangens hiperboliczny. Po obserwacji krzywej uczenia reprezentującej dokładność oraz wartość funkcji straty zauważono zjawisko przeuczenia. Dokładność dla zbioru uczącego rosła wraz z malejącą dokładnością dla zbioru walidacyjnego. Z tego powodu zdecydowano się na użycie metody przerywania (ang. *dropout*), zdefiniowanej w tej warstwie. Pomogło to zmniejszyć wpływ zjawiska przeuczania się modelu.

¹<https://nlp.stanford.edu/projects/glove/>



Rysunek 5.1. Graf przedstawiający budowę jednowarstwowej architektury LSTM.

Model: "Jednowarstwowa architektura LSTM"		
Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 200, 100)	1683200
lstm (LSTM)	(None, 120)	106080
dense (Dense)	(None, 4)	484
Total params: 1,789,764		
Trainable params: 106,564		
Non-trainable params: 1,683,200		

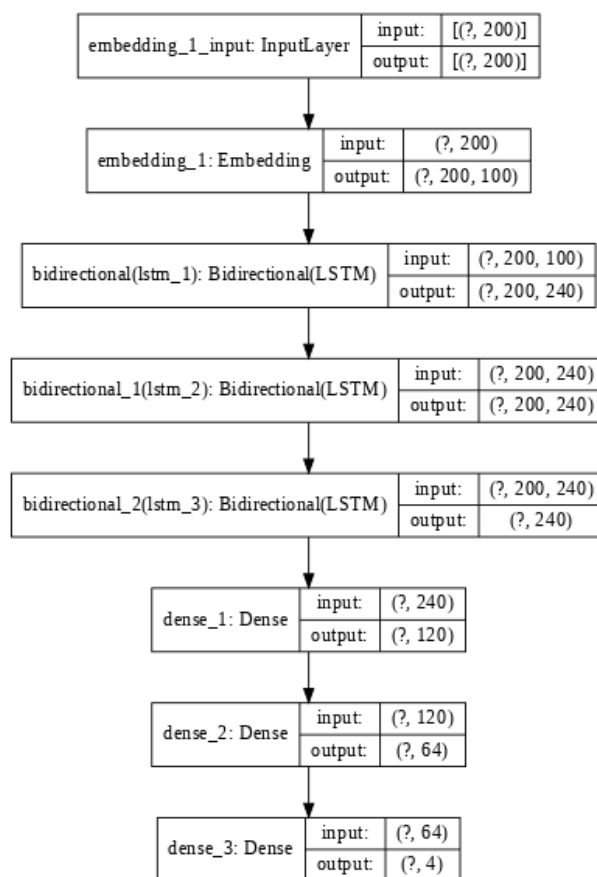
Rysunek 5.2. Tabela przedstawiająca szczegóły budowy jednowarstwowej architektury LSTM.

Ostatnim etapem jest wybór odpowiedniej etykiety emocji za pomocą warstwy gęstej. Wyjście z poprzedniej warstwy nazywane stanem ukrytym (ang. *hidden state*) o szerokości 120 jest gęsto połączone z 4 neuronami decydującymi odpowiednio o każdej z kolejnych etykiet emocji. Jako funkcję aktywacji tej warstwy, zgodnie z sugestią organizatorów konkursu użyto funkcję sigmoidalną. Zauważona jednak została niekompatybilność tej funkcji z konkretnym zastosowaniem dla klasyfikacji wieloklasowej co zostało poprawione w kolejno zaprezentowanych modelach.

Końcowym elementem budowy tego modelu jest wybór algorytmu optymalizacyjnego. Wybrano algorytm RMSprop [10] (ang. *Root Mean Square Propagation*), który dobrze radzi sobie z wygasającymi wskaźnikami uczenia oraz przeciwdziała obliczeniowym problemom numerycznym. Jako funkcję straty użyto *Categorical Cross-Entropy loss*, która bardzo dobrze radzi sobie w problemach klasyfikacji wielu klas.

5.2 Głęboka architektura LSTM

Głęboka architektura LSTM jest rozszerzeniem architektury jednowarstwowej opisanej w punkcie 5.1. Głównym celem było porównanie złożoności architektury płytkiej i głębokiej oraz wpływ rozszerzenia modelu o kolejne warstwy na wynik. Modyfikacje polegały przede wszystkim na dodaniu kolejnych warstw oraz lekkie modyfikacje parametrów oraz



Rysunek 5.3. Graf przedstawiający budowę głębokiej architektury LSTM.

funkcji wewnątrz modelu. Pierwsze dwie warstwy, wejściowa (*Input Layer*) oraz mapująca słowa na gęstą reprezentację wektorową (*Embedding*) zostały bez zmian.

Do rozszerzenia jednowarstwowej części sieci rekurencyjnej zamiast zwykłych komórek LSTM użyto dwukierunkowych komórek LSTM. Z założenia rozszerzenie to powinno poprawić wydajność modelu przy problemach z klasyfikacją sekwencji [2]. W tym przypadku można było użyć tego rozszerzenia ponieważ od razu są dostępne wszystkie sekwencje czasowe sekwencji wejściowej. W tym momencie dwukierunkowe komórki LSTM łączą dwie ukryte warstwy o przeciwnych kierunkach. Dzięki tej formie uczenia warstwa wyjściowa może jednocześnie uzyskiwać informacje z przeszłych jak i przyszłych stanów, co nie było możliwe przy użyciu podstawowych komórek LSTM. Zabieg ten pomaga lepiej zrozumieć kontekst gdyż znaczenie danego słowa może zależeć także od słów, które są przed danym słowem. Dodatkowo zamiast jednej warstwy sieci rekurencyjnej użyto łącznie trzy warstwy używające dwukierunkowego LSTM. Pierwsze dwie warstwy na wyjściu oprócz ukrytego stanu zwracają także sekwencję o długości 200, którą przekazują na wejście kolejnej warstwy sieci LSTM. Widać to na rysunku 5.3, który przedstawia graf głębokiej architektury LSTM.

Kolejną modyfikacją było dodanie kolejnych warstw sieci gęstej. Wcześniej użytą jedną warstwę zawierającą 4 neurony wyjściowe rozszerzono o kolejne dwie warstwy gęste. Pierwsza z nich zawiera 120 neuronów a druga zawiera 64 neurony. Szczegóły tej budowy

Model: "Głęboka architektura LSTM"

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 200, 100)	1683200
bidirectional_1 (Bidirectional)	(None, 200, 240)	212160
bidirectional_2 (Bidirectional)	(None, 200, 240)	346560
bidirectional_3 (Bidirectional)	(None, 240)	346560
dense_1 (Dense)	(None, 120)	28920
dense_2 (Dense)	(None, 64)	7744
dense_3 (Dense)	(None, 4)	260
Total params: 2,625,404		
Trainable params: 942,204		
Non-trainable params: 1,683,200		

Rysunek 5.4. Tabela przedstawiająca szczegóły budowy głębokiej architektury LSTM.

Tablica 5.1
Tabela porównująca szczegóły budowy poszczególnych modeli.

model	parametry trenowalne	parametry stałe	SUMA
Jednowarstwowy LSTM	106,564	1,683,200	1,789,764
Głęboki LSTM	942,204	1,683,200	2,625,404
BERT todo	xx	xx	xxx

przedstawiono na rysunku 5.4, który przedstawia tabelę prezentującą szczegóły budowy głębokiej architektury LSTM. Do dodanych warstw użyto innej funkcji aktywacji jaką jest jednostronnie obcięta funkcja liniowa (ang. *ReLU*), która stała się standardem do użycia w wewnętrznych warstwach gęstej sieci neuronowej [12]. W ostatniej warstwie także zamieniono funkcję aktywacji na funkcję *softmax*, która lepiej nadaje się do klasyfikacji wieloklasowej.

Porównując dokonane rozszerzenia modelu z jedną warstwą LSTM, oraz wykonując opisane modyfikacje warstw zwiększyła się złożoność modelu. Prostym sposobem na porównanie modeli jest sprawdzenie liczby parametrów, które definiują zachowanie sztucznych neuronów, inaczej nazywane wagami neuronu. W tabeli 5.1 przedstawione są liczby parametrów podzielonych na dwie grupy. Parametry trenowalne modelu to są wagi, które ulegają modyfikacji, liczba tych parametrów zwiększyła się około dziewięciokrotnie co ilustruje skalę zmian. Parametry stałe to są wagi użyte do generacji gęstych reprezentacji wektorowych, które były zamrożone na czas nauki modeli.

Rozdział 6

Ewaluacja modeli

6.1 Nauka modeli

6.1.1 Dobór parametrów

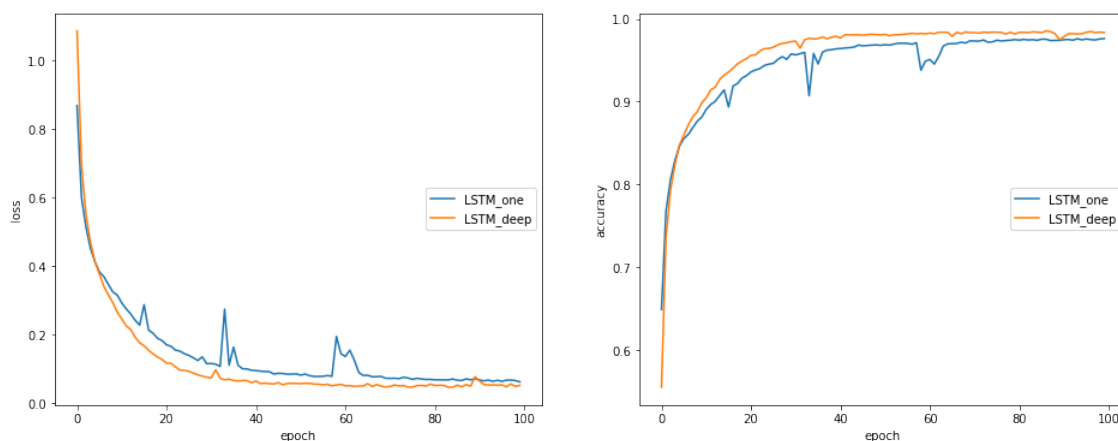
Proces nauki poszczególnych modeli rozpoczęto od zdefiniowania *hiper-parametrów* [9]. Są to parametry, których wartości są ustawione przed rozpoczęciem procesu uczenia się i definiują one zachowanie poszczególnych warstw modelu. Początkowe wartości tych parametrów zdefiniowano zgodnie z powszechnymi wartościami używanymi w podobnych architekturach oraz na podstawie załączonej literatury. Dalsze kroki polegały na dostosowaniu tych parametrów do osiągnięcia najlepszego wyniku. Dobrym sposobem na znalezienie optymalnego doboru parametrów są metody typu wyszukiwanie w siatce (ang. *grid search*) oraz wyszukiwanie losowe (ang. *random search*) [5]. Polegają one na porównywaniu działania modelu z przyjętymi parametrami wybranymi z przestrzeni parametrów. Podjęto próby użycia tych metod, jednak ze względu na ograniczone zasoby obliczeniowe oraz długi czas przetwarzania ograniczono przestrzeń parametrów do minimum. Pozwoliło to na przeszukanie bardzo małego zbioru wartości co i tak dało satysfakcjonujące wyniki.

Dobre wartości dla modeli jednowarstwowego LSTM 5.1 oraz głębokiego LSTM 5.2 przedstawiono w tabeli 6.1. Założenie użycia tych samych hiper-parametrów dla obu modeli było sprawdzenie wpływu głębokości (liczby warstw) modelu na wynik. Rozmiar LSTM dotyczy liczby komórek LSTM w jednej warstwie, dobór wartości 120 zależał od wielkości zbioru danych oraz od ograniczeń zasobowych. Parametr przerywania (ang. *dropout*) użyty do przeciwdziałania przeuczenia się modelu został wybrany z pośród dwóch wartości 0.2 oraz 0.5, okazało się że 0.2 ma lepszy wpływ na wynik. Rozmiar wektora słów dotyczy szerokości gęstej reprezentacji wektorowej każdego słowa, czym większa wartość tym więcej informacji przenoszonych przez te wektory, jednak ograniczenia zasobów pozwoliły na szerokość 100, gdzie maksymalna dostępna szerokość to 300. Pozostałe parametry dotyczą sposobu nauki modeli oraz szybkości i długości tego procesu, wartości te pozostały domyślne.

Tablica 6.1

Tabela prezentująca wybrane *hyper-parametry* modeli korzystających z LSTM.

<i>hyper-parametr</i>	wartość
rozmiar LSTM	120
wskaźnik uczenia się	0.003
przerywanie (ang. <i>dropout</i>)	0.2
liczba epok	100
rozmiar wektora słów	100
wielkość grupy (ang. <i>batch</i>)	200

**Rysunek 6.1.** Wykresy przedstawiające przebieg nauki dla modeli jednowarstwowego LSTM oraz głębokiego LSTM

6.1.2 Przebieg nauki

Do procesu nauki użyto platformę *Colab*¹, która umożliwia korzystanie z zasobów karty graficznej (GPU) umożliwiającej dużo wydajniejszy proces nauki modeli. Dla jednowarstwowego LSTM czas przetwarzania jednej epoki to około 9 sekund, a dla modelu głębokiego LSTM to 64 sekundy. Przy użyciu laptopa czasy te były prawie dziesięciokrotnie większe co uniemożliwiło by tak sprawne projektowanie i dostrajanie modeli.

Proces nauki rozpoczęto od podstawowego modelu jednowarstwowego LSTM. Po doborze hiper-parametrów związanych z architekturą modelu dobrane były odpowiednie wartości dla liczby epok oraz wskaźnika uczenia się. Zauważono że po dwudziestej epoce dokładność przekroczyła 90%, a po setnej epoce poprawa wyników była na tyle niewielka że przzerwano proces nauki. Ostateczne wyniki są przedstawione na wykresach 6.1 przedstawiających przebieg nauki dla modeli jednowarstwowego LSTM oraz głębokiego LSTM. W początkowych etapach nauki wartości funkcji straty (ang. *loss*) oraz dokładności (ang. *accuracy*) były bardzo podobne, jednak po dziesiątej epoce widać przewagę modelu głębokiego LSTM oznaczonego kolorem pomarańczowym. Wynikać to może z dużo większej złożoności tego modelu oraz możliwości zapamiętania przez ten model większej liczby cech przechowywanych w tekście.

¹<https://colab.research.google.com/>

6.2 Metryki

6.2.1 Macierz pomyłek

W dziedzinie sieci neuronowych oraz oceny jakości modeli uczenia maszynowego, a w szczególności w problemach klasyfikacji często pomocna jest macierz pomyłek, inaczej nazywana macierzą błędów. Jest to specyficzny układ tabel, który pozwala na wizualizację wydajności algorytmu. Składa się on z dwóch wymiarów: rzeczywistym i przewidywanym. Podaje on liczbę wyników prawdziwie dodatnich (TP), prawdziwie ujemnych (TN), fałszywie dodatnich (FP) oraz fałszywie ujemnych (FN).

Tablica 6.2

Tabela ukazująca macierz pomyłek.

Przewidziana etykieta	Prawdziwa etykieta	
	Positive	Negative
	Positive	Negative
Positive	TP	FP
Negative	TN	FN
Łącznie	P	N

Pozwala to na bardziej szczegółową analizę oraz formułowanie wielu metryk na podstawie wartości z tej macierzy. Niektóre z tych miar zastosowano do ewaluacji modeli i zaprezentowano w kolejnych podsekcjach.

6.2.2 Miara jakości nauki

Główną metryką analizowaną w trakcie nauki modelu była dokładność (ang. *accuracy*), zdefiniowana w poniższym równaniu 6.1.

$$dokladnosc = \frac{TP + TN}{P + N} \quad (6.1)$$

Miara ta odwzorowuje efektywność nauki. Podczas jej obserwacji można stwierdzić czy każda kolejna epoka nauki poprawia wynik, czy też go pogarsza. Zbiór danych treningowy miał zrównoważony rozkład klas na tyle, że zastosowanie tej metryki miało sens. Niestety jeśli zestaw danych mocno niezbalansowany, tak jak w przypadku zbioru testowego miara ta może wprowadzić w błąd. Z tego powodu zastosowanie tej metryki ograniczono tylko do obserwacji procesu nauki, a ostateczną ocenę modelu wykonano na podstawie innych miar.

6.2.3 Ostateczna ocena modelu

Ostateczną ocenę modelu dokonano za pomocą obliczenia mikro średniego wyniku F1 dla trzech klas emocji *Happy*, *Sad*, *Angry*. Wybór ten wynika z niezbalansowania klas w zbiorze testowym oraz dominacją występowania etykiety *Others*, którą pominięto w ostatecznej ocenie. Wynik F1 jest średnią harmoniczną precyzji i czułości, zdefiniowanych w poniższych wzorach. Końcowy wynik uzyskany jest poprzez uśrednianie w skali mikro, na podstawie częstotliwości klas występujących w zbiorze treningowym. Poniższe wzory przedstawiają poszczególne składniki oraz samą funkcję F1.

$$\text{precyzja} = \frac{TP}{FP + TP} \quad (6.2)$$

$$\text{czulosc} = \frac{TP}{FP + TP} \quad (6.3)$$

$$F1 = \frac{2 \cdot \text{precyzja} \cdot \text{czulosc}}{\text{precyzja} + \text{czulosc}} \quad (6.4)$$

6.3 Wyniki

Jakość wyuczonych modeli na zbiorze treningowym porównano poprzez ewaluację predykcji tych modeli na zbiorze testowym. Głównymi założeniami porównań było sprawdzenie wpływu liczby warstw sieci neuronowej na wynik oraz porównanie tradycyjnych rekurencyjnych warstw do bardziej złożonych architektur jakim jest BERT.

Porównanie jednowarstwowej architektury LSTM z głęboką architekturą LSTM pokazuje przewagę dla modelu głębokiego. Większa liczba warstw pomogła osiągnąć lepszy wynik miary mikro F1, jednak jest to niewielka poprawa. Można spekulować że jednowarstwowa sieć LSTM dla tego zbioru danych była wystarczająca aby osiągnąć maksimum możliwości tej architektury. Z drugiej strony użycie gęstej reprezentacji wektorowej każdego słowa mogło ograniczyć możliwości warstw LSTM i spowodować że użycie kolejnych warstw sieci pomogło na zniwelowanie tylko małych błędów. Niestety analiza wpływu na wynik poszczególnych elementów sieci neuronowej jest utrudniona co uniemożliwia wyciągnięcie jasnych i sprawdzonych wniosków.

TODO BERT

Tablica 6.3

Tabela pokazująca wyniki poszczególnych modeli.

model	dokładność	mikro precyzja	micro czułość	mikro F1
Jednowarstwowy LSTM	0.85	0.49	0.70	0.58
Głęboki LSTM	0.87	0.51	0.71	0.60
BERT todo	xx	xx	xxx	xxx

Rozdział 7

Podsumowanie

Zakończenie pracy zwane również Uwagami końcowymi lub Podsumowaniem powinno zawierać ustosunkowanie się autora do zadań wskazanych we wstępie do pracy, a w szczególności do celu i zakresu pracy oraz porównanie ich z faktycznymi wynikami pracy. Podejście takie umożliwia jasne określenie stopnia realizacji założonych celów oraz zwrócenie uwagi na wyniki osiągnięte przez autora w ramach jego samodzielnej pracy.

Integralną częścią pracy są również dodatki, aneksy i załączniki zawierające stworzone w ramach pracy programy, aplikacje i projekty.

Literatura

- [1] Robin Brochier, Adrien Guille, and Julien Velcin. Global vectors for node representations, 2019.
- [2] Zixiang Ding, Rui Xia, Jianfei Yu, Xiang Li, and Jian Yang. Densely connected bidirectional lstm with applications to sentence classification, 2018.
- [3] Paul Ekman. Facial expression and emotion, 1993.
- [4] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification, 2018.
- [5] Petro Liashchynskyi and Pavlo Liashchynskyi. Grid search, random search, genetic algorithm: A big comparison for nas, 2019.
- [6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.
- [7] Robert Plutchik. A psychoevolutionary theory of emotions, 1982.
- [8] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances, 2019.
- [9] Philipp Probst, Bernd Bischl, and Anne-Laure Boulesteix. Tunability: Importance of hyperparameters of machine learning algorithms, 2018.
- [10] Sebastian Ruder. An overview of gradient descent optimization algorithms, 2016.
- [11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.
- [12] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network, 2015.
- [13] Peixiang Zhong, Di Wang, and Chunyan Miao. Knowledge-enriched transformer for emotion detection in textual conversations, 2019.

Dodatek A

Płyta CD

Płyta CD z elektroniczną wersją pracy, stworzonym projektem oraz danymi wejściowymi.



© 2020 inż. Jakub Zdanowski

Instytut Informatyki, Wydział Informatyki i Telekomunikacji
Politechnika Poznańska

Skład przy użyciu systemu \LaTeX na platformie Overleaf.