Eötvös Lóránd University

Faculty of Informatics

Department of Information Systems

Social media sentiment analysis based on COVID-19

Supervisor:
Attila Kiss Phd.
Associate Professor

Author:
László Nemes
Computer science MSc

Budapest, 2022

EÖTVÖS LORÁND TUDOMÁNYEGYETEM

INFORMATIKAI KAR

DIPLOMAMUNKA TÉMABEJELENTŐ

Hallgató adatai:

Név: Nemes László

Neptun kód: RVR55V

Képzési adatok:

Szak: programtervező informatikus, mesterképzés (MA/MSc)

Tagozat: Nappali

Belső témavezetővel rendelkezem

Témavezető neve: *Dr. Kiss Attila Elemér*

munkahelyének neve, tanszéke: Eötvös Loránd Tudományegyetem Informatikai Kar, Információs Rendszerek Tanszék

munkahelyének címe: 1117 Budapest, Pázmány Péter sétány 1/C

beosztás és iskolai végzettsége: Tanszékvezető egyetemi docens, ELTE Informatika Doktori Iskola

A diplomamunka címe: *Közösségi média érzelmi elemzés COVID-19 alapján*

A diplomamunka témája:

(A témavezetővel konzultálva adja meg 1/2 - 1 oldal terjedelemben diplomamunka témájának leírását)

A diplomamunka során, az ehhez készítendő szoftver segítségével a szociális/közösségi média, elsősorban a Twitter közösségi oldal felhasználóinak bejegyzései, tweetjei kerülnének felhasználásra. Különböző hangulatelemző modellek segítségével történő vizsgálatok során, a megadott érdeklődési témában kerülne sor adatbányászatra, tisztításra és előkészítésre, amely elsősorban a COVID-19 témájával foglalkozna. A munka során a Twitter közösségi oldalról, az általuk biztosított lehetőséggel, API-al történik az adatok bányászata. Ezt követően az adott témában összeállított adatokon kerül sor a hangulatelemzésre, így megállapítva, hogy az adott témához az emberek „mennyire állnak" pozitívan, negatívan, kevésbé pozitívan, kevésbé negatívan esetleg semlegesen. A hangulatelemzést elsősorban egy Ismétlődő neurális háló (RNN) biztosítja, emellett további eszközök is felhasználásra kerülnének, mint például a BERT, NLTK, TextBlob. Így egy átfogó képet kapva az adott futtatáskor az adott témában egy megadott számú tweet alapján hogyan is vélekednek az emberek. A különböző modellek által kapott eredmények pedig összehasonlíthatóak, így képet kapva a különböző modellek egymáshoz viszonyított eredményeiről. Ezen hangulatelemzések mellett további adatelemzések kerülnének elvégzésre, hogy még átfogóbb és részletes képet lehessen kapni a különböző hangulati csoportokba tartozó bejegyzésekről, mint például milyen igék és tulajdonnevek fordulnak elő gyakran a különböző kategóriákba tartozó bejegyzésekben, mindezt további elemzésekkel is kiegészítve. A futtatások eredményei a hosszú távú vizsgálatok elvégzését támogató és jövőbeli elemzésekhez is felhasználható formában kerülnének elmentésre NoSQL adatbázisban, így biztosítva egy esetleges webes platformba történő integrációt is a jövőben. A munka az elmúlt félévekben készített "Social Media Sentiment Analysis Based on COVID-19" és "Prediction of Stock Values Changes using Sentiment Analysis of Stock News Headlines" cikkekre alapozva történne.

Budapest, 2021.02.22.

Contents

Abstract

Social media platforms are increasingly take the place of communication of information, which intensified even more during the pandemic. News portals and governments are also increasing attention to digital communications, announcements and for the response or reaction monitoring. Twitter as one of the largest social networking sites, which became even more important in the communication of information during the pandemic, provides space to lot of different opinions and news, with many discussions as well. In this thesis, we look at the sentiments of people where we use tweets to determine how people relate to Covid-19 over a given period of time. These sentiment analyzes are augmented with information extraction and named entity recognition to get an even more comprehensive picture. The sentiment analysis is based on the'Bidirectional encoder representations from transformers' (BERT) model, which is the basic measurement model for the comparisons. We consider BERT as the baseline and compare the results with the RNN, NLTK and TextBlob sentiment analyzes. The RNN results are significantly closer to the benchmark results given by BERT, both models are able to categorize all tweets without a single tweet fall into the neutral category. Then, by a deeper analysis of these results, we can get an even more concise picture of people's emotional state in the given period of time. The data from these analyzes further support the emotional categories and provide a deeper understanding that can provide a solid starting point for other disciplines as well, such as linguistics or psychology. Thus, the sentiment analysis, supplemented with information extraction and named entity recognition analyzes, can provide a supported and deeply explored picture of specific sentiment categories and user attitudes.

Chapter 1

Introduction

Social media has become the number one channel of communication for people. Here they share their thoughts, opinions on different topics, and also share what articles they have read etc., shaping their narrow community with these activities.

These activities intensified even more during the pandemic, people spent more time online during lockdown and home office periods, therefore, their news consumption has changed and social media portals became their primary communication channel. We cannot announce the end of the epidemic yet, but we can already say that this displacement to this online space will be lasting in the coming periods, both in terms of work and news consumption, communication and different entertainments.

We definitely need to address these manifestations on different platforms (in this case focusing on Twitter) and as machine learning becomes more popular and important, as does natural language processing (NLP), we need to address, analyze and research emotions on these platforms.

1.1 Sentiment analysis

There are many options for executing sentiment analyzes, from 'human categorization' to 'dictionary based 'and 'deep learning' methods. In the field of tools, we can choose from fully ready-to-use tools, development kits and completely custom-developed models. One such tool is 'TextBlob'[1], which is fully ready to be inte-

[1]TextBlob documentation: `https://textblob.readthedocs.io/en/dev/`

grated into any analysis, just import the library and it is ready to use. As mentioned earlier, there are also options that allow us to create our own models, build and train them based on our own data.'Using Bidirectional encoder representations from transformers' (BERT)[2] for sentiment analysis is one of the most powerful tool what we can use, but we can also create a 'Recurrent neural network'[3] (RNN) or use the 'Natural Language Toolkit' [4] (NLTK) with the VADER lexicon and SentimentIntensityAnalyzer.

The main goal is to train a model to sentiment prediction by looking correlations between words and tag it to positive or negative sentiment. Thus, we created the RNN, BERT, NLTK - Vader lexicon models and imported the TextBlob tool into our analysis. We compared these primarily with the results of BERT. For the sentiment analyzes, we also expanded the usual 'positive', 'negative', and 'neutral' categories with 'strongly positive or negative' and 'weakly positive or negative' options for deeper analysis and to explore differences between models.

1.2 Further analysis

By performing further analysis on the data labeled by the RNN model obtained in this way, it is possible to determine even more precisely what emotions the given topic evoked from people in a given time period, in the 'covid' theme in this case. For these results we used 'Information extraction' (IE) and 'Named entity recognition' (NER) analyzes.

Today is a information overload age, the way we read stuff has changed. Most of us tend to skip the entire text, whether that is an article or a book and just read the 'relevant' bits of text. Journalists are also increasingly striving to highlight the most relevant information in their articles so only reading these highlights and the headline, can we have a "frame or the knowledge of the most valuable information parts" about this subject. The task of Information extraction involves extracting meaningful information from unstructured text data and presenting it in a structured format. Simplified, 'Named entity recognition' provides a solution

[2]BERT: `https://github.com/google-research/bert`

[3]RNN: `https://www.ibm.com/cloud/learn/recurrent-neural-networks`

[4]NLTK documentation: `https://www.nltk.org/`

for understanding text and highlighting categorized data from it. Where we can be defined different methods of the Named entity recognition extraction like 'Lexicon approach' or 'Rule-based systems' or even 'Machine learning based system'. By performing these analyzes, we can get deeper, information-supported sentiment results that can provide the foundation for many other researches.

In the IE area, 'Part of Speech' (POS) tagging based analyzes and 'Dependency Graph' generation were performed, followed by NER analysis. With the POS tagging, we determined which words people use most often in positive and negative tweets, and also we examined what 'stopwords' occur in these cases. With the help of the 'Dependency Graph' we looked at what was the most positive tweet in the given analysis, how this tweet is structured. Then, in the NER analysis, we expanded all of this and tried to get a picture of what the differences were in the case of positive and negative tweets. What people, places, and more were mentioned in their tweets related to that topic.

The 'spacy'[5] ibrary provided the basics for the analyzes. Like the NER analysis, which based on default trained pipelines from 'spacy', which can identify a variety of named and numeric entities, including companies, locations, organizations and products.

The RNN model was built and taught using the libraries and capabilities provided by 'Tensorflow'[6] and 'Keras' [7]. The DataSet is created and cleaned by a our written scraper script which use the Twitter API. This script always providing the most up-to-date data is possible in a given time period in a given topic. (covid)

[5]Spacy:`https://spacy.io/`

[6]Tensorflow: `https://www.tensorflow.org/`

[7]Keras: `https://keras.io/`

Chapter 2

Related Works

2.1 Previous works in similar themes

The *Social media sentiment analysis based on COVID-19* [1] paper can be considered the basis and starting point of this work, this is why this title was chosen.

Where we conclude and analyse the sentiments and manifestations (comments, hastags, posts, tweets) of the users of the Twitter social media platform, based on the main trends (by keyword, which is mostly the 'covid' and coronavirus theme in this article) with Natural Language Processing and with Sentiment Classification using Recurrent Neural Network. Where we analyse, compile, visualize statistics, and summarize for further processing. The trained model works much more accurately, with a smaller margin of error, in determining emotional polarity in today's 'modern' often with ambiguous tweets.

The other basis and starting point is the *Prediction of stock values changes using sentiment analysis of stock news headlines* [2] paper. Where we cover the topic of the stock value changes and predictions of the stock values using fresh scraped economic news about the companies. We are focussing on the headlines of economic news. We use numerous different tools to the sentiment analysis of the headlines. We consider BERT as the baseline and compare the results with three other tools, VADER, TextBlob, and a Recurrent Neural Network, and compare the sentiment results to the stock changes of the same period. The BERT and RNN were much more accurate, these tools were able to determine the emotional values without neutral sections, in contrast to the other two tools. Comparing these results with

the movement of stock market values in the same time periods, we can establish the moment of the change occurred in the stock values with sentiment analysis of economic news headlines.

By combining, rethinking and supplementing these two works, we can analyze people's emotional attitudes and manifestations on a given topic in a whole new approach. An excerpt from the thesis is published as *Information Extraction and Named Entity Recognition Supported Social Media Sentiment Analysis during the COVID-19 Pandemic* [3], which provides an opportunity for further work on the topic.

2.2 Sentiment analysis in social media

Due to the great popularity of Twitter, it can provide data for many researchers. Like in *Sentiment Analysis for Social Media* [4] and *Deep Learning for Information Triage on Twitter* [5] where the authors works on the scope of information exchange or triaging on Twitter in a variety of situations. Which is based on the need for different types of information after different events have occurred. In terms of events, we can think of disasters or political events, and so on. This information is then classified according to credibility and then classified into primary and secondary information category. Where the first is from the first hand and the secondary category is retweet, etc. The classification will be presented including the proposed one based on convolutional neural networks.

The authors in *Deep Convolution Neural Networks for Twitter Sentiment Analysis* [6] introduce a word embedding method obtained by unsupervised learning based on large twitter corpora, this method using latent contextual semantic relationships and co-occurrence statistical characteristics between words in tweets, with the integration into a deep convolutional neural network.

Many people use different social media platforms as news sources, which is a significant reason to analyze them. By relying on this data, people may run the risk of drawing erroneous conclusions when reading the news or planning to purchase a product. Therefore, there is a need for systems that are able to detect and classify emotions and help users find the right information on the web. Therefore, in the

A Domain-Independent Classification Model for Sentiment Analysis Using Neural Models [7], the authors propose a general approach to sentiment analysis that able to classify the sentiments of different datasets robustly. The model is trained on the IMDb dataset and then tested on three different datasets.

There are a numerous ways to measure public opinion on social platforms, one approach users might have various degrees of influence depending on their participation in discussions on different topics. In the *Combining Post Sentiments and User Participation for Extracting Public Stances from Twitter* [8], the authors combining sentiment classification and link analysis techniques for extracting stances of the public from social media (Twitter). The authors also look into the participation of popular users in social media by adjusting the weight of users to reflect their relative influence on interaction graphs, and used deep learning methods such as Long Short-Term Memory (LSTM) to learn the long-distance context.

The authors of the *Twitter Sentiment Analysis Using Hybrid Cuckoo Search Method* [9] used the following approach, they proposed a novel metaheuristic method (CSK) which was based on K-means and cuckoo search. The method provides a new way to find optimal cluster heads based on the sentimental content of the Twitter dataset.

For companies, it may be worthwhile to perform sentiment analysis to assess the effects based on financial texts written by different news portals just like foreign currency exchange rate movements in the paper of *An Intelligent Event-Sentiment-Based Daily Foreign Exchange Rate Forecasting System* [10]. In the case of English texts, this is clearly more common and produces fairly accurate results. The authors of *Financial Context News Sentiment Analysis for the Lithuanian Language* [11] perform a similar sentiment analysis on texts provided by Lithuanian portals. They performed this analysis using two of the most commonly used traditional machine learning algorithms, Naive Bayes and support vector machine (SVM), and one deep learning algorithm, a long short-term memory (LSTM). Plus they used the optimization of the hyperparameters which was performed by grid search to find the best parameters for each classifier. The results of the applied machine learning algorithms show that the highest accuracy is obtained using a non-balanced dataset, via the multinomial Naive Bayes algorithm.

In the *Sentiment Analysis of Social Images via Hierarchical Deep Fusion of Content and Links* [12], the authors combined the visual content with different semantic fragments of textual content through a three-level hierarchical LSTMs (H-LSTMs) to learn the inter-modal correlations between image and text at different levels. To exploit the link information effectively, the linkages among social images are modeled by a weighted relation network and each node is embedded into a distributed vector. The authors demonstrated the effectiveness of our approach on both machine weakly labeled and manually labeled datasets.

Sentiment analysis plays / can play a significant role in improving service and product quality, and can help develop marketing and financial strategies to increase company profits and customer satisfaction. In the *Sentiment Classification from Unstructured Reviews Using Ensemble Classifier* [13] we can find out a voting classifier Gradient Boosted Support Vector Machine (GBSVM) which is constituted of gradient boosting and support vector machines.

Polarity detection is key for applications such as sentiment analysis. The problem with existing word embedding methods is that they often do not differentiate between synonymous, anonymous, and unrelated word pairs. In *A Polarity Capturing Sphere for Word to Vector Representation* [14], the authors propose an embedding approach that solves the problem of polarity. The approach is based on embedding the word vectors in a sphere, where the point product between the vectors represents the similarity.

The sentiment analysis can be represented by the supporting vector machine. In *Application of Support Vector Machine (SVM) in the Sentiment Analysis of Twitter DataSet* [15] the authors proposes a Fisher kernel function method based on probabilistic latent semantic analysis that improves the kernel function of the support vector machine. With this method, latent semantic information including probabilistic characteristics can be used as classification characteristics and to improve the effect of classification on support vector machines. The authors of *Emotion AI-Driven Sentiment Analysis: A Survey, Future Research Directions, and Open Issues* [16] and *Gender Classification Using Sentiment Analysis and Deep Learning in a Health Web Forum* [17] give an overview of emotion AI-driven sentiment analysis in various domains. In the considered sample data, the aspect-based ontology approach, Support

Vector Machine, and term frequency achieved high accuracy and provided better sentiment analysis results in each category. In addition, we can get to know about the ensemble learning model of sentiment classification which was presented in *Deep Learning Application to Ensemble Learning—The Simple, but Effective, Approach to Sentiment Classifying* [18], also known as CEM (classifier ensemble model). The experiments conducted based on different real datasets found that they sentiment classification system is better than traditional machine learning techniques, such as Support Vector Machines.

In the *Lexicon-Enhanced Attention Network Based on Text Representation for Sentiment Classification* [19] they propose a lexicon-enhanced attention network (LAN) based on text representation to improve the classification of sentiments. Combining the sentiment lexicon with attention mechanism in the word embedding module, they can obtain the sentiment-aware word embeddings as the input of deep neural network, which bridges the gap between sentiment linguistic knowledge and deep learning methods.

BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding [20] present a model which became one of the most significant tool of the natural language processing. BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks.

The authors of the *Using Social Media to Mine and Analyze Public Opinion Related to COVID-19 in China* [21] analyzed the public opinion related to COVID-19 in China based on social media. The number of Weibo (a Twitter-like microblogging system in China) texts has changed over time for different themes and subthemes that correspond to different developmental stages of the event. The spatial distribution of Weibo for COVID-19 was mainly concentrated in the urban agglomerations of Wuhan, Beijing-Tianjin-Hebei, Yangtze River Delta, Pearl River Delta, and Chengdu-Chongqing. There is a synchronization between frequent daily discussions on Weibo and the trend of the COVID-19 outbreak in the real world. The reaction of the population is very sensitive to the epidemic and major social events,

especially in urban agglomerations with convenient transport and large populations.

2.2.1 Sentiment analysis in COVID-19

For the authors of *Sentiment Analysis and Its Applications in Fighting COVID-19 and Infectious Diseases: A Systematic Review* [22] the aim of the research was to review and analyze the incidence of different types of infectious diseases such as epidemics, pandemics, viruses or outbreaks over the last 10 years to understand the application of sentiment analysis and to obtain key literature findings.

In addition, analyzing social media, in the *COVID-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA.* [23] We can learn from 1.2 million tweets which was collected across five weeks of April–May 2021, what emotions and attitudes have evoked from the people about different vaccines to Covid-19 as response. Where they deploy natural language processing and sentiment analysis techniques to reveal insights about Covid-19 vaccination awareness among the public. Where there is a clear positive attitude of people towards vaccinations, despite the negative news that initially appeared. In addition, in the case of the security measure, people were more positive about the various topics. The authors also use TextBlob and VADER to the sentiment classification.

The other huge social platform of our time is undoubtedly the Instagram. In the *Mining Textual and Imagery Instagram Data during the COVID-19 Pandemic* [24] where the authors examined the instagram entries of three major vaccine manufacturers. In the comments under the posts of these companies, the users' intention to comment was mainly to make general statements, communicate facts and share experiences, which in this context meant their post-vaccination experience. In most cases, users do not ask help or advice about COVID-19 or the vaccination process. The best performing algorithms for intent classification were Support Vector Machines and Random Forest, and the polarity analysis showed a highly polarized - more neutral and negative result. Similarly, in the *Classification of Cyber-Aggression Cases Applying Machine Learning* [25] where the authors applied Random Forest and OneR to classify of offensive comments, or in the *Identifying Polarity in Tweets from an Imbalanced Dataset about Diseases and Vaccines Using a Meta-Model Based*

on Machine Learning Techniques [26] where the authors analyze the polarity of tweets with a particular vaccine and related diseases.(The set of tweets retrieved for a study about vaccines and diseases during the period 2015–2018.) The results are showed that the highest accuracy was achieved with the Random Forest model.

The authors of the *COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model* [27] analyze the Sina Weibo popular Chinese social media site posts, where the BERT model is adopted to classify sentiment categories and TF-IDF (term frequency-inverse document frequency) model is used to summarize the topics of posts. The analyses provide insights on the evolution of social sentiment over time and the topic themes connected to negative sentiment on the social media sites.

2.3 Information extraction and Named entity recognition

Automating clinical de-identification through deep learning techniques has been shown to be less effective in languages other than English due to dataset scarcity. Therefore, a new Italian identification data set was created from the COVID-19 clinical records provided by the Italian Radiological Society (SIRM). Two multilingual in-depth learning systems have been developed for this low-resource language scenario: the objective of *Crosslingual Named Entity Recognition for Clinical De-Identification Applied to a COVID-19 Italian Data Set* [28] is to investigate their ability to transfer knowledge between different languages while maintaining the necessary features to correctly perform the Named Entity Recognition task for de-identification.

The development of COVID-19 automated detection systems based on natural language processing (NLP) techniques can be a huge help to support clinicians and detect COVID-19-related abnormalities in radiological reports. In *COVID-19 Detection in Radiological Text Reports Integrating Entity Recognition* [29], the authors propose a text classification system based on the integration of different sources of information. The system can be used to automatically predict whether or not a patient has radiological findings consistent with COVID-19 on the basis of radiolog-

ical reports of chest CT. To train the text classification system they apply machine learning approaches and named entity recognition.

The authors of *Comprehensive Named Entity Recognition on CORD-19 with Distant or Weak Supervision* [30] created this CORD-NER dataset with comprehensive named entity recognition (NER) on the COVID-19 Open Research Dataset Challenge (CORD-19) corpus, which covers many new entity types related to the COVID-19. CORD-NER annotation is a combination of four sources with different NER methods.

Free-text clinical notes can contain critical information to address different issues. So, we need data-driven, automatic information extraction models to use this text-encoded information in large-scale studies. *Extracting COVID-19 Diagnoses and Symptoms from Clinical Text: A New Annotated Corpus and Neural Event Extraction Framework* [31] introduces a new clinical corpus, called the COVID-19 Annotated Clinical Text (CACT) corpus, which contains 1472 notes with detailed notes describing the diagnosis, examination, and clinical presentation of COVID-19. The authors presented a span-based event extraction model that collectively extracts all observed phenomena and achieves high performance in identifying COVID-19 and symptom events with associated assertion values.

The huge amount of unstructured free-form text in medical records is a major barrier. An information extraction based approach has been described by the authors of the *Text Mining of the Electronic Health Record: An Information Extraction Approach for Automated Identification and Subphenotyping of HFpEF Patients for Clinical Trials* [32], which automatically converts unstructured text into structured data, which is cross-referenced against eligibility criteria using a rule-based system to determine which patients qualify for a major HFpEF clinical trial.

With X-ray images from patients with common bacterial pneumonia, confirmed Covid-19 disease, and normal incidents, was utilized for the automatic detection of the Coronavirus disease. The aim of the authors of *Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks* [33] is to evaluate the performance of state-of-the-art convolutional neural network architectures proposed over the recent years for medical image classification.

Besides the works discussed above, there are many other methods of sentiment

analysis and data analysis. In this thesis, we compare the results of sentiment analysis models, which was listed earlier (TextBlob, NLTK, RNN, BERT), and then perform further analyzes on the labeled data from RNN model to explore and explain this result in more depth. Such a comparison and further analysis had not been discussed in the related works.

Chapter 3

DataSet building and usage

There are several possible directions for providing data for analysis, from manual work to fully automated options, or created and released data by others for free usage. These all have advantages and possibly disadvantages.

3.1 Human effort

Perhaps one of the most accurate options for creating a dataset is to collect data (tweets) on a specific topic by human effort. The probability of mismatched tweet will be included into the dataset is minimal, of course, the human error factor still exists. That is the reason, why this solution option already questionable, it is really worthwhile to create the dataset this way. This method is one of the slowest and most expensive, arguably obsolete method for dataset creation. Due to the existing human error factor and cost, it is definitely worth moving in a different direction in data preparation and dataset creation.

Furthermore, our main goal is to create the fastest and most up-to-date dataset as possible, where we can perform immediate analyzes on up-to-date data after specific events or news.

3.2 Existing datasets

You can find many pre-made and maintained datasets on the internet, you can even think of the possibilities provided by Kaggle. In the case, we must take into

account that, this datasets only consisting of a larger number of tweets without any specific topics. Of course, there are also topic-specific datasets of tweets, but here the speficield topic and the given time period of the datasets causes the problems.

As we wrote earlier, the goal is to use the most up-to-date datasets as possible for analysis, so that if there is any news or announcement on the topic, we can immediately run new analyzes using these fresh tweets, which were written as response to this new event on social media. Nowadays, things change very quickly, one announcement can change a lot, especially in the field of covid, traditional polls are slow and outdated. Furthermore, waiting for someone else to compile and publish a dataset that includes the period of time, which relevant to us, the results of the analysis may already be completely irrelevant or outdated. This is where the various scraping and api options open up to create datasets covering a given topic in a structured way as quickly as possible. Which provides that, we can really analyze the "average user's" reaction and emotional attitude to certain announcements and news, what effects it has had.

3.3 Scraping and APIs

With the help of APIs provided by companies and various web scraper and helper libraries, the dataset creation can be greatly accelerated and simplified. In the case of scraping, we should definitely mention the 'BeautifulSoup', 'urlopen' and 'Request' libraries, which makes easier to write dataset building scripts. In addition to these solutions, various APIs are available, such as the Twitter API, what we use to create fresh datsets. Twitter[8] provides an opportunity to create a dataset in this way, in a completely simple and legal way. (These libraries, what we have mentioned before can be linked to python, but there are many other languages with similar useful libraries.)

This allows us to create a fully automated fast dataset creation method, which is cost effective, optimized for our logic and the error factor is minimal too.

[8]Twitter Developer Platform: `https://developer.twitter.com/en/docs`

3.4 Dataset building with Twitter API

Our script (Python) only needs the given topic as a keyword, (which is the 'covid') a start and end date, finally a limit number for number of tweets to compile a dataset on the topic we specify. In the field of language, we use English, but this can also be changed as a parameter. The 'tweetpy'[9] library was used to write the script. While creating the dataset, we also perform a simple cleaning task on the dataset as well.

```
def dataset_building(self, tag, limit, begin_date, end_date, lang):
    with open('result.csv', mode='wt', encoding='UTF-8', newline='') as file:
        w = csv.writer(file)
        w.writerow(['Time', 'UserName', 'Tweet_text', 'All_Hashtags', 'Followers_count'])


        for tweet in tweepy.Cursor(self.api.search, q=tag + ' -filter:retweets', lang=lang,
        tweet_mode='extended', since=begin_date, until=end_date).items(limit):
            w.writerow([tweet.created_at,
                        tweet.user.screen_name,
                        self.clean_tweet(tweet.full_text),
                        [e['text'] for e in tweet._json['entities']['hashtags']],
                        tweet.user.followers_count])


def clean_tweet(self, text):
    return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t]) | (\w +:\ / \ / \S +)", " ", text).split())

```

Code 3.1: Part from the dataset builder script

The dataset consists of the following values: 'Time' - as the time, when the tweet was written, 'UserName' - the name of the user who wrote the tweet, 'Tweet text' - the text of the tweet, the most important data for us, 'All Hashtags' - a list of hashtags used in the tweet and finally 'Followers Count' - the number of followers of the user who wrote the tweet.

It is noticeable, in addition to the text of the particular tweet, we also saved additional data such as the follower count of the users and used hashtags. The main reason for this is, when we use information extraction after sentiment analysis, we can analyze the most positive and most negative tweets separately, what were the two most extreme opinions, and how many people was reached with this opinions

[9]tweetpy: `https://docs.tweepy.org/en/stable/`

based on only the users follower count, without any retweet. (Providing an option as a basis for further research.)

Thus, it is ensured that the most up-to-date dataset is available for each analysis in a fully controlled manner. On the given topic, within specified time interval, with the specified size of the dataset.

Chapter 4

Methodology

4.1 Analysis Diagram

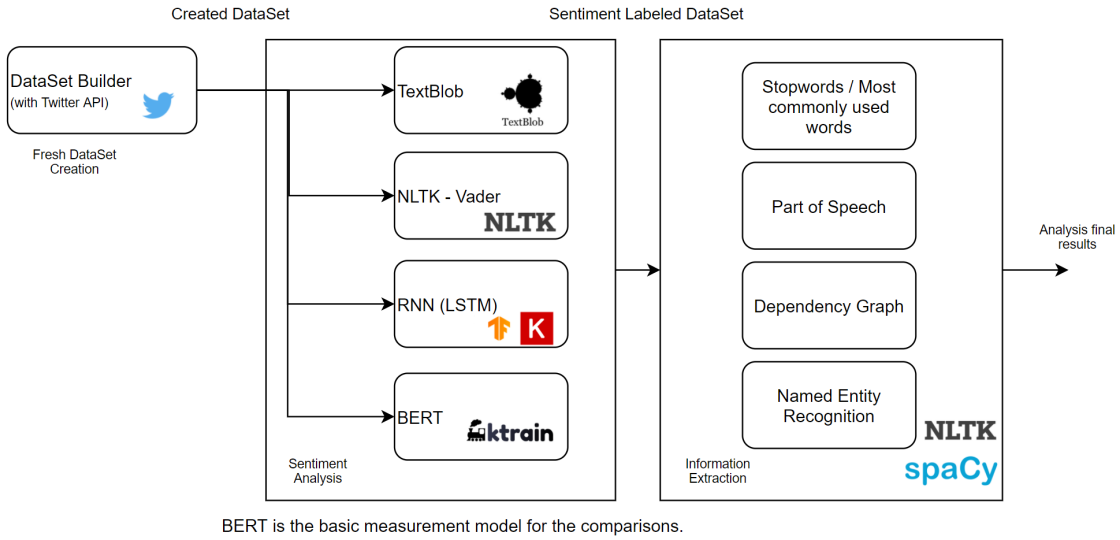


Figure 4.1: Analysis process

The Figure 4.1 shows the whole process of analysis. In each case, we perform the sentiment analyzes on the freshly created dataset. As mentioned earlier, BERT provides a kind of comparative result. (BERT uses the transformer mechanism, which is an outstanding achievement and a remarkable breakthrough of the current NLP.) Then, we continue the analyzes on the dataset labeled by the "X" model, which was the closest to the BERT results. By a deeper analysis (Information extraction) of these results, we can get an even more concise picture of people's emotional state in the given period of time.

4.2 Sentiment analysis

There are several options for performing sentiment analysis. The scale extends from labeling with human work to machine and deep learning. The natural language processing (NLP) is a very interesting topic, that can even be mentioned as a separate or unique part of artificial intelligence.

As mentioned earlier, we perform analyzes on covid themed tweets from different time intervals using TextBlob, NLTK-VADER, RNN and BERT models. The results of BERT are used as a kind of benchmark against the other models. TextBlob and NLTK - VADER are third-party easy to integrate solutions. The RNN model is our model, what we have built with 'Tensorflow' and 'Keras' frameworks. For the implementation of BERT we used the 'ktrain'[10] library to simplify this model implementation.

4.2.1 TextBlob

TextBlob is a powerful NLP library for Python that builds on NLTK and provides an easy-to-use interface to the NLTK library. With this tool, we can perform variety of NLP tasks, from tagging parts of speech to sentiment analysis, and from language translation to different text classifications, but we focus on sentiment analysis. TextBlob is a lexicon-based approach and offers two emotional metrics, polarity and subjectivity, it ignores the words that does not belongs to the lexicon and focuses only to the known words to produce a score for polarity and subjectivity measures. If we perform an sentiment analysis, we actually determine the polarity value of the sentences, where this value can be between -1 and 1. The data can be labeled with the appropriate sentiment value (positive, negative, or neutral). Here, we have expanded the given scale for a more detailed result with 'strongly positive and negative' and 'weakly positive and negative' options, and also adjusted accordingly the polarity categories. Where the polarity value is closer to +1 that means more 'strongly' positive sentiment, if this value is closer to -1 that means more or 'strongly' negative sentiment, 0 can be defined as neutral sentiment on this extended slate.

[10]ktrain: `https://pypi.org/project/ktrain/`

4.2.2 Natural Language Toolkit (NLTK) - Valence Aware Dictionary and sEntiment Reasoner (VADER)

NLTK stands for Natural Language Toolkit. This toolkit is one of the most powerful NLP libraries which contains packages to make machines understand human language and reply to it with an appropriate response. Now, we use VADER Lexicon and focus on sentiment analysis with the 'SentimentIntensityAnalyzer'. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a part of the Natural Language Toolkit (NLTK) packages, it is a lexicon and rule based sentiment analysis tool commonly used to analyze the sentiments expressed in social media, but it works well on texts from other domains as well.

VADER takes into account the polarity and intensity of emotions expressed in context, and performs particularly well when analyzing unique characters used in tweets, such as emoticons or slang. This tool produces a compound score, which scales between -1 and +1 just like in TextBlob.

4.2.3 Recurrent Neural Network (RNN)

When we talk about traditional neural networks, all outputs and inputs are independent of each other. But in the case of recurrent neural networks, the hidden layer on the previous run become part of the input to the same hidden layer in the next run.

What is Recurrent Neural Network (RNN)[11] - A neural network that is intentionally run multiple times, where parts of each run feed into the next run. Specifically, hidden layers from the previous run provide part of the input to the same hidden layer in the next run. Recurrent neural networks are particularly useful for evaluating sequences, so that the hidden layers can learn from previous runs of the neural network on earlier parts of the sequence.

For example, the following figure of Google shows a recurrent neural network that runs four times. Notice that the values learned in the hidden layers from the first run become part of the input to the same hidden layers in the second run. Similarly,

[11] `https://developers.google.com/machine-learning/glossary/#recurrent_neural_network`

the values learned in the hidden layer on the second run become part of the input to the same hidden layer in the third run. In this way, the recurrent neural network gradually trains and predicts the meaning of the entire sequence rather than just the meaning of individual words.

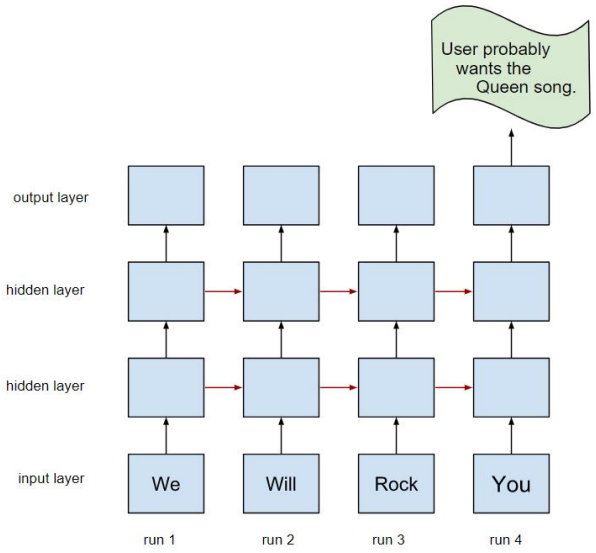


Figure 4.2: Recurrent neural network

We can mention as an advantage of RNN models:

- RNN can process inputs of any length.
- RNN model is modeled to remember each information throughout the time which is very helpful in any time series predictor. Even if the input size is larger, the size of the model does not increase.

As disadvantage we can mention:

- Due to its recurrent nature, the computation can be slow.
- The training can be difficult.

Model

We have used tools provided by Keras and Tensorflow to build the model. Where we created a Sequential model by passing a list of layer instances. (A Sequential model is appropriate for a plain stack of layers where each layer has exactly one input tensor and one output tensor.) The first layer was the Embedding layer which,

can be used for neural networks on text data. Embedding layer enables us to convert each word into a fixed length vector of defined size. It requires that the input data be integer encoded, so that each word is represented by a unique integer.

Then we used Bidirectional layer[12], which is a layer wrapper. This wrapper takes a recurrent layer as an argument. It also allows us to specify the merge mode, that is how the forward and backward outputs should be combined before being passed on to the next layer. The default mode is to concatenate, and this is the method often used in studies of bidirectional LSTMs. (We used the default mode.) We used LSTM layers with the Bidirectional layers. The Long Short-Term Memory (LSTM) is an RNN 'architecture', this networks are type of recurrent neural network, and capable of learning order dependence in sequence prediction problems.

Next is the Dense and Dropout layers. A dense layer is a classic fully connected neural network layer, each input node is connected to each output node. A dropout layer is similar except that when the layer is used, the activations are set to zero for some random nodes. This is a way to prevent overfitting. We used a Dense layer with 'relu' activation, then a Dropout layer, and again a Dense layer with 'sigmoid' activation.

Difference between RNN and LSTM

All RNN has a feedback loop in the recurrent layer. This allows them to maintain information in "memory" over time. However, it can be difficult to train standard RNNs to solve problems that require learning long-term temporal dependencies. This is because the gradient of the loss function decays exponentially with time, this is called the disappearing gradient problem. LSTM networks are a type of RNN that uses special units in addition to standard units. LSTM units contain a "memory cell" that can maintain information in memory for a long time. A set of gates is used to control when information is written into memory, when it is output, and when it is forgotten. This architecture allows them to learn longer-term dependencies.

[12]Bidirectional layer: `https://keras.io/api/layers/recurrent_layers/bidirectional/`

Trained model information

The RNN model was trained based on an IMDB review dataset.[13] The dataset comes from the official tensorflow catalog, which provide 25,000 highly polar reviews for training, and 25,000 for testing. We used the "subwords8k" option with 8185 vocab size. (The data consists of labels and texts.) In the test and train dataset sections we used shuffle method as well.

The accuracy of our model was 84.7% on the test dataset. The model is not overfitting and it is more generalized and can make good predictions for new data. Furthermore, we can mention that the Buffer Size was 10000 and the Batch Size was 64. In the 'compile' the loss argument was "binary crossentropy" with the "Adam" optimizer.

We also save our trained models in '.h5' format. This previously mentioned model was used to analyze further tweets.

The 'positive', 'neutral', 'negative' labels were expanded in this case as well,just like in the previous models with 'strongly positive and negative' and with 'weakly positive and negative' labels. Furthermore, it should be mentioned, unlike the previous models, the sentiment value (predicted compound value) here scales between 0 and +1, instead of -1 and +1 values.

4.2.4 Bidirectional Encoder Representations from Transformers (BERT)

Unlike traditional NLP models, which follow a one-way approach, i.e. reading the text from left to right or right to left, BERT reads the entire word sequence at once. BERT makes use of a Transformer, which is essentially a mechanism for building relationships between words in a dataset. In a simplest form, BERT consists two processing models - an encoder and a decoder. The encoder reads the input text and the decoder generates the predictions. However, since the main purpose of BERT is to create a pre-trained model, the encoder takes precedence over the decoder. BERT is a remarkable breakthrough in NLP.

[13]Dataset: `https://www.tensorflow.org/datasets/catalog/imdb_reviews`

As we have mentioned earlier, the BERT model was implemented with the capabilities provided by the 'ktrain' library, which is a lightweight wrapper for Tensorflow and Keras. The full concept of BERT was developed and published by Google, which has made significant progress in many areas of NLP. The significant development of the google translate can be attributed to this as well.

Model

In the case of BERT, the model was created using the ktrain "text .text classifier" method and then the "get learner" method. The "get learner" parameter received the "text .text classifier", train and validation data and the batch size which was 6.

About the data: 25,000 labeled reviews were used as a train dataset and also 25,000 labeled reviews were used as a validation dataset for the model, where the text column was 'Reviews', and the label columns was 'Sentiment'.

The training was done with the help of the "fit onecycle" method where the value of the learning rate parameter was 2×10^{-5}. (lr = 2e-5)

4.3 Information extraction

Information extraction is the process of extracting information from unstructured textual sources to enable finding entities and classifying or storing them in a database or preparing this information for further analysis so, the task of information extraction (IE) is to extract meaningful information from unstructured textual data and present it in a structured form.

In general, extracting structured information from unstructured texts involves the following main subtasks:

- Pre-processing of the text - where the text is prepared for processing with the help of computational linguistics tools such as tokenization, sentence splitting, morphological analysis, etc.

- Finding and classifying concepts - where mentions of people, things, locations, events and other predefined concepts are perceived and classified.

- Connecting the concepts - the task of identifying relationships between the extracted concepts.
- Unifying - this task is presenting the extracted data into a standard form.

4.3.1 Part of speech (POS)

We all know that sentences consist of words belonging to different parts of speech (POS). Some of these POS are: noun, pronoun, verb, adjective, adverb, preposition, conjunction, and intersection.

POS determines how a particular word works in the meaning of a particular sentence. For example, the word 'right'. In the sentence, "The boy was awarded chocolate for giving the right answer", 'right' is used as an adjective. While, in the sentence, "You have the right to say what you want," 'right' is treated as a noun.

POS tag of a word carries a lot of significance when it comes to understanding the meaning of a sentence. But, sometimes extracting information purely based on the POS tags is not enough. If we would like to extract the subject and object from a sentence, we cannot do that based on POS tags. For that, we need to look at how these words are related to each other.

There are several methods of the POS such as 'Rule-Based POS tagging', which method use contextual information to assign tags to unknown. Like, if an ambiguous/unknown word X is preceded by a determiner and followed by a noun, tag it as an adjective. Or 'Transformation-based tagging', where the tagger based on transformations or rules, and learns by detecting errors. Or even 'Stochastic (Probabilistic) tagging', which based on probability of certain tag occurring.

For the analyzes, we used the methods of the 'spacy' and 'nltk' libraries to perform the analyzes. The choice of 'spacy' was conscious, to use a library which is a popular choice in the industry as well, in addition to the scientific approach.

Spacy can make predictions of which tag or label most likely applies in this context, which based on trained pipeline and its statistical models. A trained component includes binary data that is produced by showing a system enough examples for it to make predictions that generalize across the language.

4.3.2 Dependency graph

Dependency parsing is the process of analyze the grammatical structure of a sentence and find out the related words and the type of relationship between them.

Again, we used the tools provided by 'spacy' library. Spacy has a syntactic dependency parser. The parser powers the sentence boundary detection, and lets us iterate over base noun phrases, or "chunks".

4.3.3 Named entity recognition

Named entity recognition (NER) is an information extraction task, which identifies mentions of various named entities in unstructured text and classifies them into predetermined categories, such as person names, organisations, locations, date/time, monetary values, and so on.

Terms that represent specific entities are more informative and have a unique context. Furthermore, they represent real world objects like people, places, organizations, etc., which are often proper names. Thus, NER is a prominent factor in information extraction that identifies named entities and segmenting them into appropriate classes.

Based on this, we can define the task of NER in these three steps: Detect a named entity, Extract the entity and Categorize the entity.

In the case of NER, several implementations can be used. 'Lexicon approach', which relies on a knowledge base called ontology and contains all terms related to a particular topic, grouped in different categoiries. The system looks for matches with named entities. 'Rule-based systems', which is a series of grammatical rules hand crafted by computational linguists. Where we can get results of high precision but low recall. Last but not least, 'Machine learning based systems', which builds an entity extractor and feeds the model with large volume of annotated training data. Here we need tagged and clean training data.

For named entity recognition we used 'spacy' library and 'displacy' visualizer.

Chapter 5

Sentiment analysis results

The analyzes were run in late August and early September. Accordingly, we defined time intervals (29 August 2021 and 31 August 2021, 2 September 2021 and 4 September 2021) and defined the topic keyword, which was the "covid" and set the dataset size to 500 tweets to build the datasets of 500 tweets from both September and August time intervals using the Twitter API Standard option.

We would like to present the methods of this analysis flow in the first place, we expect similar results with a larger amount of data as well. The reason for this period, this is the period of starting school in many countries. School may have already started or will start soon. It is a particularly important period in the knowledge of the next, fourth wave of covid.

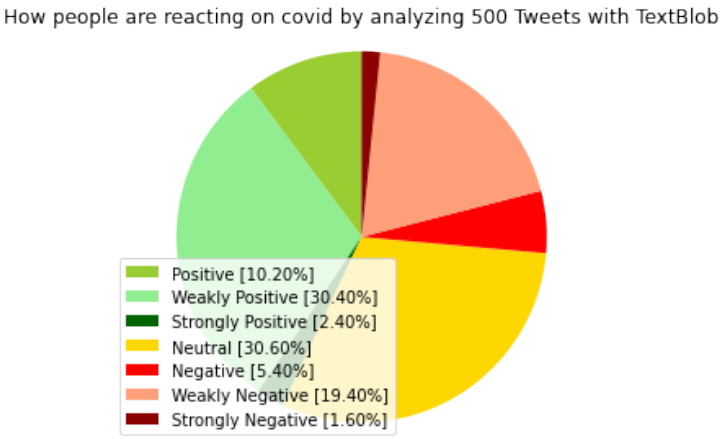
The classification of the tweets was based on the polarity and compound values, which was obtained from the different models. The models were used here as described in the methodology section. In the case of TextBlob and NLTK-VADER, the appropriate methods of the library were parameterized and used, in the case of RNN and BERT it was taught and used according to their previous descriptions.

The basic result is determined using the BERT transformer mechanism. We do not aim to compare all models with all other models, we would like to present and explain the methodological differences of the TextBlob, NLTK-VADER and RNN models, and then analyze the results of the model that best approached the results of BERT in more depth.

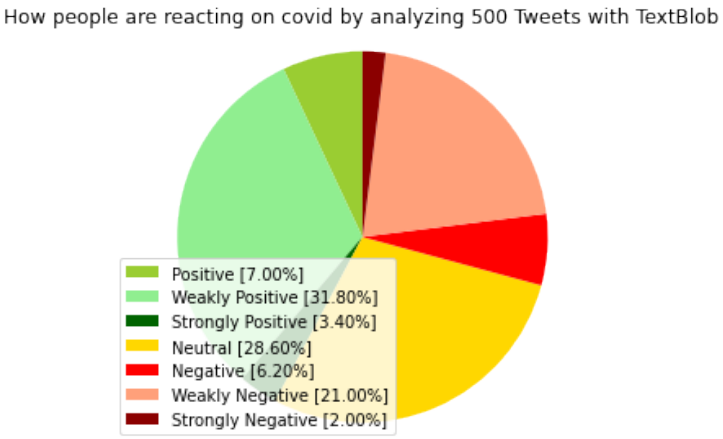
We expect the results of the RNN to be the closest to the results provided by BERT due to its methodological sophistication.

The interval for each category was properly defined, including the extended ('strongly', 'weakly') categories as well. Based on the values, the tweets were categorized and labeled in the appropriate category. In the case of BERT, the positive and negative categories were not further subdivided due to the role.

5.1 TextBlob



(a) TextBlob results from August



(b) TextBlob results from September

Figure 5.1: Analysis of sample of 500 tweets by TextBlob, using "covid" keyword. The time periods stands between 29 August 2021 and 31 August 2021, and 2 September 2021 and 4 September 2021.

In the Figure 5.1, the neutral value dominates in both examined periods, which significantly distorts the result. The August results in Figure 5.1 (a) shows a

30.60 percent of neutral value, which is significant. The results from September in Figure 5.1 (b) also shows that, the neutral value is 28.60 percent. A small shift can be seen in the case of the neutral values of the two studied periods, which was rather in the negative direction.

In both August and September, 'weakly positive' values dominated their category with 30.40 and 31.80 percentages. In the negative section we can see a similar 'weakly negative' dominance. Due to the significant neutral values, the results are not exactly the most favorable for further analysis.

5.2 Natural Language Toolkit (NLTK) - Valence Aware Dictionary and sEntiment Reasoner (VADER)

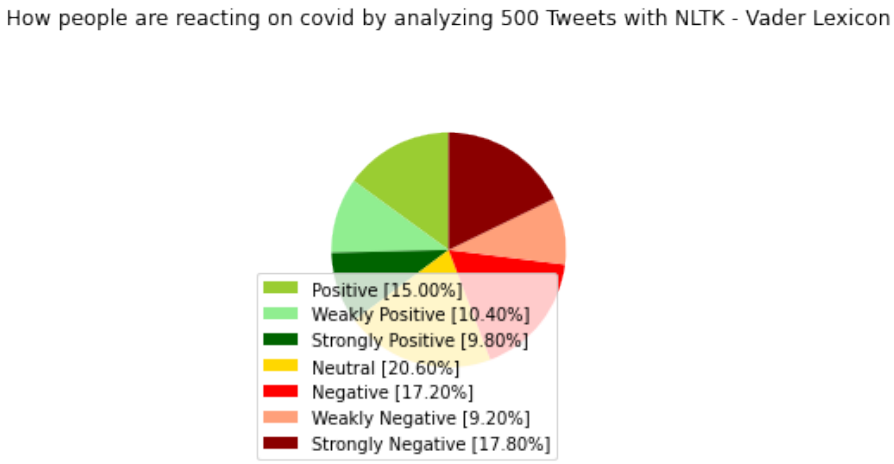
In the Figure 5.2, the results of NLTK - VADER show a significant improvement over the results of the previous TextBlob. It is enough to look only at the values of the neutral categories and see significant differences in the stages of the positive and negative parts.

In the case of the August result, which can be seen in part (a) of the Figure 5.2, the neutral value decreased significantly, now it is only 20.60 percent. Similarly, in part (b) of the Figure 5.2, the neutral value is 19.20 percent, compared to previous results, which reached 30 percent or it was very close to this value.

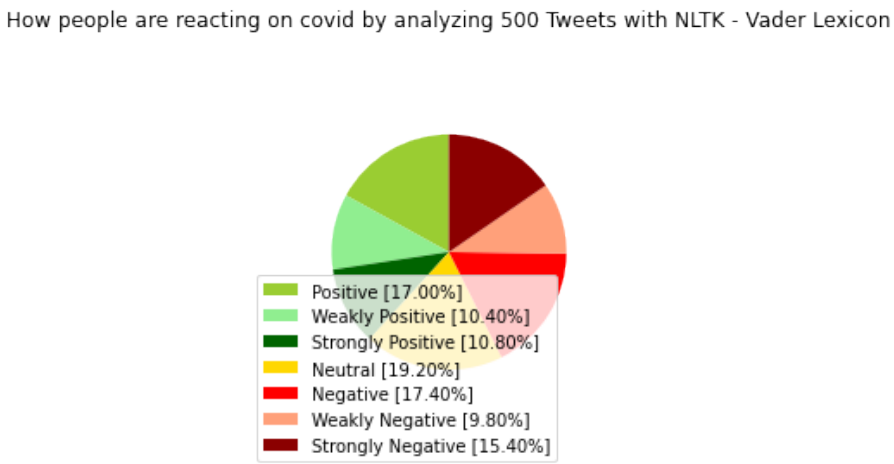
In the case of the August results, there are also significant differences within the positive parts, there is no longer such a 'weakly positive' dominance, due to the technological changes we can assume a more accurate result on the same datasets as we used in the case of TextBlob. Here, we can see 15 percent 'positive', 10.40 percent 'weakly positive' and 9.80 percent 'strongly positive' sentiment values. In September, 17 percent 'positive', 10.40 percent 'weakly positive' and 10.80 percent 'strongly positive' sentiment values were observed.

Similar movements can be observed in the negative sections, with 9.20 percent 'weakly negative', 17.20 percent 'negative' and 17.80 percent 'strongly negative' in August. In September, 9.80 percent were 'weakly negative', 17.40 percent were

'negative' and 15.40 percent were 'strongly negative' sentiment values. Despite a significant decrease in the neutral section, there is still too much data in this category, although we can definitely report an improvement over previous TextBlob results. The goal is to eliminate or considerably minimize of the neutral values in order to confirm the results with subsequent analyzes. A neutral value still makes the result little bit uncertain.



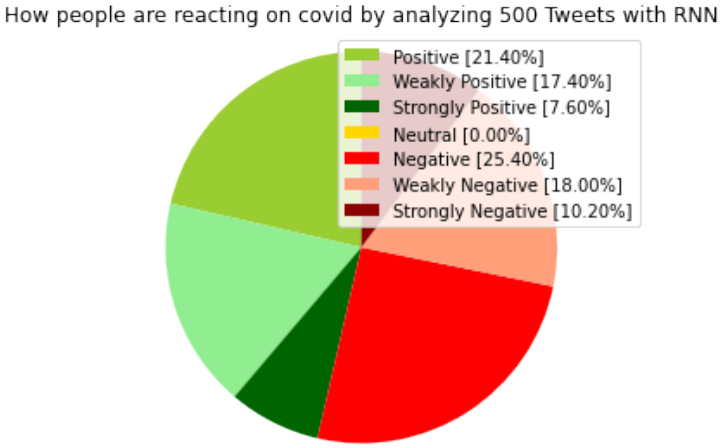
(a) NLTK-VADER results from August



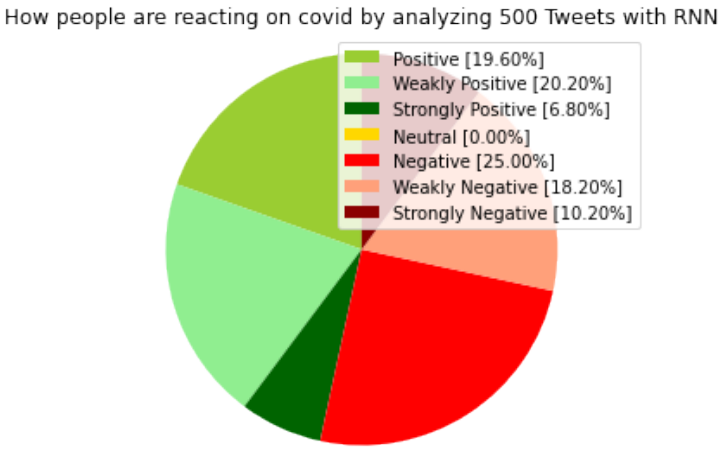
(b) NLTK-VADER results from September

Figure 5.2: Analysis of sample of 500 tweets by NLTK - VADER, using "covid" keyword. The time periods stands between 29 August 2021 and 31 August 2021, and 2 September 2021 and 4 September 2021.

5.3 Recurrent Neural Network (RNN)



(a) RNN results from August



(b) RNN results from September

Figure 5.3: Analysis of sample of 500 tweets by RNN, using "covid" keyword. The time periods stands between 29 August 2021 and 31 August 2021, and 2 September 2021 and 4 September 2021.

In the Figure 5.3, the results of RNN, compared to the previous two (TextBlob and NLTK-VADER), the neutral section is 0 percent in both August and September results, which is a significant improvement. In addition, small changes in distribution were observed in both the positive and negative sections compared to the previous models. In the case of the previous models, especially in the case of the NLTK-VADER results, there is a similarity in the result categories, both in positive and negative sections, the huge difference, of course is the neutral category, our model

was able to place all tweets in some category, as we expected, which significantly increases the establishment of a clearer picture of this specific periods.

The value of ‘strongly positive’ was 7.60 percent in August, down from 6.80 percent in September. The ‘positive’ section was 21.40 percent in August, but it was 19.60 percent in September, the ‘weakly positive’ values rising from 17.40 percent in August to 20.20 percent in September. Overall, in addition to the changes in ratios, the positive section increased by 0.2 percent overall, but there was a shift toward the ‘weakly positive’ section.

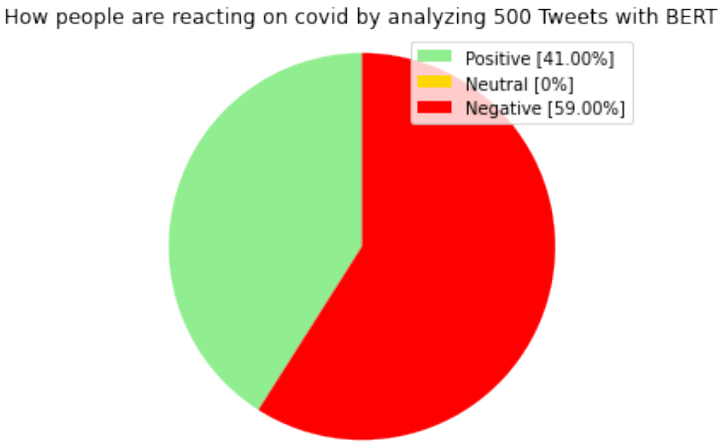
For the negative sections, the ‘strongly negative’ value was unchanged at 10.20 percent in both August and September. The ‘negative’ value fell from 25.40 percent in August to 25 percent in September. The ‘weakly negative’ value rose from 18 percent in August to 18.20 percent. The small 0.2 percent increase in positive values, and even in the case of minimal movements inside of negative section, still the negative sections represent a larger overall section, plus in the case of positive values, a shift toward a ’weakly’ value should be highlighted.

In summary, the results of the RNN model and the results of previous models shows a strong division, there is some kind of "boundary line" based on the studied periods, which is very difficult to move. People have their opinion about the pandemic, that has lasted for almost two years. Due to the significant neutral result seen in the TextBlob result, it is difficult to write a conclusion, but the results of the subsequent NLTK-VADER and then the RNN results, where the neutral values decreased significantly and then disappeared already give some picture. They show a shift in the negative direction, during the period under review the negative sections provided the higher percentage value overall, and in the case of the RNN model the shift to the already mentioned ‘weakly positive’ section can be highlighted again.

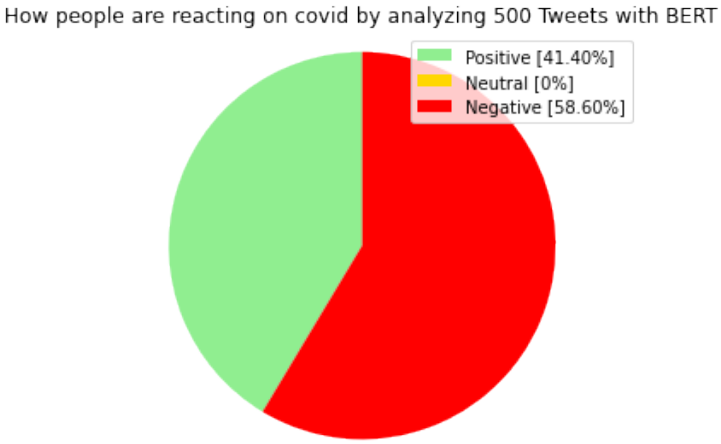
Vaccinations, and the relatively ’free summer’, also provide the basis for the positive parts in the studies, and the uncertainties of starting school and the fourth wave continue to maintain a more negative attitude.

5.4 Bidirectional Encoder Representations from Transformers (BERT)

As we have mentioned earlier, BERT was used as a kind of comparative result. Figure 5.4 shows the results obtained by BERT.



(a) BERT results from August



(b) BERT results from September

Figure 5.4: Analysis of sample of 500 tweets by BERT, using "covid" keyword. The time periods stands between 29 August 2021 and 31 August 2021, and 2 September 2021 and 4 September 2021.

Of course, without the neutral category, in the case of BERT, in contrast to the previously presented models, we did not further categorize the positive and negative categories, because we only consider the results of BERT as a benchmark

/ comparative result for comparison to the other models, so we obtained a classic, ‘positive’, ‘neutral’, ‘negative’ result in the same time periods as in the previous models.

For the BERT model, the ‘positive’ section was 41 percent in August, which rise to 41.40 percent in September. The ’neutral’ section was 0 percent according to our expectations. The ‘negative’ section was 59 percent in August, down from 58.60 percent in September. The results of BERT are mostly approximated by the results of the RNN model, which met our expectations. The aggregated positive result for RNN in August was 46.40 percent and the negative result was 53.60 percent. Similarly, in September where the aggregated positive score was 46.60 percent and the negative was 53.40 percent. Here, we can see a slight shift in the positive direction too, but overall the negative section dominates. This confirms the effectiveness of our RNN model, where we could also see a more detailed statement by further categorizing in positive and negative sections.

Based on the comparative results by BERT, we will perform further analyzes on the results of the RNN model, to gain more insight into about the sentiment results in this periods. To do this, we perform Information extraction (IE) and Named Entity recognition (NER) analyzes. For the TextBlob and NLTK models, due to the significant neutral categories, we did not include a comparison with the results of BERT.

Our goal with the help of these analyzes, is to give a comprehensive picture of this periods, what sentiment states people are in and what characterizes the tweets, which was written at that time. How the tweets were structured, what was mostly mentioned in them, what can be said about these tweets.

Chapter 6

Information extraction results

As we have mentioned earlier, these analyzes are performed on the results of the RNN model. After the sentiment analyzes, we have aggregated the extended sentiment categories, so the analyzes were performed on separate positive and negative datasets.

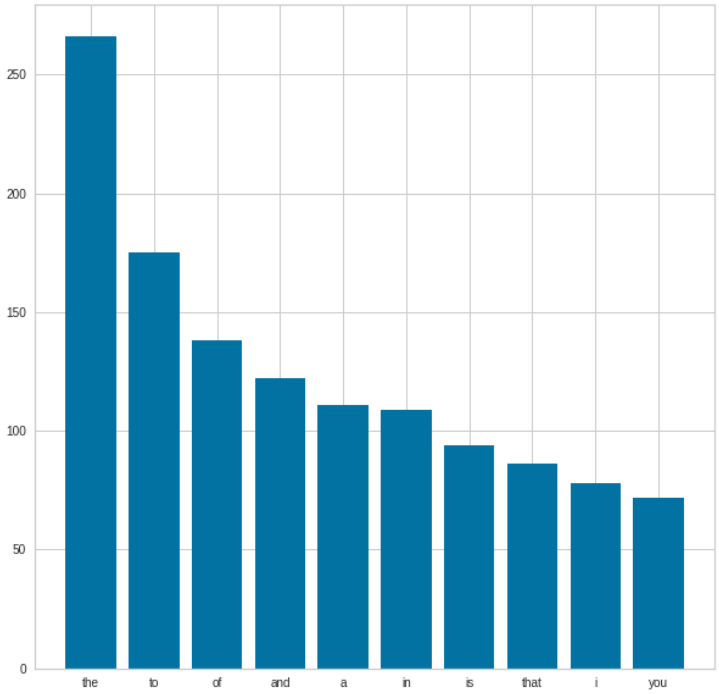
We started the POS analysis by comparing the 'stopwords' (what words occur in a positive and negative attitude) and then, we followed this with the most commonly used words in the same categorization approach. The "nltk.corpus" ('stopwords' download and inclusion in the analysis) and "nltk.tokenize" libraries were used.

This was followed by 'stopwords' removals and re-tokenization of tweets, with the entire POS analysis, which covering the positive and then the negative category. Finally, for the most followed positive tweets we built dependency graphs. The 'spacy', 'spacy - en core web sm' pipeline and the 'displacy' visualization option were used for these analyzes.

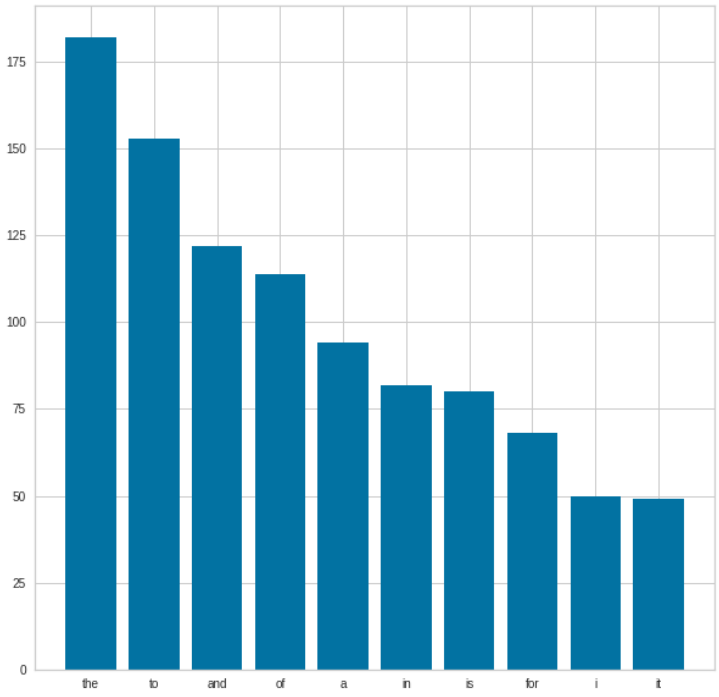
6.1 Stopwords and most commonly used words

6.1.1 August

Stopwords are the most common words in any natural language. For the purpose of analyzing text data and building NLP models, these stopwords might not add much value to the meaning of the document.



(a) 'Stopwords' from negative tweets

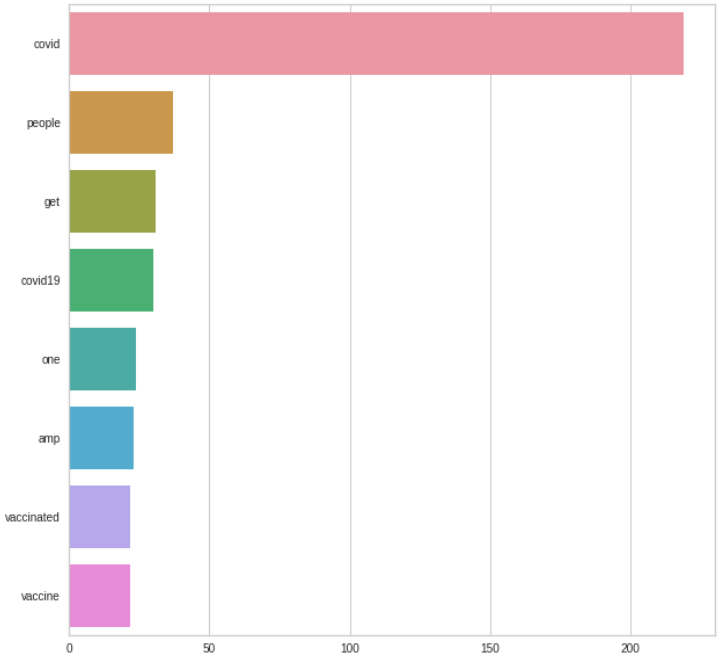


(b) 'Stopwords' from positive tweets

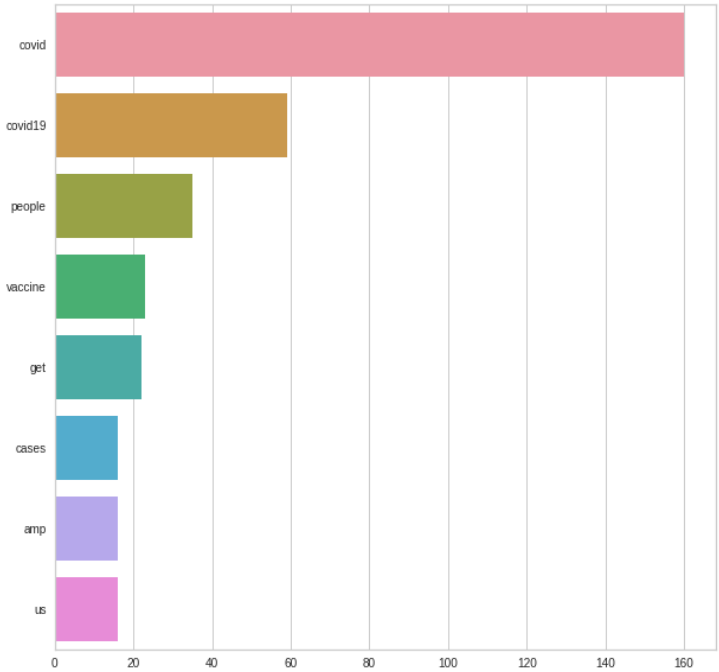
Figure 6.1: Negative and positive 'Stopwords'. The time periods stands between 29 August 2021 and 31 August 2021.

Figure 6.1 shows that 'stopwords' were very similar in both positive and negative tweets, in some cases we see changes in positions, such as "and" and "of". In addition,

in the negative case, the number of "the" can be highlighted.



(a) Most commonly used words from negative tweets



(b) Most commonly used words from positive tweets

Figure 6.2: Most commonly used words in negative and positive tweets. The time periods stands between 29 August 2021 and 31 August 2021.

The Figure 6.2 shows that for the most commonly used words, the word of

"covid" completely dominates in both negative and positive tweets. After that, there are differences, such as in the negative case, the word of "covid" is followed by the following words: "people", "get", "covid19" as opposed to the positive case, where the next three words are: "covid19", "people", "vaccine". In the negative case, the "vaccine" or "vaccinated" words appears only at the very end of the figure, in contrast in positive tweets, the "vaccine" word is the fourth most common word.

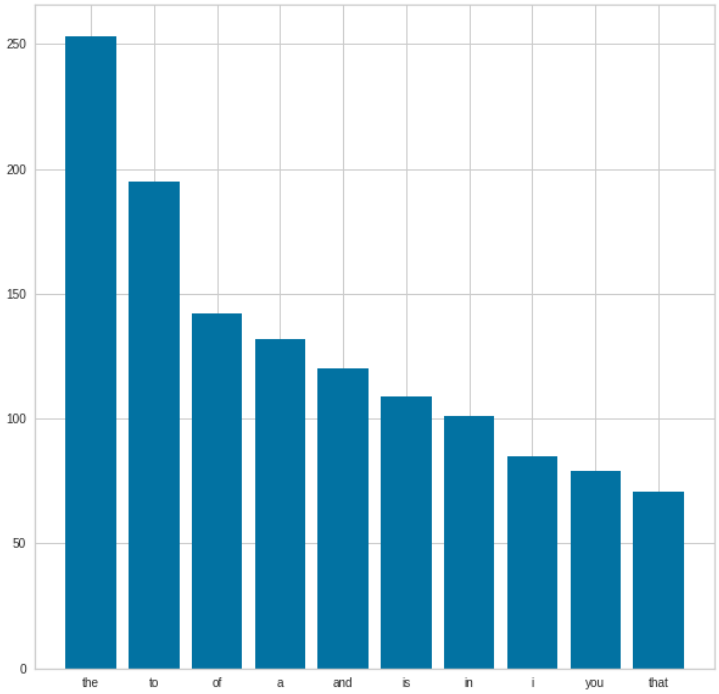
6.1.2 September

The Figure 6.3 shows what 'stopwords' occurred in September for negative and positive tweets. In the case of negative tweets, the first three 'stopwords' are the same as in August. In the case of positive tweets, the number of "the" 'stopwords' are increased, compared to the number of August. The third place of "a" can be mentioned, which was at the fifth place in August.

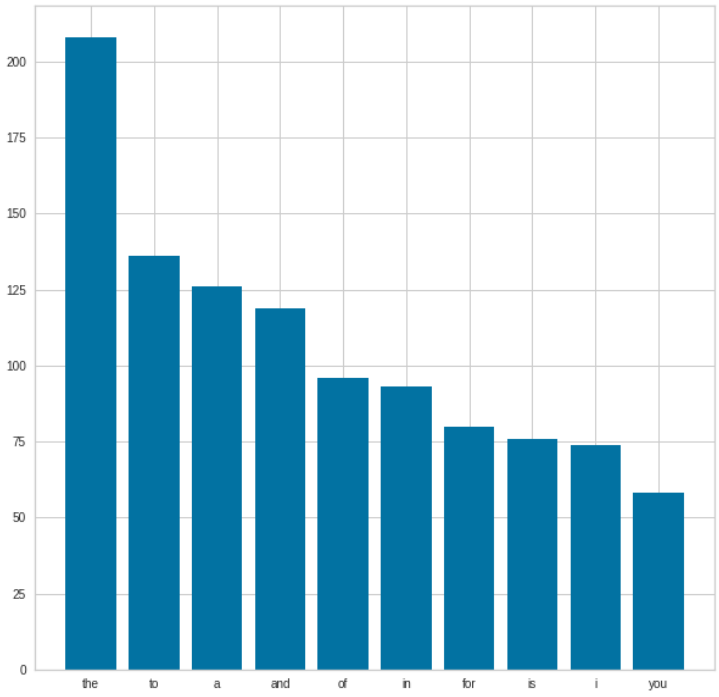
The Figure 6.4 shows that even in September, the word of "covid" completely dominated the tweets as well. In the case of negative tweets, the "covid" is followed by the following three words: "people", "covid19" and "get". In positive tweets, after the "covid" these three words coming: "covid19", "people" and "get". For both negative and positive words, the three most common words following the word of "covid" are the same. There is a difference in the order, for negative tweets the word of "people" is the first after the "covid" word, in positive words the "people" is the second in the queue after the "covid" word, the first is the "covid19".

In the case of negative tweets, it should be noted that the word of "vaccine" was significantly ahead compared to the August results. In contrast to the positive words, the word of "vaccine" slid significantly backwards, and the word "cases" moved forward, plus the word of "health" appears on the plot, which was not displayed previously.

Compared to August, only small changes are seen, the plots describe what words occur in a tweets on covid topic, and we can get an idea of about the topics people are interested in, and how they describe their opinions about it.

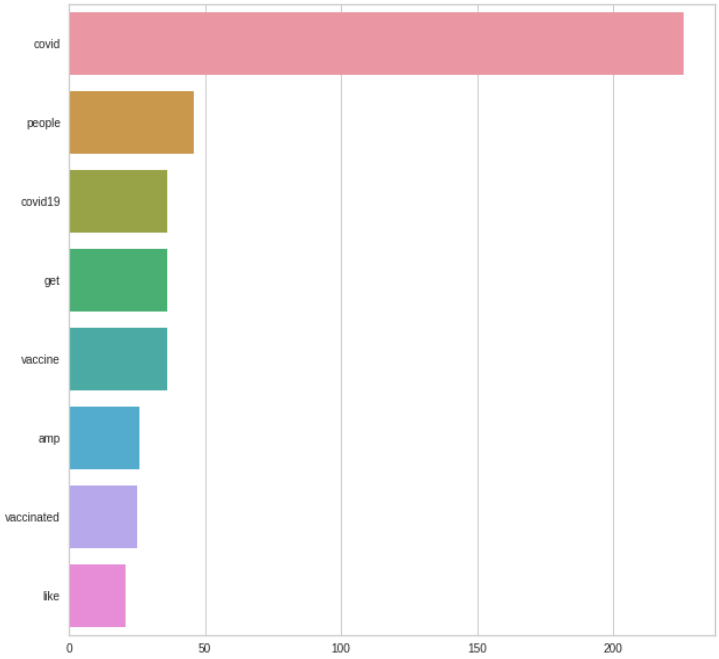


(a) 'Stopwords' from negative tweets

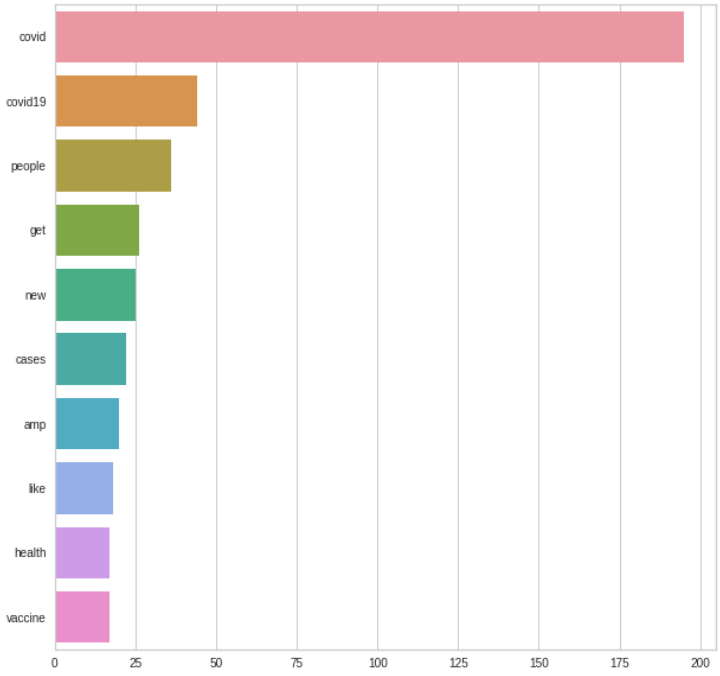


(b) 'Stopwords' from positive tweets

Figure 6.3: Negative and positive 'Stopwords'. The time period stands between 2 September 2021 and 4 September 2021.



(a) Most commonly used words from negative tweets



(b) Most commonly used words from positive tweets

Figure 6.4: Most commonly used words in negative and positive tweets. The time period stands between 2 September 2021 and 4 September 2021.

6.2 Part of Speech Tags and Dependency graph

After analyzing the different words, for both negative and positive tweets, it is definitely worth to make a full Part of speech analysis of what elements build up

the negative and positive tweets.

As we have mentioned earlier, in some cases, a dependency graph can be used to see the actual relationships between words and to draw conclusions from them. Therefore, for the tweets with the most followers, we created a dependency graph from the datasets.

6.2.1 August

The Figure 6.5 shows that the analysis was done with 4269 token corpus in the case of negative tweets, where the number of nouns exceeds two thousand. This is followed by verbs, adjectives and adverbs. The number of digits can also be highlighted.

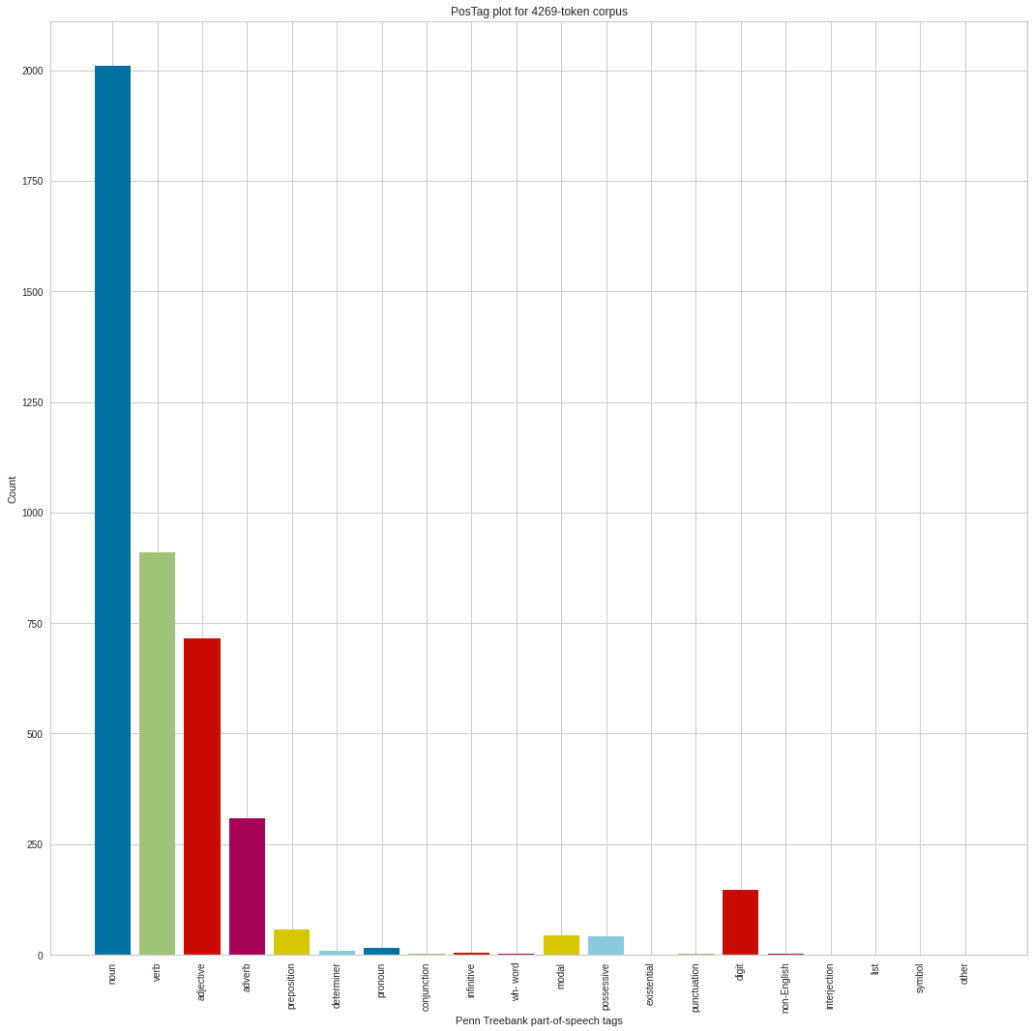


Figure 6.5: Part of speech tagging for negative tweets. The time period stands between 29 August 2021 and 31 August 2021.

The Figure 6.6 shows the part of speech analysis results from August on the positive tweets, which contains 3553 token corpus. Of course, the number of nouns is the most prominent here as well, followed by verbs, adjectives and adverbs. We cannot see unusual results here either. Comparing the negative and positive POS analyzes in August, we can mainly see the differences in the proportions, both in each POS groups and in the number of tokens that can be analyzed.

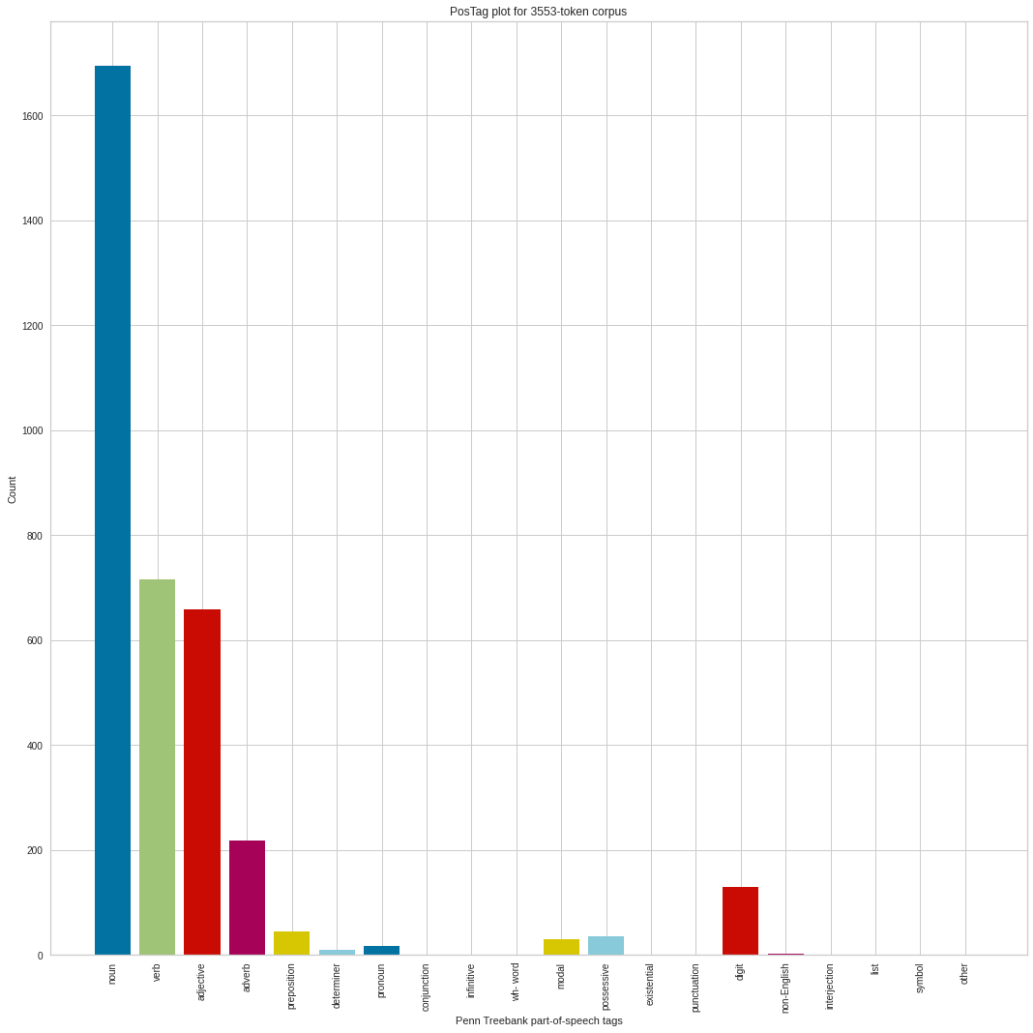


Figure 6.6: Part of speech tagging for positive tweets.The time period stands between 29 August 2021 and 31 August 2021.

Following the POS analyzes, let's look at the results of the dependency graph (Figure 6.7 shows the structure of the tweet.), using the positive twitter post with the most followers from the August dataset. There are two links at the end of the tweet, this is covered in the figure.

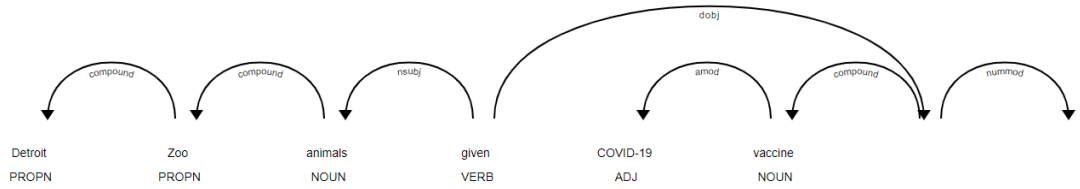


Figure 6.7: Most followed user's tweet (positive). The time period stands between 29 August 2021 and 31 August 2021.

6.2.2 September

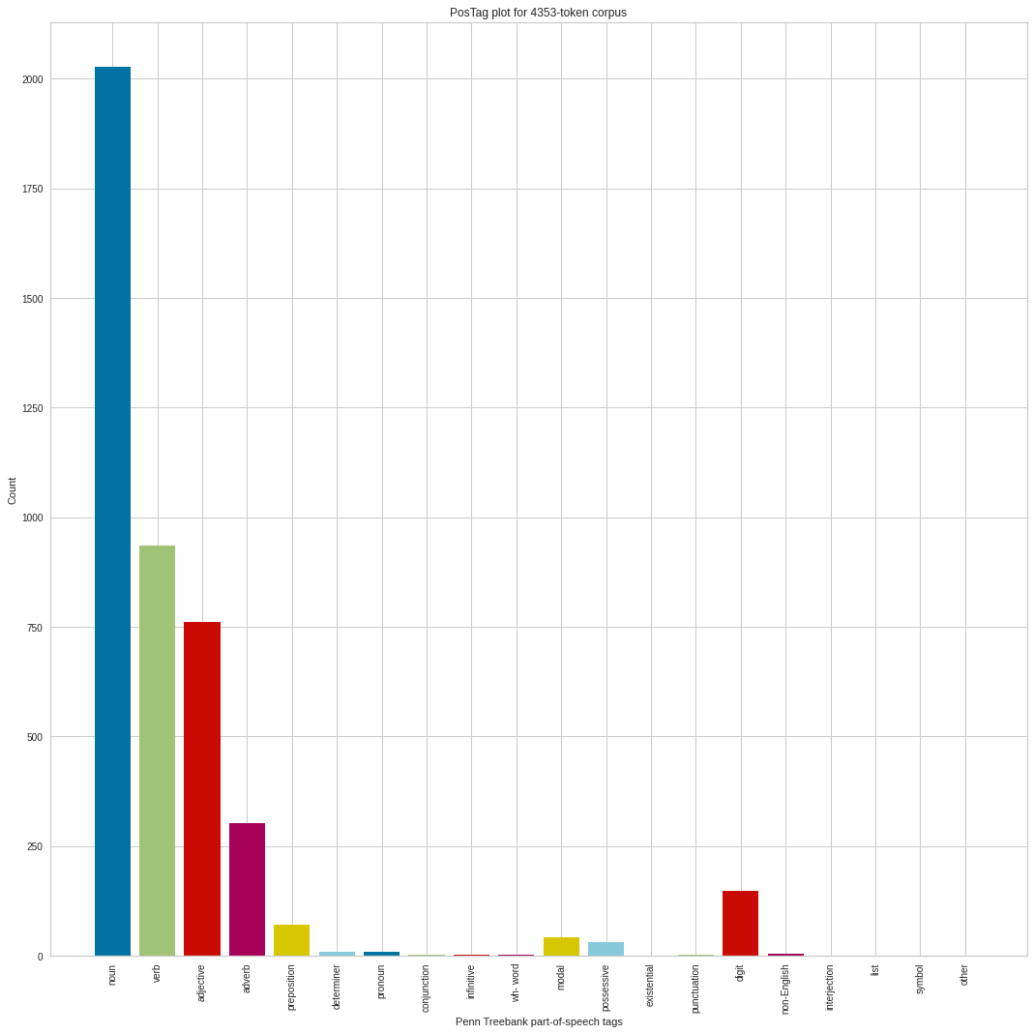


Figure 6.8: Part of speech tagging for negative tweets. The time period stands between 2 September 2021 and 4 September 2021..

The Figure 6.8 shows the POS analysis of the negative tweets in the September dataset, which includes 4353 token corpus. The structure of the analysis, of course,

similar to previous analyzes, in the same way the noun dominates, followed by verbs, adjectives and adverbs. If we compare the August negative POS results with the POS analysis results of the September negative tweets, we can see shifts. In addition to the increase in the number of nouns, the number of adjectives produced a more serious increase. In addition, minimal movements are noticeable in the other POS categories as well.

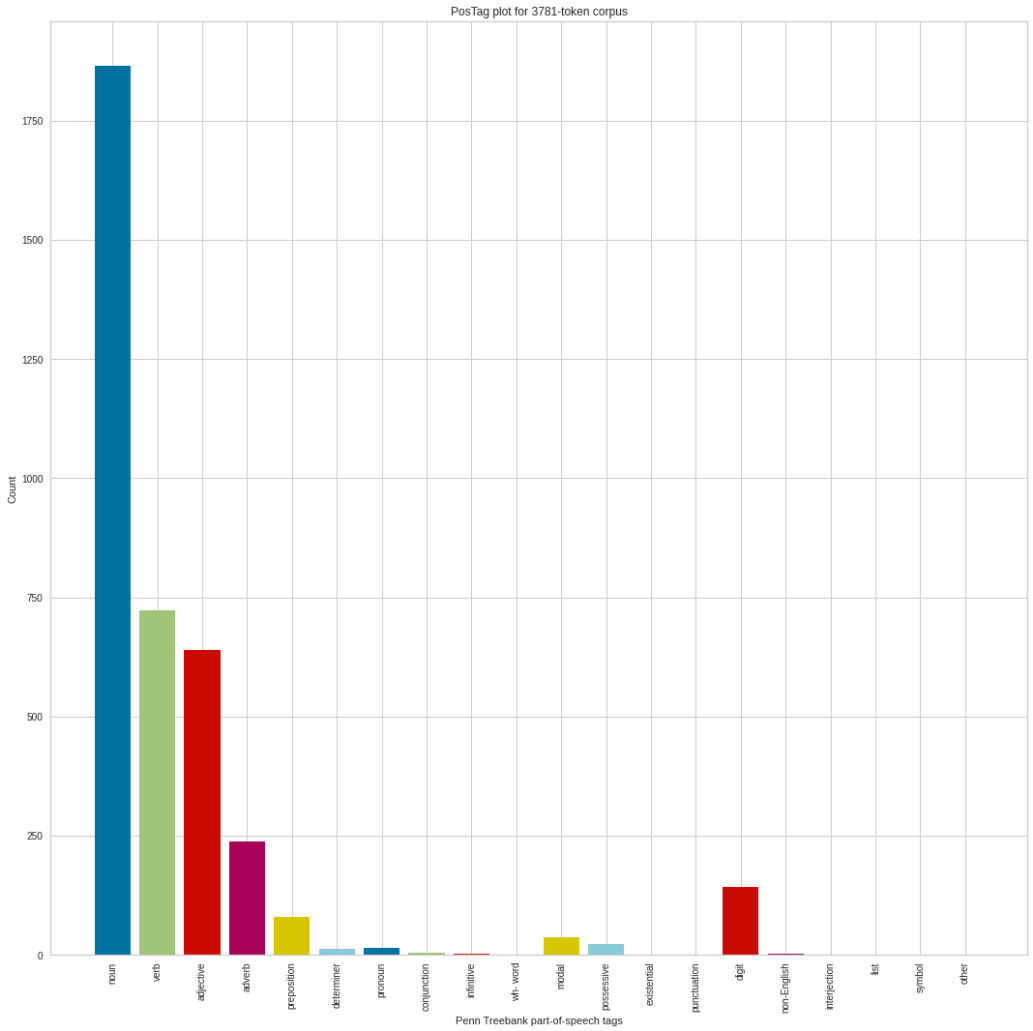


Figure 6.9: Part of speech tagging for positive tweets. The time period stands between 2 September 2021 and 4 September 2021.

The Figure 6.9 shows the POS analysis of the positive tweets in September, where 3781 token corpus were identified. Compared to the POS results of negative tweets, the order of the POS categories is the same. In addition to the decrease in the number of nouns, we can also see a significant decrease in the case of verbs, adjectives and adverbs. Of course the smaller number from the tokenization process

also plays a role in this, which is again a change or difference in the structure of tweets.

Comparing the positive POS results in August and the positive POS results in September, it can be seen that the number of tokens were similarly reduced compared to the results obtained in the negative cases. Which already draws attention to significant differences in the words of the texts of negative and positive tweets. Comparing the POS categories for the positive tweets in August and September, we can see decreases again in verbs and an increase in the number of nouns and adjectives.

Following the POS analyzes, let's look at the results of the dependency graph. (Figure 6.10) In this case, a fairly long tweet has reached the most people directly, so here we would like to illustrate that the method can be used for large and aggregate sentence, sentences. There are two links at the end of the tweet, this is covered in the figure.

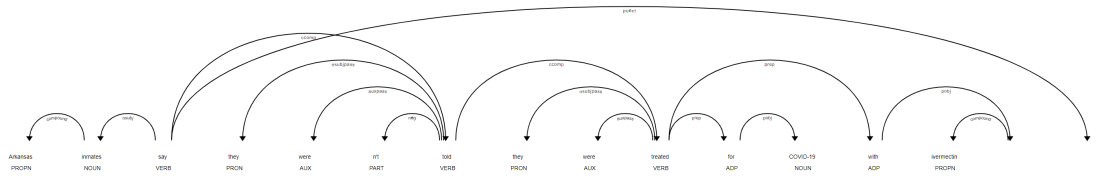


Figure 6.10: Most followed user's tweet (positive). The time period stands between 2 September 2021 and 4 September 2021.

With the help of Part of speech and word analyzes, which examine a deeper structure following the sentiment analysis, we already have a picture of the tweets, which were written during the given periods. What characterizes the negative and positive tweets, what differences appear between positive and negative tweets in a given period. We could see what words occurred most often in the periods for both positive and negative tweets, and what differences appear in the tweets written on the same topic in the two periods. The POS analysis even showed the structure of the tweets, and how much differences there are between the texts of the positive and negative tweets, which occurred in the case of tokenization first, the number of tokens in positive cases is significantly lower.

Based on the Information extraction analyzes and results, it may be worthwhile to include other disciplines such as psychology or linguistics in future work and expand the analyzes purposefully.

In the next section, we explore the results with Named Entity Recognition to gain more detailed information.

6.3 Named entity recognition results

We continue to use the RNN results, continuing the analyzes what we have started in the Information extraction section. Thus, the RNN results still aggregate to the the positive and negative parts.

6.3.1 August

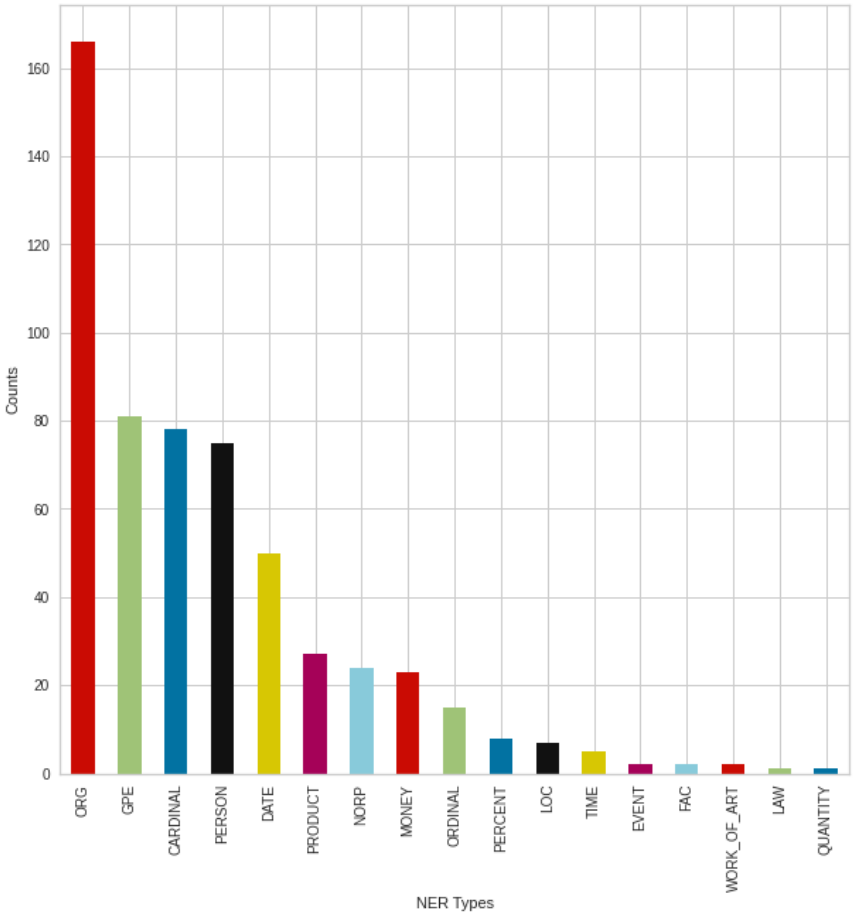


Figure 6.11: NER types of the negative tweets. The time period stands between 29 August 2021 and 31 August 2021.

The Figure 6.11 shows the negative tweets posted in August broken down into NER types to see how these posts are structured, what people mention primarily on the topic of covid. In most cases, various organizations, agencies, institutions were mentioned ('ORG'). This is followed by countries, states, cities ('GPE'). In addition, numbers ('CARDINAL' - Numerals that do not fall under another type.) and people / persons ('PERSON') followed these types before dates ('DATE'). After different organizations, which is an outstanding result, the types that follow are very close results. Based on the results, money ('MONEY') and various products ('PRODUCT') were less mentioned at the time.

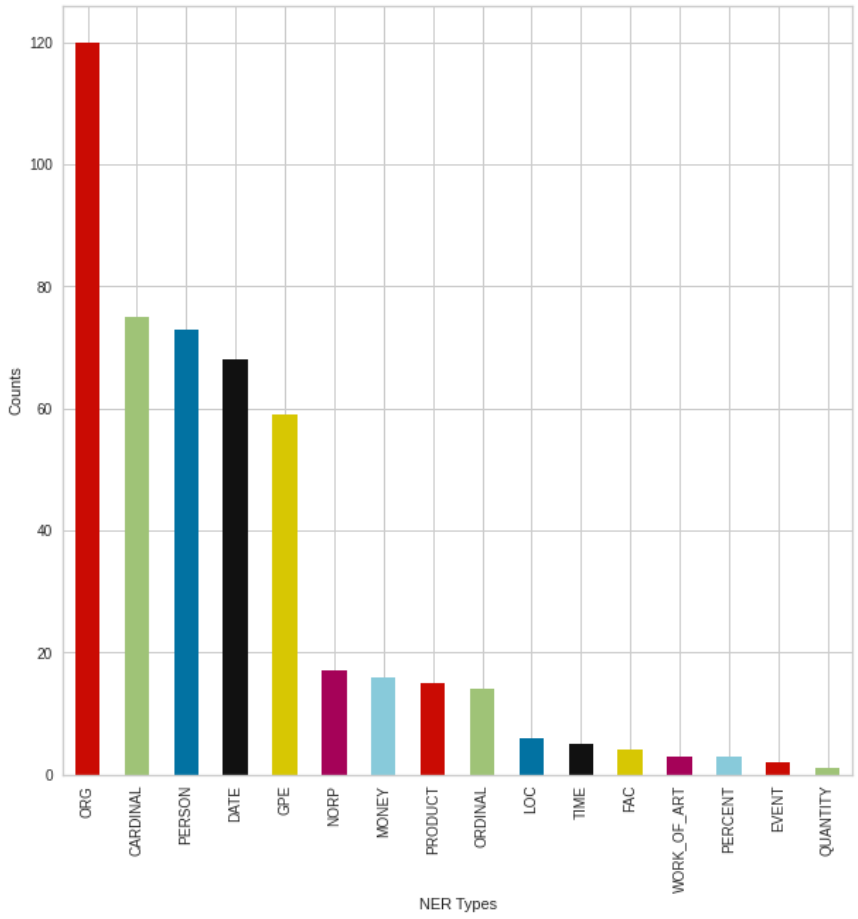


Figure 6.12: NER types of the positive tweets. The time period stands between 29 August 2021 and 31 August 2021.

The Figure 6.12 shows the breakdown of August positive tweets into NER types. In this case, the organizations, companies, institutions, etc. ('ORG') produced an outstanding result, just like in negative tweets. This is followed by a more significant rearrangement. While in the case of negative tweets the type of countries, states,

cities ('GPE') was the second strongest NER type, in positive cases the numbers type ('CARDINAL') was the second strongest NER type, and the countries, states, cities were only the fifth, which is a significant difference. Furthermore, for positive tweets, the third strongest was the 'PERSON' type, followed by the dates ('DATE').

These results suggest that people are actively talking about news, events, sharing what they have read about the topic and arguing for their opinions, which they are also trying to support, confirm, their information.

6.3.2 September

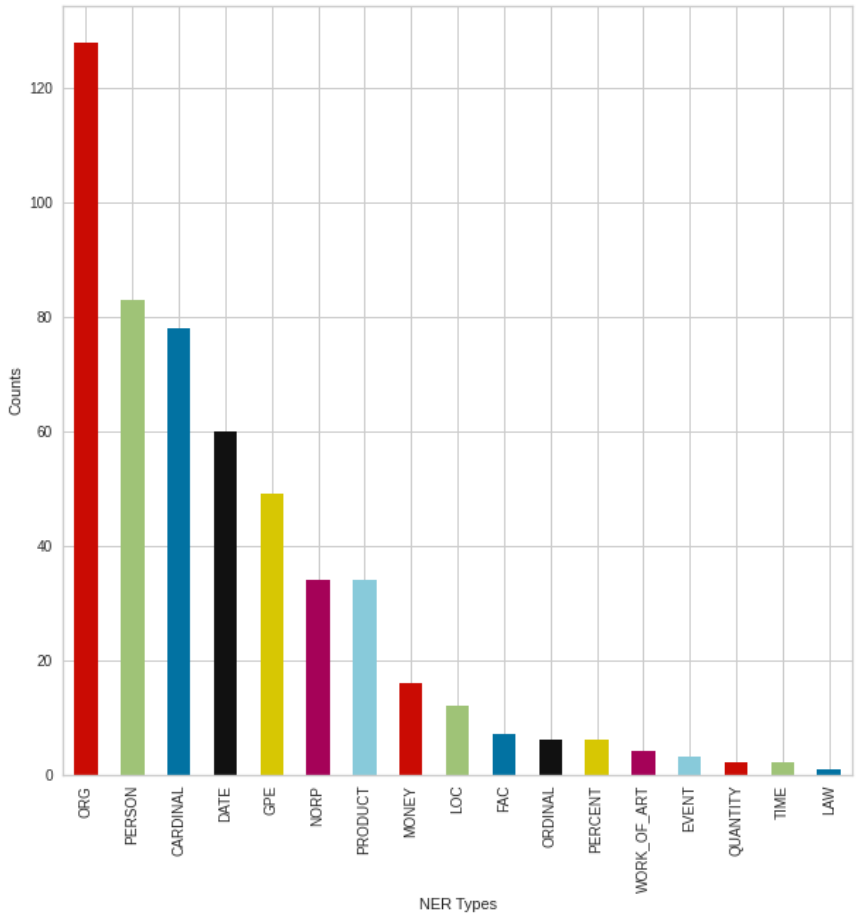


Figure 6.13: NER types of the negative tweets. The time period stands between 2 September 2021 and 4 September 2021.

The Figure 6.13 shows the result of the negative tweets posted in September, broken down into NER types, where once again an outstanding result from organizations, companies, institutions ('ORG') can be seen. Followed by the types of

persons ('PERSON') and numbers ('CARDINAL'). Contrary to previous August results, there was an increase in the type of nationalities or religious or political groups ('NORP'), similar to the type of products ('PRODUCT'). But the trend form August can still be seen with minimal changes in the strongest types.

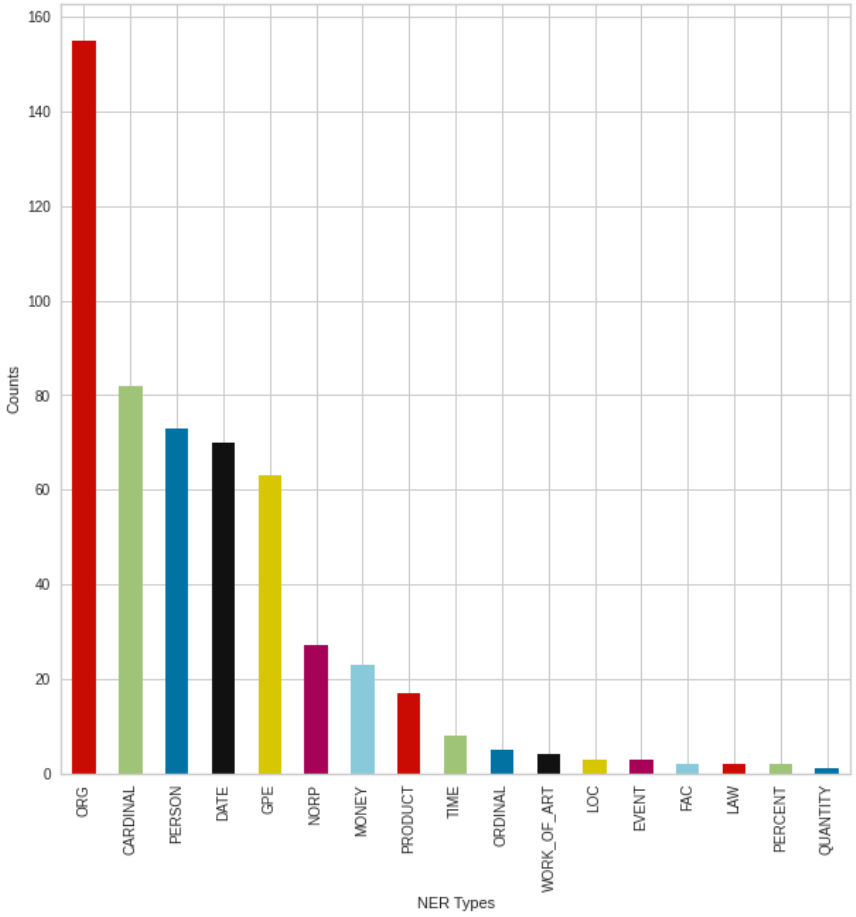


Figure 6.14: NER types of the positive tweets. The time period stands between 2 September 2021 and 4 September 2021.

The breakdown into NER types of the positive tweets shown in the Figure 6.14. In the case of the formation of types, this is the same as the previous August trend, especially in the case of the strongest types. If we compare the negative and positive results in September, we can see a rearrangement in the case of the less mentioned types, and a setback of the nationalities or religious or political groups ('NORP') type. But mainly the setback of products type ('PRODUCT') in the positive case, which can be highlighted.

6.3.3 Supplement

Extended explanations of types. All of the types mentioned in the description have been explained, and these types are summarized in more detail here.

The NER types: PERSON - People, including fictional. NORP - Nationalities or religious or political groups. FAC - Buildings, airports, highways, bridges, etc. ORG - Companies, agencies, institutions etc. GPE - Countries, cities, states. LOC - locations, mountain ranges, bodies of water etc. PRODUCT - Objectives, vehicles, foods, etc. EVENT - Named hurricanes, battles, wars, sports events, etc. WORK OF ART - Titles of books, songs. LAW - Named documents made into laws. LANGUAGE - Any named language. DATE - Absolute or relative dates or periods. TIME - Times smaller than a day. PERCENT - Percentage. MONEY - Monetary values. QUANTITY - Measurements, as weight or distance. ORDINAL - first, second, etc. CARDINAL - Numerals that do not fall under another type.

6.3.4 NER Type 'GPE' - deep analysis

In the case of NER types, the elements of the GPE (countries, states, cities) type were mentioned the second most often in the case of negative tweets in August, which was only the fifth most often mentioned in the case of positive tweets. Therefore, we supplement the analysis with the words mentioned in the GPE type in the August in both negative and positive tweets to see what might have resulted in this. (It is possible to extend any type shown in the figure.)

The Figure 6.15 shows the top 20 GPE for negative tweets. In the other Figure 6.16, we can see the GPEs mentioned in the case of positive tweets. In a negative case, the most mentioned country was Afghanistan, which may come as a surprise at first, but at the time, all media platforms were dealing with the Afghan withdrawal and the consequences, which also had an impact on covid themed tweets. Afghanistan was followed by the United States, China and the state of Florida. In positive tweets Afghanistan was the second after the United States, the third was Florida state. The coronavirus is different in countries, states and this creates a different situation, not surprisingly these are mentioned in the tweets, the unique situation is given by the situation in Afghanistan in this case - which was a unique

situation at the end of the summer.

With further analyzes, it was possible to explore explanations, details and information in addition to the sentiment analysis, which gives a much deeper picture of the real sentiment results of the given period, and what shaped these sentiment results.

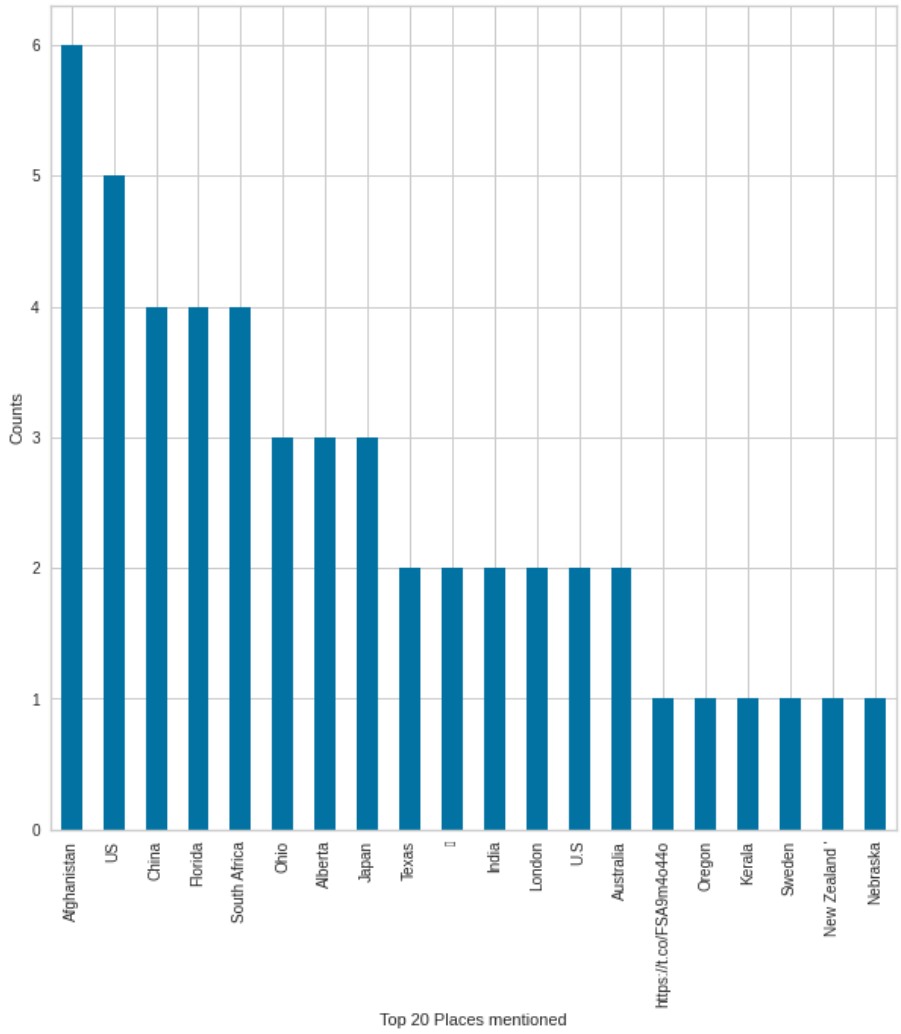


Figure 6.15: GPEs mentioned in negative tweets. The time period stands between 29 August 2021 and 31 August 2021.

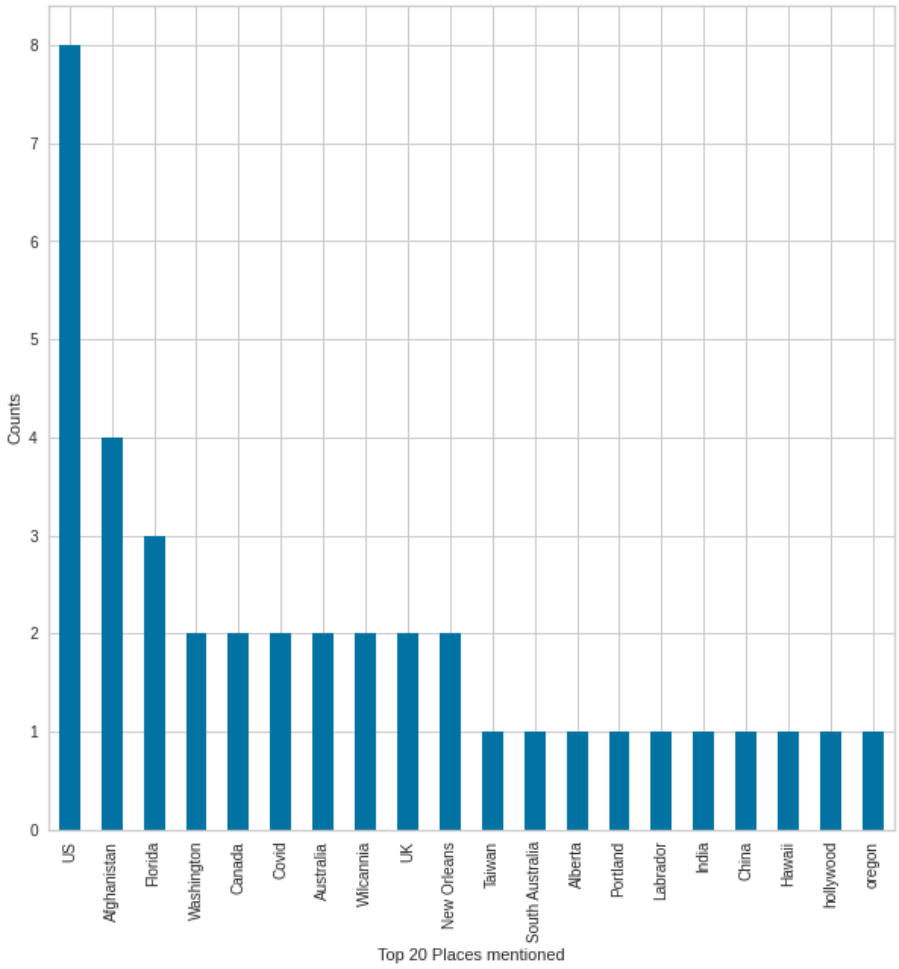


Figure 6.16: GPEs mentioned in positive tweets. The time period stands between 29 August 2021 and 31 August 2021.

Chapter 7

Discussion and Conclusions

7.1 Conclusion

In this work, we used different models for sentiment analysis to determine how people relate to the topic of covid in social media, primarily Twitter. We have created several models: BERT, RNN, NLTK - VADER and TextBlob to analyze "fresh" datasets. The primary goal was to work with the latest data for the period under study, so we always created the datasets according to a given limit number with the covid keyword and the given time period of the analyzes.

The sentiment analysis was extended. In addition to the usual 'positive', 'neutral', 'negative' categories, we extended that with 'strongly positive and negative' and 'weakly positive and negative' categories to detect smaller sentiment movements within the positive and negative categories when comparing the sentiment results of different time intervals.

BERT provided a comparison result for our other models, where the results of the RNN model was the most approximated to the results of BERT. Thus, we performed additional Information extraction and Named entity recognition analyzes on the sentiment categorized and labeled results by RNN to get a deeper picture of sentiment analysis. How people write / build their tweets, what is characteristic of their writing, what is the word usage of positive and negative tweets , what places, people and more were mentioned and what events may affect their tweets. Thus, we obtained a detailed analytical result on how the result of the emotional analysis developed.

The sentiment outcomes of the late August and early September period what we examined and extended by Information extraction and Named entity recognition analyzes, explained some of the sentiment changes between the two study periods, examined and provided a detailed picture of tweets. These analyzes also give a whole new picture to traditional sentiment analysis.

7.2 Future work

As future work, very interesting and valuable results could be achieved by involving additional disciplines such as linguistics or psychology and expanding the research with targeted further analyzes.

By introducing new classifications, analyzes, and keeping the current analyzes up to date, a new extended sentiment analysis library or wrapper could be created. This could extends and simplifies sentiment analysis using multiple models, and it could also provides additional analyzes to interpret and management the data. This can even provide specialized analyzes for different areas as well.

7.3 Acknowledgments

I would like to thank Attila Kiss, associate professor and head of the Department of Information Systems, for his joint work during my studies, which made it possible to publish three joint papers.

7.4 Thesis links

All links from the thesis were accessible on 1 October 2021.

Bibliography

[1] László Nemes and Attila Kiss. "Social media sentiment analysis based on COVID-19". In: *Journal of Information and Telecommunication* 5.1 (2021), pp. 1–15.

[2] László Nemes and Attila Kiss. "Prediction of stock values changes using sentiment analysis of stock news headlines". In: *Journal of Information and Telecommunication* (2021), pp. 1–20.

[3] László Nemes and Attila Kiss. "Information Extraction and Named Entity Recognition Supported Social Media Sentiment Analysis during the COVID-19 Pandemic". In: *Applied Sciences* 11.22 (2021), p. 11017.

[4] Carlos A. Iglesias and Antonio Moreno. "Sentiment Analysis for Social Media". In: *Applied Sciences* 9.23 (2019), p. 5037. ISSN: 2076-3417. DOI: 10.3390/app9235037. URL: http://dx.doi.org/10.3390/app9235037.

[5] Michal Ptaszynski et al. "Deep Learning for Information Triage on Twitter". In: *Applied Sciences* 11.14 (2021), p. 6340. ISSN: 2076-3417. DOI: 10.3390/app11146340. URL: http://dx.doi.org/10.3390/app11146340.

[6] Zhao Jianqiang, Gui Xiaolin, and Zhang Xuejun. "Deep convolution neural networks for twitter sentiment analysis". In: *IEEE Access* 6 (2018), pp. 23253–23260.

[7] Nour Jnoub, Fadi Al Machot, and Wolfgang Klas. "A Domain-Independent Classification Model for Sentiment Analysis Using Neural Models". In: *Applied Sciences* 10.18 (2020), p. 6221. ISSN: 2076-3417. DOI: 10.3390/app10186221. URL: http://dx.doi.org/10.3390/app10186221.

[8] Jenq-Haur Wang, Ting-Wei Liu, and Xiong Luo. "Combining Post Sentiments and User Participation for Extracting Public Stances from Twitter". In: *Applied Sciences* 10.22 (2020), p. 8035. ISSN: 2076-3417. DOI: 10.3390/app10228035. URL: http://dx.doi.org/10.3390/app10228035.

[9] Avinash Chandra Pandey, Dharmveer Singh Rajpoot, and Mukesh Saraswat. "Twitter sentiment analysis using hybrid cuckoo search method". In: *Information Processing & Management* 53.4 (2017), pp. 764–779.

[10] Muhammad Yasir et al. "An Intelligent Event-Sentiment-Based Daily Foreign Exchange Rate Forecasting System". In: *Applied Sciences* 9.15 (2019), p. 2980. ISSN: 2076-3417. DOI: 10.3390/app9152980. URL: http://dx.doi.org/10.3390/app9152980.

[11] Rokas Štrimaitis et al. "Financial Context News Sentiment Analysis for the Lithuanian Language". In: *Applied Sciences* 11.10 (2021), p. 4443. ISSN: 2076-3417. DOI: 10.3390/app11104443. URL: http://dx.doi.org/10.3390/app11104443.

[12] Jie Xu et al. "Sentiment analysis of social images via hierarchical deep fusion of content and links". In: *Applied Soft Computing* 80 (2019), pp. 387–399. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2019.04.010. URL: https://www.sciencedirect.com/science/article/pii/S1568494619302017.

[13] Madiha Khalid et al. "GBSVM: Sentiment Classification from Unstructured Reviews Using Ensemble Classifier". In: *Applied Sciences* 10.8 (2020), p. 2788. ISSN: 2076-3417. DOI: 10.3390/app10082788. URL: http://dx.doi.org/10.3390/app10082788.

[14] Sandra Rizkallah, Amir F. Atiya, and Samir Shaheen. "A Polarity Capturing Sphere for Word to Vector Representation". In: *Applied Sciences* 10.12 (2020), p. 4386. ISSN: 2076-3417. DOI: 10.3390/app10124386. URL: http://dx.doi.org/10.3390/app10124386.

[15] Kai-Xu Han et al. "Application of Support Vector Machine (SVM) in the Sentiment Analysis of Twitter DataSet". In: *Applied Sciences* 10.3 (2020),

p. 1125. ISSN: 2076-3417. DOI: 10.3390/app10031125. URL: http://dx.doi.org/10.3390/app10031125.

[16] Priya Chakriswaran et al. "Emotion AI-Driven Sentiment Analysis: A Survey, Future Research Directions, and Open Issues". In: *Applied Sciences* 9.24 (2019), p. 5462. ISSN: 2076-3417. DOI: 10.3390/app9245462. URL: http://dx.doi.org/10.3390/app9245462.

[17] Sunghee Park and Jiyoung Woo. "Gender Classification Using Sentiment Analysis and Deep Learning in a Health Web Forum". In: *Applied Sciences* 9.6 (2019), p. 1249. ISSN: 2076-3417. DOI: 10.3390/app9061249. URL: http://dx.doi.org/10.3390/app9061249.

[18] Khai Tran and Thi Phan. "Deep Learning Application to Ensemble Learning—The Simple, but Effective, Approach to Sentiment Classifying". In: *Applied Sciences* 9.13 (2019), p. 2760. ISSN: 2076-3417. DOI: 10.3390/app9132760. URL: http://dx.doi.org/10.3390/app9132760.

[19] Wenkuan Li et al. "Lexicon-Enhanced Attention Network Based on Text Representation for Sentiment Classification". In: *Applied Sciences* 9.18 (2019), p. 3717. ISSN: 2076-3417. DOI: 10.3390/app9183717. URL: http://dx.doi.org/10.3390/app9183717.

[20] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[21] Xuehua Han et al. "Using Social Media to Mine and Analyze Public Opinion Related to COVID-19 in China". In: *International Journal of Environmental Research and Public Health* 17.8 (2020), p. 2788. ISSN: 1660-4601. DOI: 10.3390/ijerph17082788. URL: http://dx.doi.org/10.3390/ijerph17082788.

[22] A.H. Alamoodi et al. "Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review". In: *Expert Systems with Applications* 167 (2021), p. 114155. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2020.114155. URL: https://www.sciencedirect.com/science/article/pii/S0957417420308988.

[23] Naw Safrin Sattar and Shaikh Arifuzzaman. "COVID-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA". In: *Applied Sciences* 11.13 (2021), p. 6128. ISSN: 2076-3417. DOI: 10.3390/app11136128. URL: http://dx.doi.org/10.3390/app11136128.

[24] Dimitrios Amanatidis et al. "Mining Textual and Imagery Instagram Data during the COVID-19 Pandemic". In: *Applied Sciences* 11.9 (2021), p. 4281. ISSN: 2076-3417. DOI: 10.3390/app11094281. URL: http://dx.doi.org/10.3390/app11094281.

[25] Guadalupe Obdulia Gutiérrez-Esparza, Maite Vallejo-Allende, and José Hernández-Torruco. "Classification of Cyber-Aggression Cases Applying Machine Learning". In: *Applied Sciences* 9.9 (2019), p. 1828. ISSN: 2076-3417. DOI: 10.3390/app9091828. URL: http://dx.doi.org/10.3390/app9091828.

[26] Alejandro Rodríguez-González et al. "Identifying Polarity in Tweets from an Imbalanced Dataset about Diseases and Vaccines Using a Meta-Model Based on Machine Learning Techniques". In: *Applied Sciences* 10.24 (2020), p. 9019. ISSN: 2076-3417. DOI: 10.3390/app10249019. URL: http://dx.doi.org/10.3390/app10249019.

[27] Tianyi Wang et al. "COVID-19 Sensing: Negative sentiment analysis on social media in China via Bert Model". In: *Ieee Access* 8 (2020), pp. 138162–138169.

[28] Rosario Catelli et al. "Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set". In: *Applied Soft Computing* 97 (2020), p. 106779. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2020.106779. URL: https://www.sciencedirect.com/science/article/pii/S1568494620307171.

[29] Pilar López-Úbeda et al. "COVID-19 detection in radiological text reports integrating entity recognition". In: *Computers in Biology and Medicine* 127 (2020), p. 104066. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2020.104066. URL: https://www.sciencedirect.com/science/article/pii/S0010482520303978.

[30] Xuan Wang et al. "Comprehensive named entity recognition on cord-19 with distant or weak supervision". In: *arXiv preprint arXiv:2003.12218* (2020).

[31] Kevin Lybarger et al. "Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework". In: *Journal of Biomedical Informatics* 117 (2021), p. 103761. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2021.103761. URL: https://www.sciencedirect.com/science/article/pii/S1532046421000903.

[32] Siddhartha R Jonnalagadda et al. "Text mining of the electronic health record: an information extraction approach for automated identification and subphenotyping of HFpEF patients for clinical trials". In: *Journal of cardiovascular translational research* 10.3 (2017), pp. 313–321.

[33] Ioannis D Apostolopoulos and Tzani A Mpesiana. "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks". In: *Physical and Engineering Sciences in Medicine* 43.2 (2020), pp. 635–640.

List of Figures