"Sentiment Analysis on English and Greek Twitter Data towards vaccinations"

Chrysoula Dontaki

SID: 3308210010

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

JANUARY 2023

THESSALONIKI – GREECE

"Sentiment Analysis on English and Greek Twitter Data towards vaccinations"

Chrysoula Dontaki

SID: 3308210010

Supervisor:	Assoc. Prof. Christos Tjortjis
Supervising Committee Members:	Dr. Paraskevas Koukaras
	Dr. Aikaterini Tzafilkou

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

JANUARY 2023

THESSALONIKI – GREECE

Acknowledgments

The dissertation began to be prepared in July 2022 and was completed in January 2023. Its source and inspiration consisted of the courses Programming for Data Science, Advanced Database Systems and Data Mining.

First and foremost, I would like to express my deep gratitude to my supervisor, Dr. Christos Tjortjis, for his patient guidance and his useful suggestions for this research work. His willingness to give his time so generously combined with our frequent scheduled meetings was very much appreciated.

I would also like to extend my thanks to the IHU premises for their availability and their support in any requests.

I am very grateful to my family and all the people inside and outside of the academic circle whose support was catalytic in getting me where I am today.

Abstract

This dissertation focuses on Twitter sentiment analysis related to COVID-19 vaccines in English and Greek language. This dissertation was written as part of the MSc in Data Science at the International Hellenic University.

The COVID-19 pandemic caused by the coronavirus SARS-CoV-2 originated in China in December 2019 [1]. The virus has infected and killed thousands of people according to the World Health Organization (WHO) has announced the COVID-19 outbreak as a pandemic that has hit the world [2]. An end to this pandemic can bring a worldwide vaccination campaign. However, vaccines have traditionally been met with public fear and hesitancy. During the lockdown imposed to many countries, people spent hours every day on social media platforms sharing their opinions and expressing their feelings. As a result, Twitter has become a valuable main resource for gathering information about people's emotions towards SARS-CoV-2 vaccination. Extracting useful knowledge from naturally written texts is important for governments and health experts to understand people's beliefs and establish effective campaign ideas, to increase vaccination acceptance. Therefore, the sentiment analysis process of classifying opinions towards vaccines like "positive", "negative" or "neutral" can yield remarkable findings.

To be more precise, the goal of this study is to classify people who are in favor or against vaccination, as well as people's preferences for the three types of vaccines (Pfizer, Moderna, AstraZeneca) that are available today. Luckily, this task can be automated with the power of Machine Learning (ML) and Natural Language Processing (NLP). Twitter data have been retrieved in portions at different points of time during a period of seven months using Python programming language. After data preprocessing, the sentiment analysis was conducted using TextBlob, Valence Aware Dictionary and sEntiment Reasoner (VADER), AFINN and NRC tools. Graphical representation and performance analysis with state-of-the-art models (Logistic Regression, Decision Tree, Random Forest, XGBoost, and SVM Classifier) have been conducted on the tweets.

Our results indicate that when using **English** '**summer**' tweets from Twitter with **TextBlob** as a sentiment analysis tool, **DT** is the ML algorithm that gives the highest accuracy equal to **97.99%** and F1-Score equal to 97.98%. In the **autumn** period, **DT** demonstrates again the best performance with an accuracy equal to **97.94%**. The accuracy rate was slightly reduced to 0.05%. When examining the classification performance

results for the **Greek** language dataset, it is observed that the algorithms have the ability to distinguish better in the Greek language when a tweet has a positive, negative or neutral mood. **DT** was again the winner with **99.89%** accuracy and 99.88% F1-Score. Regarding the **autumn** period, the performance of DT **improved** by 0.03% reaching **99.92%**.

Keywords: COVID-19, Pandemic, CoronaVirus, Twitter, Sentiment Analysis, Data Mining on Twitter, Machine Learning

I hereby declare that the work submitted is mine and that where I have made use of another's work, I have attributed the source(s) according to the Regulations set in the Student's Handbook.

Table of Contents

Tables

Figures

1. Introduction

Twitter is one of the most popular social media platforms which has over 300 million accounts and the numbers are rapidly increasing daily [3]. Millions of people give their opinion on different topics on social media platforms, still, Twitter is the richest source to learn about people's opinions and sentiment [4]. Sentiment-aware systems these days have many applications from business to the social sciences [5]. Coronavirus, known as COVID-19 has been one of the most discussed spreading diseases worldwide.

To control this treatment, several vaccines have been developed and approved. Less than 1 year after the declaration of the pandemic, the Pfizer vaccine was the first to get approved for the widespread use and more specifically it was authorized for use in the United Kingdom on 2 December 2020 [6]. Marcec, and Likic investigated that the Western world relies mostly on messenger RNA (mRNA) vaccines developed by Pfizer and Moderna, as well as on the ChAdOx1 vaccine from AstraZeneca/ Oxford. However, the low acceptance and worries about the efficacy of vaccines are also present due to the poor and insufficient information that people have, which is significantly influenced by social media use [7].

Therefore, it is vital to learn and analyze users' comments and reviews on vaccines to help health staff and government organizations to get benefit from those data. Even though vaccines have been tested and their safety and effectiveness are confirmed by medical scientists, a significant number of people are still hesitant about them.

Hence, we decided to look at the feelings of people, derogatorily called anti-vaxxers, separately. For each tweet, it is important to determine whether its emotion is positive, negative, or neutral. Also, an issue that creates difficulties in its analysis is the limitation of 280 characters, which in addition to words, letters, numbers and symbols, may also contain icons, the so-called emojis, which complicate the analysis of the text.

As social networks, especially Twitter, contain small texts and people may use different words and abbreviations which are difficult to extract their sentiment by current NLP systems easily, thus some researchers have used deep learning and Machine Learning (ML) techniques to extract and mine polarity of the text [8].

Text mining is an important issue that analysts are constantly asked to address. Some ML techniques, such as the various supervised and unsupervised algorithms, are common

tactics. Summary approaches are plentiful. One approach could be to rank the meaning of the sentences within the text and then create a summary for the text based on the significant numbers. A different approach is called end-to-end productive models. In some areas such as image recognition, and question-and-answer, the second method works better [9].

In this study, sentiment analysis on Twitter data has been conducted for monitoring public opinion, regarding COVID-19 vaccines. **Four NLP lexicon-based** approaches (TextBlob, the Valence Aware Dictionary and sEntiment Reasoner VADER, Afinn and NRC, along with **five ML models** (Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), XGBoost Classifier (XGB) and Support Vector Machine (SVM)) have been used for sentiment analysis.

TextBlob is a successful tool in which we can implement NLP easily and quickly. On the other hand, **Vader** is a lexicon and rule-based sentiment analysis tool which is very good when dealing with social media data. **Afinn** is also the simplest yet most popular lexicon used to perform sentiment analysis. It would be interesting to detect and classify emotions in tweets, such as happiness, fear and trust, which have the potential to improve citizens' and society's understanding. That's why we chose the **NRC Sentiment and Emotion Lexicon**.

We would like to test which of the four classifiers gives the best result. The **idea behind choosing the five ML models** is that we wanted to try all the classifiers on the dataset ranging from simple ones to complex models and then try to find out the one which gives the best performance among them. The results revealed people's feelings to devise relevant policies to increase the acceptance of COVID-19 vaccination.

1.1 Aims and Objectives

The purpose of this thesis is to apply ML techniques and sentiment analysis to predict people's sentiments about vaccination. The main goal is to identify **the best-suited model** to perform a sentiment analysis on our Twitter data regarding the Covid-19 vaccines and find out the sentiment's polarity from the data to show the distribution of the sentiments as follows: positive, negative and neutral. In this thesis, multiple ML techniques were applied to 246.626 English tweets and 99.315 Greek tweets, which were extracted from

Twitter between 25–05–2022 and 30–11–2022, divided into summer and autumn periods. A literature study and an experiment were set. Time-series analysis was performed to obtain daily sentiments.

1.2 Contributions

The key contributions of this thesis are summarized as follows:

i) A collection of two COVID-19 vaccination datasets, one in English and one in the Greek language, pre-processing and feature extraction were conducted as detailed in Chapter 3.

ii) Sentiment analysis on Twitter data was performed using four NLP approaches (TextBlob, VADER, Afinn and NRC) as described in Chapter 4.

iii) Training and testing five ML models and comparing them, as presented in Chapter 5.

iv) Evaluation of the five ML models for sentiment prediction and fine-tuning of the best-performing model as discussed in section 5.8.

1.3 Thesis Outline

Chapter 1 contains an introduction to the main subject along with the problem definition. Some fundamental concepts and tools, needed for the comprehension of the rest of the dissertation, are explained in Chapter 2. An overview of the sentiment analysis literature with a focus on vaccines and related published work of other scientists is also provided. Chapter 3 elaborates on the proposed research design, as well as the preprocessing part of the study. The implementation of the sentiment analysis is presented in Chapter 4, while Chapter 5 analyzes the experimental results and the trend of sentiments expressed on Twitter. Finally, Chapters 6 and 7 contain conclusions, challenges and future directions.

2. Background

2.1 Literature Review

Knowledge Discovery in Databases (KDD) is a dynamic field of research, which combines the use of databases with Statistics, Artificial Intelligence, Expert Systems and Data Visualization [10]. The purpose of the KDD is to find the knowledge that we wish to be neither expected nor used to and whose existence we did not even know until the moment we discovered it. This knowledge can be relationships, trends and patterns that concern all data. Its ownership and use in guiding decisions concerning future activities can be decisive in creating some benefit or advantage that will bring greater financial gains, security, etc. «The secret of success is to know something nobody else knows – Aristotle Onassis» [11].

The areas of activity in which the knowledge discovery is applied are many such as:

Banks: Promotion of banking products to customers.

Supermarket: Installation of products on the shelves with such combinations to increase sales.

Financial companies: Stock progress forecast from historical data [12].

Medical: Finding new factors that affect diseases [13].

Meteorology: Better prediction of tomorrow's weather. etc.

"**Sentiment analysis or opinion mining** is the computational study of opinions, sentiments and emotions expressed in the text" [14]. Opinion and sentimental mining are important to research areas considering the huge number of daily posts on social networks, extracting people's opinions is an ambitious task. About 90 percent of today's data has been provided during the last decade and getting insight into this large-scale data is not negligible [15].

Sentiment analysis has many **applications** for different areas in business, for example, for receiving feedback on products and services through which companies can learn user comments and reviews on social media, and rapidly improve their products or services

offered to consumers. For large texts, one solution could be to understand the text, its summary and its weight whether its content is positive or negative, or neutral. Two fundamental approaches are the extraction method and the subtraction method. In the export method, words and word phrases are extracted from the original text to create a summary. In the subtraction method, an attempt is made to create an internal linguistic representation and then to create a summary that is more like a human summary.

On Twitter, bots are accounts that are operated automatically to post, retweet, or reply and may vary in sophistication from simply reposting links to certain (often malicious) Web pages to more sophisticated masquerading of humans aiming to alter the discourse of a topic [16]. Therefore, the emergence of **Twitter bots** influencing social media conversations about public health threats requires researchers and practitioners to develop new communication strategies considering their influence.

Studies have shown that bots on social media (i.e., social bots) can influence opinion trends by posting a substantial number of automated messages. Zhang, M., Qi, X., Chen, Z., & Liu, J. [17] revealed that 8.87% of the users were social bots, with 11% of tweets in the corpus. A total of 314,342 vaccine-related tweets from 11 December 2020 to 31 August 2021, were obtained. Bot analysis discovers that 1.45% of the corpus users were identified as likely bots which produced 4.59% of all tweets [18].

Mønsted, Sapieżyński, Ferrara, & Lehmann [19] conducted a large-scale information diffusion experiment on Twitter using social bots to deliver intervention-related messages, including a message to encourage users to vaccinate. Broniatowski et al. [20] analyzed the messaging patterns and effects of known Russian troll accounts and limited samples of detected social bots within various online vaccination conversations from July 2014 through September 2017. It discovered that accounts classified as bots posted vaccine-related content at higher rates than average users but provide equal attention to both pro- and anti-vaccination sentiment.

Lexicon-Based Approaches

The sentiment analysis operation regarding the ongoing vaccinations around the globe was done by using three different Lexicon-Based approaches: **TextBlob**, **Vader**, **Afinn** and **NRC**. The goal is to find out the sentiment's polarity from the data to show the

distribution of the sentiments as follows: **positive**, **negative** and **neutral**. NRC extracts the emotions behind the words and categorizes them into the following: **fear**, **anger**, **anticipation**, **trust**, **surprise**, **positive**, **negative**, **sadness**, **disgust** and **joy**. They have differences in the way they calculate the polarity scores of a text, which makes them have different results. An exhaustive comparison of these four methods was applied.

TextBlob

assigns scores to each word and it calculates the overall sentiment by taking **the average of these scores**. For each word, there is **a sentiment score** (how positive/negative/neutral are they) and **a subjectivity score** (how opinionated are they). A positive polarity score indicates that the tweet has a positive sentiment, while a negative polarity score shows that the tweet has a negative sentiment. In the case, that the polarity score equals zero, the tweet is considered neutral [21]. The subjectivity range is 0 to 1, where 1 is the most subjective and 0 is the most objective. We stored the sentiment of our cleaned data in a field in our database called "sentiment_TextBlob" and the subjectivity in a field called "subjectivity".

Vader (Valence Aware Dictionary for Sentiment Reasoning)

is another popular model for sentiment extraction in the domain of social media that takes into account the intensity of the sentiment [22]. The idea of this model is that it returns a sentiment rating to each word of the text. Then, summing and normalizing these ratings the compound score that indicates the sentiment score of the text is provided. The compound score range lies between **-1 (intense negative sentiment)** and **+1 (intense positive sentiment)**. The sentiment is considered negative if the compound score **<= -0.05**, neutral if the compound score is in the range (-0.05,0.05) and positive if the compound score **>=0.05** [23].

Afinn

is based on the **Affective Norms for the English Words** lexicon (ANEW) proposed by Bradley and Lang [24]. It is a lexicon of English terms labeled manually by Finn Arup Nielsen [25] with a score in the **range [-5,5]**. A positive score specifies a positive emotion,

a zero score is a sign of neutral sentiment, and a negative score indicates a negative emotion.

NRC Sentiment and Emotion Lexicon

proposed by Mohammad and Turney (2013) and originally called **EmoLex**. It considers each word to be one or a combination of multiple moods, including **eight basic emotions** (anticipation, sadness, disgust, joy, anger, surprise, fear and trust) and **two sentiments** (positive and negative). The NRCLex library measures the **emotional effect** of a body of text. These emotions can offer additional insights. The library is built using approximately 27,000 words (Affect dictionary) and is based on the **National Research Council Canada** (NRC) affect lexicon and NLTK library's WordNet synonym sets [26].

Supervised Learning or Inductive Learning in which, given the existence of a finite set of examples of entry that are characterized by some price (continuous or distinct), is called for the construction of a function. This function not only approaches the input data correctly but also generalizes as well as possible to new examples.

Classification Algorithms:

Logistic Regression (LR)

The supervised ML classification algorithm Logistic Regression is used to **predict the likelihood of a target variable**. The logistic function, also called the **sigmoid** function is an S-shaped curve that can take any real-valued number and map it into a value **between 0 and 1**. In general, Logistic Regression refers to **binary logistic regression** with binary target variables, but it can also predict two additional types of target variables [27].

Decision Tree (DT)

Decision Tree is a **rule-based supervised** ML algorithm used in both regressions, as well as classification problems [28]. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. While the node at the top of the decision tree is the root

node [29]. The advantages of Decision Trees are that they are **simple to understand and interpret** and require **little data preprocessing**. However, Decision tree learners can create over-complex trees that do not generalize the data well and can be unstable.

Random Forest (RF)

Random Forest is **an ensemble learning method for classification**. It builds decision trees on different samples of the given dataset and takes the **average** to improve the predictive accuracy of the dataset. The greater number of trees in the forest leads to higher accuracy and **prevents the problem of overfitting**. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing categorical variables and usually performs better results [30].

XGBoost Classifier (XGB)

Extreme Gradient Boosting (XGBoost) is **a distributed Gradient Boosted Decision Tree** (GBDT) ML algorithm that is scalable. It is very efficient for classification and includes **parallel tree boosting**. The XGBoost creates new models to predict from previous models. Then, the models are combined to establish a final prediction and the loss is minimized. It can assist in **tuning the model** and in **algorithm enhancement** [31].

Support Vector Machine (SVM)

Support-vector machine (SVM) model is **a supervised learning model** that analyzes data for classification. Original training data is **transformed to higher dimensions** using a non-linear mapping function. In higher dimensions, it searches for the linear optimal separating hyperplanes (i.e. **decision boundary**) using support vector margins. It is capable of handling both linear and nonlinear data [32].

Performance Evaluation Metrics

To analyze and compare the results of our five classification models, different metrics were used as evaluation criteria. In this work, we incorporated Accuracy and F-score to assess the algorithms, but we also provided the short definitions of Precision, Recall or

Sensitivity and Specificity, since they are essential for the computation and definition of Accuracy and F-score [33].

Accuracy, as described by Kotsiantis [34], is "the fraction of the number of correct predictions over the total number of predictions".

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

Precision shows the proportion of how many instances the model classified correctly to the total number of true positive and true negative examples. In other words, precision shows the **exactness** of the classifier with respect to each class.

$$\text{Precision} = \frac{TP}{TP+}$$

Recall or Sensitivity represents the proportion of how many instances the model classified correctly as positive to the total number of true positives and false negatives. Recall shows the **completeness** of the classifier for each class.

$$\text{Recall} = \frac{TP}{TP+FN}$$

Specificity represents the proportion of how many instances the model predicted correctly as negative to the total number of true negatives and false positives. While Sensitivity measure is used to determine the proportion of actual positive cases, which got predicted correctly, Specificity measure is used to determine the proportion of actual negative cases, which got predicted correctly.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

F-Score is defined as the harmonic mean of precision and recall.

$$\text{F1-Score} = 2\,\frac{\text{precision x recall}}{\text{precision + recall}}$$

where TP stands for True Positive, FP for False Positive, FN for False Negative and TN for True Negative [35].

2.2 Related Work

During the pandemic, many researchers tried to find and understand the attitude of people associated with vaccines which seemed to have many aspects. They also tried to find out if there is a preference for any of the available vaccines. Therefore, it is worth looking at these studies and finally comparing our results with them. It makes sense to discover if over time and as the coronavirus is eliminated, people's feelings about the vaccine change. Another issue that is important to explore in related papers concerns ML (ML) algorithms and Lexicon-Based approaches.

Sentiment analysis on tweets written in the English language is more common. However, Kapoteli et al., [36] identified the emotions of both English and Greek tweets during the period between May 19, 2021, and November 19, 2021. In the first step, to represent the text as numerical vectors three different techniques were applied: (i) Term Frequency - Inverse Document Frequency (TF-IDF), (ii) Word2Vec and (iii) BERT, which is a new language representation model introduced in 2018 [37]. The BERT model performed better, so their analysis was then based on that. Studying English tweets, they noticed that most of the emotions were neutral in the first months of the period they researched, but as time went on, negative emotions increased. Negative emotions also prevailed in Greek tweets.

It is also worth considering the work of Asderis [21], who explained in detail the two most common Lexicon-Based approaches: TextBlob and VADER. His study is based on Twitter data related to Covid and specifically used the hashtags of the most common vaccines: Pfizer, Moderna, AstraZeneca and Johnson&Johnson. In addition, he used hashtags that refer to anti-vaxxers. TextBlob has classified most tweets as neutral, regardless of the hashtag and are followed by positive emotions. Tweets related to anti-vaxxers show more negative sentiments compared to other hashtags daily. Taking into account the VADER approach, interesting conclusions emerge. Anti-vaxxers have a high rate of negative emotions. Only AstraZeneca seems to have neutral emotion as dominant. The rest of the vaccines seem to be more likable as they have more positive emotions than

negative or neutral ones. With all this in mind, it was confirmed that TextBlob and VADER produce different results. As the author of the paper found in further analysis, about 50% of the tweets were ranked with the same sentiment by the two Lexicon-Based approaches.

Marcec and Likic [38] conducted a Twitter sentiment analysis to identify differences between the AstraZeneca, Pfizer and Moderna vaccines. The tweets were collected from December 1, 2020, to March 31, 2021. The data was annotated using the Affin lexicon-based approach. The results of this study showed that most tweets related to Pfizer and Moderna were labeled as positive. This trend did not change during the four months. On the other hand, the negative feeling corresponds to the AstraZeneca vaccine, which has increased greatly during this period. The side effects of this vaccine that appeared at that time caused negative reactions in people.

Shamrat and Chakraborty [39] also implemented a study aimed at understanding people's feelings about the three most popular vaccines. Using the Twitter API, they exported 30,000 tweets with the corresponding hashtags (#Pfizer, #Moderna, #AstraZeneca). After pre-processing the data, they used KNN (K Nearest Neighbors) algorithm to classify it into positive, neutral and negative. Most of them were classified as negative and only a small percentage were classified as neutral. To be more specific, most of the tweets that mentioned Pfizer or Moderna had positive content. On the contrary, most of the tweets that referred to AstraZeneca had negative content. Considering the previous study, it is confirmed that people do not support AstraZeneca like other vaccines.

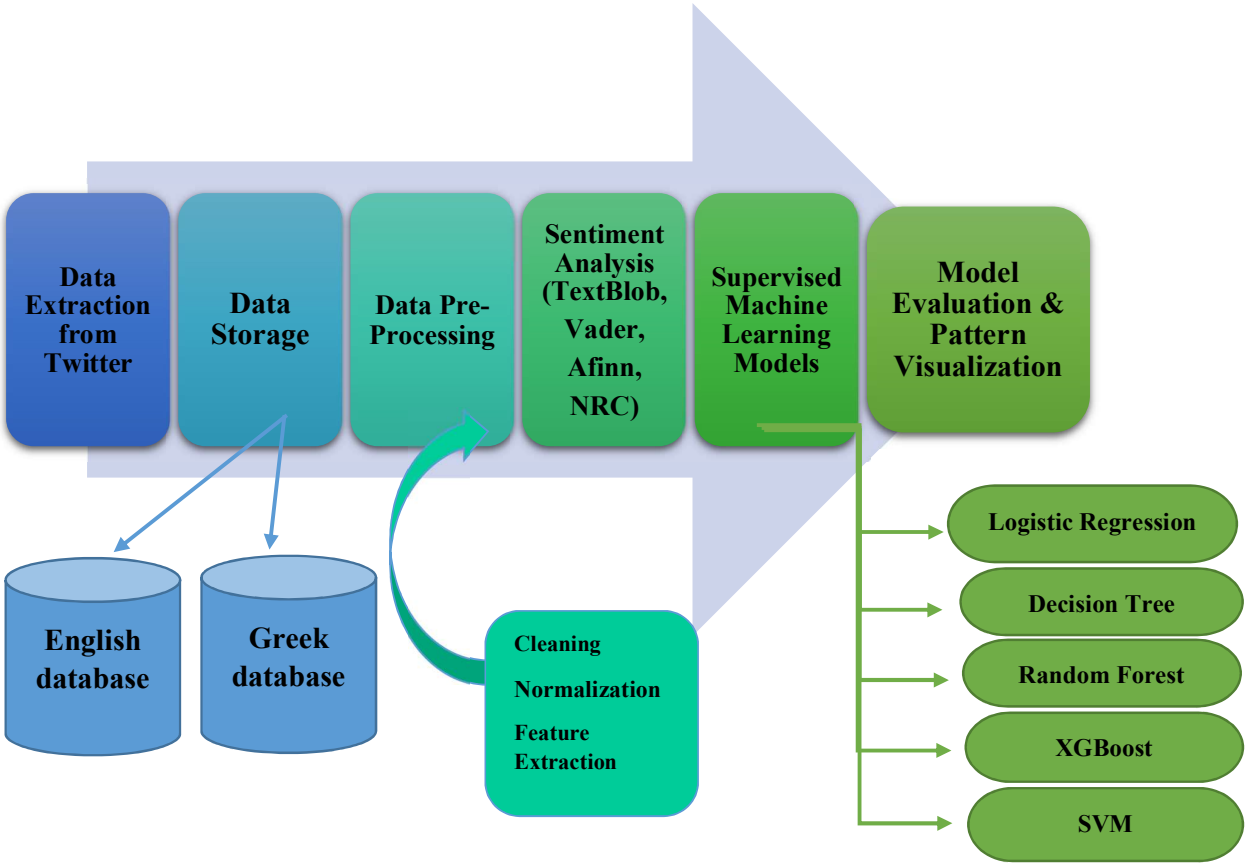
A very exhaustive sentiment analysis was accomplished in the paper [40]. This study was based on data from 2021 and early 2022. The authors sought to observe how people's feelings about vaccination changed during this period. Evaluating the three Lexicon-Based approaches mentioned above, TextBlob was the one that produced the best results. Subsequently, Logistic Regression and Decision Tree achieved 93% accuracy. The performance of the models was not so good when they were trained in data annotated by Vader or Afinn. They concluded that regardless of the season, most tweets had a neutral sentiment. However, in 2022 the number of negative emotions was higher than in 2021. It will be interesting to consider if this trend continues, in other words, if negative emotions continue to increase during 2022.

Another notable work was implemented in a paper [41] based on Twitter data from the first Covid-19 vaccination announcement over one month. The aim was to determine the general opinion of the people about the start of the vaccination campaign. To have a sample for the training set, they manually labeled 1,00% of the dataset with labels: favor, against and neutral. Most tweets were annotated as neutral. ML and Deep Learning models were trained in this dataset and their performance was evaluated with precision, recall, F1-score and accuracy. BERT was the winner and suggested that the highest number of tweets were classified as neutral. Since this study was conducted at the beginning of vaccine implementation, it is worth comparing these findings with our results and looking at how emotions have changed now that two years have passed.

From November 23, 2019, to May 15, 2020, Rahim & Rafie [42] conducted another Twitter emotion analysis of the Covid vaccine. Using the Twitter API, a total of 105,965 tweets were collected and after cleaning and preprocessing the data, TextBlob classified them as positive, negative and neutral. The majority class was negative as 41% of the tweets were labeled as negative. Only 20% were annotated as positive. Next, SVM models with RBF (Radial Basis Function) and polynomial kernel classified the data. SVM with RBF kernel achieved the highest accuracy of 91% in contrast to the polynomial which reached 87%.

3. Data & Methodology

Figure 1: Flowchart of research design



This view was inspired by typical database systems, data warehousing, ML and statistics communities.

3.1 Connection to Twitter and API key generation

The tweets are imported from Twitter using the **API keys** provided by the Twitter Developer Account we created. More specifically, **elevated access** is needed for the retrieval of many up-to-date tweets for the Covid-19 vaccines.

3.2 Data Collection

Data in the form of raw tweets is extracted by specifying **keywords** to search for in the tweets. We defined two functions named get_en_Tweets(keyword) and get_el_Tweets(keyword) for the collection of English and Greek tweets respectively. Calling them several times with the following keywords, we managed to download **246.626** English and **99.315** Greek tweets. The choice of the respective hashtags was made based on the hashtag's popularity on Twitter. For each search, the more recent tweets are selected for the collection.

Table 1: Hashtags list for tweet search

COVID-19 Vaccination Topic	#vaccine, #COVID19Vaccination, #Antivax, #antivaxx, #antivaxxers, #GetVaccinated, #CovidVaccine, #Covidiots, #VaccineSideEffects, #VaccineDeath, #Pfizer, #AstraZeneca, #Moderna, #εμβόλιο (=vaccine), #ανεμβολίαστοι (=unvaccinated), #εμβολιασμένοι (=vaccinated), #κορονοιος (=coronavirus), #κρούσματα (=cases), #πανδημία (=pandemic)

From 25 May 2022 to 30 November 2022 we acquired them in portions at different points in time. In addition to the posted text, the following **15 fields** are included:

Table 2: The fields of the retrieved tweets data frame

Fields	Information
created_at	The created Timestamp of tweet
tweet_id	The unique URL/ID of tweet
screen_name	Twitter user's name with @user
name	Username of person
description	A small description of user
account_creation_date	Created date of user's account
location	User's location: city, country
urls	A list with some URLS of tweet
n_followers	Number of followers of user

n_retweets	Number of retweets of this tweet
hashtags	A list with hashtags & each indices (character's position inside tweet)
source	Source of tweet
favourites	Likes/favourites of tweet
account's tweets	The number of tweets (including retweets) issued by the user
keyword	Keyword to search relevant tweets

3.3 Data Storage

After collecting these data and merging them into two different panda data frames, one for English and one for Greek tweets, we saved them in **pickle files**. Then, we connected with **MySQL** and created a database with two tables ("english_tweets" and "greek_tweets") to store my Twitter data. A cursor object was created allowing me to execute SQL statements. We used **df.to_sql()** command to insert the data into the correct table in my database using the **sqlalchemy library**. The columns "hashtags" and "URLs" were converted from lists to strings, to be inserted correctly in the database table. The next step was to export the Twitter information as an **Excel file**.

3.4 Data Preprocessing

The preprocessing of the data is a crucial step as it elects the efficiency of the other steps down the line. It involves syntactical correction of the tweets as desired. The steps involved should aim for making the data more machine-readable to reduce vagueness in feature extraction [43]. The steps that were applied to **English tweets** are given below:

1) As there is no limit to the number of times a tweet can be retweeted, we removed retweets because they contain duplicate content that could skew our analysis. A **retweet** is when someone shares someone else's tweet. It is similar to sharing on Facebook (reposts). So, retweets were filtered out. Another way we could do this would be if we set tweet_id as the primary key to prevent duplicates. In this way, **204.533 English** tweets and **61.109 Greek tweets** remained.

2) Converting **datetime to date** for the columns "created_at" and "account_creation_date" of our data frame.
3) **Expand Contractions**. They are words or combinations of words that are shortened by dropping letters and replacing them with an apostrophe. It is a useful preprocessing step as the words play an important role in sentiment analysis [44].
4) Applying **lowercase** which means converting all letters to lowercase.
5) Removing **URLs, user's mentions, hashtags, punctuations, digits & emojis** using **regular expressions**, since those terms don't provide meaningful context for discovering inherent topics from the tweet.
6) Removing **stop words**, which are commonly used words that do not contribute much to the ML model. This allows us to focus on the important words instead.
7) A necessary NLP technique that we used is **Tokenization**. It is the process of breaking down a tweet into words [45].
8) Another significant NLP technique was **Lemmatization**. It is a method that converts words to their lemma or dictionary form by using vocabulary and morphological analysis of words. To achieve an effective lemma or root meaning of the word using WordNetLemmatizer, the input word must be passed in lowercase to the WordNetLemmatizer algorithm to achieve accuracy [46].

For **Greek language tweets**, a different **smaller preprocessing** was applied to the tweet_text. We didn't remove punctuations, hashtags and emojis because they might hide some important different emotions if we removed them. Therefore, it is deemed necessary to keep the Greek tweets in the text as they are, as the Greek language may suggest more hidden meanings and inferences leading to the more efficient categorization of the tweets into positive, negative or neutral. The pre-processing stage of Greek tweets should not be as strict as that of English ones. We used a stopwords list provided by the Natural Language Toolkit (NLTK) library and a list of Greek stopwords [47].

Although another attempt was to implement the **same pre-processing steps**, which were applied to English tweets, as it was useful to compare these two different preprocesses and draw our conclusions about the correct classification of Greek tweets. This stricter pre-processing also helped us find the most common words and hashtags in Greek tweets. If we retained punctuation marks, hashtags, and emojis, the word clouds we would create would contain words that wouldn't add any important information to our charts.

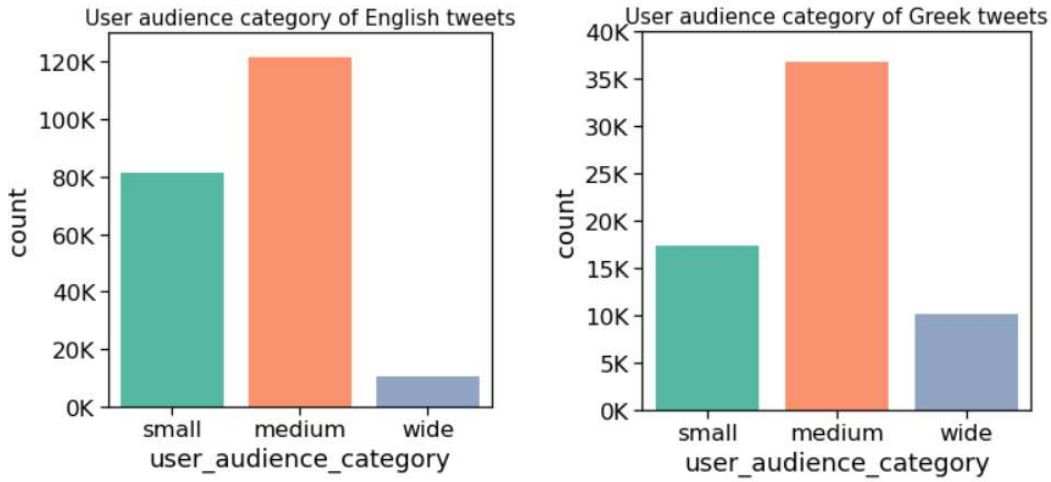
3.5 Feature Extraction - Story Generation and Visualization from Tweets

Choosing an appropriate feature set is the most critical part of any ML algorithm. We provided a short description of all the features used, as well as our intuitive argument behind each feature.

From the number of followers of every user, we could understand which of our tweets belongs to a user that has a **low**, **medium**, or **high** twitter audience. Thus, we included a new variable called 'user_audience_category'. Dividing the users into three buckets: **less than 300** followers, **between 300 and 10.000** followers and **over 10.000** followers, also allowed us to understand the size of the audiences the messages reach. This breakdown was made after observing the number of followers of our tweets and searching on Twitter for the average number of users' followers. As Carly Fiorina said, "The goal is to turn data into information, and information into insight". Hence, we should extract features from data and transform them into formats that are suitable for ML algorithms.

From Figure 2 we notice that the size of influence each user has is medium as most users who post tweets have **a medium number of followers** (between 300 and 10,000). It is a challenging task to analyze the potential social audience because it will help, for example, a company to spend resources efficiently by sending offers to the appropriate audience and maximizing its profits [48].

Figure 2: User audience category of **English and Greek** text tweets



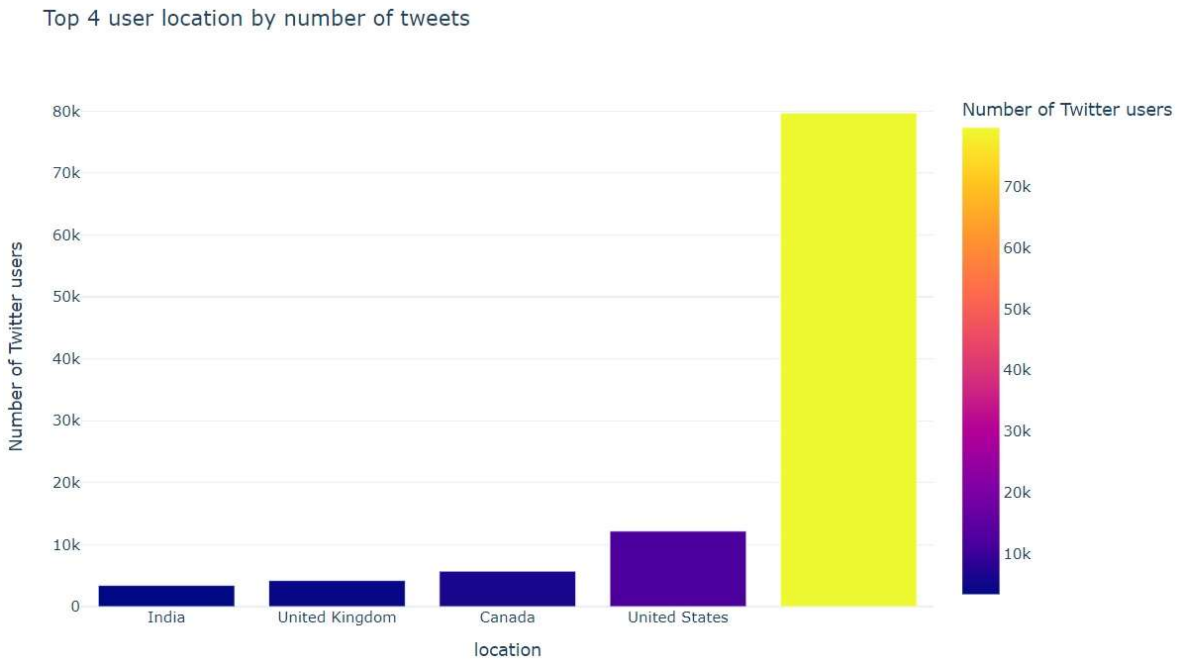
After observing the audience following the tweets, the next step was determining **which places they originate from**. We looked at the four most common user-declared locations for our dataset. The results indicate that **Canada** is the country with the most active Twitter presence regarding vaccines, as shown in Table 3 below. This feature can also let the user understand the public opinion about an issue in a specific location [49]. For some tweets, however, the location label was less specific and in the form of 'State, Country' (e.g., 'Texas, USA') or even worse in the form of a unique value (e.g., 'On The Interwebs'). We wanted to have as granular, a label as possible for each tweet [50].

Therefore, we **aggregated** all these tweets with **single labels** that were **in the form of 'Country'** (e.g., 'United States') since we wanted to achieve a country-level location of Twitter users. After **creating a list of each sub-location merged with the main location**, we found that the **United States of America** has the highest number of tweets (12.190) as shown in Figure 3 below. It is followed by **Canada**, the **United Kingdom**, and finally **India** with 3.422 tweets.

Table 3: Geo-locating tweets **before** and **after merging** the "location" column

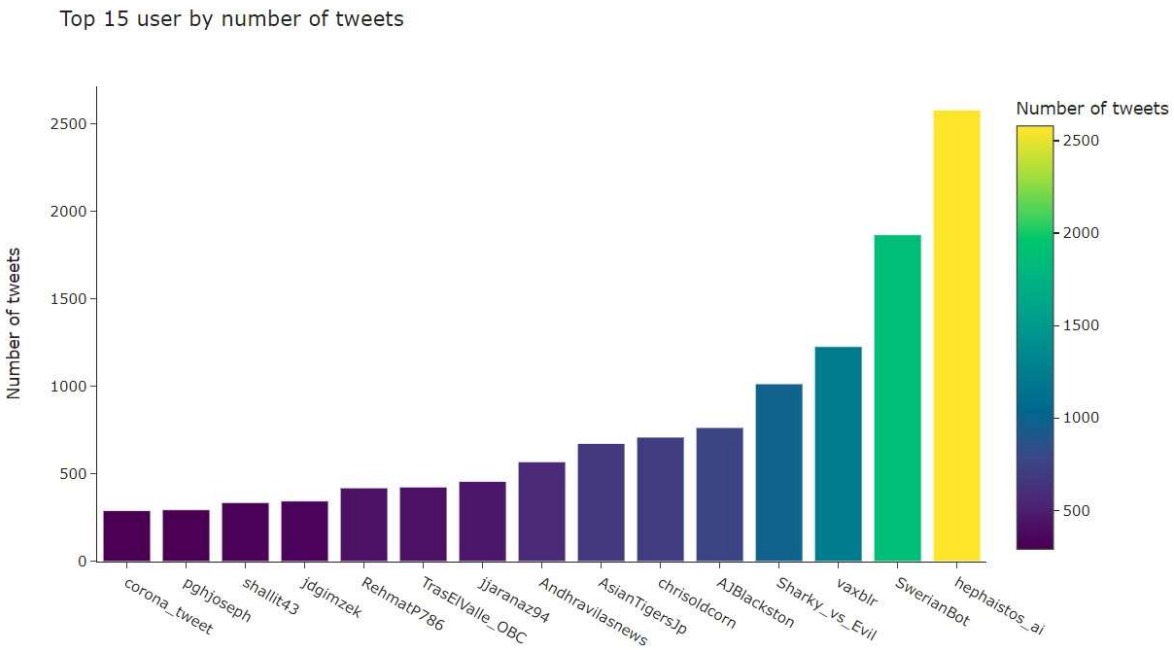
Before		After	
Canada	**3.064**	**United States**	**12.190**
Paris	2.601	Canada	5.726
United States	2.421	United Kingdom	4.209
USA	1.828	India	3.422

Figure 3: The **top four countries** with the most English tweets



Another feature worth looking at is the **top users** who post daily on Twitter about vaccines and Covid-19. The results are presented in Figure 4 and it can be seen that "hephaistos_ai" and "SwerianBot" are **the most frequent usernames** with **2.580** and **1.867 tweets** respectively. As a result, these users have a greater influence on people about what they write and can guide public opinion, especially if we take into account the large number of followers they have. However, there is the **danger of social bots or fake accounts**. We don't know if these users are **valid or not** and what purposes they have by posting on Twitter.

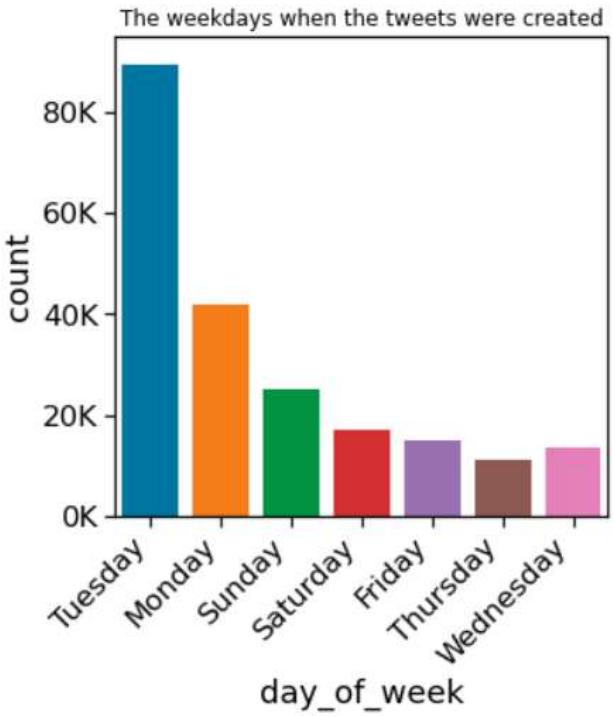
Figure 4: The **top fifteen users** with the most active presence on Twitter



Two other features created are '**word_counts**', which is the number of words tweets have and '**word_counts_cleaned**', which is the number of words tweets have after preprocessing. So, we can see that each tweet has been reduced in size after preprocessing.

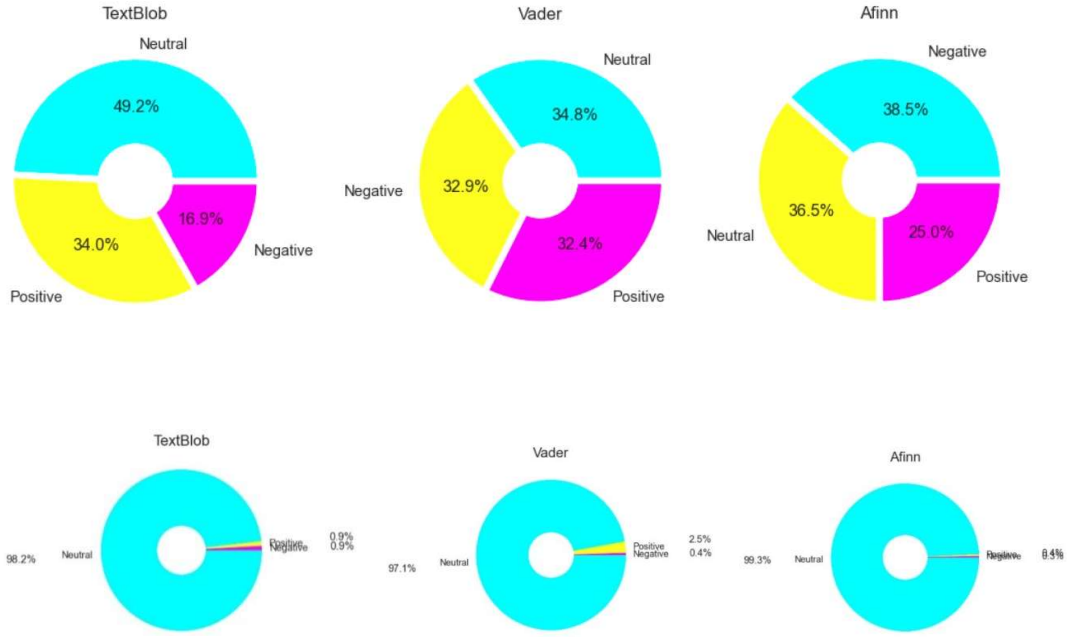
Also, from the timestamp of the tweet that was created, we got the **weekdays that the tweets were created**. The histogram below shows us that most people expressed their views on COVID-19 on **Tuesdays** and **Mondays**. According to the **European Center for Disease Prevention and Control (ECDC)**, which is a public health agency of the European Union (EU) (operational since 2005), the cases and the new measures for the coronavirus are announced at the end of the week on **Thursday** [51]. Therefore, it is reasonable for users to "tweet" more after the weekend as those days are for relaxation from everyday life. Monday and Tuesday are therefore the days with the most tweets about vaccinations.

Figure 5: The **weekdays** when the **English tweets** were created



4. Lexicon-Based Approaches

Figure 6: Pie Charts representing sentiment polarity scores using **TextBlob, Vader** and **Afinn** classifiers in **English** and **Greek** tweets



In Figure 6 the results for the **TextBlob** approach showed that the **neutral polarity** in our keywords was significantly **high** (49.2%). Vader-assigned positive polarity scores were lower as compared to TextBlob. TextBlob gave 34% positive tweets, while VADER set a positive polarity score at 32.4% indicating 1.6% lower positive tweets than TextBlob. Afinn assigned a less positive polarity score (25%) compared to VADER. Although **Afinn** classifies most tweets as **negative** (38.5%). Thus, in English tweets, there are different public sentiments of the public, while **most Greek tweets** are **neutral** using all three classifiers.

Figure 7: Tweets sentiment score for the **top four countries** using **TextBlob**, **Vader** and **Afinn** classifiers

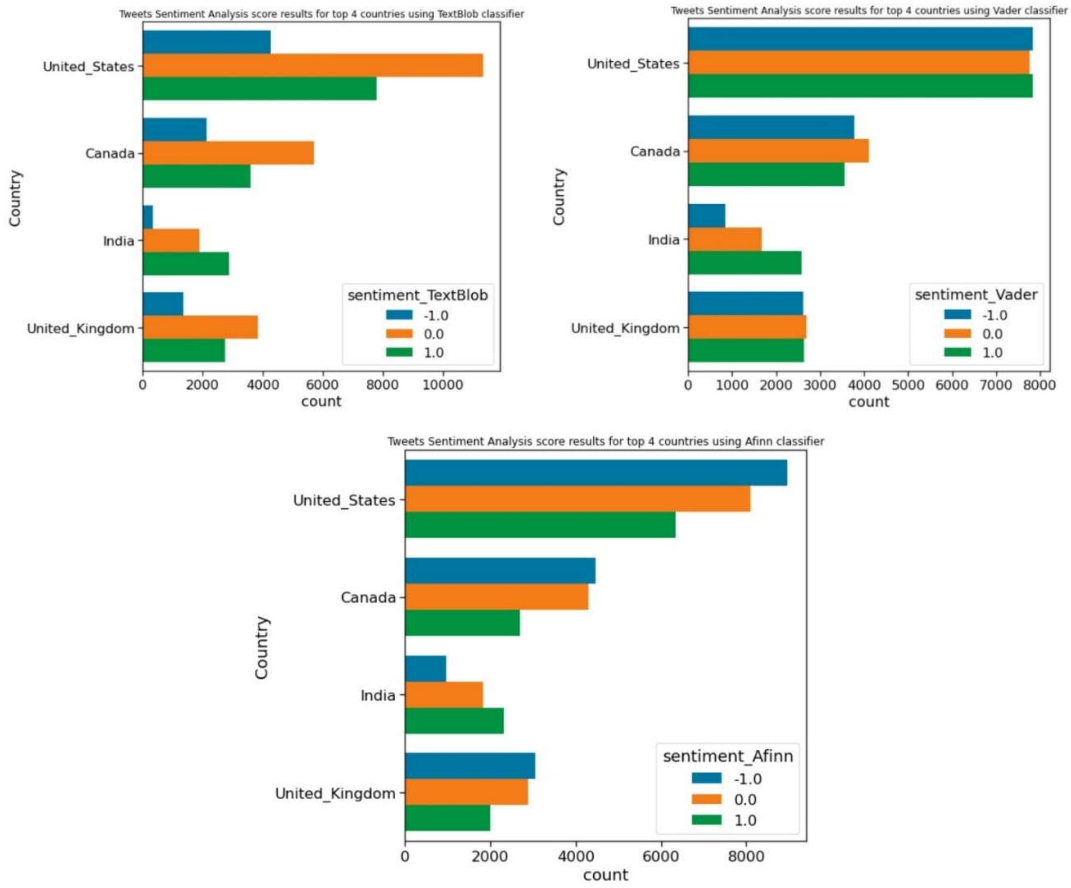


Figure 7 illustrates **Twitter users' opinions** classified by **TextBlob**, **Vader** and **Afinn** in the **top four countries**. These three classifiers differ in the way they categorize tweets, so it is logical to note that there are differences in the feelings of people living in the **United States of America** (USA), **Canada** and the **United Kingdom** (UK). The results of this study agree only in **India** which has a **positive** attitude towards vaccinations. This country also has a **very limited set of negative tweets**.

Figure 8: Pie Charts depicting Sentiment Polarity distribution for **each vaccine using the three different Lexicon-Based approaches**

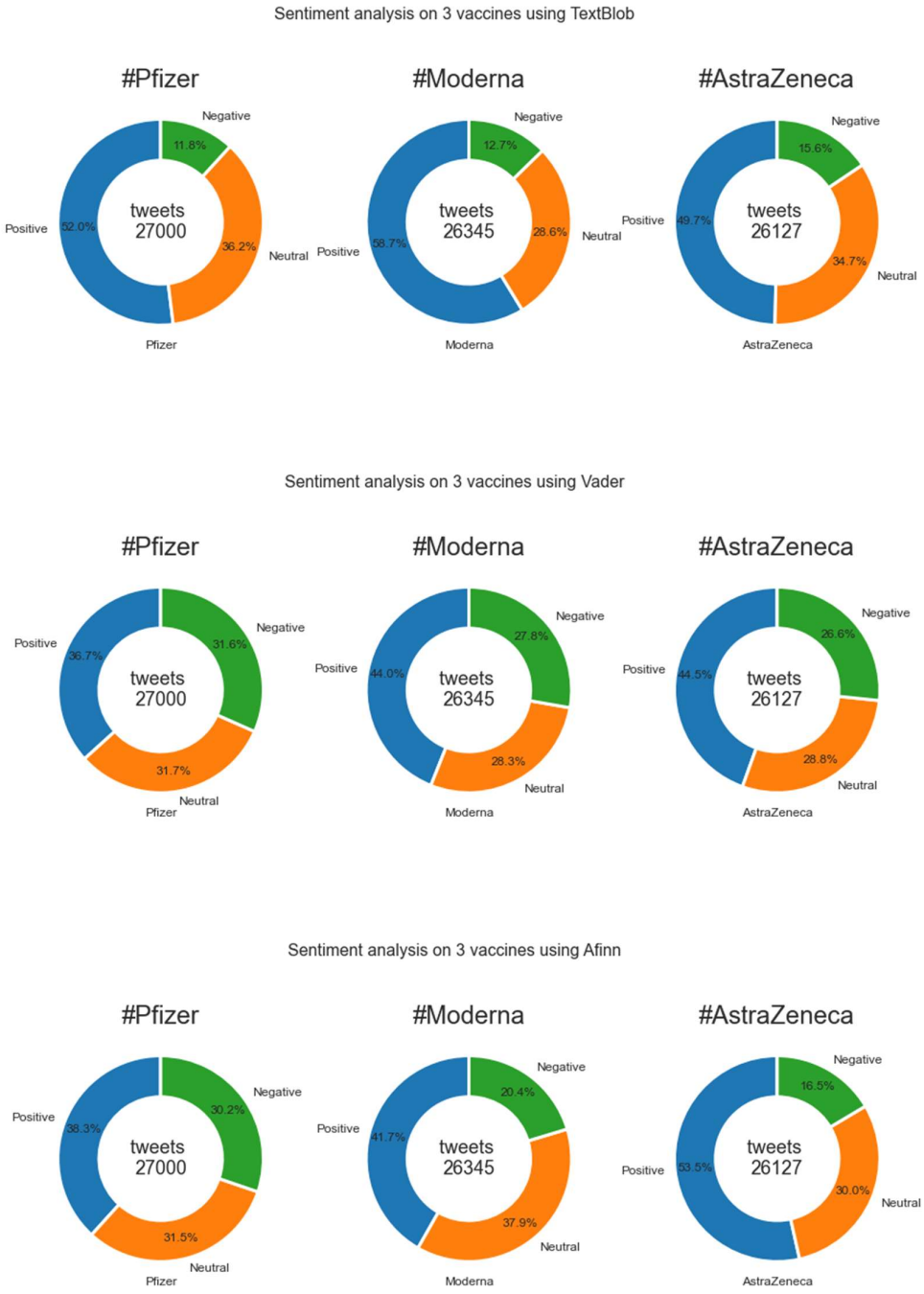
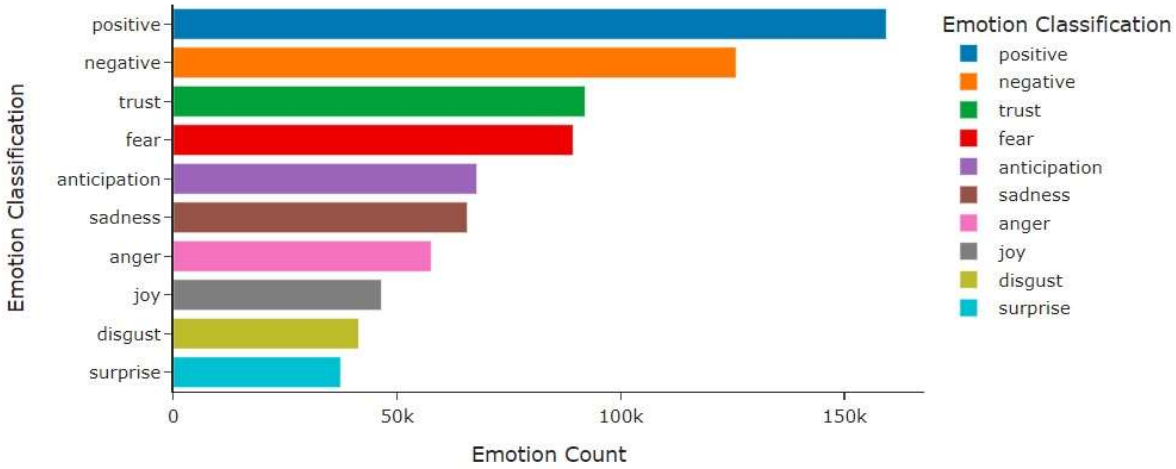


Figure 8 shows three graphs representing the ratios of positive, neutral and negative tweets for each vaccine (Pfizer, Moderna, AstraZeneca) given by the TextBlob, Vader and Afinn approaches. The displayed results show that **the majority of tweets belong to the positive class**, followed by the neutral tweets, while negative tweets are the lowest.

Thus, people in general have published **optimistic views** about the three vaccines. More specifically, **TextBlob** gave **Moderna more positive polarity** scores compared to **Vader**, which gave **AstraZeneca** the highest positive scores. **Afinn** also believed that **AstraZeneca** is the most reliable and famous vaccine followed by Moderna.

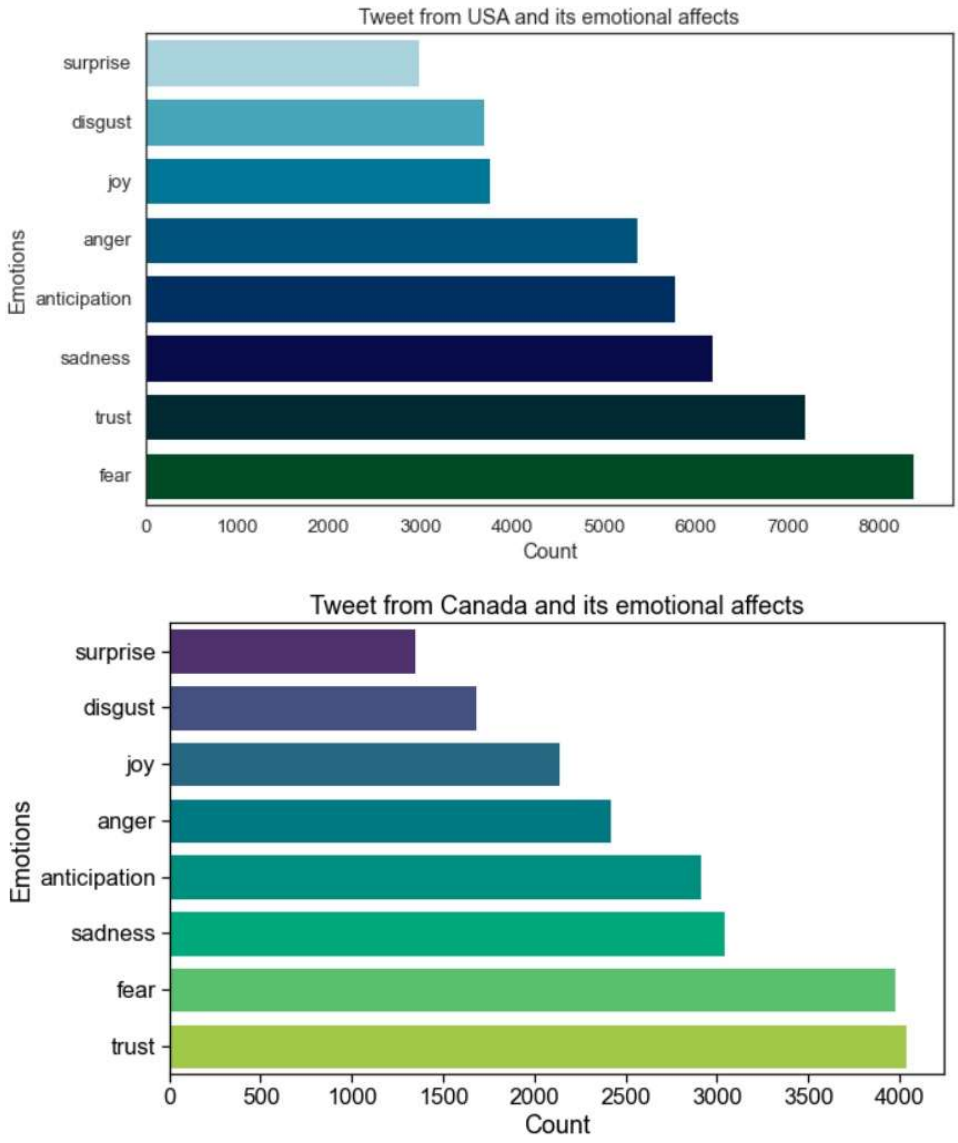
Using the **NRC Sentiment and Emotion Lexicon**, we extracted the raw emotion scores from our tweets. We further compute the **frequencies of each emotion** within our tweets. For each word, the emotion is represented by a ten-dimensional vector to reflect the 10 different moods, ranging from **0 (extremely lack of this emotion)** to **1 (extremely full of this emotion)** [52].

Figure 9: The **emotion scores** from the **English** tweets using the **NRC Lexicon**



In Figure 9 the sentiment analysis results are provided using the **NRC Sentiment and Emotion Lexicon** in our **English** dataset. From the above histogram, we can see **a substantially larger count of "positive"** (159.361k) compared to **"negative" emotions** (125.802k), followed by **"trust"** mood (92.075k) from our tweets. This is a sign that people support vaccines, believe in their productivity and encourage other people to get vaccinated through their tweets.

Figure 10: The **emotional effects** of tweets from the **USA** and **Canada** using **NRC Lexicon**



We found that the top two countries with the most tweets are the **United States of America (USA)** and **Canada** with **12.190** and **5.726** tweets respectively (Table 3). Thus, it's interesting to look at these people's feelings about vaccines in these two countries. Figure 10 shows that in both countries the prevailing emotions of people are **"fear"** and **"trust"**. This shows that people trust vaccines and consider them safe, but there is always a feeling of fear because it is something new and can cause several side effects.

Figure 11: Word clouds with **the most common hashtags** of **English** and **Greek** tweets

We also created word clouds to find out **which hashtags are used most often** by Twitter users. The first word cloud in Figure 11 above refers to **English tweets**. The hashtags **"Pfizer"**, **"COVID19"**, **"vaccine"**, **"Moderna"** and **"CovidVaccine"** are used the most, which makes sense. The second word cloud concerns **Greek tweets**. The use of the hashtags **"εμβόλιο"** (=vaccine), **"κρούσματα"** (=cases), **"κορονοιός"** (=coronavirus) and **"πανδημία"** (= pandemic) also confirm the previous word cloud with the most common words in the tweets.

One of the goals of this study is to examine the emotions of the vaccinated and "anti-vaxxers". Since "anti-vaxxers" are against vaccines it is worth considering whether their feelings about Covid vaccines confirm this approach. To proceed with this analysis, we assumed that hashtags: «#GetVaccinated», «#COVID19Vaccination» refer to **vaccinated** people, and hashtags: «#Covidiots», «#antivaxxers», «#VaccineSideEffects», «#antivax», «#VaccineDeath» and «#antivaxx» refer to "**anti-vaxxers**". So, by combining these hashtags we classified tweets into "anti-vaxxers" and vaccinated.

Figure 12: Time series of sentiments for **Vaccinated** and **Antivaxxers** using three different Lexicon-Based Approaches

Figure 12 illustrates **how the average sentiment score changes over time** for **vaccinated** and "**anti-vaxxers**" people using the TextBlob, Vader and Afinn methods. From May 25, 2022, until November 30, 2022, we collected tweets. So we chose this period to form my time series. At first glance, we noticed that in all approaches **vaccinated sentiments** are always **more positive** than those of "**anti-vaxxers**". There are great fluctuations in people's feelings in daily tweets **within seven months**.

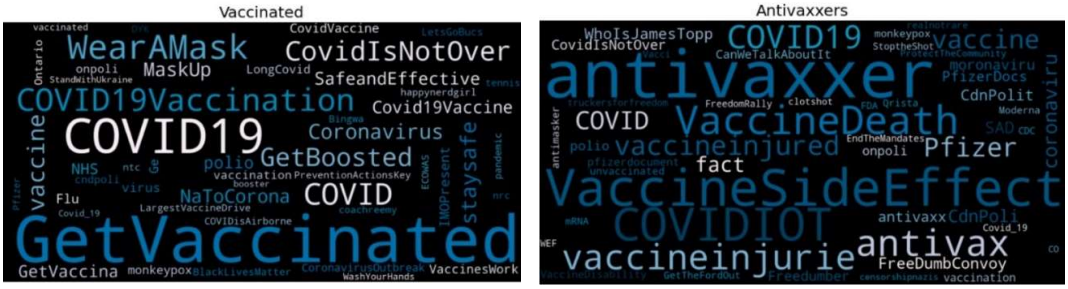
First, with **TextBlob** there are two peaks in the sentiment score for vaccinated people in mid-July, October and November 10th. Only two days with very negative tweets on the 13th of October and 15th of November round to -1%. The rest of the days the tweets were positive for vaccinated. For anti-vaxxers, the worst day was on the 14th of June with -0.4%, while on other days most tweets had a neutral score.

In the **Vader** plot, the highest sentiment average was noted on the 10th of October and November reaching 1% representing a positive day for vaccinated, something that is also confirmed by TextBlob. Before this day, the sentiment was maintained at high levels, with small drops on the 7th and 18th of July and the 12th of September. Big drops were seen on October 13th and November 15th around -1%, similar to TextBlob. Some of the most intense decreases for anti-vaxxers were noted on the 14th of June, 13th and 29th of July, 12th of August and 17th of September where the sentiment average reached approximately -0.5%.

On the 10th of July, October and November, the feelings of the vaccinated became the most positive with the **Afinn** method, as we previously found with TextBlob and Vader. For anti-vaxxers there are few negative sentiments and especially in the tweets that have been created on June 6th and 14th, August 12th and September 17th similar to Vader.

Finally, these graphs clearly show the positive sentiments of the vaccinated, as well as the negative sentiments of the anti-vaxxers, regardless of which Lexicon-Based approach will be used. However, there are differences in the days of most positive or negative tweets in each method.

Figure 13: Word clouds with the hashtags that are used by **Vaccinated** and "**Anti-vaxxers**"



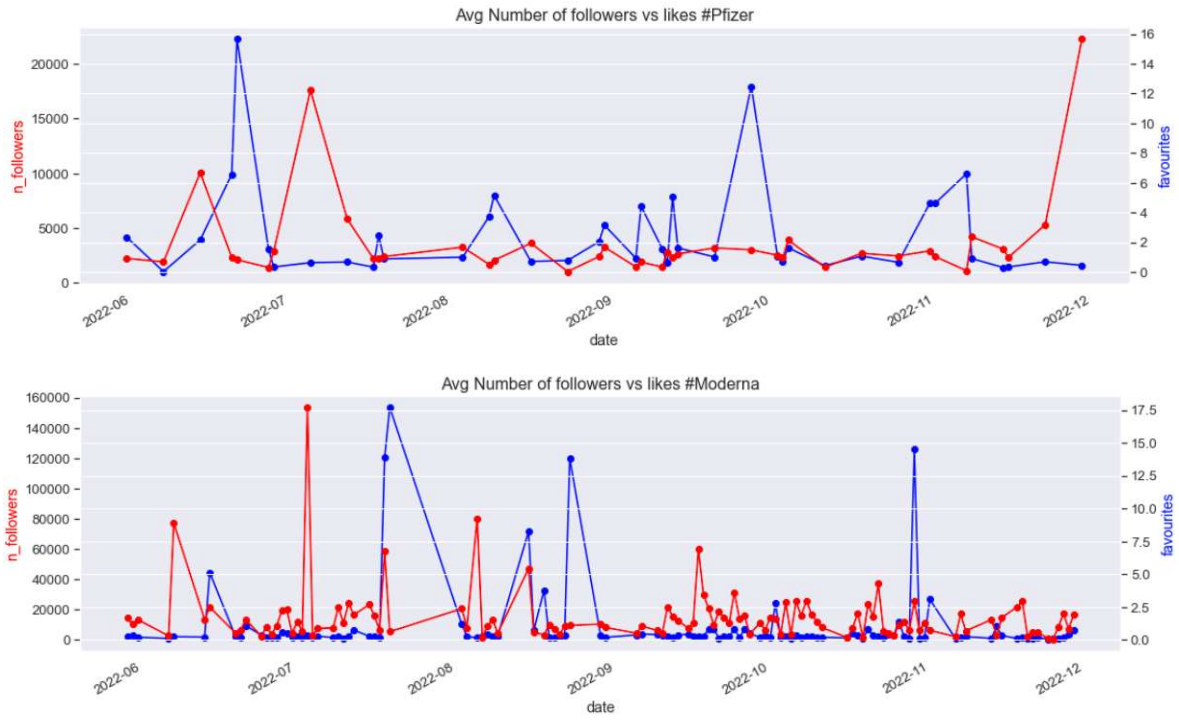
Following the previous analysis, we created word clouds to find out **which hashtags are most commonly used by vaccinated and** "**anti-vaxxers**", in addition to the ones we used to collect tweets. The first word cloud in Figure 13 refers to the vaccinated.

The hashtags "**WearAMask**", "**MaskUp**" and "**CovidIsNotOver**" suggest the fear that although Covid-19 is still shrinking, it has not disappeared, and the masks must remain in our lives. Also, the hashtag "**GetBoosted**" is another way of expressing someone who has been vaccinated.

The second word cloud concerns anti-vaxxers. The use of the hashtags "**Vaccineinjuries**", "**VaccineSideEffect**" and "**VaccineDeath**" confirm that these people believe that vaccines can cause several side effects. Some "anti-vaxxers" use the hashtag "**FakeNews**", as many of them refute a lot of news about Covid.

"**Pfizer**" appears in both word clouds. It makes sense since we discovered that tweets related to the Pfizer vaccine have either positive or negative emotions.

Figure 14: Comparison of **average number of followers vs likes per vaccine**



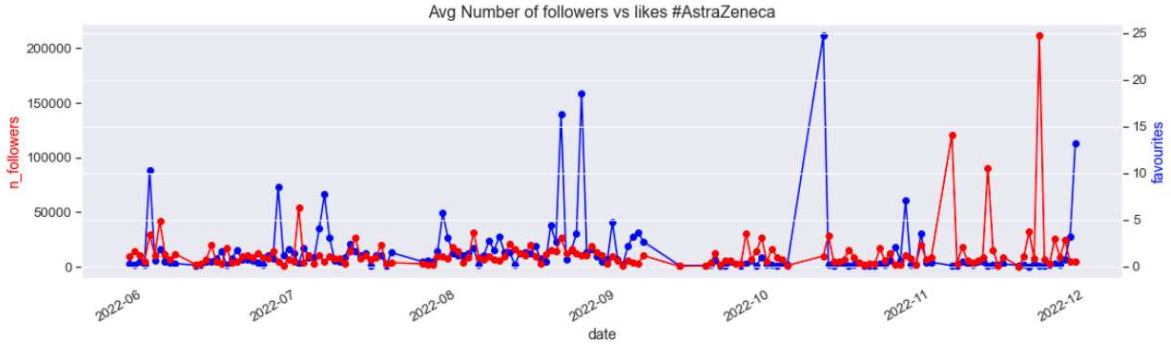


Figure 14 shows **how the average number of followers vs. likes changes over a given set of dates per vaccine** (Pfizer, Moderna, AstraZeneca). The average number of followers is denoted by the red line and likes by the blue line. The spikes in the number of followers correspond to **almost the same spikes** in the number of likes, meaning that tweets written by people with many followers also received a lot of likes. Thus, the number of users' followers is **somehow correlated** with the number of likes. The two lines in each graph are **somewhat consistent**. There is indeed some correlation between followers and likes. Certain spikes that arise in the likes also correspond to spikes that arise in the followers as well.

5. Model Training & Evaluation – Classifier Selection

Before our data can be fed into a ML model, it needs to be converted into a format that the model can understand. For the training of the ML models, the labeled tweets dataset, which was created from the three different Lexicon-Based approaches, was used. The features were extracted from the labeled dataset using the "**Count Vectorizer**" and the "**TF-IDF Vectorizer**". We evaluated both standard Bag-of-Words and TF-IDF features. The performance of the following prominent classifiers was reviewed in our research. More specifically, the experiments were performed using TextBlob, Vader and Afinn sentiments as target classes with the selected ML models to determine the best method in terms of accuracy and F1-Score. We used various parameter settings for these models in our experiments.

5.1 Extracting Features from Clean Tweets

A) Bag-of-Words (BOW)

For the vectorization of the tweets, we first performed **Count Vectorization** on the data and created a bigram model. Vectorization is the process of converting a collection of text documents into a matrix of token counts. A document term matrix is generated where each cell is the count corresponding to the title indicating the number of times a word appears in a document, also known as the term frequency. We adjusted the "CountVectorizer" to 1, 2 (unigrams and bigrams). Increasing the "ngram_range" means that the vocabulary is expanded from single words to short phrases of our desired length [53].

B) TF-IDF

Similar to the Count Vectorization method, in the **Term Frequency-Inverse Document Frequency (TF-IDF)** method, a document term matrix is generated and each column represents a single unique word. The difference in the TF-IDF method is that each cell does not indicate the term frequency, but the cell value represents a weighting that

highlights the importance of that particular word to the document. TF-IDF works by penalizing the common words by assigning them lower weights while giving importance to words that are rare in the entire corpus [54].

TF = (Number of times term t appears in a document) / (Number of terms in the document).

IDF = log(N/n), where, N is the number of documents and n is the number of documents a term t has appeared in.

TF-IDF = TF*IDF

Thus, the TF-IDF word embedding approach was also tested to decide which method is faster and more efficient. The selected parameter for the TF-IDF Vectorizer is the "max_features" which indicates the number of most frequently occurring words to convert, which is 5000 in our case. We also removed stop words such as an, is, are, we and at, since they do not provide much information for classification.

The next step was to **split** the dataset into **training** and **test sets**. 80% for the training set and 20% for testing was chosen after running multiple variations of these percentage thresholds and checking the output results.

To perform the ML algorithms, we selected a subset from our data frame. More specifically, we decided to divide our dataset into two periods, the **summer** and the **autumn**. **From May 25 to August 25** was the **summer season**, so we created two datasets: **100.382 English tweets** and **36.068 Greek tweets**. **From August 26 to November 30** was the **autumn season**, so we created two other datasets: **104.151 English tweets** and **25.041 Greek tweets**. As a result, four different data frames generated from equal-sized months were produced. It would therefore be interesting to examine whether, over time, people's views on vaccination have changed or remain the same.

5.2 Algorithmic performance of tweets over the Summer

A) Results Interpretation of English tweets

Table 4: Performance results of the **English tweets** using the **TextBlob** approach

	Vector-Space Representation	
Model	**Bag-of-Words (BOW)**	**TF-IDF**
LR	**97.29** **97.28**	**94.05** **94.05**
DT	**97.99** **97.98**	**95.76** **95.74**
RF	**97.76** **97.75**	**96.60** **96.57**
XGB	**95.73** **95.69**	**95.02** **94.96**
SVM	**97.79** **97.79**	**94.81** **94.75**

Table 5: Performance results of the **English tweets** using the **Vader** approach

	Vector-Space Representation	
Model	**Bag-of-Words (BOW)**	**TF-IDF**
LR	**95.30** **95.30**	**89.51** **89.51**
DT	**95.47** **95.46**	**92.46** **92.46**
RF	**95.47** **95.46**	**94.16** **94.16**

XGB	**91.00** **91.02**	**90.58** **90.60**
SVM	**95.92** **95.92**	**90.29** **90.29**

Table 6: Performance results of the **English tweets** using the **Afinn** approach

	Vector-Space Representation	
Model	**Bag-of-Words (BOW)**	**TF-IDF**
LR	**95.84** **95.85**	**90.69** **90.73**
DT	**96.84** **96.84**	**93.22** **93.22**
RF	**96.05** **96.06**	**94.53** **94.54**
XGB	**92.00** **92.05**	**91.34** **91.41**
SVM	**96.47** **96.48**	**91.60** **91.63**

Tables 4, 5 and 6 present the results of the ML models for accuracy and F1-score evaluation metrics using the **TextBlob**, **Vader** and **Afinn** classifiers. Regardless of the classifier we used, **DT** and **SVM** received the highest level of accuracy and F1-Score applying the **BOW** approach, while according to the **TF-IDF** approach, **RF** is the optimal algorithm.

More specifically, using the **TextBlob** classifier and the BOW approach, **DT** produced the highest accuracy (97.99%) calculated taking into account the number of correctly classified samples and the total number of samples. The output shows that we can successfully classify a tweet as positive, negative, or neutral with **97.99%** accuracy. The F1-Score of our model was **97.98%** considering both precision and recall, for times when

we want a compromise between the two. It represents the harmonic mean of precision and recall and will be high if both are high. In our model, both precision and recall were high. However, in the TF-IDF method, **RF** outperformed the other models but did not manage to reach better scores than BOW.

In the **Vader** method with BOW, **SVM** was the winner with **95.92%** accuracy and F1-Score followed by DT and RF which achieved similar performances. Using the TF-IDF representation, **RF** was the strongest with a score of 94.16%, something that is also confirmed by TextBlob. In addition, the performance of the models **decreased significantly** when the dataset was changed from TextBlob to annotated Vader. For example, DT accuracy dropped to 95.47% from 97.99%, while LR remarkably reduced to 89.51% from 94.05% when trained with a Vader annotated dataset. The F1 scores of the models had similar reductions. When the dataset changed from Vader to **Afinn**, all models were slightly improved, such as **SVM** accuracy increased from 90.29% to 91.60%.

Experimental results revealed that the models perform better when used with TextBlob annotated data compared to Vader and Afinn. Previous studies [55, 56] show that models perform better when trained on TextBlob labeled data and this study confirms the same.

B) Results Interpretation of Greek tweets

Table 7: Performance results of the **Greek tweets** using the **TextBlob** approach

	Vector-Space Representation	
Model	**Bag-of-Words (BOW)**	**TF-IDF**
LR	**99.72** **99.72**	**97.08** **97.69**
DT	**99.79** **99.79**	**99.43** **99.43**
RF	**99.71** **99.70**	**99.65** **99.64**
XGB	**99.78** **99.77**	**99.65** **99.64**

SVM	**99.76** **99.76**	**99.60** **99.57**

Table 8: Performance results of the **Greek tweets** using the **Vader** approach

	Vector-Space Representation	
Model	**Bag-of-Words (BOW)**	**TF-IDF**
LR	**99.28** **99.25**	**91.42** **93.76**
DT	**99.45** **99.43**	**98.70** **98.66**
RF	**99.31** **99.26**	**99.00** **98.92**
XGB	**99.15** **99.09**	**98.89** **98.79**
SVM	**99.38** **99.35**	**98.45** **98.20**

Table 9: Performance results of the **Greek tweets** using the **Afinn** approach

	Vector-Space Representation	
Model	**Bag-of-Words (BOW)**	**TF-IDF**
LR	**99.78** **99.76**	**94.50** **96.59**
DT	**99.89** **99.88**	**99.20** **99.16**
RF	**99.63** **99.56**	**99.40** **99.27**

XGB	**99.75** **99.72**	**99.45** **99.31**
SVM	**99.79** **99.77**	**99.39** **99.16**

Regarding Tables 7, 8 and 9, it can be noted that the maximum accuracy (**99.89%**), as well as the maximum F1-Score (**99.88%**), are again obtained with the **DT** model for the **BOW** approach. Nevertheless, **SVM** had similar high F1 scores and accuracy. In the **TF-IDF** method, **RF** also outperformed the other models, as we previously found with English tweets. Furthermore, we observed that the models performed worse on Vader, while with TextBlob and Afinn performed equally well. In the **Greek tweets**, all models had very good performances, approaching 100%.

Comparing the English and Greek tweets, the Greek ones managed to gather higher scores. Specifically, the highest score achieved in English tweets was 97.99% while in Greek 99.89%. The difference is 1.9%, which means that **the algorithms can distinguish better in the Greek language when a tweet has a positive, negative or neutral mood**. This is because the pre-processing stage of the Greek tweets was not as strict as that of the English ones. These results confirm that we did not effectively remove punctuations, hashtags and emojis in Greek tweets.

C) Results Interpretation of all tweets – model comparison over the Summer

To address feelings about ongoing vaccinations around the world, three NLP lexicon-based approaches, including **TextBlob**, **Vader** and **Afinn**, were developed along with two vector space representations of tweets, including **Bag-of-Words** and **TF-IDF** and five ML models, including **LR**, **DT**, **RF**, **XGB** and **SVM**. The following visualizations aim to present and analyze the performance of these methods for sentiment analysis and draw meaningful conclusions about our data. With the use of charts, pies, graphs and other visual elements, there is a better understanding and an easier identification of patterns and outcomes in large datasets [57].

Figure 15: Compare accuracy scores of **English tweets** for each ML model using three Lexicon-based approaches

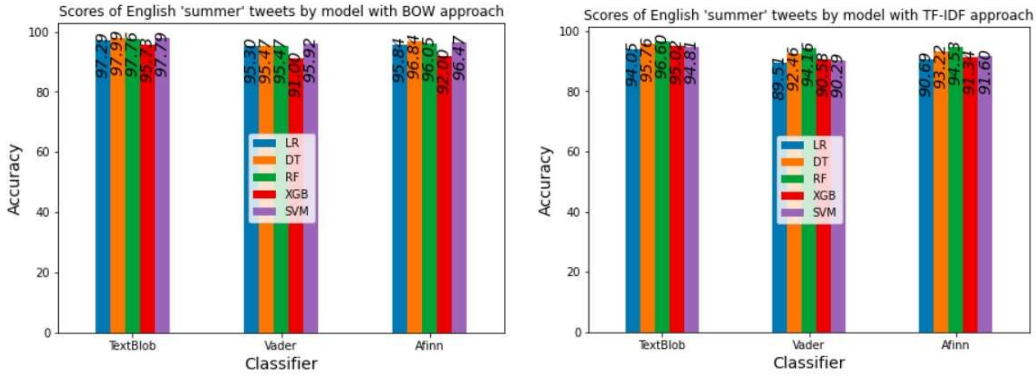
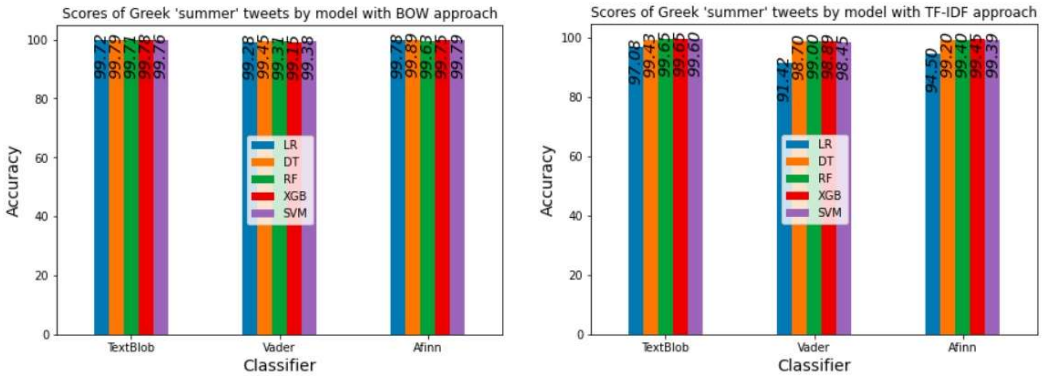


Figure 16: Compare accuracy scores of **Greek tweets** for each ML model using three Lexicon-based approaches

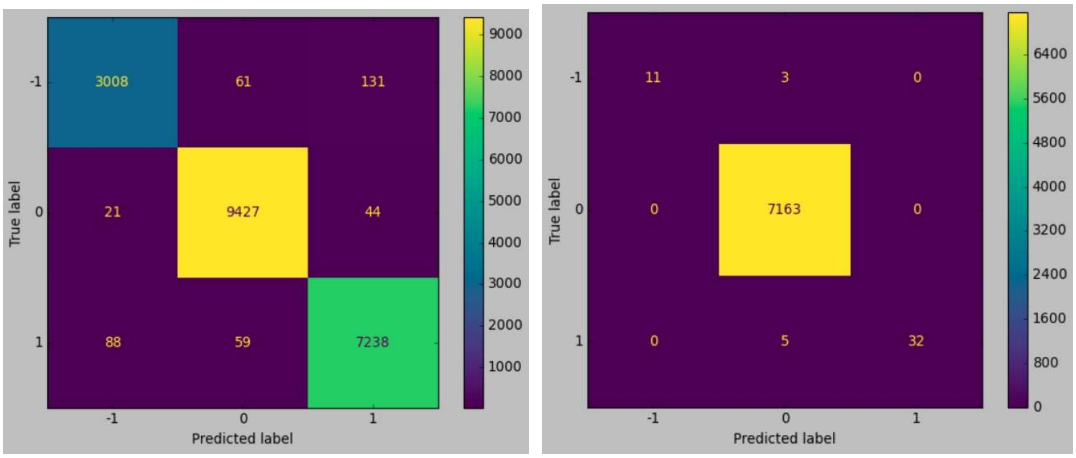


Results of the experiments on **English** and **Greek Twitter data** are provided in Figures 15 and 16. Due to space reasons, we only report accuracy scores and not F1-Scores. The best-performing model for English and Greek tweets is the **DT** which exploits the **BOW** approach, and the worst is the XGB. However, we found that **RF** had the highest accuracy rate for predicting sentiments when analyzed with the **TF-IFD** method, while LR had the lowest. The gap with BOW's approximation results for English tweets is **quite large** (6.99%, from 91% to 97.99%), while for Greek tweets the accuracy ranges from 99.15% to 99.89% (only a 0.74% difference).

By comparing the effectiveness of the different Lexicon-Based approaches, it emerges that **TextBlob** performs significantly **better** than both Vader and Afinn. In addition, models with TF-IDF features took longer to run than models with BOW features. To sum

up, the analysis performed on Twitter data showed that **DT** has the best performance on such data.

Table 10: Confusion Matrices of **Decision Tree** model with **BOW** representation for **English** and **Greek** '**summer**' tweets



To evaluate the performance of a model, many different metrics can be used. Two obvious measures of performance are accuracy and F1-Score, as we have already seen. But in the case of classification, this deserves further analysis. The above confusion matrices are performance measurement techniques for the **Decision Tree** model with the **BOW** approach and show the distribution of model errors. The first confusion matrix is for English tweets, while the second one is for Greek. DT yielded **97.99% accuracy** for **English tweets** with **TextBlob** sentiment classifier and **99.89%** for **Greek** ones with **Afinn** classifier. The target variable has three classes to be predicted. In the case of this project, the classes are either "positive", "negative" or neutral. Each row represents the instances in an actual class, while each column represents the instances in a predicted class.

From the **first** confusion matrix in Table 10, we observed that our model almost perfectly predicted the neutral tweets (9.427), while only 65 (21+44) tweets were misclassified as negative or positive. However, for tweets with negative sentiment, our model got wrong having 131 incorrect predictions, considering them as positive. Respectively regarding the positive sentiment, the model's errors reached 88 tweets. Thus, an important observation from this table was that DT finds it difficult to distinguish positive from negative tweets.

In the **second** 3 by 3 confusion matrix, out of 14 actual negative tweets, the model predicted that 3 were neutral, and out of 7.163 neutral, it predicted all correctly. Of the 37 positive tweets, only 5 were predicted to be neutral. The diagonal line is the true actual values for each class. So, DT made some minor mistakes in categorizing some positive or negative tweets as neutral, although it handled these perfectly.

5.3 Algorithmic performance of tweets over the Autumn season

A) Results Interpretation of English tweets

Table 11: Performance results of the **English tweets** using the **TextBlob** approach

	Vector-Space Representation	
Model	**Bag-of-Words (BOW)**	**TF-IDF**
LR	**97.45** **97.43**	**94.49** **94.47**
DT	**97.94** **97.93**	**95.68** **95.66**
RF	**97.62** **97.61**	**96.60** **96.57**
XGB	**95.72** **95.68**	**95.15** **95.10**
SVM	**97.83** **97.81**	**95.06** **95.00**

Table 12: Performance results of the **English tweets** using the **Vader** approach

	Vector-Space Representation	
Model	**Bag-of-Words (BOW)**	**TF-IDF**

LR	**95.13** **95.13**	**89.92** **89.91**
DT	**95.00** **94.99**	**91.74** **91.73**
RF	**94.75** **94.75**	**93.59** **93.58**
XGB	**91.02** **91.02**	**90.89** **90.89**
SVM	**95.66** **95.66**	**90.61** **90.59**

Table 13: Performance results of the **English tweets** using the **Afinn** approach

	Vector-Space Representation	
Model	**Bag-of-Words (BOW)**	**TF-IDF**
LR	**95.60** **95.61**	**90.60** **90.63**
DT	**96.02** **96.02**	**92.66** **92.66**
RF	**95.32** **95.33**	**94.34** **94.35**
XGB	**91.86** **91.90**	**91.18** **91.22**
SVM	**96.24** **96.24**	**91.25** **91.28**

Tables 11, 12 and 13 show the accuracy and F1-Score, comparing the initial sentiment values of the records with the predicted ones from the models. The results provided an accuracy percentage of **97.94%**, achieved by **DT** using **TextBlob** as the sentiment analysis tool. Other classifiers including LR, RF, XGB and SVM were also tested but

attained lower accuracy rates for the selected dataset. As in the summer dataset, applying the **BOW approach** achieved an **F1-Score** of **97.93%**, which was slightly lower compared to the 97.98% of the summer period, leading to **a decrease of 0.05%**. The fairly high metrics prove the validity of the classifiers and models.

B) Results Interpretation of Greek tweets

Table 14: Performance results of the **Greek tweets** using the **TextBlob** approach

	Vector-Space Representation	
Model	**Bag-of-Words (BOW)**	**TF-IDF**
LR	**99.60** **99.58**	**93.51** **95.67**
DT	**99.68** **99.68**	**99.14** **99.10**
RF	**99.44** **99.40**	**99.36** **99.28**
XGB	**99.68** **99.67**	**99.36** **99.28**
SVM	**99.60** **99.58**	**99.30** **99.21**

Table 15: Performance results of the **Greek tweets** using the **Vader** approach

	Vector-Space Representation	
Model	**Bag-of-Words (BOW)**	**TF-IDF**
LR	**99.42** **99.39**	**89.54** **92.64**
DT	**99.40** **99.39**	**98.54** **98.53**
RF	**99.46**	**99.08**

	99.43	99.01
XGB	99.12 99.05	98.92 98.82
SVM	99.46 99.43	98.30 98.05

Table 16: Performance results of the **Greek tweets** using the **Afinn** approach

	Vector-Space Representation	
Model	**Bag-of-Words (BOW)**	**TF-IDF**
LR	99.88 99.87	94.43 96.58
DT	99.92 99.92	99.48 99.46
RF	99.84 99.83	99.62 99.57
XGB	99.84 99.83	99.62 99.57
SVM	99.88 99.87	99.62 99.55

From the experimental results presented in Tables 14, 15 and 16, **Afinn** has the highest accuracy in comparison with TextBlob and Vader. The accuracy and F1-Score of sentiments are **99.92%** using the **BOW** vectorization method and the **DT** model. Keeping in mind the results of our analysis, we can say that TextBlob and Afinn are far better than the Vader approach for adoption in the analysis of Twitter sentiments about vaccination. In conclusion, in the autumn it seems that the predictions of the emotions of the Greek tweets became more accurate than in the summer. In summer an accuracy of 99.89% was achieved while in autumn 99.92%, leading to **an increase of 0.03%**.

C) Results Interpretation of all tweets – model comparison over the Autumn

Figure 17: Compare accuracy scores of **English tweets** for each ML model using three Lexicon-based approaches

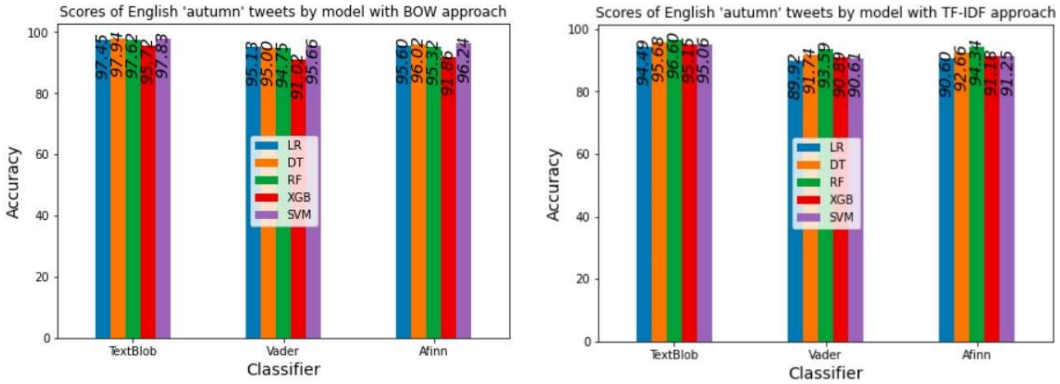
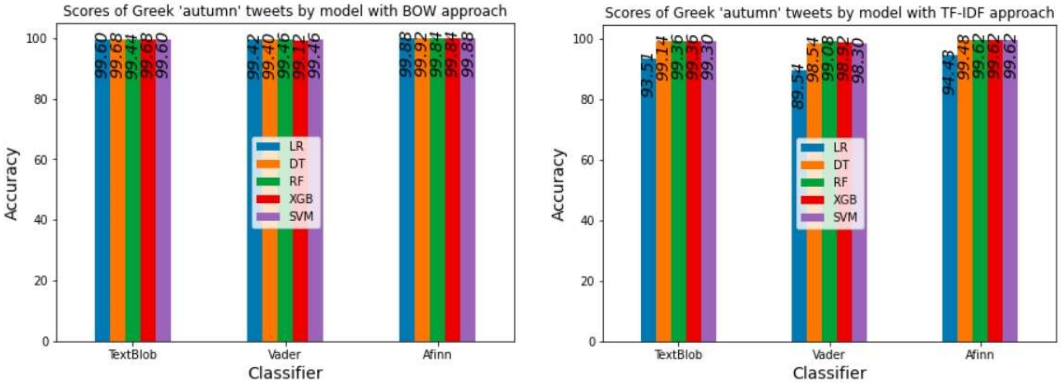
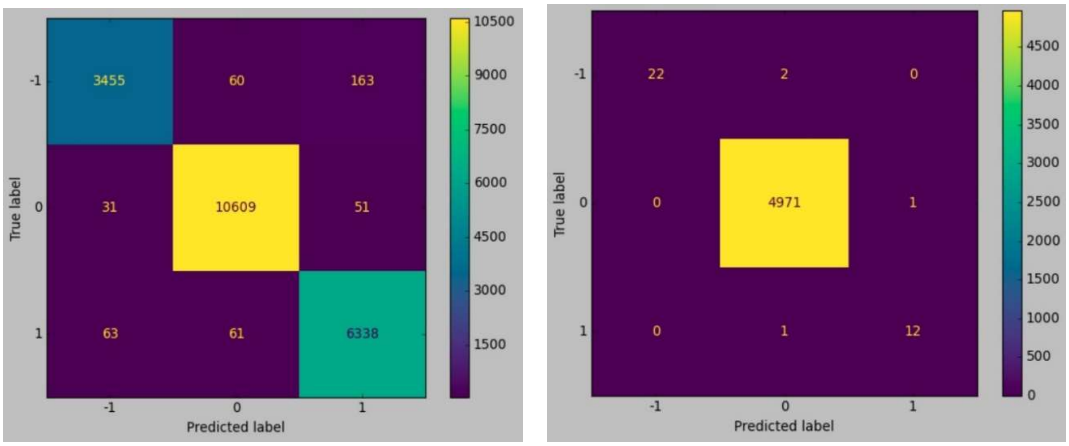


Figure 18: Compare accuracy scores of **Greek tweets** for each ML model using three Lexicon-based approaches



In Figures 17 and 18 we combined the results of each classifier (TextBlob, Vader, Afinn) and compared the best models across the different approaches (BOW, TF-IDF). It is indicated that the performance of **TextBlob and Afinn surpassed** that of **Vader** in both accuracy and F1-Score. Using **TF-IDF** to get vector representations showed **worse performance** than the BOW approach for all models. Moreover, it is noticeable that the **DT** model **outperformed** all other models trained in this study.

Table 17: Confusion Matrices of **Decision Tree** model with **BOW** representation for **English** and **Greek** '**autumn**' tweets



In Table 17, the first confusion matrix is for English tweets, while the second one is for Greek. **DT** yielded **97.94%** accuracy for **English** tweets with the **TextBlob** sentiment classifier and **99.92%** for **Greek** ones with **Afinn** classifier.

From the **first** confusion matrix, we observed that our model almost perfectly predicted the neutral tweets (10.609), while only 82 (31+51) tweets were misclassified as negative or positive. However, for tweets with negative sentiment, our model got wrong having 163 incorrect predictions, considering them as positive. Respectively regarding the positive sentiment, the model's errors reached 63 tweets.

In the **second** 3 by 3 confusion matrix, out of 24 actual negative tweets, the model predicted that 2 were neutral and out of 4.971 neutral, it predicted all correctly, except for 1 positive. Of the 13 positive tweets, only 1 was predicted to be neutral. As a result, the observations of the previous confusion tables in Table 9 were confirmed here as well. DT **misclassified some positive or negative tweets as neutral**, thus increasing the number of neutral tweets.

5.4 Model Fine-Tuning

After finding the predictions for the models, Grid Search was performed on the two most accurate models: **Decision Tree** with Bag-of-Words features and **Random Forest** with TF-IDF features. Grid Search with Cross-Validation (**GridSearchCV**) is a brute-force method for finding the best hyperparameters for our dataset and our model. More

specifically, it is the process of splitting the same dataset in K-partitions and for each split, we search the whole grid of hyperparameters to an algorithm (Decision Tree or Random Forest) in a brute-force manner of trying every combination [58]. Cross-validation is a very important method used to create better-fitting models by training and testing on all parts of the training dataset. The one drawback experienced while incorporating GridSearchCV was the runtime. Cross-validation and grid tuning lead to longer training times given the repeated number of iterations a model must train through [59]. So, we tried parallel computations, setting the parameter "n_jobs = -1" where is the number of jobs running in parallel ("-1" means that all CPU cores will be used, which drastically reduces the training time).

However, my parameter tuning gave me worse results than with the default parameters. We used the Random Forest model with default parameters which gave me **96.60%** accuracy. When we used parameter tuning with GridSearchCV, we got **a very low accuracy (63.83%)** after trying different combinations. This can happen if the customized parameters we have chosen for tuning are worse than the default parameters. Parameter tuning only works if a set of customized parameters makes a better setup than the default setup. The search can only test the parameters that we fed into the "param_grid" argument. There could be a combination of parameters that further improves the performance of the model. But we did not include more values for each parameter, because conducting an exhaustive search of all parameters is an incredibly time-consuming task, as mentioned before. Therefore, the only way to find the best possible hyperparameters for our dataset was by trial and error, which is the main concept behind hyperparameter optimization.

Then, we tried to tune the hyperparameters of the Random Forest model using RandomSearch with Cross-Validation (**RandomizedSearchCV**). It implements a randomized search over parameters, where each setting is sampled from a distribution over possible parameter values. This has two main benefits over an exhaustive search. A budget can be chosen independent of the number of parameters and possible values and adding parameters that do not influence the performance does not decrease efficiency [60]. We defined the hyperparameter space, which had a bigger range of values than the one we built for grid search since random search does not try out every single combination of hyperparameters. It randomly samples hyperparameters to find the best ones, which means that, unlike grid search, random search can look through a large number of values

quickly. The advantage of this approach is the accuracy in prediction which has been improved to **92.83%**. Optimizing the RF model, gave us better accuracy than GridSearchCV but did not reach the optimal accuracy of 96.60%, achieved with the default parameters.

5.5 Model Validation

Sentiment analysis uses NLP to make sense of human language, and ML to automatically deliver accurate results. But, many times, the TextBlob, Vader and Afinn classifiers perceive for example a positive comment as a negative one, thus producing a wrong sentiment score. It would be more valid if we checked people's sentiments using our judgment and compared it with the results of our models. It would be desirable to be consistent with the ground truth. However, it is clear that one's judgment can be subjective and vary from person to person. Therefore, by selecting a sample of tweets from all the hashtags we used to extract our data and assigning different people to classify the tweets as positive, negative and neutral, the occurrence of misclassification was limited.

More specifically, we selected ten tweets for each hashtag. Regarding the English tweets, we collected 130, since there were 13 hashtags. While for the Greek ones, we collected 60, since there were 6 hashtags. Then, we merged all the English tweets and exported them to an Excel file without the sentiment of the tweets, so that we could assign them to people to manually categorize the sentiment for us. In this way, the validation of the sentiment of our tweets became more objective.

In Table 18, the results showed that **69 of the 130 English** tweets are correctly classified according to the model, yielding **a 53.07% algorithmic-to-human compatibility rate**. The lowest score was 41.53%, meaning that 54 out of the 130 English tweets are correctly categorized. Regarding the **Greek** tweets, the classifier considered most of them with a neutral feeling, which is **68.33% consistent** with the sentiment of people who manually identified them. The person who gathered the least number of correct tweets does not keep up with the algorithm. In Table 19 it appears that just 27 out of 60 i.e. 45% compatibility rate of the algorithm with the human. As a result, it is not easy to recognize the sentiment of the people, as it is something not so stable and we have seen opinions differ from person to person. Many tweets were caustic, ironic and funny

making it difficult to categorize them into three strict categories: positive, negative and neutral.

Table 18: Compatibility score for **English** tweets

Person	Correctly classification of **English** tweets according to the model	Algorithmic-to-human compatibility rate
1st	54/130	41.53%
2nd	59/130	45.38%
3rd	59/130	45.38%
4th	61/130	46.92%
5th	65/130	50%
6th	69/130	**53.07%**

Table 19: Compatibility score for **Greek** tweets

Person	Correctly classification of **Greek** tweets according to the model	Algorithmic-to-human compatibility rate
1st	27/60	45%
2nd	32/60	53.33%
3rd	34/60	56.66%
4th	38/60	63.33%
5th	39/60	65%
6th	41/60	**68.33%**

In conclusion, we would say that the results of the algorithms are not so encouraging for English tweets, while they are for Greek ones. They do not fully correspond to reality, however, this is a drawback of any research. Given that subjectivity is unavoidable, we asked more than one person to check the same tweets to ensure that the coding was consistent with the classifiers, although the sample of 130 and 60 tweets for the English and Greek respectively is not representative. If we evaluate another 500 tweets, these percentages may change. However, this attempt helped us to set some guidelines for what is included in each emotion category. Table 20 shows the result of the classifier as well as the result of the people who manually classified the tweets according to their opinion.

Table 20: Algorithmic to human compatibility

created_at	tweet_text	keyword	sentiment_ classifier	sentiment_ person
2022-08-23	Instead of paying student debt how about paying back the Americans who lost their jobs due to mandating a #vaccine that doesn't work	#vaccine	-1	-1
2022-08-22	@WalesOnline @demonscythsynth Only when you're a #mRNA scam victim. I feel sorry for you people, honestly. #VaccineSideEffects	#Vaccine SideEffects	1	-1
2022-08-23	RT @darmamar: I'm not #AntiVax, I'm #AntiSlavery. You don't own me.	#antivax	0	-1
2022-05-31	RT @ChapCareOrg: As #COVID19 cases rise again locally and throughout the nation, the best protection against the virus is masking and vacc…	#Get Vaccinated	1	1

6. Discussion - Threats to validity

More and more people express their opinions on Twitter, making it a huge data source. During the global COVID-19 outbreak, many individuals, as well as organizations and government agencies post their viewpoints regarding the coronavirus. Even though vaccines are considered a weapon against COVID, and millions of people are vaccinated every day, there are still several doubts about the safety of vaccines. A significant number of people worldwide believe that vaccines are harmful.

This research work aimed to **identify the emotional state of people about coronavirus**. A total of **246.626 English tweets** and **99.315 Greek tweets** were retrieved and stored in two different tables in MySQL Database. **Data preprocessing** was then performed, which includes cleaning and normalization techniques. It was not as strict in the Greek language as in English.

Important features were extracted that helped to better understand our data. For example, the tweets were **geo-located** inferring that the USA and Canada are the countries with the highest number of tweets. Also, users who tweet more have a medium number of followers.

Sentiment analysis was implemented using four well-known Lexicon-Based techniques. When the **English** dataset was studied, the prevailing sentiment with the TextBlob and Vader classifiers was **neutral**, while with Afinn it was **negative**. **Neutral** sentiment dominated the **Greek** dataset with the TextBlob, Vader and Afinn approach. That means that the majority of users expressed themselves **more objectively** without using personal opinions and emotions. Using the **NRC** Sentiment and Emotion Lexicon in our **English** dataset, it is observed that there is a large number of **positive** tweets, followed by a sentiment of "**trust**".

The results of this study suggest that people living in the USA and Canada have different emotions. Considering that every country reacted differently in its battle with the virus, personalized analysis for a specific country, with their government's response to the pandemic waves is more easily managed and manipulated.

The attitude of the people towards the **three vaccines** (Pfizer, Moderna, Astrazeneca) was **positive**, as confirmed by Marcec and Likic [35] and Shamrat and Chakraborty [36]. More specifically, TextBlob gave Moderna more positive polarity scores, while Vader and Afinn gave AstraZeneca the highest positive scores. The fourth classifier used only in our **English** dataset was the NRC Sentiment and Emotion Lexicon. According to this, there is a large number of **positive** emotions, which contrasts with the other classifiers. Therefore, people are positive for the course of the pandemic, supporting the work of scientists, governments and the WHO. Although, in both the USA and Canada people are overwhelmed with mixed feelings of "**fear**" and "**trust**".

Time-based sentiment analysis was also performed to analyze the change in trends of people regarding COVID-19 sentiments. People's reactions vary day to day from posting their feelings on Twitter. It is confirmed that "**anti-vaxxers**" are expressing their **displeasure** with the COVID-19 vaccine, as the majority of their sentiments are classified as **negative** regardless of the date, as previously found by Asderis [19]. Thus, it can be assumed that anti-vaxxers are tired of the pandemic and are not so confident in expressing their personal beliefs but prefer to share already published thoughts. Based on this, the viral spread of misinformation and the creation of fear can be explained.

For the vectorization of the tweets, **BOW** (Bag-f-Words) and **TF-IDF** (Term Frequency-Inverted Document Frequency Features) embedding approaches were used. Afterward, the experimentation with **five classification models** (Logistic Regression, Decision Tree, Random Forest, XGBoost Classifier and SVM) was done using various parameter settings for these models in our experiments and the evaluation of their performance was comparatively analyzed.

We have divided our data into two periods, **summer and autumn**. As a result, four data sets generated from equal-sized months were produced. The classification of emotions using **TextBlob** and the vectorization of tweets using **BOW** contribute to the better performance of the ML models.

For the **summer** season, the best results for both **English** and **Greek languages** were achieved with the **Decision Tree**, which was also proven in the paper [37]. Top accuracy was **97.99%** and the top F1-Score was 97.98%. For the **Greek** language, Afinn was the winner with **99.89%** accuracy and 99.88% F1-Score.

As for the **autumn** period, the performance of the DT model for the **English** tweets was slightly reduced to 0.05%, reaching **97.94%**. While the accuracy rate of the DT for **Greek** tweets improved by 0.03%, leading to the highest accuracy score of **99.92%**.

By taking into consideration the **threats to the validity** of the present study, it should be referred that Twitter data contains a lot of noise. Although they have handled it properly and implemented some techniques, there is still the possibility of imperfect data. The lack of good data can cause our algorithms to perform poorly and hence limit the capabilities of our model. Additionally, the tweets collected for this study were in English and Greek and their users may not be representative of the general public. As a result, their tweets simply reflected the views and feelings of Internet users about vaccination, which could serve as a limitation of the study. Another threat to the validity of our results is spam accounts, fake accounts and bots that are involved in Twitter data.

7. Conclusions and Further work

Twitter sentiment analysis falls into the realm of text and opinion mining. It focuses on analyzing the emotions of tweets and feeding the data to a ML model to train it and then check its accuracy so that it can be used in future analyses.

Considerable work has been done in the field of sentiment analysis either from sentiment lexicons or from ML techniques. In this study, the importance of analyzing emotions on social media was presented. There was a focus on Twitter and python programming language was the principal tool for the implementation of emotion analysis.

This research focused on providing **a comparison between sentiment lexicons** (TextBlob, Vader, Afinn and NRC) so that the best can be adopted for sentiment analysis. We **validated three of the sentiment analysis lexicons** with **five ML algorithms using two different vectorization methods** (Bag-of-Words and TF-IDF). As a result, we calculated sentiments from three analyzers named TextBlob, Vader and Afinn. Additionally, we tested their results with five supervised ML classifiers, Logistic Regression, Decision Tree, Random Forest, XGBoost Classifier and SVM, **dividing** the dataset **into two periods**, **summer and autumn**. This is a novel approach to this research which was not present in any previous work.

We realized through the findings about the emotion of the posts, that the use of English tweets, and particularly with the use of **TextBlob** as a sentiment analysis tool and **BOW** approach, produces the greatest predictive power. This means that our model can correctly predict the true positive and true negative points. In terms of accuracy, **DT** presents **the best performance** (97.99%) while in terms of F1-Score, it demonstrates the best discriminatory power (97.98%). Finally, the sentiment of Greek tweets was predicted 99.89% correctly for most days. In the autumn period, the DT model achieved a bit lower accuracy (97.94%) compared to that of the summer period. Meanwhile, in Greek tweets, the performance of the model increased reaching 99.92%.

Moreover, the traditional ML model performed better than the tuned model, even after extensively tuning the hyperparameters of this model. More advanced tuning is required

to try and produce better performance with this model. However, despite tuning being a very laborious task, it still does not guarantee that the model will then outperform DT.

This research has provided some useful insights into people's emotions and psychology. Lexicon-based Twitter sentiment analysis is a **valuable and easily implemented tool** to track the sentiment regarding SARS-CoV-2 vaccines. Our study may serve as a proof of concept demonstrating that using **a simply implemented method** it is possible to track the sentiment towards vaccines almost **in real-time**. Such insight may prove valuable in enabling the **planning and implementation of healthcare interventions** aimed at **increasing the uptake of COVID-19 vaccines** and **fighting vaccine hesitancy**. Governments and health experts could use this **on-the-fly approach obtained from sentiment lexicons**, to develop effective crisis management strategies and better inform the public. It is encouraging that feelings about the Pfizer, Moderna and AstraZeneca vaccines seem to **become positive over time**, as this may increase vaccine acceptance rates.

In future research, we should conduct the analysis with a larger dataset, acquired from multiple social network sites. Furthermore, we could evaluate the impact of bots posting misinformation and thereby influencing social media sentiment towards vaccination. Other Lexicon-based approaches, as well as vector representation methods, such as Word2Vec can be further investigated.

To take our initiative to next level, we would like to zoom in the combination of different preprocessing steps and the effects of these steps on the performance of the models would be interesting research to explore in the future. Additionally, a more in-depth analysis of the models can be conducted, such as Artificial Neural Networks (ANN) and Deep Learning (DL) models, as this might greatly affect further generalization on new datasets. Thus, several different approaches could be used and compared with the lexicon-based Twitter sentiment analysis model results as a baseline.

References

[1] Wang, H., Wang, Z., Dong, Y., Chang, R., Xu, C., Yu, X., ... & Cai, Y. (2020). Phase-adjusted estimation of the number of coronavirus disease 2019 cases in Wuhan, China. Cell discovery, 6(1), 1-8.

[2] WHO, C. O. (2020). World health organization. Responding to the Community Spread of COVID-19. Reference WHO/COVID-19/Community_Transmission/2020.1.

[3] Statista, 2022, https://www.statista.com/statistics/303681/twitter-users-worldwide/

[4] Scott, J. (2011). Social network analysis: developments, advances, and prospects. Social network analysis and mining, 1(1), 21-26.

[5] Poria, S., Cambria, E., & Gelbukh, A. (2015). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 2539-2544).

[6] Marcec, R., & Likic, R. (2021). Using twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. Postgraduate Medical Journal.

[7] 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.

[8] Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior, 31, 527-541.

[9] Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research, 50, 723-762.

[10] C. Tjortjis, Data Mining - Lecture notes, 2022.

[11] Ramakrishnan, R., & Gehrke, J. (2002). Συστήματα διαχείρισης βάσεων δεδομένων: Β΄ τόμος.

[12] P. Koukaras, C. Nousi and C. Tjortjis, “Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning”, Telecom, MDPI, 3(2), 358-378, 2022.

[13] V. Chouliara, E. Kapoteli, P. Koukaras and C. Tjortjis, “Social Media Sentiment Analysis related to COVID-19 Vaccinations”, Artificial Intelligence and Machine Learning for Healthcare. Intelligent Systems Reference Library Vol. 229, pp. 47–69, Springer, 2023.

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

[15] Bagheri, H., & Shaltooki, A. A. (2015). Big Data: challenges, opportunities and Cloud based solutions. International Journal of Electrical and Computer Engineering, 5(2), 340.

[16] Dunn, A. G., Surian, D., Dalmazzo, J., Rezazadegan, D., Steffens, M., Dyda, A., ... & Mandl, K. D. (2020). Limited role of bots in spreading vaccine-critical information among active twitter users in the United States: 2017–2019. American Journal of Public Health, 110(S3), S319-S325.

[17] Zhang, M., Qi, X., Chen, Z., & Liu, J. (2022). Social Bots’ Involvement in the COVID-19 Vaccine Discussions on Twitter. International Journal of Environmental Research and Public Health, 19(3), 1651.

[18] Yuan, X., Schuchard, R. J., & Crooks, A. T. (2019). Examining emergent communities and social bots within the polarized online vaccination debate in Twitter. Social media+ society, 5(3), 2056305119865465.

[19] Mønsted, B., Sapieżyński, P., Ferrara, E., & Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using Twitter bots. PLoS ONE, 12, e0184148.

[20] Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., & Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. American Journal of Public Health, 108, 1378–1384.

[21] Asderis G.–A., “Sentiment Analysis on Twitter Data: a Detailed Comparison of TextBlob and VADER”, MSc Dissertation. IHU, 2022.

[22] E. Tsiara, C. Tjortjis, 'Using Twitter to Predict Chart Position for Songs', 16th Int'l Conf. on Artificial Intelligence Applications and Innovations (AIAI 20), 2020.

[23] Razzaq, A., Abbas, T., Hashim, S., Qadri, S., Mumtaz, I., Saher, N., ... & Nawaz, S. A. (2022). Extraction of Psychological Effects of COVID-19 Pandemic through Topic-Level Sentiment Dynamics. Complexity, 2022.

[24] Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings (Vol. 30, No. 1, pp. 25-36). Technical report C-1, the center for research in psychophysiology, University of Florida.

[25] Gan, Q., & Yu, Y. (2015, January). Restaurant Rating: Industrial Standard and Word-of-Mouth--A Text Mining and Multi-Dimensional Sentiment Analysis. In 2015 48th Hawaii International Conference on System Sciences (pp. 1332-1340). IEEE.

[26] Mohammad, S. M., & Turney, P. D. (2013). Nrc emotion lexicon. National Research Council, Canada, 2, 234.

[27] Rustam, F., Mehmood, A., Ahmad, M., Ullah, S., Khan, D. M., & Choi, G. S. (2020). Classification of shopify app user reviews using novel multi text features. IEEE Access, 8, 30234-30244.

[28] Brijain, M., Patel, R., Kushik, M. R., & Rana, K. (2014). A survey on decision tree algorithms for classification.

[29] Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (2018, October). Sentiment analysis about E-commerce from tweets using decision tree, K-nearest neighbor, and naïve bayes. In 2018 international conference on orange technologies (ICOT) (pp. 1-6). IEEE.

[30] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

[31] Al-Qudah, D. A., Ala'M, A. Z., Castillo-Valdivieso, P. A., & Faris, H. (2020). Sentiment analysis for e-payment service providers using evolutionary eXtreme gradient boosting. IEEE Access, 8, 189930-189944.

[32] Huang, Q., Chang, S., Liu, C., Niu, B., Tang, M., Zhou, Z., 2015. An evaluation of fake fingerprint databases utilizing SVM classification. Pattern Recognit. Lett. 60, 1-7.

[33] Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian joint conference on artificial intelligence (pp. 1015-1021). Springer, Berlin, Heidelberg.

[34] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160(1), 3-24.

[35] Sutoyo, E., & Almaarif, A. (2020). Twitter sentiment analysis of the relocation of Indonesia's capital city. Bulletin of Electrical Engineering and Informatics, 9(4), 1620-1630.

[36] E. Kapoteli, P. Koukaras and C. Tjortjis, "Social Media Sentiment Analysis Related to COVID-19 Vaccines: Case studies in English and Greek language", Proc. 18th Int'l Conf. Artificial Intelligence Applications and Innovations (AIAI 22), 2022.

[37] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018), http://arxiv.org/abs/1810.04805

[38] Marcec, R., & Likic, R. (2021). Using twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. Postgraduate Medical Journal.

[39] Shamrat, F. M. J. M., Chakraborty, S., Imran, M. M., Muna, J. N., Billah, M. M., Das, P., & Rahman, O. M. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. Indones. J. Electr. Eng. Comput. Sci, 23(1).

[40] Reshi, A. A., Rustam, F., Aljedaani, W., Shafi, S., Alhossan, A., Alrabiah, Z., ... & Ashraf, I. (2022, February). COVID-19 Vaccination-Related Sentiments Analysis: A Case Study Using Worldwide Twitter Dataset. In Healthcare (Vol. 10, No. 3, p. 411). MDPI.

[41] Cotfas, L. A., Delcea, C., Roxin, I., Ioanăş, C., Gherai, D. S., & Tajariol, F. (2021). The longest month: Analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. IEEE Access, 9, 33203-33223.

[42] Abd Rahim, N., & Rafie, S. M. (2020). Sentiment analysis of social media data in vaccination. Int J, 8(9).

[43] M. Karagkiozidou, P. Koukaras, C. Tjortjis, Sentiment Analysis on COVID-19 Twitter Data: A Sentiment Timeline, Proc. 18th Int'l Conf. Artificial Intelligence Applications and Innovations (AIAI 22).

[44] P. Koukaras, C. Tjortjis and D. Rousidis, “Mining Association Rules from COVID-19 Related Twitter Data to Discover Word Patterns, Topics and Inferences”, Information Systems, p. 102054, Elsevier, 2022

[45] Baker, Q. B., Shatnawi, F., Rawashdeh, S., Al-Smadi, M., & Jararweh, Y. (2020). Detecting epidemic diseases using sentiment analysis of Arabic tweets. J. Univers. Comput. Sci., 26(1), 50-70.

[46] Patel, R., & Passi, K. (2020). Sentiment analysis on Twitter data of world cup soccer tournament using machine learning. IoT, 1(2), 218-239.

[47] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

[48] Dastanwala, P. B., & Patel, V. (2016, March). A review on social audience identification on twitter using text mining methods. In 2016 international conference on wireless communications, signal processing and networking (WiSPNET) (pp. 1917-1920). IEEE.

[49] Gautam, J., Atrey, M., Malsa, N., Balyan, A., Shaw, R. N., & Ghosh, A. (2021). Twitter data sentiment analysis using naive bayes classifier and generation of heat map for analyzing intensity geographically. In Advances in Applications of Data-Driven Computing (pp. 129-139). Springer, Singapore.

[50] Cheng, Z., Caverlee, J., & Lee, K. (2010, October). You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 759-768).

[51] European Center for Disease prevention and Control. (2021). Maps in support of the Council Recommendation on a coordinated approach to travel measures in the EU.

[52] Mohammad, S., Salameh, M., & Kiritchenko, S. (2016, May). Sentiment lexicons for Arabic social media. In Proceedings of the tenth international conference on language resources and evaluation.

[53] Barry, J. (2017). Sentiment Analysis of Online Reviews Using Bag-of-Words and LSTM Approaches. In AICS (pp. 272-274).

[54] Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. Expert systems with applications, 38(3), 2758-2765.

[55] Saad, E., Din, S., Jamil, R., Rustam, F., Mehmood, A., Ashraf, I., & Choi, G. S. (2021). Determining the Efficiency of Drugs under Special Conditions from Users' Reviews on Healthcare Web Forums. IEEE Access, 9, 85721-85737.

[56] Nousi, C., & Tjortjis, C. (2021, September). A Methodology for Stock Movement Prediction Using Sentiment Analysis on Twitter and StockTwits Data. In 2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 1-7). IEEE.

[57] World Health Organization. (2021). Tools for making good data visualizations: the art of charting (No. WHO/EURO: 2021-1998-41753-57181). World Health Organization. Regional Office for Europe.

[58] Ranjan, G. S. K., Verma, A. K., & Radhika, S. (2019, March). K-nearest neighbors and grid search cv based real time fault monitoring system for industries. In 2019 IEEE 5th international conference for convergence in technology (I2CT) (pp. 1-5). IEEE.

[59] Guo, J., Nomura, A., Barton, R., Zhang, H., & Matsuoka, S. (2018, March). Machine learning predictions for underestimation of job runtime on HPC system. In Asian Conference on Supercomputing Frontiers (pp. 179-198). Springer, Cham.

[60] Agrawal, T. (2021). Hyperparameter optimization using scikit-learn. In Hyperparameter Optimization in Machine Learning (pp. 31-51). Apress, Berkeley, CA.

Appendix A

In Figures 19 and 20, two correlation matrices are provided, measuring the associations between the variables and the strength of their relationship. The correlation matrix summarizes the data and allows us to see which variable pairs have the highest correlation. However, the observable pattern is that **not all variables are highly correlated with each other**. More specifically, the highest correlation with a value of **17%** was observed between the "**sentiment_TextBlob**" and "**subjectivity**" variables in the **English tweets** and **-0.03%** respectively in the **Greek tweets**. We excluded the field "tweet_id" because it does not provide useful information in our analysis.

Figure 19: Correlation Matrix for all numeric attributes of our data frame "**English tweets**"

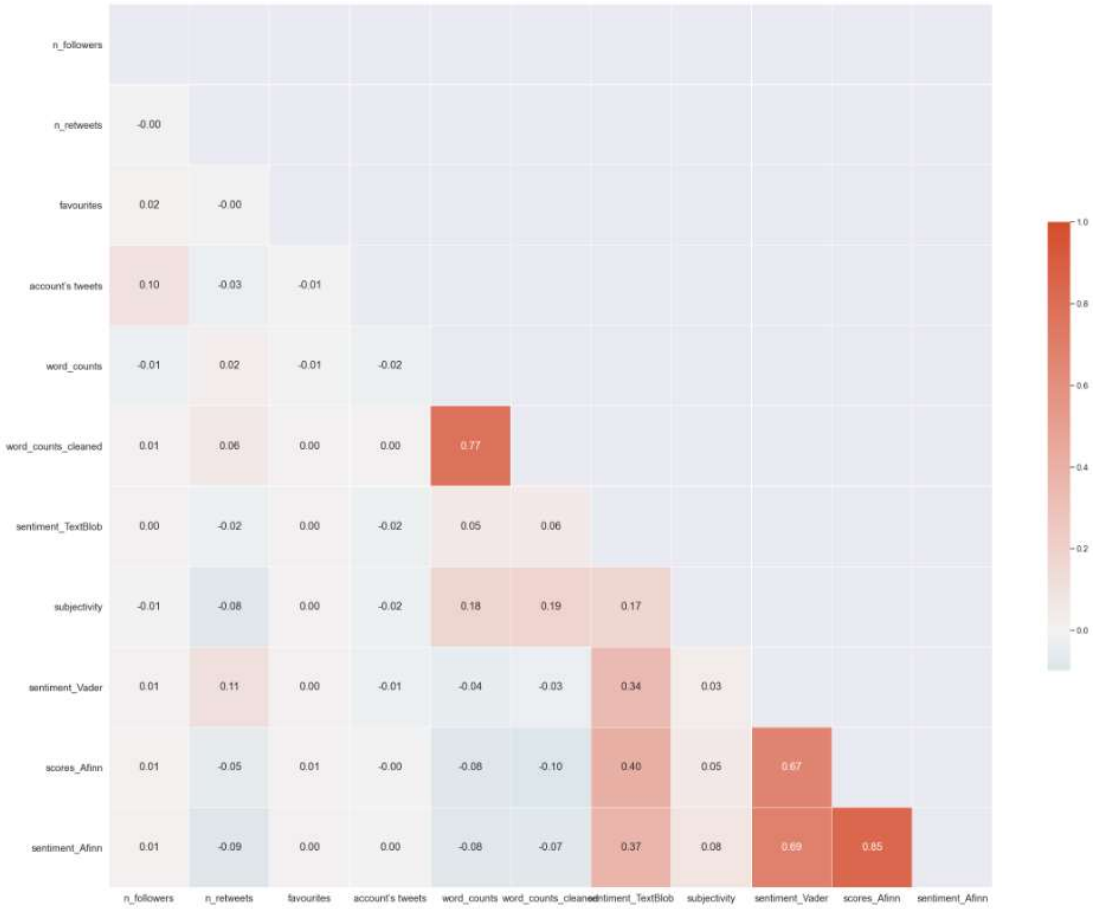


Figure 20*:* Correlation Matrix for all numeric attributes of our data frame "**Greek tweets**"