Sentiment Analysis on COVID19 Twitter data: A sentiment Timeline

Makrina Karagkiozidou

SID: 3305200021

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in eBusiness and Digital Marketing

JANUARY 2022

THESSALONIKI – GREECE

Sentiment Analysis on COVID19 Twitter Data: A Sentiment Timeline

Makrina Karagkiozidou

SID: 3305200021

Supervisor:	Assoc. Prof. C. Tjortjis
Supervising Committee Members:	Dr D. Filippidou
	Dr D. Karapiperis

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in eBusiness and Digital Marketing

JANUARY 2022

THESSALONIKI – GREECE

Abstract

This dissertation was written as a part of the MSc in e-Business and Digital Marketing at the International Hellenic University. The main purpose of this research is to create a sentiment timeline of Twitter users, regarding the COVID19 vaccines, aiming at the same time to extract significant information based on the conducted sentiment analysis regarding the dominance of each sentiment and their influential power based on their engagement rates. For the successful implementation of the analysis, several datasets were examined for the creation of the training model and the external complementary dataset. Following, various algorithms were recruited with the Random Forest algorithm to perform better and therefore be selected for the model training, achieving an accuracy of 91.5%. The findings of the analysis, show that Twitter users are positive in their majority for the vaccines and are aligned with the WHO's recommendations. In the timely analysis, it occurs that the more doubtful period was when complications and side-effects of the vaccines were reported, as well as when new restrictions started to be applied due to new variants. The minority of the tweets, represented by negative tweets appear, however, to have a higher influential power with their retweet rates to outperform positive and neutral sentiments.

Makrina Karagkiozidou

01/2022

Acknowledgements

Given the opportunity, I would most like to express my gratitude to my primary supervisor, Christos Tjortjis, who guided me throughout this project. His advice and guidance through this period were insightful and of significant importance for the successful implementation of this dissertation I would also like to thank the International Hellenic University, for their support during my studies. Last but not least I would like to show my gratitude to my family and friends who supported me and kept me motivated during the most stressful and difficult moments.

Contents

1 Chapter 1: Introduction

Humans are social creatures and tend to share their opinion on various matters. Since the end of the ‘00s, when social media were created and became widely used, their users were offered the means to post and publish their opinion to the rest of the world. As a result, not only companies but also the organizations started facing a new challenge, as each member of their community could share their opinion publicly and influence the rest users. This new challenge emerged the need for the organizations to gather the information that is included in the published posts. However, with the rapid growth of the microblogs and the forums where users can share their opinion on every possible matter, from product reviews to their aspect on a serious social and political aspect. Such is the COVID19 virus and the polarization of the society, on a worldwide level, regarding the actions and measurements that each government is taking to prevent its spread.

Covid19 is a new virus, a member of the family of coronaviruses, that started its spread in the December of 2019 in a city called Wuhan in China [67]. Since then, humanity has changed a lot in every aspect of life. COVID19, due to its rapid spread among humans and the unpredictable rates of mortality, has been characterized by both the media and the academic community as a pandemic. People have become more conscious of healthcare matters. People were masks, keep distance from each other and eliminate social activities to the minimum. The governments of all the countries have taken various measurements from mandatory mask use to local and nationwide lockdowns, aiming to stop the spread of the virus, as it affects not only everyone’s life but also the whole socio-economical situation of each country.

Based on the findings of Atalan (2020) [6], the citizens of each community were not only afraid of their health, but they were also anxious and insecure about their future, increasing the anxiety and depression levels of every society. Many small businesses were forced to close, due to lockdowns and minimization of outdoor activities. Therefore, many people lost their jobs facing unemployment as well as the constant threat of the virus. All these unexpected events created a social and political situation, that enhanced the gap among the governments and their citizens.

The creation of the vaccine was considered to be the solution to all these problems. And yet, when the time had come, it occurred that a part of each country's community was not only unwilling to be vaccinated but was also against it expressing its negative opinion freely on social media like Facebook.com and Twitter.com. In this study, sentiment analysis on Twitter Data will take place, investigating the current situation regarding the COVID19 vaccines, aiming to identify patterns, swifts in users' public opinion and beliefs throughout time, based on milestone events and important announcements of the World Health Organization

1.1 How sentiment analysis started

Sentiment Analysis, also known as opinion mining, is, based on Liu [34], the scientific field that analyzes among others, people's opinions, sentiment and evaluation on specific entities such as products, companies, organizations, ideas and events.

Liu [34], believes that the first academics to use the term sentiment analysis were Nasukawa and Yi in 2003 in their paper "Sentiment analysis: Capturing favorability using natural language processing." [43], while Dave et al. [16] were the ones to introduce the term opinion mining. Another term that has been used to describe sentiment analysis techniques and methodologies is affective computing, as mentioned by Cambria et al.[12]. Since then, a lot of research has been done on this field indicating the importance of the knowledge that can be gained through this process. The main applications of Sentiment Analysis according to Feldman [22] contain reviews on products and services, by automating text analysis and providing valid feedback on the actual opinion of the users and potential customers. In extension, the same monitoring can happen apart from specific products for the general reputation and image of the company's brand. On the same way of thinking, not only companies have brands. People like politicians, celebrities and public figures are their brand. Therefore, sentiment analysis is important for them not only to monitor their reputation online, but it can also work as a very powerful means of information for them when running for a position, to understand their potential voter's needs, wills and feelings on various matters discussed online. Public opinion especially when it is posted on social media, also affects stock market rates [31]. Sentiment analysis allows

researchers to predict, or even notice those changes in the public opinion early enough providing them with a very important competitive advantage in the stock market.

1.2 What is Sentiment Analysis

Based on the current bibliography, there is not one specific definition that describes the term sentiment analysis or opinion mining. Balahur & Steinberger (2009) [7] in their work, gathered several definitions that were given in the past years for the term, concluding that there was not a definition generic and descriptive enough to cover their needs, leading them to propose their definition based on their field. One of the definitions describes the term as follows.

"Sentiment analysis or opinion mining is the computational study of opinions, sentiments and emotions expressed in the text" [33].

In a further explanation, it is declared that sentiment analysis is a process followed by researchers and scientists that are aiming to extract insights and information for a specific topic or product from the opinion that has been expressed in a written format, like short texts, essays or letters.

1.3 Fields of implementation

Sentiment analysis can be applied in almost all the industries and the research fields, as long as there is written feedback, reviews and generally data on the matter of analysis [71] [50]. Beneficiaries of this type of analysis can use the extracted knowledge for better decision making, smarter strategy implementation and efficient risk management since they have the opportunity to learn the actual sentiment and feelings of not only their clients but also possible future customers leading them to make intelligent decisions, plan more effective marketing plans and in case of a crisis it can be used as a tool to minimize the risk and the losses of the situation. Opinion mining allows organizations to monitor their feedback on social media in real-time and make better decisions, while at the same time, they can enjoy superior returns. Furthermore, he considers aspect-level sentiment

analysis as the most fine-grained method. One of the results of his study was that most organizations and companies insist on creating simplistic techniques, aiming to avoid the facing of open challenges resulting in a less efficient analysis [22].

Business, politics, public actions and finance are all beneficiaries of sentiment analysis methods and with its implementation be able to result in the overall success of each one of these industries. To start with, Business through sentiment analysis can monitor their consumer's voices, brand reputation, focus on online advertising with the use of bloggers and then identify possible dissatisfaction oriented with this method and they can manage the overall aspect of online commerce. In politics, voting-advise applications, as well as clarification of politicians' positions, can be implemented with the use of sentiment analysis. Respectively, public actions can monitor real-world events, legal matter blogs (also known as blawgs), policy or government regulation proposals and intelligent transportations systems. Lastly, the finance applications of sentiment analysis contain Prices of commodities and shares evolution and financial risk individuation [1]

1.4 Benefits of Sentiment Analysis

Overall, with the use of Sentiment Analysis, every counterpart that can take part in any way in this process can be benefited. As it was already mentioned in the previous section, businesses, organizations and public administration offices increase their income, improve their performance and complete their long-term goals more easily. At the same time, customers and potential clients or users, that are the receivers of their actions also benefit, since their voice is meaningfully heard. Therefore, the products are now aligning with their true needs, the government have real-time and honest feedback on each one of their actions and correct or further enhance their decisions.

1.5 Open Challenges

There are still many open challenges in sentiment analysis and opinion mining methods. A number of these challenges regards the technical part of the analysis affected by the development aspect, while the rest are related to the general application of the model. The first type of challenges includes the heterogeneous characteristics of the data, incomplete, uncertain and sparse data, and challenges in semantic relations on multi-source data

fusion. Respectively, application challenges consist of doubts of whether sentiment analysis is indeed helping the enterprise strategies, the influence of a post and the impact that social bots have on their analysis [63].

Taking into account the informal language and the slang, used on social media, affected by human interaction to technology, as well as its constant evolution due to this relation, implemented models should be able to adjust to these changes [52].

Additionally, other challenges include the development of multilingual classifiers, building common user profiles by integrating the same user data from different social media applications, and enhancing Stanford Treebank by adding the ability to be applied at aspect level or document level instead of sentence level. In addition to handling implicit word meaning and indirect text, building domain-independent lexicon or classifiers, building real-time sentiment analysis systems which can dynamically capture new data and enhance results according to feedback [63].

1.6 Dissertation Scope

The main scope of this dissertation is to present the main terms and knowledge about the field of sentiment analysis, while at the second part of this study a sentiment analysis experiment on Twitter data will take place analyzing the sentiment and the opinion of Twitter users throughout the main events that are happening during COVID19 pandemic, examining the alignment of the users with the public measurements and government policies as well as the polarity among the users themselves.

1.7 Dissertation Outline

This dissertation contains an extended description and analysis of sentiment analysis and opinion mining. In the introduction, an outline of the first mentions, as well as the definition of the term is presented. Additionally, the main applications and benefits that result from the use of opinion mining are mentioned. To continue, we record open challenges that exist in the field of sentiment analysis, we further define the scope of this dissertation and, in this final paragraph, we present a brief description of the dissertation's outline and

contents. The second chapter includes that analysis of the main types of sentiment analysis methods and techniques. Furthermore, we describe the main evaluation and accuracy measures that are used for the comparison of the systems and provide an overview of the most important factors that are used to test and indicate the performance of the methods used. In chapter three, a literature review is conducted presenting relative work from scientists that are researching in the field of sentiment analysis. However, we do not only present their work, we identify at the same time their weak points, scientific gaps and similarities with our approach. In chapter four, we present and analyze the dataset as well as the preprocessing method that has been implemented in our research aiming to clean and make the selected data easier to manipulate and retrieve insights. The fifth chapter of our dissertation contains the experimental work that has been done providing the workflow and code used for the extraction of the results. In chapter six we present the results and discuss their significance to the objective of the research, while in the seventh and last main chapter the conclusions of the work are provided while future work ideas are proposed.

2 Chapter 2: Background

Sentiment analysis even though is a concept that exists among the communities and businesses and could be easily identified in the first human communities, is a relatively recent field in the scientific communities. One of the first attempts to describe and attempt to explain the term sentiment analysis and opinion mining was found in 1931, in a Sociology journal that presented various methods for the measurement of public opinion [20]. However, it was not until 2004, that scientific articles on sentiment analysis started to trend in academic communities, making it one of the most rapidly growing research areas [39]. That can be explained by the expansion of Web 2.0, which allowed users to generate their content themselves and express their personal beliefs publicly, challenging businesses and organizations. This challenge mainly contained their need to better understand their potential clients, offering products capable to satisfy their requirements. Web 2.0 and social media, forums, reviews and even in the earlier days surveys and questionnaires allowed them to listen to every user's need and adapt their products in alignment with their research findings.

Deep Learning, Machine Learning and Data Mining are technologies that were introduced to companies that offered them the advantage to manipulate big data and automate the process. In the next paragraphs of this chapter, the main methods of sentiment analysis and opinion mining will be presented, offering a brief overview, for further understanding of the following work.

2.1 Machine Learning Sentiment Analysis

Sentiment analysis can be divided into three main categories. Supervised, Unsupervised and Hybrid based on the methods use of labelled datasets. Each method can be more suitable for different tasks and domains [36].

2.1.1 Supervised Methods

Supervised Sentiment Analysis methods are in their majority used in classification problems. By dividing a dataset into two parts, the training and the testing part, the model is first trained to fit the labelled attribute and then with the use of the testing set, it is evaluated [60]. Supervised methods can be prone to bias and insufficiency as they are affected by the quality and quantity of the training data [36]. Some of the most popular Supervised Sentiment Analysis techniques, are Support Vector Machine (SVM), Naïve Bayesian Classifiers and Decision Trees [54].

Support Vector Machine

Support Vector Machine Classifier is a computer algorithm created to assign labels to the objects of a dataset. As with every supervised method, SVM can be trained based on an example training set and then predict the values of non-labelled records. For the completion of the task, the SVM algorithm transforms the data for each record into a data point in an n-dimensional graph, where n is the number of parameters that are included in the analysis. Then the algorithm produces a hyperplane, which represents the decision boundary. In a two-dimension example, a hyperplane is a straight line. Therefore, each record that falls into one side of the graph, will be classified as the respective label. However, the optimized outcome of the algorithm is the best hyperplane. The main characteristic of this hyperplane is that it has the maximum margins from all the label groups [18] [45].

Naïve Bayesian Classifiers

Naïve Bayesian Classifiers are usually preferred by analysts for their simplicity and speed. It appears to perform better than more complex and sophisticated classifiers, in a wide range of application fields, even in domains and problems that are not characterized by independent features. The formula of the classifier is

$$c^* = arg\ max_c\ P(c/w) \tag{1}$$

where

$$P(c/w) = [P(w/c)P(c)]/P(w) \tag{2}$$

[24]. In eq. 2, P(c/w) represents the joint probability of w and c. In other words, it calculates the probability of the object w to happen given that w is true. Their algorithms are

based on the Bayesian Theorem, which assumes that each variable, which represents the features, is independent. Therefore, Naive Bayes calculates the probability that an object will align in a specific class, and assigns it to the one with the highest given probability. This type of classifier tends to perform better when the dataset contains independent features, but also in functionally dependent features [56].

Decision Tree Classifiers

Another Supervised Sentiment Analysis type is Decision Tree Classifiers. As all the trees in the computing society, the decision Tree contains nodes, edges and leaf nodes. Nodes represent the features participating in the analysis, edges represent the tests that are done in the features, including changing the feature weights, and lastly the leaf nodes that represent the final category which resulted from the executed tests [68]. Decision Tree Classifiers start from the root node and move in depth successfully passing the tests in between. Based on the results the leaf node is then assigned the final label value. This kind of classifier is preferred by scientists, as they tend to perform better than other classifiers. While they have seen extensive use in many applications of speech and language processing [54].

2.1.2 Unsupervised Methods

Unsupervised sentiment analysis methods, also known as clustering methods, are characterized by the absence of human action. In the real-world, most classification cases, do not contain labelled data. Unsupervised learning overcomes this problem [10]. Unsupervised Methods are divided into two main categories based on the followed approach, Partition Clustering and Clustering. In the first type clusters are not allowed to overlap each other while in the latter one, clusters can be sub-clusters of bigger ones, creating nests [58]. In the category of the Unsupervised methods, algorithms like k-Means, Fuzzy C-means, Divisive methods, Bottom-Up or Agglomerative, are included [27].

k-Means

One of the most common and popular unsupervised partitioning classification methods is the use of the k-Means algorithm. K-Means is an algorithm that even though it does not require labelled data for its successful performance it requires a predefined number of clusters to produce, as well as some random initialization values for the cluster centroids [66]. The k-Mean algorithm can be described into five steps:

1. Initialization of the centroids and predefine the number of clusters.
2. Group in the clusters the data points.
3. Assign each object to its nearest cluster centre based on the selected distance function.
4. Recalculate or update the position of each cluster's centroids.
5. Repeat steps 3 and 4 until the centroids no longer move [5].

Fuzzy C-Means

Fuzzy C-Means is an algorithm that is mostly used in clustering tasks. Applications of this method are found in feature analysis, pattern recognition and image processing [74]. This algorithm examines every data point and its proximity to each cluster, for a better group of the points. At the end of each iteration, both the centroids and the memberships are shifted [30]. The performance of this algorithm depends on the selection of the initial values and the membership value. The closer the initialization values are to the optimal solution, the better the algorithm will perform, as it will require fewer iterations to create the final clusters [28].

Divisive methods (Top- Down)

In the Top-Down or Divisive hierarchical clustering methods, all the data points are part of one bigger cluster. Then objects that are more different than the rest will create independent clusters that will fall in the hierarchy. The process continues until there is one cluster for each one of the data points. Therefore, there are as many clusters as the number of data points participating in the analysis [58].

Agglomerative (Bottom-Up)

In contrast to the previous method Agglomerative or Bottom-Up clustering techniques resemble a reverse tree-like structure. In the bottom of the tree, or else the leaves of the

tree, there are clusters, that are equal in number and consist of the individual objects, while in the root of the tree there is one single cluster that contains all the objects [17]. In other words, in the Agglomerative methods, the clusters start with each data point being part of their personal cluster, and at the end of the process, all data points are part of one bigger cluster [58].

2.1.3 Semi-Supervised

As its name indicates, Semi-Supervised methods are used on a combination of supervised and unsupervised methods. Therefore, the training set consists mainly of unlabeled and a smaller amount of labelled data. The main purpose of Semi-Supervised classifiers is to exploit unlabeled data to produce better prediction procedures [73]. These methods are preferred for sentiment analysis as they provide high accuracy when labelled data are lacking [15]. Semi-Supervised methods include Graph-Based Methods, Wrapper-Based Methods and Topic-Based Methods [65].

Graph-Based

Graph-Based methods contain the use of graphs for the sentiment of a text. Each labelled and unlabeled record are represented as nodes, while the edges represent the similarity between the nodes [78]. Their main task is to assign a label to all the unlabeled data based on their similarities with the rest instances. The main prerequisite for these methods to work properly is the existence of a similarity measure that will eventually classify each instance to the appropriate label [65]

Wrapper-Based

Wrapper – Based methods are characterized by a big number of iterations in the training process. In every iteration, the algorithm is labelling a specific amount of unlabeled records, based on the decision function that is affected by the prediction power of each feature. Then the model is re-trained for the next iteration. of It is one computational demanding process. Self-Training and co-training are the most popular branches of this method [65]. Even though it has been characterized as one computationally demanding process, it usually produces higher accuracy than other feature selection methods [37].

Topic-Based

The main difference of the Topic-based method, with the methods described above, is that it does not focus only on the given dataset of labelled and unlabeled records, but it also takes into consideration the general context of the analysis topic. For instance, the word ‘shot’ is considered a negative word, since it usually describes the injury of a person. In the medical community, on the other hand, ‘shot’, can be used to describe the vaccination of a person against a virus. Therefore, it is clear, that words’ meanings can differ based on the topic. In most cases, the topic. Proposed models are built in such a way, that the classification model is using, specific classifiers for each cluster that occurs [77].

2.2 Lexicon-based Sentiment Analysis

The lexicon-based approach of sentiment analysis is not part of the machine learning methods that have been presented till now. It is considered a more easily understandable approach, as the analysis is implemented with the use of sentiment lexicons that calculate the polarity of a text to characterize the sentiment of the text into positive, neutral and negative, requiring however high human factor involvement [59]. The used semantic lexicons can be generated either automatically or manually and are used to calculate the polarity of each sentence. There are two main approaches of Lexicon-based analysis, Dictionary-based approach and Corpus-based approach. Their main differences are that the first one requires some seed words given by the user, that are then enriched automatically, incorporating synonyms. The latter one uses context-specific words to indicate the sentiment of each word. It contains two main methods: The statistical approach and the Semantic approach [26].

2.3 COVID19 Pandemic

Since the December of 2019 humanity started its battle with a new virus that later would be characterized as a pandemic, that will have a high impact on everyone’s lives, changing the way, customers, companies and governments work and function. The official name

of this virus is SARS-CoV-2, however, there are a lot of other widespread names such COVID19, Coronavirus, Corona and many others depending on the language of each country. The infectious characteristics of the virus increased the threats for viral and rapid expansion worldwide [64]. Since the first cases in Wuhan of China were found, a constant battle against the spread of the virus began.

2.3.1 The Timeline of the Pandemic

Based on the information of the World Health Organization (WHO) response in their COVID-19 pandemic website [69], in the current paragraph a timeline of the expansion and the fight of the WHO is presented.

Date	Event
31 Dec 2019	The WHO becomes aware of the 'viral pneumonia' that Wuhan, People's Republic of China is facing. This would be then characterized as the beginning of the pandemic.
04 Jan 2020	WHO posted their first tweet informing the world of the cluster of pneumonia cases in Wuhan.
10-12 Jan 2020	WHO published guidance documents for countries regarding among others, Infection prevention and control, Travel Advice, Clinical Management, Laboratory Testing and Surveillance case definitions.
11 Jan 2020	The announcement of the first death from the Coronavirus by the Chinese media.
13 Jan 2020	The first case outside of Wuhan, People's Republic of China was reported in Thailand.
19 Jan 2020	The WHO Western Pacific Regional Office tweeted that there was evidence of limited human to human transmission.
13 Mar 2020	Europe has become the epicentre of the pandemic with more reported cases and deaths than the rest of the world.
04 Apr 2020	WHO reports that over 1 million cases of coronavirus are confirmed worldwide.

16 Apr 2020	WHO introduces ‘lockdowns’ as suggested health and social measure.
10-14 May 2020	The WHO releases new suggestions on health and social measures for workplaces, schools, mass gatherings and public health criteria.
05 Jun 2020	The WHO published new guidelines regarding the use and quality of masks.
22 Sep 2020	WHO issues the use of a quality antigen-based rapid diagnostic test for detecting the SARS-CoV-2 virus.
Oct 2020	Another new variant is reported from Indian authorities. This variant in May 2021 will be called the Delta Variant (WHO, 2021)
14 Dec 2020	The first vaccination shot took place in the United States [57]. United Kingdom reported a SARS-CoV-2 variant to the WHO.
18 Dec 2020	A new variant of SARS-CoV-2 is rapidly spreading in three provinces in South Africa.
31 Dec 2020	The WHO issued the emergency use validation for a COVID-19 vaccine, focusing on equitable global access.
5 Jan 2021	One of WHO representatives met to review the vaccine data for Pfizer’s vaccine. It was the first to receive an emergency use validation from WHO.
9 Jan 2021	Another variant was reported from the Japanese Authorities in the samples of Brazilian Travellers
25 Jan 2021	WHO released recommendations for the use of the Moderna vaccine against COVID-19.
29 Jan 2021	The WHO publishes their recommended COVID-19 tests (PCR and Antigen).
15 Feb 2021	The WHO released two versions of the AstraZeneca COVID-19 vaccine for emergency use.
4 Mar 2021	Vaccination Data are now published on the WHO Coronavirus Dashboard (https://covid19.who.int/)
12 Mar 2021	Johnson & Johnson COVID -19 vaccine is now listed for emergency use by WHO.

17 Mar 2021	A statement was made by the WHO regarding the AstraZeneca vaccine safety. The reason was reports of rare blood coagulation disorders in people that had recently received the vaccine.
7 Apr 2021	The WHO COVID-19 subcommittee for Vaccine Safety made a statement noting that the AstraZeneca vaccine safety risks indicate further research, as the cooccurrence of the vaccine and the blood clots was considered plausible but was not confirmed.
14 Jun 2021	Lockdown extended in England by 4 weeks due to Delta Variant [41]
22 Jun 2021	Updated recommendations for Pfizer, Moderna and Janssen vaccines against COVID-19 [75]
12 July 2021	WHO considers the use of Booster vaccines to maintain protection against the virus [75]
18 Aug 2021	US government announces the initiation of booster doses from September [13]
23 Aug 2021	The Pfizer 2-dose COVID-19 vaccine receives full FDA approval [47].
31 Aug 2021	The WHO releases a statement on booster doses and to whom they shall apply [75]
22 Sep 2021	FDA approved boosters doses for certain populations [48].
19 Nov 2021	FDA Authorizes boosters of the Pfizer and Moderna COVID-19 vaccines for adults [49].
22 Nov 2021	Austria is the first country in Europe to impose a 'lockdown' both for vaccinated and not [8].
26 Nov 2021	The latest variant called Omicron is detected in South Africa and Botswana [70] [76]

Table 1Coronavirus timeline

It is clear from the timeline that WHO and each government following always the best-believed strategy acted accordingly for the elimination of the new virus. However, new variants complicated their work. The battle against Coronavirus is not a simple process, given the human factor that affects the final results of each action. People after two years

of isolation and strict measurement seem to live a sentimental rollercoaster with their anger increasing with every wave of the pandemic [3]. Therefore, this work is implemented to monitor Twitter users' sentiment on the announcement of each event from the WHO.

3 Chapter 3: Literature Review

In this chapter an extensive description of the term sentiment analysis will be presented, including the definition, techniques and methodologies used throughout its execution.

3.1 Related Work

3.1.1 Sentiment Analysis of Twitter Data

Agarwal et al., in 2011 [2] published their paper "Sentiment Analysis of Twitter Data". The main objective of their research was to create models of classification for Twitter posts. They focused on mainly three models, the unigram model, a feature-based model and a tree kernel-based model. Additionally, they proposed and analyzed 100 features that can be indicative of each Tweets sentiment. It was noticed that Twitter-specific features could add value to the classifiers, but most importantly, they found that using tools of standard natural language processing, can be very efficient even when used in different genres than the one they were trained on. Finally, they concluded that the feature-based and the tree- kernel models tend to outperform the unigram model, and that sentiment analysis on twitter data does not differ from the rest of social media platforms and the models used for their sentiment analysis. However, their research was maintained in a theoretical aspect of the field lacking the experimental part with the specialization in a specific subject.

3.1.2 Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis

Jianqiang et al. in 2018 [29], published a paper that aimed to fill in the gap of previous researchers. They noticed that most of the scientific papers focusing on this topic, fail to give the required attention and ignore the part of pre-processing and cleaning the data. For this reason, in their paper, they attempted to compare text-preprocessing methods. These methods contain replacing negative mentions, removing URL links in the corpus, reverting words that contain repeated letters to their original English form, removing

numbers, removing stop words and expanding acronyms to their original words. Their experimental comparison of the methods, using four classifiers, on five Twitter datasets, showed that cleaning the data on URL, stop words and numbers is appropriate to reduce noise, but do not affect the performance of each classifier. Meanwhile, the replacement of negations and the expansion of acronyms can improve the accuracy of the classifier used in the model. Lastly, they also stress that when using SVM classifiers the removal of numbers is found to improve the accuracy of the model. Jiangiang et al. even though they also specialized their research on Twitter Data sentiment analysis, still kept their work in a more theoretical and generic aspect.

3.1.3 Social media sentiment analysis based on COVID-19

Nemes and Kiss, in 2021 [44], in alignment with many scientists of the field conducted as well a COVID19 related sentiment analysis from data gathered from social media, and more especially posts, comments and retweets from the Twitter social media platform. One of the elements in their study that is allowing their research to stand out from the rest in the field is that in their training model, they also verify their findings with an external open-source dataset. Also, they compared the results of the sentiment analysis using the RNN method and TextBlob, showing that the results can vary based on the used method. The comparison showed that RNN produced a very good percentage of accuracy while TextBlob was also efficient but had a tendency to classify posts as neutral, minimizing the qualitative outcomes of the research. Overall, they found that the was a positive sentiment in the posts at the beginning of the pandemic that was maintained in time. At the same time, however, it was also noticed an increase in the negative sentiment analysis. Indicating the polarity of the discussion. In alignment with our research, they used additionally external data, but they focused on the differences and the comparison of the methods, instead of inquiry a specific research topic.

3.1.4 Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks

Manguri et al. in their paper that was published in 2020 [38], conducted a sentiment analysis based on posts published on Twitter during the week of 09-04-2020 to 15-04-2020 and were related to the keywords “Covid19” and “coronavirus”. For the collection of the tweets, they implemented an algorithm in Python, using the Tweepy library. After

the collection of the required data, the sentiment analysis was done using the Text Blob library also using Python. The findings of their analysis showed that there is a high percentage of polarity among the users of the platform, when expressing their opinion on the situation, while they also noticed that their opinion changes daily depending on government and media actions, broadcasts and new guidelines. Lastly, they believe that since Twitter is used by more professionals and official personalities, the results of sentiment analysis on this platform, are more accurate than on the rest of social media like Facebook, Wechat and Instagram. Their work was one of those that dealt with the worldwide phenomenon of the COVID19 pandemic, using however a really small set of data minimizing the accuracy of the results that could be extracted.

3.1.5 COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis

Lyu et al., in their recent paper in 2021 [35], discussed the broader topic of COVID19 vaccines on Twitter, presenting a topic model as well as a Sentiment Analysis. More especially, they attempt to monitor the changes in concerns and emotions throughout the time that have an effect on the final goal of herd immunity. Their dataset contains data withdrawn from the first day that COVID19 was declared a pandemic, 11-03-2020 to 31-01-2021. Their findings indicate that the most popular topic regarding COVID19 is the discussion about vaccination, while the sentiment of the tweets was mostly positive, that was gradually increased, especially after the announcements of the vaccine's effectiveness. The most dominant feeling that emerged in the analysis was trust in the vaccines that indicate the mainly positive sentiment of the tweets. At the same time, the percentage for feelings of fear started to decrease in time enhancing their statement. This study shows a more topic specified approach, enhanced by a bigger data set that can be further extended by our research.

3.1.6 Twitter Sentiment Analysis during COVID19 Outbreak

In 2020, Dubey [21] conducted a Twitter Sentiment Analysis during the COVID19 Outbreak. In his analysis, he includes data from twelve countries worldwide, comparing the findings that occur in the sentiment analysis for each one of them. Tweets were collected

for the areas of Australia, China, Belgium, France, Germany, India, Italy, Netherlands, Spain, Switzerland, UK and USA. His data was gathered in the period of 11-05-2020 to 31-05-2020. His main purpose was to identify the feelings that citizens had expressed through the lockdown. The main results of the research showed that the bigger part of the users expressed mainly positive feelings, however feelings of fear, sadness and disgust were also identified. In four of the countries, higher rates of distrust and anger were noticed in comparison to the rest participating countries of the research. In his conclusion, he notes that China seems to tweet more negatively about the pandemic, while the most commonly mentioned words in the tweets of each country are pandemic, death, quarantine, hope, stay home. Respectively in the USA, it appears that the most used word is former president Trump. Their work differentiated from the rest as it was a comparison among the countries, including the language factor into the sentiment analysis.

3.1.7 Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic

The same year, Kruspe et al. in 2020 [32], conducted also sentiment analysis on Twitter data regarding the COVID19 pandemic, differencing their work from the rest by focusing not only on English as the main language but by conducting cross-language sentiment analysis. Their data were collected from various countries in Europe during the period December 2019 to April 2020, which were then correlated with the events in each country. The countries that are represented in the research were mostly Uk, Spain, Germany, Italy, France, Netherlands with the smaller countries being represented to a smaller degree. In the presentation of their results, they focus on each country separately, since each country's government acted differently, testing their citizens in various ways. Their findings show that until February 2020, there was little reference to COVID19 related keywords and topics. Additionally, with the announcement of the lockdown, the sentiments were mostly negative but improved and the positive sentiment analysis was increased with time.

3.1.8 Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA

One of the most recent papers was published in 2021 by Garcia and Berton [23] with their sentiment analysis on Twitter data, aiming to extract insights, in alignment with the rest

studies, but further focusing on the countries of Brazil and the USA. In their research, they detect and rank ten topics ranging from economic impacts, politics and case reports to anti-racism protests, online events and sports. Their sentiment analysis contains data from four months between April and August 2021. Their analysis showed that for all the COVID19 related topics for both countries, negative emotions were dominant, especially for the topics of case studies, proliferation care and statistics. In general, it showed that both USA and Brazil shared a common sentiment, which was due to the worldwide nature of the pandemic. It was also an interesting work that was purely focused on the two countries of the USA and Brazil minimizing their sample.

3.1.9 Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse

Sanders et al., in 2021 [61] also conducted a sentiment analysis on data retrieved from the social media platform of Twitter, using Natural Language Processing, clustering and sentiment analysis methods aiming to extract the sentiment of mask relevant tweets. Their dataset contains more than a million records from the first four months of the pandemic. More specifically, the data were collected in the period of 17th of March 2020 to 27th of June 2020. Therefore, it is obvious that the differential element of this study was targeting and analyzing mask-wearing posts. The output of their research indicates that there is extensive polarity regarding the discussion on mask-wearing, which is characterized by increasingly negative sentiment. At the same time, it seems that there is an association between the health topic that is mask-wearing with political events and the former president Donald Trump. Since it was conducted the previous year, further examination for the current year is needed.

3.1.10 Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study

Boon-Itt and Skunkan, in 2020 [11], completed a topic modeling and sentiment analysis study aiming to provide insights on the Public Perception of COVID19 based on data retrieved from Twitter. The purpose of their research was to extract information regarding the main topics that are discussed on Twitter about the pandemic. The main topics that we identified were tweeting about the emergency of the situation, case reports and statistics and means of controlling the virus. Their data were collected during the period 13th

of December 2019 to 9th of March 2020 and were filtered for COVID19 related keywords, establishing this research as one of the first on the topic. In their results they created also a timeline of the frequency of each symptom mentioned in the posts, differentiating their analysis from the rest. Meanwhile, the results of their sentiment analysis indicate that there was in general a negative sentiment for COVID19 and the whole pandemic that has affected the whole world.

From the papers that have been included in the literature review, it appears that there is a gap in the existing literature as there is too little work presenting the sentiment of Twitter users throughout the timeline of the pandemic. Therefore, in this study, open-source datasets that contain twitter posts will be combined with twitter data collected in the period 15-09-2021 to 30-11-2021 to identify the reactions and the feelings of users to public announcements on governments measurements, vaccination's FDA approvals, to mandatory vaccinations.

4 Chapter 4: Dataset

In this chapter, the main information of the methodology used for the collection of our main data and the information that is contained in the dataset will be presented as well as the main preprocessing techniques that are implemented in the model to clean and manipulate the data.

4.1 Dataset Description

The data that are used in the analysis are collected using the Twitter API v2. For this purpose, a Twitter Developer account was requested from Twitter. With the abilities that the Twitter Developer Account offers, a daily extraction of Tweets was possible. For the collection of the tweets, a simple workflow on Rapidminer was implemented as depicted in the picture below. Initially, the connection of the Twitter account with Rapidminer was established. Then ten 'Twitter Search' nodes were implemented for each one of the researched keywords. These are: 'covid19 vaccines', 'coronavirus', 'pandemic', 'Pfizer 'Vaccine', 'Delta Variant', 'Vaccine Certificate', 'Covidiots', 'Covidscam', 'PCR test', 'Rapid test'. For each search, the more recent tweets in English are selected for the collection. Then, the result of each one of these searches is stored with the implementation of the 'Write Excel' node, to a separate excel. To continue, with the help of 'Union' nodes, the data of each excel are combined and are finally written with another 'Write Excel' node to the another excel that keeps all daily Data.

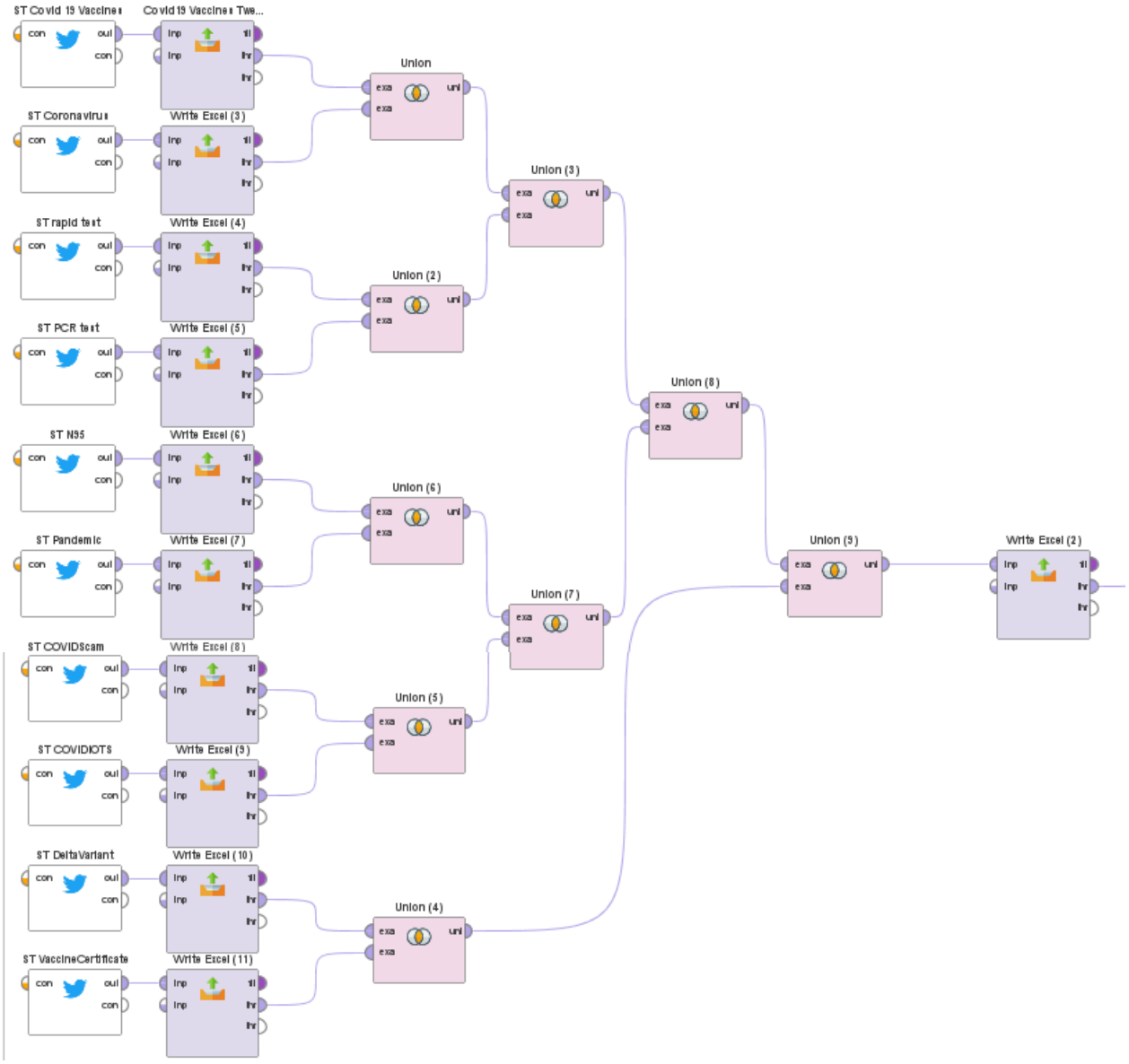

Figure 1RapidMiner Workflow for Tweet Collection

With the successful completion of this process, another workflow is executed as shown below.

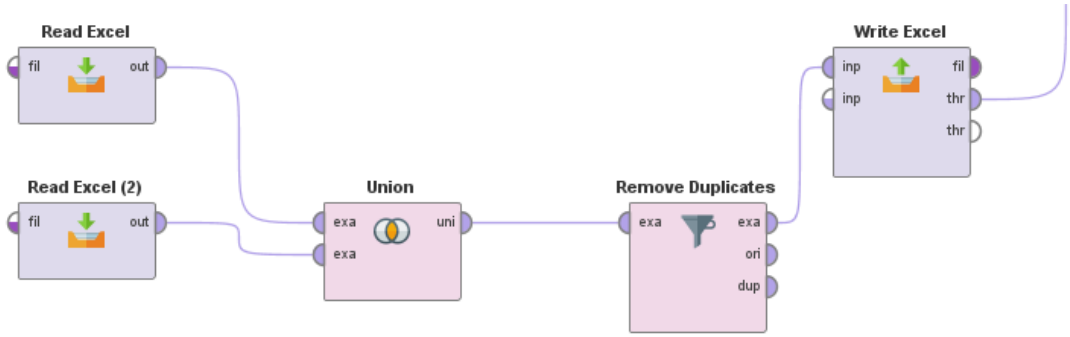

Figure 2 RapidMiner Workflow for Duplicate Removal

As it is depicted in the workflow, there are two 'Read Excel' nodes, the first one reads the data from the excel that collects the daily gathered data, while the second one the excel that contains all the data from the beginning of the research. Every day, the data of these files are combined and remove the duplicates that may occur with the use of the 'Drop Duplicates' node. Finally, the unique tweets are stored in the same excel that contains all the data.

The private data were collected during the period 15 Sep 2021 to 10 Dec 2021. However, since this is timeline research of Twitter user's there was a need to extend the dataset with tweets from the beginning of the pandemic. For these purposes, the dataset 'All COVID19 Vaccine Tweets' by Gabriel Preda (2021) [53] was selected. It contains data from 12 Dec 2020 for vaccine-related keywords.

4.1.1 Training Set

As in every sentiment analysis task, a model shall be implemented for the extraction of each tweet's sentiment. In this research, the training model was created based on the existing coronavirus labelled dataset 'Sentimental analysis of covid-19 tweets' by Dhruv Dhawan [19]. This dataset contains more than 165.000 tweets with their sentiment annotation. More specific, the tweets were notated as positive, neutral and negative. 55% of the tweets were labelled as neutral, while positive were labelled 23% and negative 22% of the total tweets in the dataset. The downside of this dataset was that there was not any description of the data, therefore, we can only assume the process followed for the annotation of each tweet. To establish the validity of their work, a manual evaluation of the data was conducted.

4.1.2 Dataset Structure

With the use of Twitter API v2 and RapidMiner for the collection of the tweets, there was a specific number of features extracted for every tweet. These features were the tweet 'ID', 'Created-At' containing information regarding the date and time of the tweet's creation, 'From-User' and 'From-User-ID' providing the username and ID of the writer, 'To-User' the user ID the tweet refers to, 'Language', 'Source', 'Text', 'Geo-Location-Latitude', 'Geo-Location-Longitude' and 'Retweet-Count'. Despite the big amount of

knowledge that can be extracted with the analysis of these data, in consideration of the main objectives of this research, the main tweet information that is required for the analysis, contains the main text of the tweet, the date and time of its creation and its retweet count as well as the name of the author. Therefore, since there was a combination of self-gathered and external datasets, they both needed to contain the same information.

4.2 Dataset Preprocessing

The preprocessing of the dataset was implemented using Python and was divided into two parts. Firstly, the training set was preprocessed, cleaning the 'text' field from elements that would complicate and reduce the accuracy of the model based on the similar research of Beleveslis et al. [9]. The 'text' field of the dataset contained the information as extracted by Twitter, meaning that they contained links, numbers, tags of other users and symbols that do not help in the building of a successful and effective training model. For instance, an original text tweet has the following form:

> RT @GlobeandMail Western Countries have been obsessed with #Covid19 since China sent Xmas Gift to the world in 2019. In #Africa, over 1.3 have survived WITHOUT the vaccines. We appreciate the #Malaria vaccine. This disease has been killing our young children. https://t.co/HV8hIa0VvK

The preprocessing function that was implemented aimed at the removal of the RT indicative, the @usernames mentioned in the main part, the # symbol, the possible link that might accompany the text and the numbers. The tweet that will occur when the function is executed will be the following:

> Western Countries have been obsessed with Covid19 since China sent Xmas Gift to the world in 2019. In Africa, over have survived WITHOUT the vaccines. We appreciate the Malaria vaccine. This disease has been killing our young children.

The volume of the tweets will reduce, and the text will contain only readable and meaningful words helping the performance of the model.

The preprocessing of the final data required some extra actions. Since the final data that were used in the analysis originated from various sources, their interoperability characteristics were of crucial importance. Therefore, before the cleansing of the data, the datasets were combined by keeping only the required features. When all the data was united under one dataset, the actual cleansing was initiated. One of the most important steps was to eliminate duplicate records. Since the collection of tweets was daily, some tweets could be gathered multiple times. Apart from that, it was noticed, that many retweets were also gathered, resulting in the dominance of viral tweet texts. As a consequence, a second level removal of duplicates was implemented based on the text of each tweet. Continuing, the function for cleansing every tweet text was executed.

In the next Chapter, the sentiment analysis methodology will be presented and described.

5 Chapter 5: Methodology

The main purpose of this research is to identify and elaborate the sentiment of Twitter users on regards to the recent topic of COVID19 vaccines. For the development of the research part of this thesis, a clear and well-defined methodology should be structured. Various tools, methods and techniques were used, aiming to complete and answer the initial objectives of this study. To identify the response of Twitter users on important COVID19 vaccine events and announcements, published tweets on the topic were collected with the use of RapidMiner. RapidMiner is a data science platform that is selected from more than 40.000 global organizations in different industries, for the execution of data-driven tasks. Apart from the tweet collection, the building of a sentiment prediction model was of crucial importance. The training model was implemented with the use of the Python programing language. Python is one of the most popular and user-friendly programming languages in the field of Data Science. The Python version used is 3.8.5. The environment in which the code was developed, is Jupyter Notebook. Its main characteristics are that it is an open-source, web-based platform that is used for scientific computing purposes, which provides data processing and visualization features.

5.1 Sentiment Prediction Model

A variety of libraries were recruited in the process of training a model to predict the sentiment of each tweet in the dataset. The libraries that were used throughout the analysis were ‘Pandas’, ‘NumPy’, ‘Re’, ‘Seaborn’ and ‘Sklearn’.

5.1.1 Pandas Library

Pandas is one of the most important Python libraries for data analysis tasks. The development of this open-source library started in 2008 and is until today supported by individuals worldwide. The highlight features of pandas contain the DataFrame object, which is ideal for the manipulation of data with integrated indexing. With the use of pandas, an analyst can read and write data from various data formats, reshape and pivot datasets, intelligent data alignment and management of missing data [51].

5.1.2 NumPy Library

NumPy is also an open-source library primary created for the easier and simpler manipulation of large arrays and matrices. Large arrays can be characterized by both big in volume datasets and multidimensional datasets. In general, it is a library that offers numerical computing tools, including random number generators, linear algebra routines, vectorization and indexing [46].

5.1.3 Re Library

The name of this library stands for Regular Expression. The main purpose of this library is to specify whether or not a given set of strings in Unicode or 8-bit format, matches the regular expression and reverse. Another important function of this library is the substitution of a specific string with another value. Functions like 're.search()', 're.split()', 're.sub()' are common for the search, separate and replace strings in the datasets [55].

5.1.4 Sklearn Library

Based on the official site of the Scikit-learn library, the use of this library is exclusively for machine learning tasks like classification and prediction. In alignment with the rest libraries, this is also an open-source, accessible and reusable code. It contains algorithms for classification, regression, clustering, dimensionality reduction, model selection and preprocessing tasks. In this study, algorithms of classification like SVM (Support Vector Machine), Nearest neighbours and random forest were tested, with the last one being selected for the training model [62].

5.1.5 Seaborn Library

Seaborn Library offers analysts the opportunity to create statistical graphics in the Python programming language. It is a complementary library based on the matplotlib and pandas data structures. Visualization of complex information is made easier and simpler. Its main advantage is that it allows the user to focus on the elements of the plots rather than the customization elements of the graph or plot [4].

5.2 Model Development

In this paragraph, the developed code for the training model will be presented and described for a better understanding of the sentiment analysis process.

To start with, the first step was to import the dataset that will be used for the training of the model. In this case, the name of the dataset was ‘Sentimental Analysis of COVID19 Tweets.csv’ as it was extracted from the Kaggle online repository.

In the first line of the code, the panda's library is imported, for the easier manipulation of the data. Then, with the functions provided by the library, the file is read and imported.

```
import pandas as pd
train = pd.read_csv('Sentimental analysis of COVID-19 Tweets.csv', encoding = 'latin-1')
```

Figure 3 Python Code: import dataset

As shown in the figure below, the columns of the dataset are requested. As an output to this request, the header of each column is provided, named ‘tweets’ and ‘sentiment’.

```
train.columns
```

```
Index(['tweets', 'sentiment'], dtype='object')
```

```
train['sentiment'].value_counts()
```

```
neutral     98844
positive    40693
negative    40322
Name: sentiment, dtype: int64
```

Figure 4 Python Code: Selection of attributes and value counts per label

For the next line of code, the number of appearances for each unique value for the selected sentiment column is requested. It occurs, that tweets characterised as neutral are dominant with 98.844 records, while positive and negative follow with 40.693 and 40.322 respectively. It appears, that the selected dataset is unbalanced in favour of the neutral tweets. Therefore, action must be taken to balance the training set.

```
#Clean the text function

import re
def cleanTxt(text):
    text = re.sub(r'(?:\@|http?\://|https?\://|www)\S+', '', text)
    text = re.sub(r'"""', '', text) #removes the '"""' symbol
    text = re.sub(r'#', '', text) #removes the '#' symbol
    text = re.sub(r'RT[\s]+', '', text) # remove RT
    text = re.sub(r'https?:\/\/\s+', '', text) #remmove the hyoerlink

    return text

#Cleanng the data
train['tidy_tweets'] = train['tweets'].apply(cleanTxt)
train.head()
```

	tweets	sentiment	tidy_tweets
0	Chinese citizens caught faking COVID-19 tests ...	2	Chinese citizens caught faking COVID-19 tests ...
1	RT @RunesSmash: After Covid dies down, Can we ...	1	After Covid dies down, Can we please normalize...
2	RT @Neurophysik: Many COVID-19 patients recove...	3	Many COVID-19 patients recover on their own. Q...

Figure 5 Python Code: Data Preprocessing

Before that, the function for the cleaning of the data is called and executed. The output of its execution is shown in the picture above. At that point, there are three columns, the two existing in the DataFrame and the new column with the cleaned text of every tweet. Per the description of the function in the previous paragraph, the new text is lacking name mentions, the RT indicative, links and the # symbol.

```
train['sentiment'] = train['sentiment'].replace(['positive'],3)
train['sentiment'] = train['sentiment'].replace(['negative'],1)
train['sentiment'] = train['sentiment'].replace(['neutral'],2)
```

Figure 6 Python Code: replace strings with numbers

Another step that was included in the preprocessing of the dataset is that the values of the sentiment column were renamed. In more detail, positive was replaced with numerical 3, negative with numerical 1 and finally, neutral with numerical 2. This replacement happens for easier manipulation of the represented data and extraction of more meaningful knowledge.

Following this task, the handling of the imbalance in the dataset is initiated. For this purpose, the overall dataset is then divided into three subsets based on their sentiment.

```
# Divide the dataset into 3 subsets based on their label

train_c1 = train[train['sentiment']==1]
train_c2 = train[train['sentiment']==2]
train_c3 = train[train['sentiment']==3]
```

Figure 7 Python Code: Separation of the dataset into subsets

Based on the code, the subset train_c1 will be consisting of the records that are noted as negative, since their sentiment value is '1'. Respectively, train_c2 and train_c3 will contain neutral and positive tweets. By following this procedure, it will be easier to retrieve the required amount of records per label.

The methodology that will be used for the handling of the imbalance is under-sampling.

```
# Undersampling the dominant labels into the minority one

sample_rec = train_c1['sentiment'].count()
train_c2_sample = train_c2.sample(sample_rec)
train_c3_sample = train_c3.sample(sample_rec)

# And unite them again in a new df

eq_train = pd.concat([train_c2_sample, train_c1, train_c3_sample],  axis=0)
eq_train['sentiment'].value_counts()
```

```
3    40322
2    40322
1    40322
Name: sentiment, dtype: int64
```

Figure 8 Python Code: Undersample the subsets to the same number of records

In this case, each label will be represented by the same number of records with the label that is less represented. It was mentioned before that the majority of the records were neutral tweets. While the minority were negative tweets with 40.322 records. Therefore, there is a need to under-sample neutral and positive tweets to match the records of negative tweets. In the first code-line, the count of the negative subset, sentiment column is assigned in the sample_rec variable. While in the next two lines, the other two subsets using the panda's function .sample() select randomly 40.322 records out of their total number of records. When this task is completed, the new datasets are all combined again in one new DataFrame. This action is implemented with the function .concat() provided by the panda's library. In the last code-line, the value count for each label in the new dataset is requested, showing that all three labels are now having the same number of records. Since the imbalance issue was handled, the next step for the creation of the

training model was to split it into two parts, the training and the testing one. The reason behind this separation is the future evaluation of the training model that will be developed in this phase.

```
# Split the dataset into training and testing sets.

from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn import svm
from sklearn import metrics
import numpy as np


X = eq_train ['tidy_tweets']
tfidf = TfidfVectorizer(max_features = 1000, ngram_range=(1,2))
X = tfidf.fit_transform(X)

Y=eq_train['sentiment'].values

X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size=0.2, random_state = 8)
```

Figure 9 Python Code: Import libraries and split the training dataset to test and train

In the first five lines of this code, the required libraries are imported. These are, as already described, ‘sklearn’ for the machine learning activities of the model and ‘numpy’ for the manipulation of the arrays. The next three lines are used for the vectorization of the tweets. Initially, the tweets are stored in the X value. The used vectorizer is TF-IDF (Term Frequency - Inverse Document Frequency). The main characteristics of this technique are that it provides the frequency of each used term in the document, but also, considers the importance of this term for the sentiment [25]. The selected parameters for the TF-IDF vectorizer is the max_features that indicates the maximum number of terms that will be built in a vocabulary, ordered by term frequency. Ngram_range indicates the minimum and the maximum number of terms that can be considered a string in the document. Then, with the use of the vectorizer, the values of X are transformed. Meanwhile, in the Y value, the sentiment values of each tweet are stored. Finally, from the ‘sklearn’ library, the train_test_split function is imported for the separation of the dataset into the training and testing subsets. The parameters indicate that the split will be 80% for the training set and 20% for the testing, while we set a random state to be ‘8’.

```
# Train on the Random Forest Classifier
from sklearn.ensemble import RandomForestClassifier

classifier=RandomForestClassifier(n_estimators=100)
classifier.fit(X_train.toarray(), Y_train)
prediction = classifier.predict(X_test.toarray())
metrics.accuracy_score(Y_test, prediction)
```

```
0.9158469041911218
```

Figure 10 Python Code: Train the model on Random Forest Classifier

In the final part of the training model, Random Forest Classifier is recruited. The classifier is assigned in the classifier variable. Then, the model is trained to fit the train set and finally, it predicts the label of the test set, indicating the final accuracy of the model. In the last code-line, the accuracy score is requested, comparing the initial sentiment values of the records with the predicted one from the model. The results provided an accuracy percentage of 91,58%. Other classifiers including SVM (Support Vector), Naïve Bayes and kNN were also tested, but they achieved lower accuracy rates for the selected dataset.

5.3 Sentiment Analysis

The development of the training model is completed, and the model is ready to predict the sentiment of the gathered dataset. To start with, the dataset is read and is imported, responding to the request of the dataset shape and columns.

```
# Upload the main dataset

dfa = pd.read_csv('All Dissertation Tweets.csv', encoding ='latin1')
dfa.shape , dfa.columns
```

```
((181275, 12),
 Index(['Created-At', 'From-User', 'From-User-Id', 'To-User', 'To-User-Id',
        'Language', 'Source', 'Text', 'Geo-Location-Latitude',
        'Geo-Location-Longitude', 'Retweet-Count', 'Id'],
       dtype='object'))
```

Figure 11 Python Code: Import the gathered Tweets

By these requests, the size of the dataset is defined, and the header of each feature is presented, providing an overview of the information contained.

```
# Keep only the en tweets and drop duplicates
dfa = dfa[dfa['Language'] == 'en']
dfa = dfa.drop_duplicates(subset='Text', keep="last")
dfa['Language'].value_counts() , dfa.shape
```

Figure 12 Python Code:

As soon as the dataset is defined, the preprocessing phase initiates. One of the main pre-requisites of the research is that the sentiment analysis will be conducted in the English language, therefore not English tweets should be dropped and be excluded from the analysis process. Since that is a dataset that scraps tweets from Twitter every day, it is expected and proven that retweets that contain the same text are gathered. Consequently, the records should be filtered on their text keeping unique values. To be able to identify the number of tweets excluded from the analysis a request of the new dataset's shape and value counts are requested.

```
# Select only the fields of interest

dfa = dfa[['Created-At', 'Text', 'Retweet-Count']]
```

Figure 13 Python Code: Filter out non-useful attributes

In this step of the process, only the fields relevant to the research objectives are selected to eliminate the computing power required during the analysis. These features are the datestamp of each tweet, its main text for the extraction of the sentiment and the retweet count, that will be used for the influence power of each sentiment.

```
#Add second external dataset.
dfb = pd.read_csv('C:/Users/makrina/Desktop/vaccination_all_tweets.csv')
dfb = dfb[['date', 'text', 'retweets']]
dfb = dfb.rename(columns = {'date': 'Created-At', 'text': 'Text', 'retweets': 'Retweet-Count'})
dfb = dfb.drop_duplicates(subset='Text', keep="last")
dfb.shape, dfb.head()
```

Figure 14 Python Code: Import external dataset

The second external dataset is imported and is preprocessed in the same way as the first one. In the code presented above the required columns are selected, the column headers are renamed to match the format of the other dataset and the duplicate tweets are dropped.

```
#Combine the 2 datasets.
df= pd.concat([dfa, dfb],  axis=0)
dfa.shape, dfb.shape, df.shape
```

Figure 15 Python Code: Combination of the datasets

Then the two datasets are combined into one DataFrame. And the final size of the dataset is requested.

```
def PredictLabel(text):
    text = tfidf.transform([text]).toarray()
    predicted = classifier.predict(text)
    return predicted

df['label'] = df['Text'].apply(PredictLabel)
df.head()
```

Figure 16 Python Code: Sentiment Prediction per Tweet

In this step of the analysis, the sentiment of each tweet is calculated and stored in a new column named ‘label’. The function that calculates the sentiment of each tweet’s text, is transformed with the TF-IDF vectorizer and then with the use of the classifier, it stores and later returns the predicted value of the sentiment. Since this is a very time consuming and demanding task, the results of the analysis are saved in a CSV file using the following code

```
#Checkpoint --> Save the labelled Tweets in .csv file
df.to_csv('Tweets Labelled.csv')
```

Figure 17 Python Code: Save Tweets with Sentiment to a CSV file

Reading once again the new labelled dataset, the sentiment analysis tasks can be initiated and executed.

```
df = pd.read_csv('Tweets Labelled.csv')
df
```

Figure 18 Python Code: Read the final dataset

```
df['Text']= df['Text'].apply(str)
#Clean the text
# the function


def cleanTxt(text):
    text = re.sub(r'(?:\@|http?\://|https?\://|www)\S+', '', text)
    text = re.sub(r'@[A-ZA-Z0-9]+', '', text) #Removes @mentions
    text = re.sub(r'#', '', text) #removes the '#' symbol
    text = re.sub(r'RT[\s]+', '', text) # remove RT
    text = re.sub(r'https?:\/\/\s+', '', text) #remmove the hyoerlink
    return text

#Cleanng the data
df['tidy_tweets'] = df['Text'].apply(cleanTxt)
pd.set_option('display.max_colwidth', None)
df['tidy_tweets']
```

Figure 19 Python Code: Final Dataset Pre-processing

Another important step of the analysis is once again the cleaning of the tweets, in alignment with the cleaning process that was executed in the training model.

```
df['tidy_tweets'] = df['tidy_tweets'].apply(lambda x: ' '.join([w for w in x.split() if len(w)>3]))
df['tidy_tweets']
```

Figure 20 Python Code: Tokenization of Tweets

In the code shown above, the preprocessed tweets, are split into words that have a min length of three characters. The output that is generated from the execution of this code, separates each tweet into tokens of words. For instance, the tweet with the text 'Lab studies suggest #Pfizer, #Moderna vaccines' will be transformed into '[studies, suggest, Pfizer,, Moderna, vaccines]'. As was expected the word 'Lab' is not included in the output of the code as its length is not more than three characters. Following the tasks, the word cloud is produced with the following lines of code.

```
# Plot the wordcloud
from wordcloud import WordCloud
import matplotlib.pyplot as plt

allWords = ' '.join([twts for twts in df['tidy_tweets']])
wordcloud = WordCloud(width = 500, height = 300,collocations=False, random_state =21, max_font_size=119). generate(allWords)

plt.imshow(wordcloud, interpolation = 'bilinear')
plt.axis('off')
plt.show()
```

Figure 21 Python Code: Implementation of a word cloud plot

The required libraries world-cloud and matplotlib are imported. Then a list with all the words detected in the tweets is created. The next step includes the definition of the word cloud, defining the width, the height of the figure as well as the maximum used font size, the collocation status and the random state. Lastly, the plot is created through the matplotlib library and is presented in the output through the plt.show() method.

Apart from the general word cloud, the word cloud for the negative and positive tweets are also created. The results will be presented and discussed in the next chapter which will include the final findings and their evaluation. The only difference is the code, which occurs in the addition of the label criteria to be equal to the sentiment representative number. For instance, the code for the implementation of the negative word cloud is shown below. It appears that for this particular word cloud, in the 'allWords' word list, only words that are included in negatively labelled tweets are added.

```
# Plot the wordcloud for negative sentiment
from wordcloud import WordCloud
import matplotlib.pyplot as plt

allWords = ' '.join([twts for twts in df['tidy_tweets'][df['label'] == 1]])
wordcloud = WordCloud(width = 500, height = 300, random_state =21, max_font_size=119). generate(allWords)

plt.imshow(wordcloud, interpolation = 'bilinear')
plt.axis('off')
plt.show()
```

Figure 22 Python Code: Implementation of the negative word cloud plot

Apart from the word clouds produced, a pie chart with the total percentage of each label was produced aiming to provide an overview of the total sentiment extracted. For this purpose, the following code was written.

```
dfo = df['label'].value_counts()
dfo.head()
```

```
3.0    263213
2.0     89334
1.0     35673
Name: label, dtype: int64
```

```
dfo.plot.pie(y='label', autopct='%1.1f%%')
```

```
<AxesSubplot:ylabel='label'>
```

Figure 23 Python Code: Pie chart of each sentiment's tweet count

Initially, a new DataFrame object was created, that contained the value counts of each label. In alignment with that, the output showed the actual number of tweets that were labelled as positive, neutral and negative, during the prediction function based on the training model. Continuing, plotting the pie chart, was possible with the use of the selected attribute on which the plot will be built and the parameter for adding the percentage in the plot.

The main objective of this work was to present a timeline overview of Twitter users' sentiment. Therefore, a line graph, with the average sentiment value per day was created. The main prerequisite for this task was that the date information would be identified as 'date' type.

```
df['Created-At'] = pd.to_datetime(df['Created-At'])
```

Figure 24 Python Code: Transform string value to Datetime

```
df['label'] = df['label'].astype(float)
dfa = df.groupby([df['Created-At'].dt.date])['label'].mean()
dfa
```

```
Created-At
2020-12-12    2.571429
2020-12-13    2.673469
2020-12-14    2.664234
2020-12-15    2.533333
2020-12-16    2.583333
                ...
2021-12-06    2.629254
2021-12-07    2.628095
2021-12-08    2.626205
2021-12-09    2.624573
2021-12-10    2.658672
Name: label, Length: 339, dtype: float64
```

Figure 25 Python Code: Transform label values to float and calculate the daily average

As shown in the code provided, the value type of the label column was changed into float from string, making arithmetic calculations possible. Following, in a new DataFrame, the records were grouped based on the date, calculating at the same time the average value of the sentiment for each day's tweets. For instance, the average sentiment for 12 Dec 2020, was 2.57 indicating that most tweets were positive.

```
from matplotlib import dates
fig, ax = plt.subplots(figsize=(20,10))
dfa.plot(figsize = (20,10))
plt.title('Average Daily Sentiment')
plt.xlabel('Date', fontsize = 14)
plt.ylabel('Average Sentiment')
plt.grid(True)
plt.legend(loc='best', ncol=2)
plt.gcf().autofmt_xdate(rotation=90)
ax.xaxis.set_major_locator(dates.DayLocator(interval=7))
```

Figure 26 Python Code: Plot implementation for the daily average sentiment

The line graph produced occurred by the execution of the code presented above. Initially, the date feature of the matplotlib library was imported. For the configuration of the plot characteristics, the size of the plot was defined, as well as the title, the axis names, the legend, the rotation of the x-axis values and the interval gap among them.

The final part of the research objective included the influence power of each sentiment. An overview of the situation can be identified by the Retweet rate per negative, neutral

and positive tweets. In the following code, the appropriate computations for the collection of the required information are presented. For each sentiment, the total number of retweets and actual tweets per sentiment is calculated.

```
dfv = df.groupby(["label"]).agg({"Retweet-Count" : "sum"})
dfv['Label count']= df['label'].value_counts()
dfv['RT/Sentiment'] = dfv['Retweet-Count']/ dfv['Label count']
dfv
```

label	Retweet-Count	Label count	RT/Sentiment
1.0	1373865.0	35673	38.512741
2.0	1540610.0	89334	17.245506
3.0	6847643.0	263213	26.015596

Figure 27 Python Code: Calculation of the retweet rate per Sentiment

Then for each sentiment, the retweet rate per sentiment is calculated as well., representing the engagement of each sentiment. In alignment with the rest results, a visualised output is requested with the use of the code shown below.

```
dfv.plot.pie(y='RT/Sentiment', autopct='%1.1f%%')
plt.title('Retweet Rates per Sentiment')
```

Figure 28 Python Code: Implementation of a pie chart with the retweet rates per Sentiment

Once again, a pie plot is requested, with indicators of each sentiments percentage of the pie. Apart from the overview, a more detailed view throughout time is also needed. For this purpose, a more extensive group-by based also on the creation date of each tweet is important for the extraction of better results and knowledge out of the given data.

```
#Work on the retweets
dfs = df.groupby(["just_date", "label"]).agg({"Retweet-Count" : "sum"})
dfs['Tweet Number'] = df.groupby(["just_date", "label"]).agg({"Retweet-Count" : "count"})
dfs['RT power'] = dfs['Retweet-Count']/dfs['Tweet Number']
dfs = dfs[['RT power']]
dfs = dfs.groupby(level=0).apply(lambda x: 100*x/x.sum())
dfs
```

		RT power
just_date	label	
2020-12-12	1.0	37.227666
	2.0	2.045476
	3.0	60.726858
2020-12-13	1.0	34.407353
	2.0	46.208222
...	...	...

Figure 29 Python Code: Calculation of daily Retweet rates per Sentiment

Through the execution of the first few lines of code, the DataFrame is grouped by both the date and the sentiment of the tweets containing as well information regarding the retweet sum and on another column information of each label's Tweet count is stored. Additionally, a third column is created, containing the daily retweet rates per sentiment. In the three last lines of code, only the more useful feature is selected, while at the same time the data are transformed into percentages, returning more easily compared results.

```
dfs.unstack().plot(figsize = (20,10), kind='bar', stacked=True ,\
                   color=['r', 'w', 'b']).xaxis.set_major_locator(dates.DayLocator(interval=7))
plt.title('Sentiment Retweet Percentage', fontsize = 14)
plt.xlabel('Date', fontsize = 14)
plt.ylabel('Percentage',fontsize = 14)
plt.legend(loc='best', ncol=3)
plt.gcf().autofmt_xdate(rotation=90)
plt.savefig('foo.png')
plt.show()
```

Figure 30 Python Code: Implementation of the retweet rates plot

Respectively, for the plot implementation of the results returned before, once again the main characteristics of the plot shall be defined. The main difference, in this case, is that there was a change in the colours of each label, focusing mainly on the positive and negative sentiments that were assigned red and blue, while neutral sentiment was represented with white aiming to declutter the final graph.

In this chapter, only the methodology and the code developed are presented and explained. The results and the graphs of each query will be shown and analysed in the next chapter.

6 Chapter 6: Results and Discussion

The main purpose of every sentiment analysis project is to extract knowledge from the gathered data. The Tweets that were gathered, cover a one-year timeframe, from 12 Dec 2020 to 10 Dec 2021. During this year, the first COVID19 vaccine was produced, shared and the mass vaccination of the world was available. Additionally, so did the rebellions and the fight against the vaccines by a part of the crowds. Aiming to identify the shift in the sentiment of Twitter Users triggered by major milestones of this year's battle against the coronavirus pandemic. Throughout this chapter, the main findings of the sentiment analysis implemented will be presented and discussed.

6.1 Twitter Topic Discussion

In every analysis that targets opinion mining, it is important to identify the topics of discussion. For instance, people on Twitter can be discussing basketball. With topic identification tasks, an analyst can understand which aspects of basketball they are engaging with. They can be tweeting for a specific team, player or even a match. A popular means to identify each field's topic is the word cloud. Researchers are including in their analysis word clouds providing a topic overview. In this research, three in total word clouds were created for the identification of topics not only for the whole dataset but also for the positive and negative sentiment sets. To start with, the word cloud that occurred with the use of the whole dataset is the following

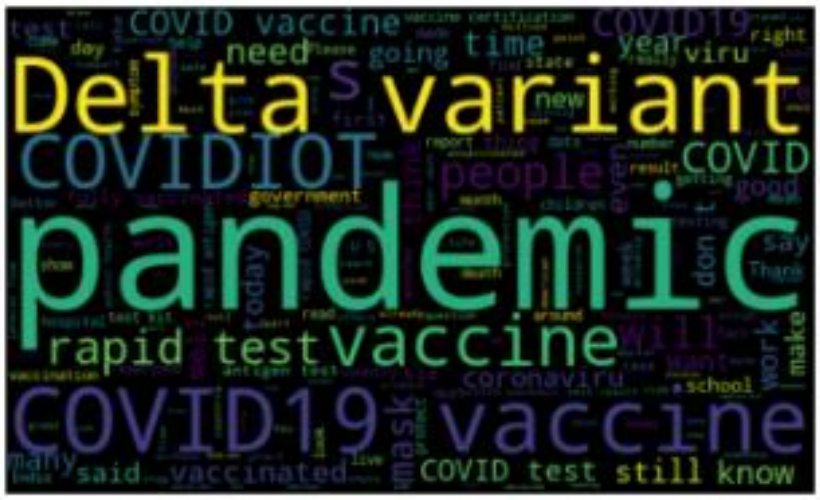

Figure 31 General word cloud

In every word cloud, the dataset the words that occur more frequently, are represented by bigger fonts. In this case, this word is 'pandemic' showing that most people are discussing the pandemic in general not only 'vaccines' which follow and 'Delta variant', which is one of the topics that gained attention the past five months. Other words that appear to dominate in the word cloud are 'COVID19' and 'COVIDIOT'. Even though the first term was expected since it is the main topic of the whole research, the second term appears to be an aggressive term that increases the polarity among the sentiments. Consequently, the word clouds of negative and positive sentiments will be presented to detect which one of the sentiments uses mainly the term 'COVIDIOT'. The word cloud in Figure 32 presents the main discussion topics of positive tweets, while Figure 33 is the negative word cloud.

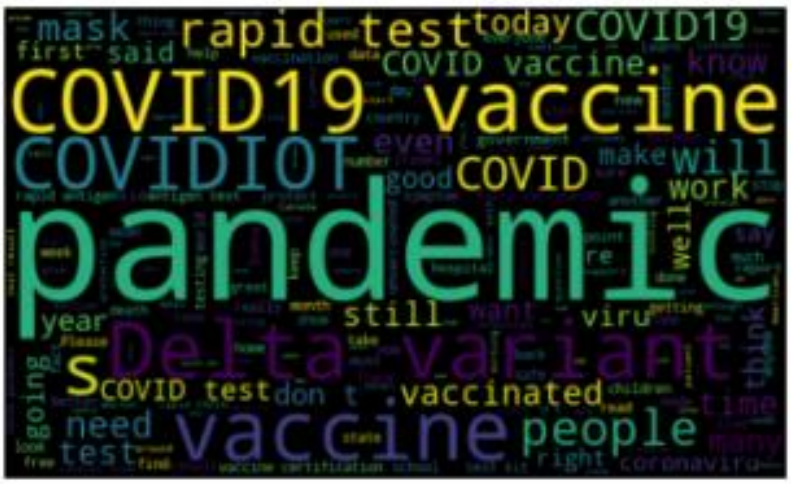

Figure 32 The positive word cloud

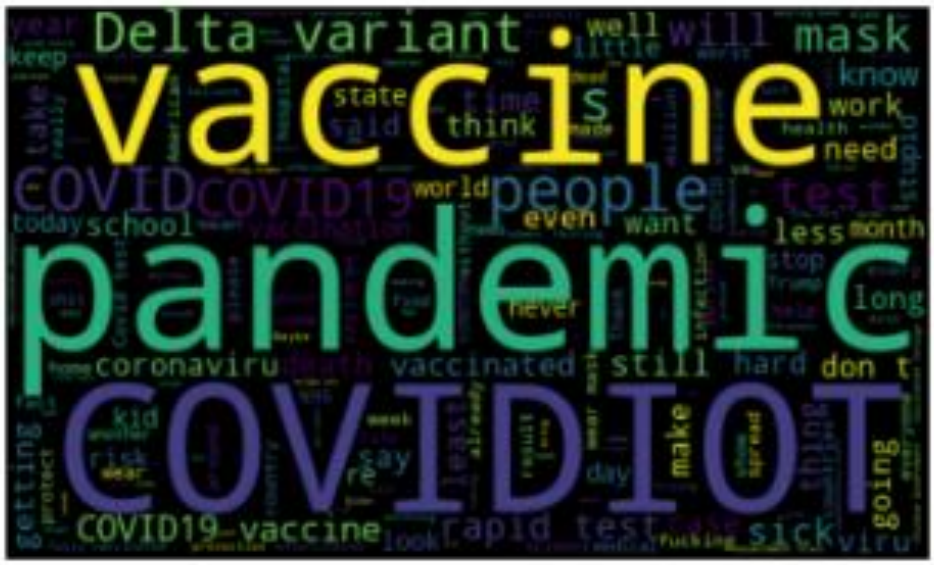

Figure 33 The negative word cloud

It is clear that in both cases, the discussion regards ‘pandemic’, ‘vaccine’ proving that all people are interested in the pandemical nature of this virus and the claimed treatment for it. In the positive word cloud, the discussion seems to be balanced among the numerous topics. They tweet about ‘vaccines’, the new variant, but apart from the topic words, there are noted some smaller very indicative words. Namely, ‘still’, ‘need’, ‘good’, ‘work’, ‘well’, ‘right’, are proof that supporters of the vaccine, try to convince the rest of users of the vaccine’s necessity. It is also noticed, that the term ‘COVIDIOT’ appears in this word cloud, leading to the assumption that a part of vaccine supporters, can be aggressive against the antivaccine group of people. On the contrary, the negative word cloud is less balanced and contains fewer terms indicating that most users are focusing on a smaller variety of topics. Respectively, for this sentiment, the most dominant terms are ‘pandemic’, ‘vaccine’, ‘COVIDIOT’. It should be noted that Twitter users of this sentiment, keep the discussion on the pandemic and vaccine aspects of the virus, however, they appear to be aggressive against the situation, using the term and characterizing the supporters of vaccines and believers of the situation, as ‘idiots’. The term ‘delta variant’ is not that dominant in this sentiment. Other less dominant terms include ‘mask’, ‘work’, ‘sick’, which indicate their focus on more practical aspects of the topic. Researchers [14] claim that most negatively driven users, use the ‘mask’ topic to express their opposition to political or ideological beliefs rather than to the actual effects of masks on health and covid19 spread prevention. Overall, each sentiment follows a different approach, whether

that is motivation to take the vaccine shot, or aggression to the whole situation, using the pandemic to express political disapproval.

6.2 Sentiment Overview

The dataset that participated in the analysis was consisting initially of 611022 Tweets in total, 384649 collected with the use of the Twitter API v2 and 226373 from the external dataset. With the preprocessing and the cleansing of the data, the total number of records was decreased to 388220. With the prediction of the label per tweet, the extraction of basic descriptive statistics is possible. It appears that the majority of the tweets are positive while negative tweets represent the minority of the tweets. Showing clearly, that Twitter users are positive and support COVID19 vaccines.

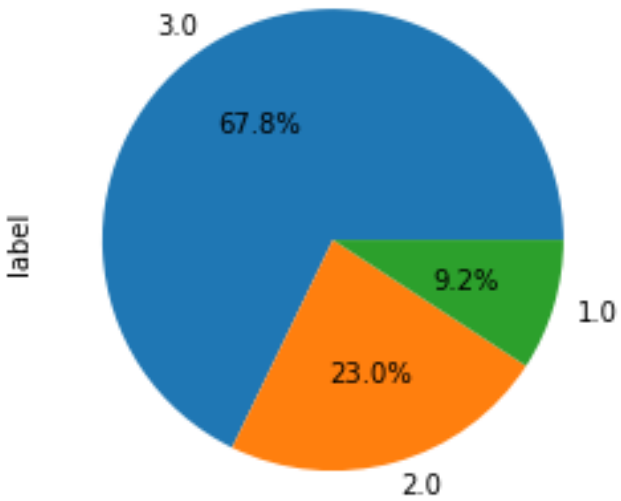

Figure 34 Sentiment Tweet Volume Pie chart

In more detail, as shown in the graph above, positive tweets that are labelled with '3,0' consist the 67,8% of the records, showing the support and belief of users for the vaccines released since last year. As Neutral were labelled the 23.0% while only 9.2% of the tweets were labelled as Negative, indicating disbelief, anger and disapproval of the vaccines.

Regarding the retweets per sentiment, some controversial findings occurred. It is found that the number of retweets compared to the actual number of each sentiment tweets show a different dynamic of each label.

label	Retweet-Count	Label count	RT/Sentiment
1.0	1373865.0	35673	38.512741
2.0	1540610.0	89334	17.245506
3.0	6847643.0	263213	26.015596

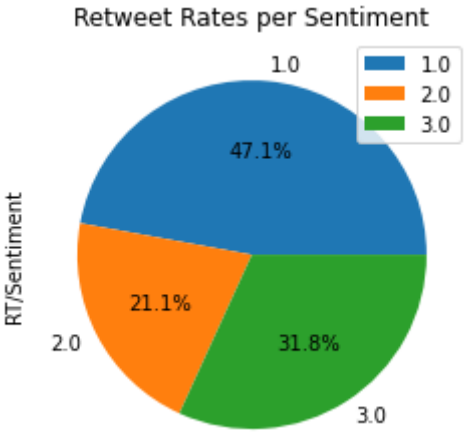

Figure 35 (a) Data of Retweet rate per Sentiment (b) Implemented Pie chart

one of the labels, the sum of Retweets is calculated, as well as the main tweets that have been assigned in each label. The third column shows the rate of retweets per tweet. For this case, it shows that Negative Tweets tend to have a higher influence and retweet rate per Tweet. These main findings require further investigation for a better understanding of the situation and more accurate knowledge regarding the data.

6.3 Twitter Daily Sentiment Analysis

The main purpose of this thesis is to create a sentiment timeline of Twitter users regarding coronavirus and the COVID19 vaccines, aiming to identify, how critical moments and events during the past year, have affected public opinion. In this paragraph, this relationship will be examined. In the following graph, the average sentiment extracted every day is presented.

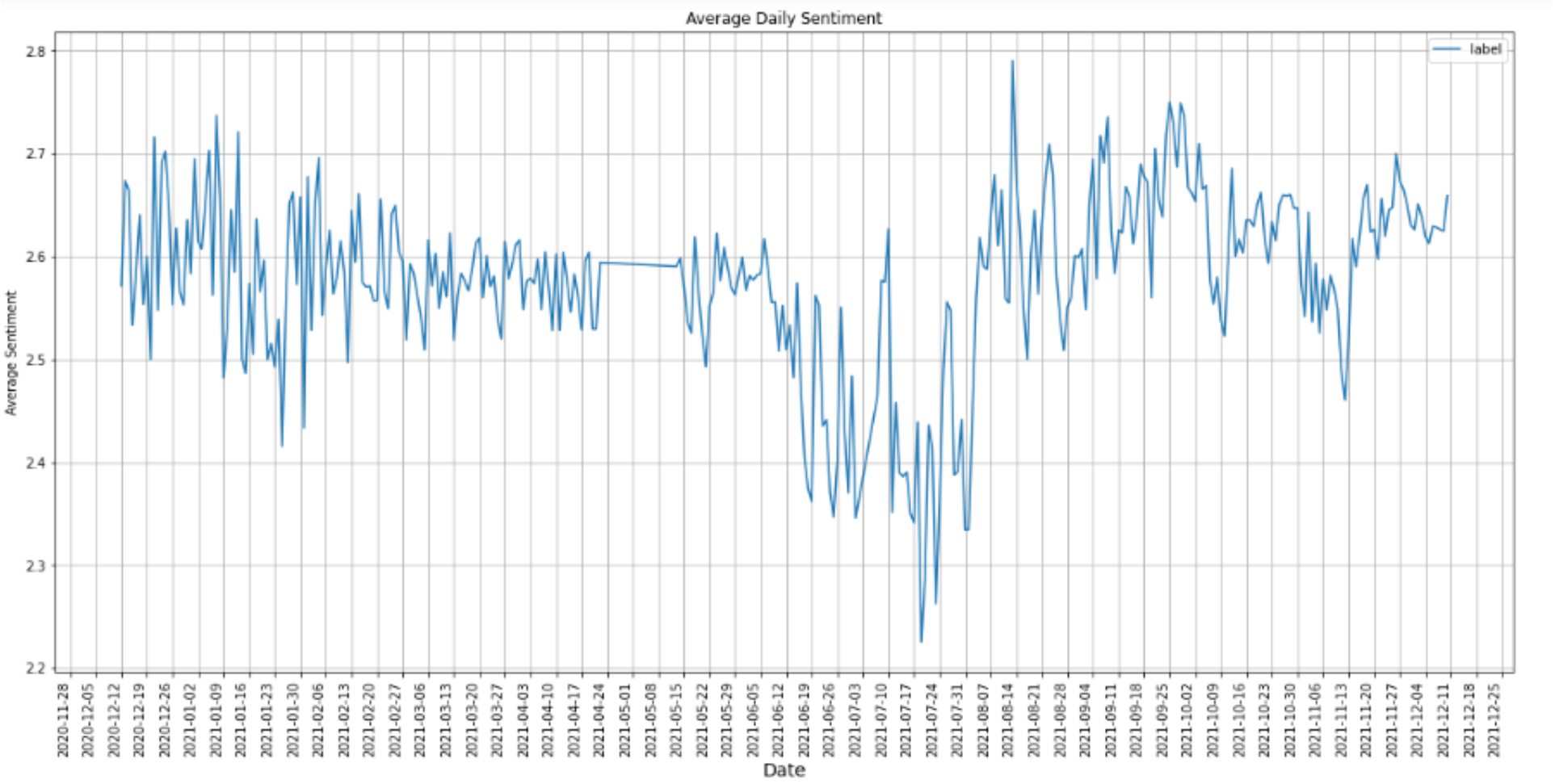

Figure 36 Daily Average Sentiment Plot

It is found that the average sentiment during the past year, is taking values from 2,25 to 2,8 indicating once again that the sentiment was in its majority neutral to positive. At a first glance, the lowest sentiment was noted in the week of 17 Jun 2021 to 24 Jun 2021, while the highest was almost a month later during the dates 07 Aug 2021 and 14 Aug 2021. Overall, it can be considered a rather neutral sentiment, which however can also indicate a big polarity of the data, considering that the majority of the tweets were labelled as positive. Based on the plot presented above, it is obvious that sentiment was shifting daily. Beginning from 12 Dec 2021, the sentiment was approximately 2,55 while until the next week the sentiment has once again in the same levels after small changes in between. The positive peak for the given period is in the first week of January. It should be noted that on 31 Dec 2020, the WHO issued an emergency use of coronavirus vaccines focusing on global access, and on 5 Jan, Pfizer was the first vaccine to receive emergency use validation. It is worth mentioning that on 9 Jan, the sentiment was dropped to 2,48 as a possible response to the new variant reported from Japanese authorities. One of the most intense decreases was noted at the end of Jan 2021, where the sentiment average reached approximately 2,42. Considering the timeline with the most important events provided in Chapter 2, on 25 Jan the WHO released recommendations for the use of the Moderna vaccine, on 29 Jan the organization publishes a list of recommended COVID19 tests. These events seem to have pressured people, leading them to show their discomfort on Twitter. The situation could be characterized as stable in the next few months with no intense rises and drops in the sentiment. During this period, most of the vaccines received emergency use validation, but, on the other hand, it was also the time that the AstraZeneca vaccine was indicated to cause blood coagulation, resulting in the spread of fear and uncertainty among people. The most intense general drop in the sentiment average through the past year was at the beginning of summer, especially in the week between 12 Jun 2021 and 19 Jun 2021. Even though the WHO continue to inform people regularly on scientific news and updates, the sentiment analysis showed that people had a drop in positivity and faith regarding the course of the pandemic. Researching further on the information published related to coronavirus, it appears that on 14 Jun 2021, the president of the United Kingdom, Boris Johnson announced the extension of the lockdown for another four weeks based on the threat of the Delta Variant. Based on the fact that the selected tweets are written in English, it is more than logical to assume that the majority of the users are located in the United Kingdom and therefore the sentiment reflects their discomfort. The United States also speak English and based on data retrieved from the WHO, they were

facing the start of the second wave of the pandemic, which showed a decrease in the middle of September, but then the cases started to increase again [72].

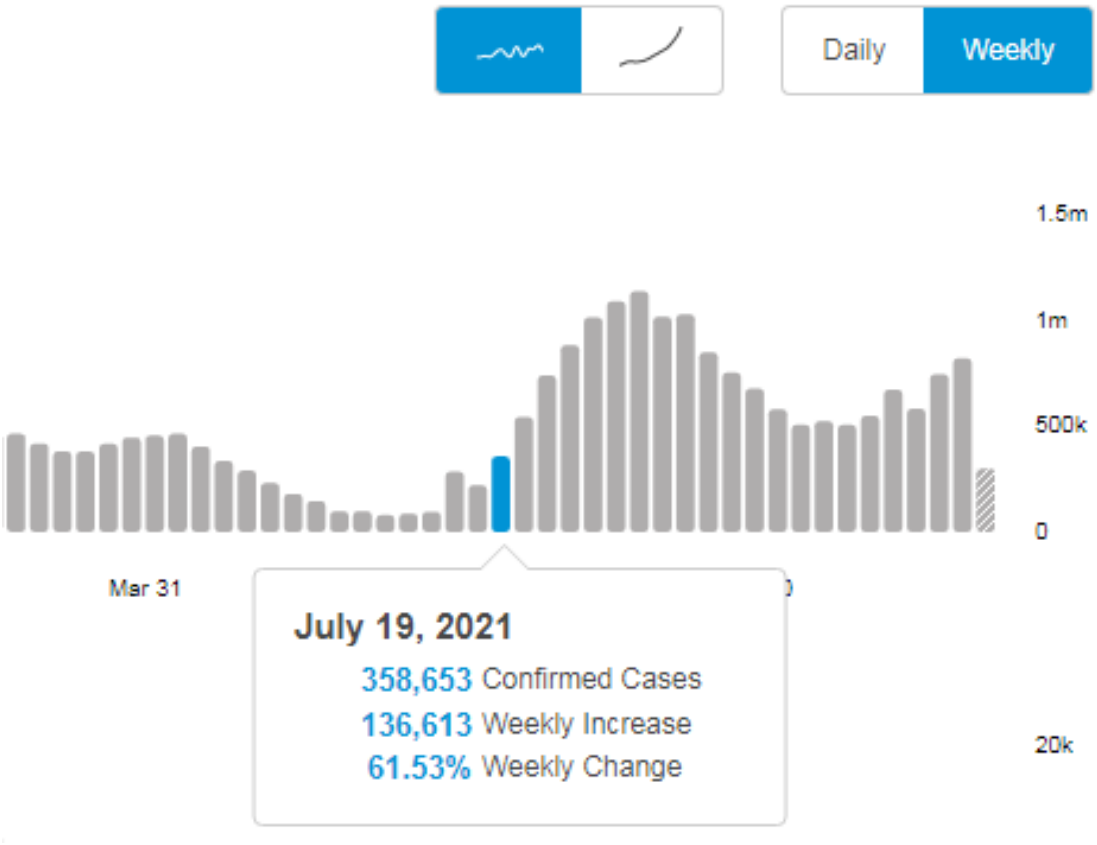

Figure 37 Weekly COVID19 cases of the United States

The fact that at the same period new recommendations for the vaccines by the WHO enhance negativity as people started to understand that vaccination is not enough to stop the pandemic, with new variants evolving each day. The lowest sentiment detected was on the week 17 to 24 July 2021, when the sentiment average dropped to approximately 2,23. During that period, the United States kept fighting against the latest wave of the COVID19 pandemic, while the United Kingdom was also facing the same problem (WHO, 2021).

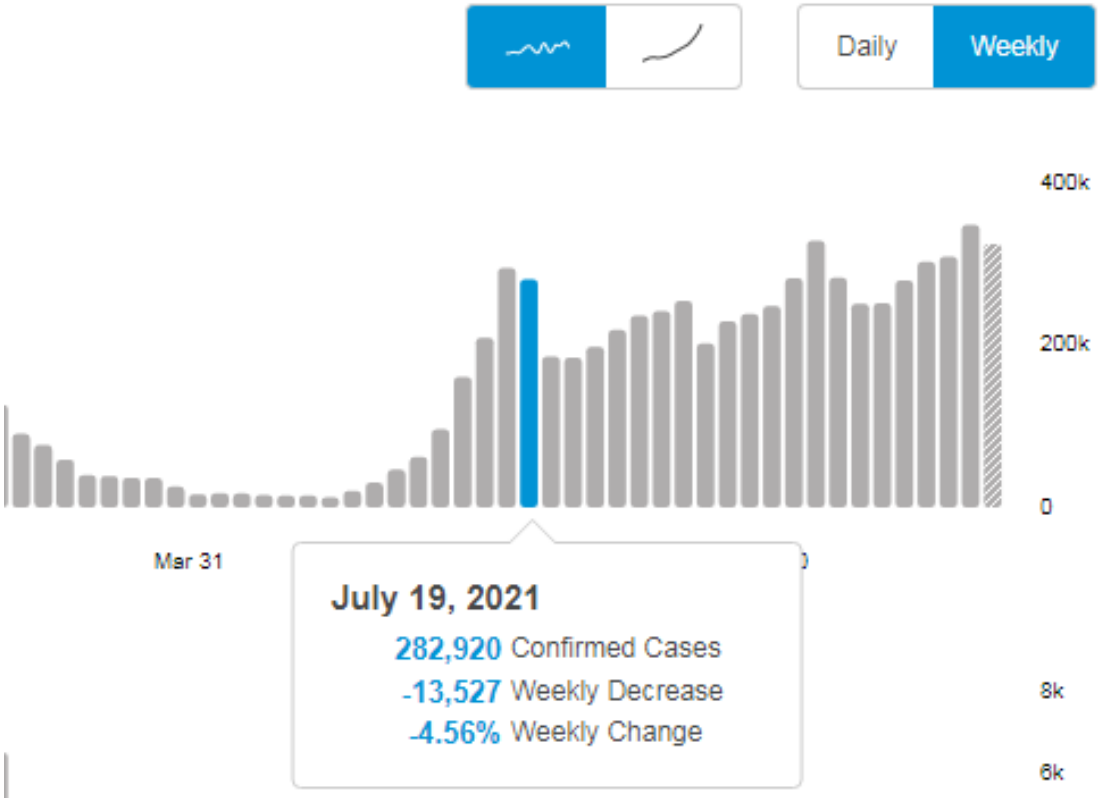

Figure 38 Weekly COVID19 cases of the United Kingdom

Since then, the sentiment has improved rapidly, showing that people are hoping for the end of the pandemic. This happened due to important events that happened in August. The higher sentiment average was noted in the middle of August reaching almost 2,8 with 3 representing a purely positive dataset. The information leading to this result could be the hope for booster doses, which were announced in the US on 18 Aug 2021 and the FDA approval of the Pfizer vaccine, on 28 Aug 2021, which allowed people to feel more certain that the vaccine is safe and effective. Until 10 Dec 2021, the Sentiment was maintained on high levels, with a small drop on 12 Nov 2021. At that period, once again cases started to increase, restrictions to non-vaccinated people are applied in Europe. A week later, Austria is the first European country to impose a total lockdown both for vaccinated and non-vaccinated. Lastly, at the end of November, the Omicron variant was announced, without affecting the total sentiment of the Twitter community.

Sentiment Influential Power

Tweets are used so that everyone can express freely their opinion. However, the power of each tweet differs depending on the influence they have on the rest users. The influence

of a tweet is described by several metrics including replies if exists link clicks, mentions and retweets [42]. In the analysed dataset rich information on retweets was contained. Therefore an attempt to extract some knowledge on the influential power of each sentiment presented. The methodology of the process was described in the previous chapter. Based on the findings of a similar approach from Medford et al. [40], the tweets with the most retweets are those with a negative context. In this research, their finding can be also be verified, from the pie chart that was presented in the previous chapter. The graph that was resulted from the analysis is the one presented on the following page. The findings are also interesting and require thorough explanation. Additionally, this is an opportunity to either verify or doubt our initial findings. As is already mentioned, neutral retweets, are represented by white, to emphasize the polarity of positive and negative tests. Even at first glance, it appears that the results are equally shared among positive and negative tweets, in contrast to the positive dominance of simple tweets. It should be noted in this point that for each tweet only the retweet count is extracted without the specific retweet number per day, offering a less precise sentiment influential power daily. However, it can indeed prove the polarity of the sentiments as well as the power of negative tweets compared to their little volume in the previous analysis. Daily analysis of the sentiments will not be implemented in this paragraph. Despite this, the main points will be discussed as well. Overall an alignment with the actual sentiment average is noted. Days that the sentiment average is decreasing, negative tweets seem to dominate against positive and negative. For instance, the higher negative percentage is achieved on the first week as well in the second week of November, when the average sentiment also dropped significantly. Respectively, positive percentages rise with the rise of the sentiment average.

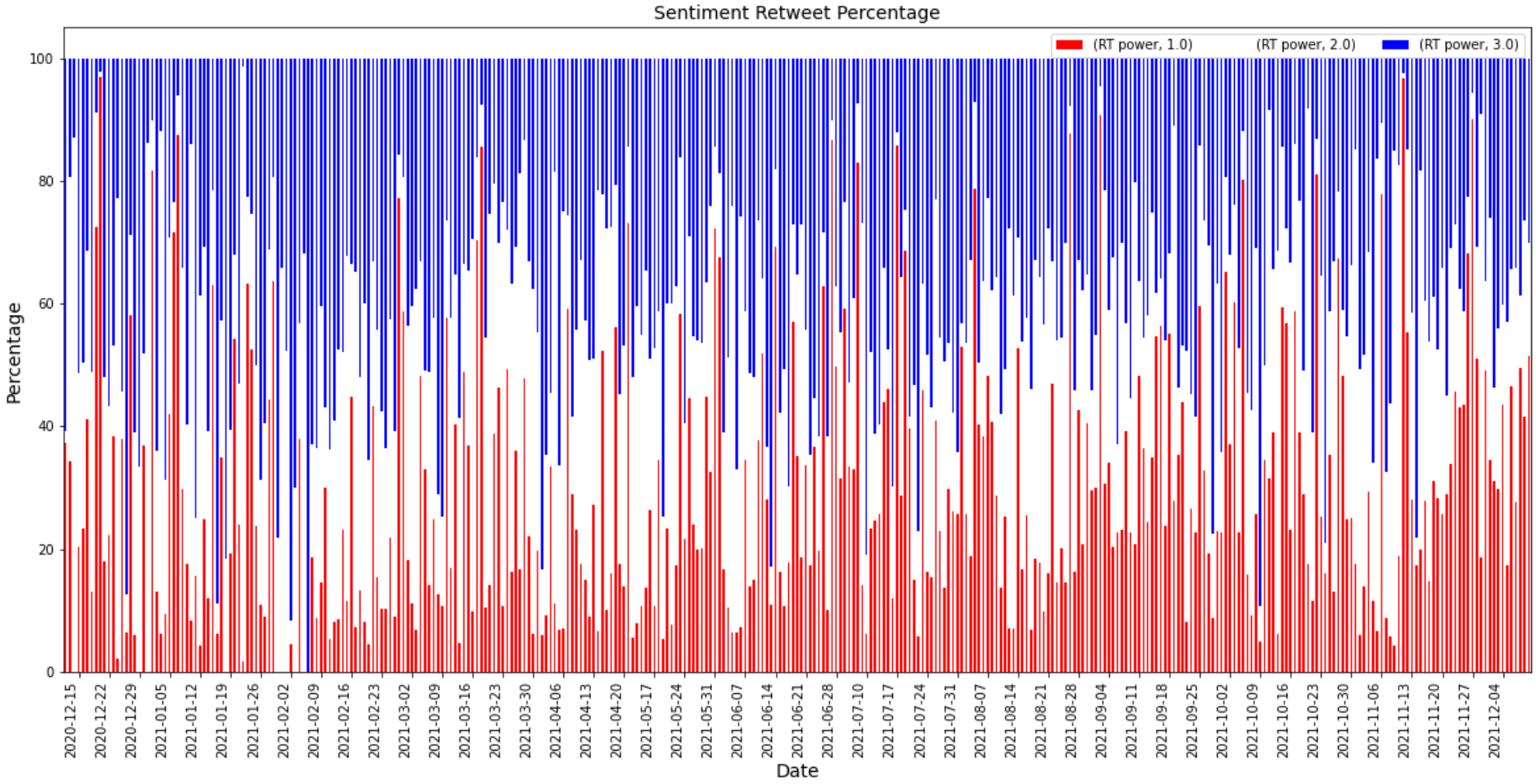

Figure 39 Daily Sentiment Retweet rates

7 Chapter 7: Conclusion and Future work

In the final chapter of this thesis, the conclusions of the research will be mentioned as well as the threats to validity that may have affected negatively the outcome of the sentiment analysis implemented.

7.1 Conclusion

COVID19 since its beginning in 2019 has become part of everyone's life, affecting every aspect of life as known till then. People faced lockdowns, restrictions and intense measures that the WHO and their governments have applied to them. Social media during this period was not only a means of communication with relatives and friends but also a very strong means for people to speak their minds and express their opinion publicly whether positive or not. In this research, the main terms and approaches are provided, offering a theoretical overview of the sentiment analysis topic.

In the second part of the dissertation, the sentiment analysis project was implemented and executed extracting the final results that reply to the main objectives of the study. First of all, word clouds were produced revealing the main discussion topic regarding the virus. It is shown that Twitter users are more interested in the pandemic, the vaccines and being aggressive for the opposite sentiment. Moving to the quantitative results, it is found that the majority of the tweets are positive showing that people are positive for the course of the pandemic, supporting the work of scientists, governments and the WHO. Furthermore, they appear to outperform the rest sentiments through time. On the contrary, the negative sentiment tends to have a higher engagement rate, since their tweets have outperformed the rest of sentiments on retweet rates. Therefore, it can be assumed that people that are against the vaccine and are tired of the pandemic, are not so confident in expressing their personal beliefs, but prefer to share already published thoughts. Based on this

characteristic of the negative sentiment the viral spread of misinformation and creation of fear can be explained.

This research has offered some useful knowledge on the timely relationship among Twitter users and the most significant events since the beginning of the pandemic and the vaccinations. However, there are some threats to the validity of this work. The lack of the dataset description, which was used for the building of the training model, can be of crucial importance for the evaluation of the research results. Nevertheless, the dataset was manually evaluated to identify whether the content was indeed valid. Additionally, the external dataset which was used for the enriching of the final data, even though it contained tweets from the December of 2020, the volume of tweets was small, restricting the accuracy of the sentiment per day. It can be concluded that researchers, that base their findings on the work of external datasets are prone to bias created from the original owners of the datasets. The sentiment of every domain is a multidimensional feature, for which the more specialized the analysis the more precise and accurate the findings will be. Therefore, since this is a more general approach, containing tweets in English, without geological restrictions, it can only be considered as an estimation of the actual situation. Considering that every country reacted differently in their battle with the virus, personalized analysis for a specific country, with their government's response to the pandemic waves would be more easily managed and manipulated.

7.2 Future work

This work can be used as a basis and inspiration for many future works. Scientists can use be influenced by this work, and conduct similar analyses, focusing either on a specific geological location or on a more precise topic.

Since the pandemic and its relationship with people is affected by various aspects like social factors, health factors, political and financial factors, researchers could attempt and extract knowledge based on this research considering as well these parameters. The main objective of such research would be a better understanding of how Twitter users are affected by all these factors.

Of course, scientists can use this research as a reference to enhance or doubt their findings on similar attempts. While an expansion of the research would be to apply the training

model to a bigger dataset, which would contain information for a more extended period or a bigger volume of data daily, in combination with a better performing algorithm providing more accurate results.

Finally, every scientist would be allowed to use the followed methodology for the extraction of information for every domain of interest apart from coronavirus.

References

1. Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications, 125(3).

2. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on language in social media (LSM 2011) (pp. 30-38).

3. Aiello, L. M., Quercia, D., Zhou, K., Constantinides, M., Šćepanović, S., & Joglekar, S. (2021). How epidemic psychology works on Twitter: evolution of responses to the COVID-19 pandemic in the US. Humanities and Social Sciences Communications, 8(1), 1-15.

4. An introduction to seaborn — seaborn 0.11.2 documentation. (2012). Retrieved December 07, 2021, from Pydata.org website: https://seaborn.pydata.org/introduction.html

5. Arzoo, M. K., Prof, A., & Rathod, K. (2017). K-Means algorithm with different distance metrics in spatial data mining with uses of NetBeans IDE 8. 2. Int. Res. J. Eng. Technol, 4(4), 2363-2368.

6. Atalan, A. (2020). Is the lockdown important to prevent the COVID-19 pandemic? Effects on psychology, environment and economy-perspective. Annals of medicine and surgery, 56, 38-42.

7. Balahur, A., & Steinberger, R. (2009). Rethinking Sentiment Analysis in the News: from Theory to Practice and back. Proceeding of WOMSA, 9.

8. BBC News. (2021, November 19). Austria to go into full lockdown as Covid surges. Retrieved December 6, 2021, from BBC News website: https://www.bbc.com/news/world-europe-59343650

9. Beleveslis, D., Tjortjis, C., Psaradelis, D., & Nikoglou, D. (2019, September). A hybrid method for sentiment analysis of election related tweets. In 2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 1-6). IEEE.

10. Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J., & Klein, D. (2010, June). Painless unsupervised learning with features. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 582-590).

11. Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. JMIR Public Health and Surveillance, 6(4), e21978.

12. Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. In A practical guide to sentiment analysis (pp. 1-10). Springer, Cham.

13. Carl O'donnell,Ahmed Aboulenein. (2021, August 18). U.S. to begin offering COVID-19 vaccine booster shots in September. Retrieved December 20, 2021, from Reuters website: https://www.reuters.com/world/us/us-begin-offering-covid-19-vaccine-booster-shots-september-2021-08-18/

14. Cerbin, L., De Jesus, J., Warnken & Gokhale, S. S. (2021). Understanding the anti-mask debate on social media using machine learning techniques. International Journal of Computers and their Applications, 150-161.

15. da Silva, N. F. F., Coletta, L. F., Hruschka, E. R., & Hruschka Jr, E. R. (2016). Using unsupervised information to improve semi-supervised tweet sentiment classification. Information Sciences, 355, 348-365.

16. Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (pp. 519-528).

17. De, A., & Kopparapu, S. K. (2013, August). Unsupervised clustering technique to harness ideas from an Ideas Portal. In 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1563-1568). IEEE.

18. Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: a comparative study on different approaches. Procedia Computer Science, 87, 44-49.

19. Dhruv Dhawan. (2020). Sentimental analysis of covid-19 tweets. Retrieved December 6, 2021, from Kaggle.com website: https://www.kaggle.com/dhruvdhawan/sentimental-analysis-of-covid19-tweets

20. Droba, D. D. (1931). Methods used for measuring public opinion. American Journal of Sociology, 37(3), 410-423.

21. Dubey, A. D. (2020). Twitter Sentiment Analysis during COVID-19 Outbreak. Available at SSRN 3572023.

22. Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.

23. Garcia, K., & Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. Applied Soft Computing, 101, 107057.

24. Goel, A., Gautam, J., & Kumar, S. (2016, October). Real time sentiment analysis of tweets using Naive Bayes. In 2016 2nd International Conference on Next Generation Computing Technologies (NGCT) (pp. 257-261). IEEE.

25. Gupta, D., & Kadakia, K. Aspect-Based Sentiment Analysis Report.

26. Gupta, N., & Agrawal, R. (2020). Application and techniques of opinion mining. In Hybrid Computational Intelligence (pp. 1-23). Academic Press.

27. Hemmatian, F., & Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. Artificial Intelligence Review, 52(3), 1495-1545.

28. Hung, M. C., & Yang, D. L. (2001, November). An efficient fuzzy c-means clustering algorithm. In Proceedings 2001 IEEE International Conference on Data Mining (pp. 225-232). IEEE.

29. Jianqiang, Z., Xiaolin, G., & Xuejun, Z. (2018). Deep convolution neural networks for twitter sentiment analysis. IEEE Access, 6, 23253-23260.

30. Kapoor, A., & Singhal, A. (2017, February). A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms. In 2017 3rd international conference on computational intelligence & communication technology (CICT) (pp. 1-6). IEEE.

31. Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., & Alfakeeh, A. S. (2020). Stock market prediction using machine learning classifiers and social media, news. Journal of Ambient Intelligence and Humanized Computing, 1-24.

32. Kruspe, A., Häberle, M., Kuhn, I., & Zhu, X. X. (2020). Cross-language sentiment analysis of european twitter messages during the covid-19 pandemic. arXiv preprint arXiv:2008.12172.

33. Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of natural language processing, 2(2010), 627-666.

34. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

35. Lyu, J. C., Le Han, E., & Luli, G. K. (2021). COVID-19 vaccine–related discussion on Twitter: topic modeling and sentiment analysis. Journal of medical Internet research, 23(6), e24435.

36. Madhoushi, Z., Hamdan, A. R., & Zainudin, S. (2015, July). Sentiment analysis techniques in recent works. In 2015 science and information conference (SAI) (pp. 288-291). IEEE.

37. Maldonado, S., Weber, R., & Famili, F. (2014). Feature selection for high-dimensional class-imbalanced data sets using support vector machines. Information sciences, 286, 228-246.

38. Manguri, K. H., Ramadhan, R. N., & Amin, P. R. M. (2020). Twitter sentiment analysis on worldwide COVID-19 outbreaks. Kurdistan Journal of Applied Research, 54-65.

39. Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review, 27, 16-32

40. Medford, R. J., Saleh, S. N., Sumarsono, A., Perl, T. M., & Lehmann, C. U. (2020, July). An "infodemic": leveraging high-volume Twitter data to understand early

public sentiment for the coronavirus disease 2019 outbreak. In Open Forum Infectious Diseases (Vol. 7, No. 7, p. ofaa258). US: Oxford University Press.

41. Morton, B. (2021, June 14). Covid: Lockdown easing in England to be delayed by four- weeks. Retrieved December 20, 2021, from BBC News website: https://www.bbc.com/news/uk-57464097

42. Muñoz-Expósito, M., Oviedo-García, M. Á., & Castellanos-Verdugo, M. (2017). How to measure engagement in Twitter: advancing a metric. Internet Research.

43. Nasukawa, T., & Yi, J. (2003, October). Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd international conference on Knowledge capture (pp. 70-77).

44. Nemes, L., & Kiss, A. (2021). Social media sentiment analysis based on COVID-19. Journal of Information and Telecommunication, 5(1), 1-15.

45. Noble, W. S. (2006). What is a support vector machine?. Nature biotechnology, 24(12), 1565-1567.

46. NumPy. (2021). Retrieved December 07, 2021, from Numpy.org website: https://numpy.org/

47. Office of the Commissioner. (2021). FDA Approves First COVID-19 Vaccine. Retrieved December 6, 2021, from U.S. Food and Drug Administration website: https://www.fda.gov/news-events/press-announcements/fda-approves-first-covid-19-vaccine

48. Office of the Commissioner. (2021). FDA Authorizes Booster Dose of Pfizer-BioNTech COVID-19 Vaccine for Certain Populations. Retrieved December 6, 2021, from U.S. Food and Drug Administration website: https://www.fda.gov/news-events/press-announcements/fda-authorizes-booster-dose-pfizer-biontech-covid-19-vaccine-certain-populations

49. Office of the Commissioner. (2021). Coronavirus (COVID-19) Update: FDA Expands Eligibility for COVID-19 Vaccine Boosters. Retrieved December 6, 2021, from U.S. Food and Drug Administration website: https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-expands-eligibility-covid-19-vaccine-boosters

50. Oikonomou, L., & Tjortjis, C. (2018, September). A method for predicting the winner of the usa presidential elections using data extracted from twitter. In 2018 South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conference (SEEDA_CECNSM) (pp. 1-8). IEEE.

51. pandas - Python Data Analysis Library. (2015). Retrieved December 07, 2021, from Pydata.org website: https://pandas.pydata.org/about/

52. Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2017). Challenges of sentiment analysis in social networks: an overview. Sentiment analysis in social networks, 1-11.

53. Preda, G. (2021). COVID-19 All Vaccines Tweets. Retrieved December 6, 2021, from Kaggle.com website: https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets

54. Raghuvanshi, N., & Patil, J. M. (2016, March). A brief review on sentiment analysis. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 2827-2831). IEEE.

55. re — Regular expression operations — Python 3.10.1 documentation. (2021). Retrieved December 23, 2021, from Python.org website: https://docs.python.org/3/library/re.html

56. Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

57. Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D., & Roser, M. (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource]

58. Saad, S., & Aref, M. (2020). A survey on sentiment analysis in tourism. International Journal of Intelligent Computing and Information Sciences, 20(1), 1-20.

59. Sadia, A., Khan, F., & Bashir, F. (2018). An overview of lexicon-based approach for sentiment analysis. In 3rd International Electrical Engineering Conference.

60. Samal, B., Behera, A. K., & Panda, M. (2017, May). Performance analysis of supervised machine learning techniques for sentiment analysis. In 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS) (pp. 128-133). IEEE.

61. Sanders, A. C., White, R. C., Severson, L. S., Ma, R., McQueen, R., Paulo, H. C. A., ... & Bennett, K. P. (2021). Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse. medRxiv, 2020-08.

62. scikit-learn: machine learning in Python — scikit-learn 1.0.1 documentation. (2021). Retrieved December 07, 2021, from Scikit-learn.org website: https://scikit-learn.org/stable/

63. Shayaa, S., Jaafar, N. I., Bahri, S., Sulaiman, A., Wai, P. S., Chung, Y. W., ... & Al-Garadi, M. A. (2018). Sentiment analysis of big data: Methods, applications, and open challenges. IEEE Access, 6, 37807-37827.

64. Shi, Y., Wang, G., Cai, X. P., Deng, J. W., Zheng, L., Zhu, H. H., ... & Chen, Z. (2020). An overview of COVID-19. Journal of Zhejiang University-SCIENCE B, 21(5), 343-360.

65. Silva, N. F. F. D., Coletta, L. F., & Hruschka, E. R. (2016). A survey and comparative study of tweet sentiment analysis via semi-supervised learning. ACM Computing Surveys (CSUR), 49(1), 1-26.

66. Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE Access, 8, 80716-80727.

67. Singhal, T. (2020). A review of coronavirus disease-2019 (COVID-19). The indian journal of pediatrics, 87(4), 281-286.

68. Swain, P. H., & Hauska, H. (1977). The decision tree classifier: Design and potential. IEEE Transactions on Geoscience Electronics, 15(3), 142-147.

69. Timeline: WHO's COVID-19 response. (2021). Retrieved December 6, 2021, from Who.int website: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline

70. Tracking SARS-CoV-2 variants. (2021). Retrieved December 6, 2021, from Who.int website: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

71. Tsiara, E., & Tjortjis, C. (2020, June). Using Twitter to predict chart position for songs. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 62-72). Springer, Cham.

72. United States of America: WHO Coronavirus Disease (COVID-19) Dashboard With Vaccination Data. (2020). Retrieved December 20, 2021, from Who.int website: https://covid19.who.int/region/amro/country/us

73. Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. Machine Learning, 109(2), 373-440.

74. Wang, W., Zhang, Y., Li, Y., & Zhang, X. (2006, June). The global fuzzy c-means clustering algorithm. In 2006 6th World Congress on Intelligent Control and Automation (Vol. 1, pp. 3604-3607). IEEE.

75. WHO news updates. (2021). Retrieved December 20, 2021, from Who.int website: https://www.who.int/news-room/news-updates

76. World Health Organization: WHO. (2021, November 28). Update on Omicron. Retrieved December 6, 2021, from Who.int website: https://www.who.int/news/item/28-11-2021-update-on-omicron

77. Xiang, B., & Zhou, L. (2014, June). Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 434-439).

78. Zhu, X. J. (2005). Semi-supervised learning literature survey.