INTERNATIONAL HELLENIC UNIVERSITY

"Twitter Sentiment Analysis related to COVID-19 Vaccinations"

Chrysoula Dontaki, Eleftheria Trigeni

UNIVERSITY CENTER OF INTERNATIONAL PROGRAMMES OF STUDIES

SCHOOL OF SCIENCE AND TECHNOLOGY

JUNE 2022

Thessaloniki – Greece

ABSTRACT

The COVID-19 pandemic caused by the coronavirus SARS-CoV-2 was originated in China in December 2019 [1]. The virus has infected and killed thousands of people according to the World Health Organization (WHO) that has announced COVID-19 outbreak as a pandemic that hit the world [2]. An end to this pandemic can bring a worldwide vaccination campaign. However, vaccines have traditionally met with public fear and hesitancy. During the lockdown, that many countries have suffered from, people spent hours every day on social media platforms sharing their opinions and expressing their feelings. As a result, Twitter become a valuable and 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, so as to increase vaccination acceptance. Therefore, the sentiment analysis process of classifying opinions towards vaccines into categories like "positive" or "negative" or "neutral" will 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 and Natural Language Processing (NLP). The twitter data has been retrieved in portions at different points of time within a month through Python programming language. After data preprocessing, the sentiment analysis operation has been done by using TextBlob, Valence Aware Dictionary and sEntiment Reasoner (VADER) and AFINN tools. Graphical representation and performance analysis with state-of-the-art models (Random Forest, Logistic Regression, Decision Tree, XGBoost Classifier) has been conducted on the tweets. The results of the study suggest that most of the well-known classification algorithms can categorize people's feelings with comparable performances, but Random Forest model outperforms all of them with an accuracy score of 96%.

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

Table of Contents

Tables

Figures

1. INTRODUCTION

Nowadays, Twitter is one of the most popular social media which has over 300 million accounts and the numbers are rapidly increasing day to day [3]. Millions of people give their opinion of different topics on a daily basis on social medias platforms, still Twitter is the richest source to learn about people's opinion and sentimental analysis [4]. The sentiment-aware systems these days have many applications from business to 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 widespread use and more specifically authorized for use in the United Kingdom on 2 December 2020 [6]. Marcec, R., & Likic, R. 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 so as to help health staff and government organizations to get benefit from those data. Even 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 these people, called antivaxxers, separately. For each tweet it is important to determine whether the 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 text. Since social networks, especially Twitter, contains small texts and people may use different words and abbreviations which are difficult to extract their sentiment by current Natural Language processing systems easily, thus some researchers have used deep learning and machine learning techniques to extract and mine the polarity of the text [8]. Text mining is an important issue that analysts are constantly asked to address. Some machine learning 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 called end-to-end productive models. In some areas such as image recognition, 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. Three NLP lexicon-based approaches (TextBlob, the Valence Aware Dictionary and sEntiment Reasoner VADER and Afinn), along with four Machine Learning models (Random Forest (RF), Logistic Regression (LR), Decision Tree (DT) and XGBoost Classifier (XGB)) have been used for sentiment analysis. The results revealed people's feelings to devise relevant policies to increase the acceptance of COVID-19 vaccination.

2. RELATED WORK – LITERATURE REVIEW

During the pandemic, many researchers sought to find and understand the attitudes of people associated with vaccines that seemed to have many facets. They also tried to find out if there was 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 work concerns Machine Learning algorithms and Lexicon-Based approaches. The most well-known Lexicon approaches are TextBlob, Vader and Afinn. It is worth noticing which of the three methods works best with such data and if our results confirm it. In the next step, prominent Machine Learning algorithms were used and compared according to their performance. So again, we would like to see which model has performed better in previous similar tasks.

Robert Marcec and Robert Likic [6] conducted a twitter sentiment analysis to identify differences between the Astrazeneca, Pfizer and Moderna vaccines. Tweets were collected from December 1, 2020 to March 31, 2021. Data were 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 Astrazeneca vaccine which has increased greatly this time period. The side effects of this vaccine that appeared at that time caused the negative reactions of the people. F. M. Javed Mehedi Shamrat and Sovon Chakraborty [10] 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 paper [11]. 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, the most tweets had a neutral sentiment. However, in 2022 the amount 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.

Sentiment analysis on tweets written in English language is more common. However, E. Kapoteli, Paraskevas Koukaras and C. Tjortjis [12] identified the emotions of both English and Greek tweets during the period between May 19, 2021 and November 19,2021. In the first step, in order to represent the text as numerical vectors applied three different techniques: (i) Term Frequency - Inverse Document Frequency (TF-IDF), (ii) Word2Vec and (iii) BERT which is a new language representation model introduced by [13] in 2018. The BERT model performed better, so the analysis was then based on that. Studying English tweets, they noticed that most of the emotions were neutral in the first months of their research, but as time went on, the negative emotions increased. Negative emotions also prevailed in Greek tweets.

It is worth considering the paper [14]. Asderis 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 antivaxxers. TextBlob has classified most tweets regardless of the hashtag, as neutral and is followed by positive emotions. Tweets related to antivaxxers show more negative sentiments compared to other hashtags on a daily basis. Taking into account the VADER approach, interesting conclusions emerge. Antivaxxers have a high rate of negative emotions. Only Astrazeneca seems to have the neutral emotion as dominant. The rest of the vaccines seem to be more likeable as they have more positive emotions than negative or neutral ones. With all this in mind, it confirms that TextBlob and VADER have 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.

Another notable work was implemented in paper [15] based on twitter data from the first Covid-19 vaccination announcement over a one-month period. 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 labelled 1,00% of the dataset with labels: favor, against and neutral. Most tweets were annotated as neutral. Machine Learning 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 classified as neutral. Since this study was conducted at the beginning of vaccines 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, Abd Rahim, N., & Rafie, S. M. [16] 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 labelled as negative. Only 20% annotated as positive. Next, Support Vector Machine models with Rbf 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

3.1 Data Collection

Tweets related to Covid-19 Vaccination collected **from 13 April 2022 to 16 May 2022**. They are imported from the Twitter using the **API keys** provided by the **Twitter Developer Account** we have created. More specifically, elevated access is needed for the retrieval of many up-to-date tweets. Python and the library Tweepy were determinant in order to retrieve them. Data in the form of raw tweets is extracted by specifying **keywords** to search for in the tweets. Defining a function named get_Tweets(keyword) and calling it thirteen times with the keywords: "#vaccine", "#GetVaccinated", "#CovidVaccine", "#Pfizer", "#antivaxx", "#Antivax", "#antivaxxers", "#Covidiots", "#VaccineSideEffects", "#Moderna", "#AstraZeneca", "#COVID19Vaccination" and "#VaccineDeath" a dataset of **54850** tweets was created with aim to implement this study. It contains the posted text along with the following 17 fields:

Table 1: 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
word counts	Number of words that tweets have

word_counts_cleaned	Number of words that tweets have after preprocessing

3.2 Data Storing

After collecting these data and merged them into a panda's data frame, we saved it in a **pickle file** just to be sure that we don't lose our data. Then, we connected with **MySQL** and created a database to store our twitter data. A cursor object is created allowing us to execute SQL statements. We used **df.to_sql()** command to insert the data into the correct table in our database using the **sqlalchemy library**. The columns "hashtags" and "urls" were converted from lists to strings, so as to be inserted correctly in the database table. The next step was to export the twitter information as **Excel file**.

3.3 Data Preprocessing

Tweets were highly unstructured and contained redundant information. So, mandatory was the preprocessing of the dataset to carry on with further analysis. By taking the following multiple steps, these issues were overcome.

1) Removing **duplicate tweets** in case the same tweet was posted multiple times accidentally or we could define the tweet_id as primary key to prevent duplicates.
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 by an apostrophe. It is a useful preprocessing step as the words play an important role in sentiment analysis [17].
4) Applying **lowercase** which means convert all letters to lower case.
5) Removing URLS, user's mentions, hashtags, punctuations, digits & emojis using **regular expressions**, since those terms don't really provide meaningful context for discovering inherent topics from the tweet.

6) Removing **stop words**, which are basically a set of commonly used words that don't contribute much to the machine learning model. This allows us to focus on the important words instead.
7) A necessary Natural Language Processing (NLP) technique that we used is **Tokenization**. It is the process of breaking down a tweet into words [18].
8) Another significant Natural Language Processing (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, it is really important that the input word must be passed in lower case to the WordNetLemmatizer algorithm to achieve accuracy [19].

3.4 Lexicon-Based Approaches

The purpose of this study was to understand people's feelings about vaccination and to determine if there is a preference for one of the existing vaccines. Three **Lexicon-Based methods** (**TextBlob**, **Vader** and **Afinn**) were used to find the polarity of the text (positive text, negative text, neutral text). 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 three methods was applied.

3.4.1 TextBlob

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). 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 polarity score equals to zero, the tweet is considered as neutral [14]. 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.

3.4.2 Vader (Valence Aware Dictionary for Sentiment Reasoning)

Vader is another popular model for sentiment extraction in the domain of social media that takes into account the intensity of the sentiment. 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 considering as negative if compound score **<= -0.05**, neutral if compound score in range **(-0.05,0.05)** and positive if compound score **>=0.05** [20].

3.4.3 AFINN

Afinn is based on the **Affective Norms for English Words** lexicon (ANEW) proposed by Bradley and Lang [21]. It is a lexicon of English terms labeled manually by Finn Arup Nielsen [22] with a score in **range [-5,5]**. Positive score specifies a positive emotion, zero score is a sign of neutral sentiment and negative score indicates a negative emotion.

3.5 Supervised Machine Learning Approaches

3.5.1 Term Frequency-Inverted Document Frequency Features (TF-IDF)

For the vectorization of the tweets TF-IDF word embedding approach was used. TF-IDF is a statistical measure that assesses **the relevance of a word to a document in a set of documents**. This is accomplished by multiplying two metrics: the number of times a word appears in a document (**TF**) and the word's inverse document frequency over a collection of documents (**IDF**) which indicates how common or rare a term is across the entire collection of documents [23].

3.5.2 Data Splitting

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

3.5.3 Models Training – Classifier Selection – Models Evaluation

For the **training of the Machine Learning models**, the labeled tweets dataset, that was created from the three different Lexicon-Based approaches, was used. The TF-IDF features were extracted from the labeled dataset and the **performance** of the following prominent **classifiers** was reviewed in our research. More specifically, the experiments were performed using TextBlob, AFINN and VADER sentiments as target classes with the selected Machine Learning models to determine the best method in terms of accuracy, precision, recall and F1 score.

3.5.3.1 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 [24].

3.5.3.2 Logistic Regression (LR)

The supervised Machine Learning 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 [25].

3.5.3.3 Decision Tree (DT)

Decision Tree is a **rule-based supervised** machine learning algorithm used in both regressions, as well as classification problems [26]. 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 [27].

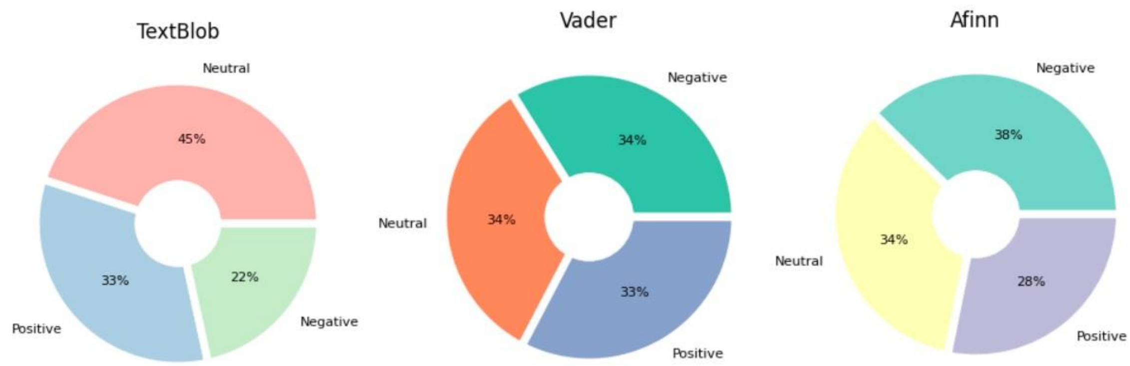
3.5.3.4 XGBoost Classifier (XGB)

Extreme Gradient Boosting (XGBoost) is **a distributed Gradient Boosted Decision Tree** (GBDT) machine learning 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** [28].

4. RESULTS

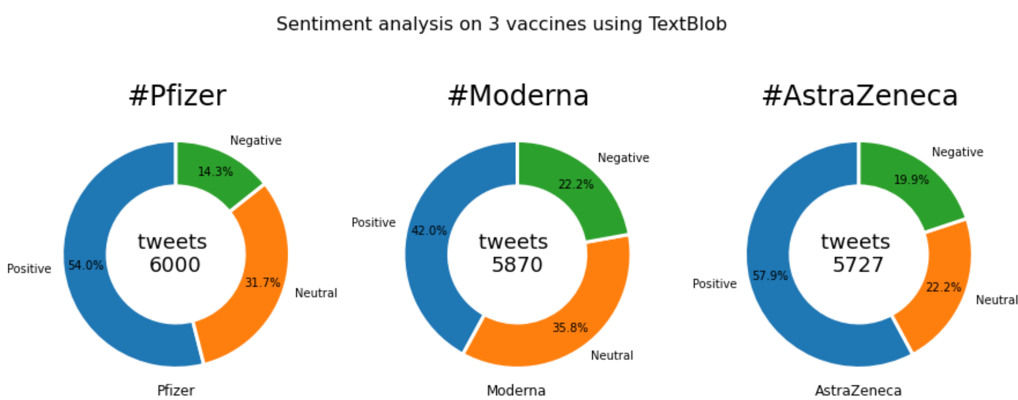
To address feelings about **ongoing vaccinations** around the world, **three NLP lexicon-based approaches** were developed, including TextBlob, Vader and Afinn, along with **five Machine Learning models**, including RF, LR, DT and XGB. The following discussions aim to present and analyze the **performance** of these methods for sentiment analysis.

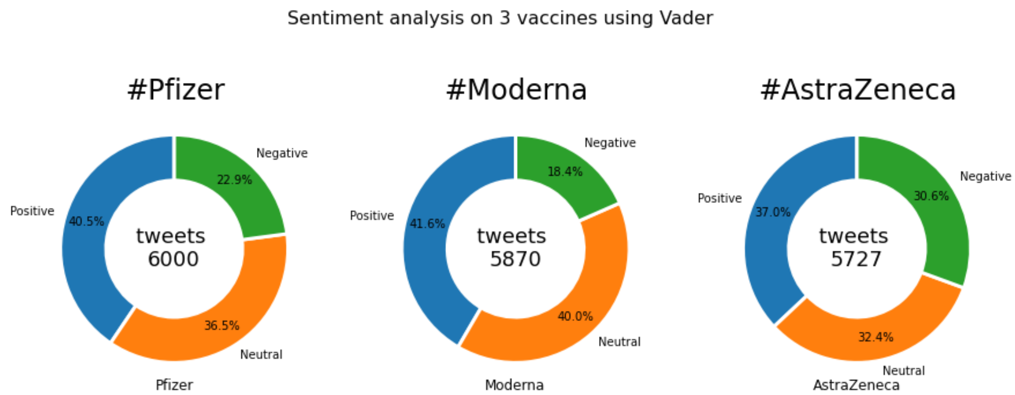
Figure 1: Pie Charts representing sentiment polarity scores using the **three different classifiers**

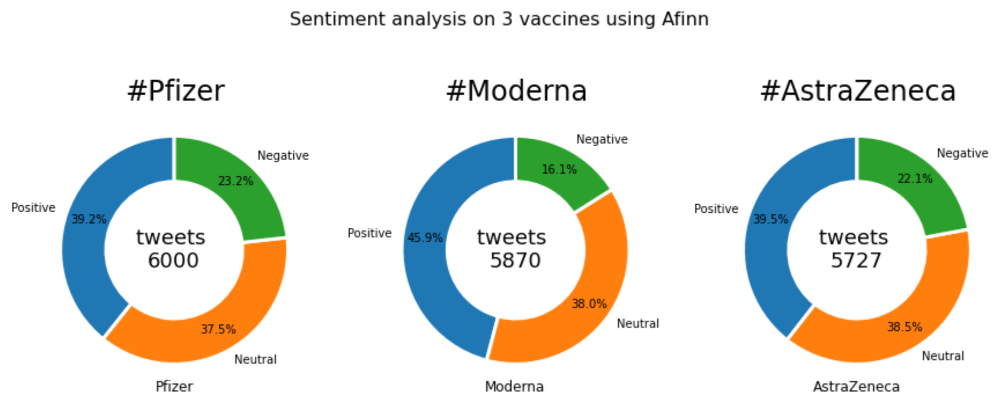


In Figure 1 the sentiment analysis results are provided using the TextBlob, Vader and Afinn methods in our collected dataset. The results for the **TextBlob** approach showed that the **neutral** polarity in our keywords was significantly high (45%). **VADER**-assigned **negative** polarity scores were higher as compared to the TextBlob. TextBlob gave 22% negative tweets, while VADER set a negative polarity score at 34% indicating 12% higher negative tweets than TextBlob. **AFINN** assigned a more **negative** polarity score (38%) compared to Vader.

Figure 2: Pie Charts depicting Sentiment Polarity distribution for **each vaccine using the three different Lexicon-Based approaches**







The above figure 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 **AstraZeneca** more positive polarity scores compared to **Vader,** which gave **Moderna** the highest positive scores. **Afinn** also believed that **Moderna** is the most reliable and famous vaccine.

Table 2: Performance results for the **four ML models** using the **three NLP approaches**

NLP Lexicon-Based approaches	Models	Accuracy (%)	F1-score (%)
TextBlob	RF	96.00	95.98
	LR	91.76	91.73
	DT	94.96	94.95
	XGB	93.70	93.68
Vader	RF	91.77	91.78
	LR	83.81	83.81
	DT	89.86	89.87
	XGB	87.60	87.64
Afinn	RF	92.24	92.27
	LR	85.42	85.51
	DT	90.50	90.51
	XGB	88.66	88.76

Table 2 presents the results for the Machine Learning models for accuracy and F1-score evaluation metrics. Regardless of the classifier we used, **Random Forest** (RF) received **the highest level of accuracy of 96%** calculated taking into account the number of correctly classified samples and the total number of samples. The output shows that we are able to successfully classify a tweet as positive, negative or neutral with 96% accuracy. The **F1-score** of our model was **95.98%** 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 [29]. In our model, both precision and recall were high.

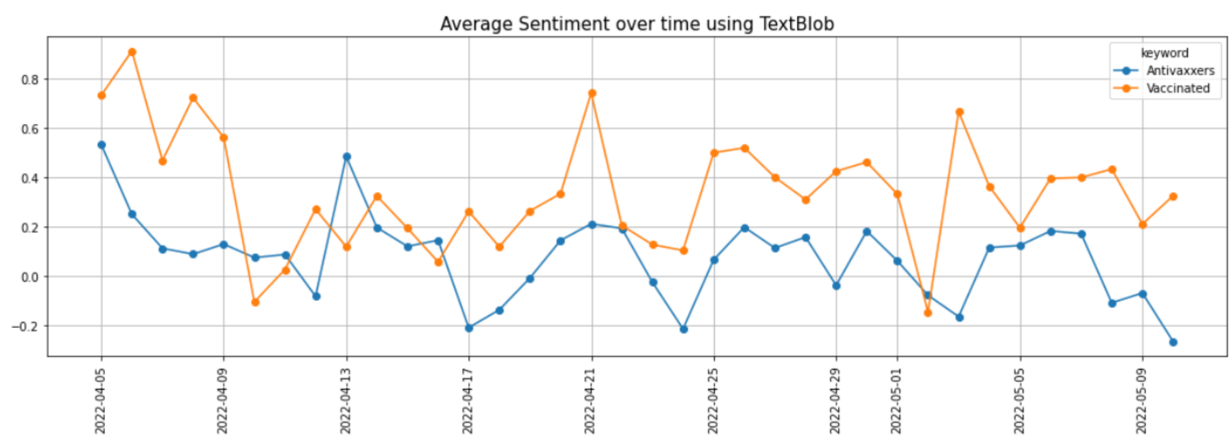
More specifically, using the **TextBlob** approach, both RF and Decision Tree (DT) produced high accuracy and F1-scores. In the **Vader** method, the RF performed better than the other models. In addition, the performance of the models **decreased significantly** when the dataset was changed from TextBlob to annotated Vader. For example, DT accuracy dropped to 89.86% from

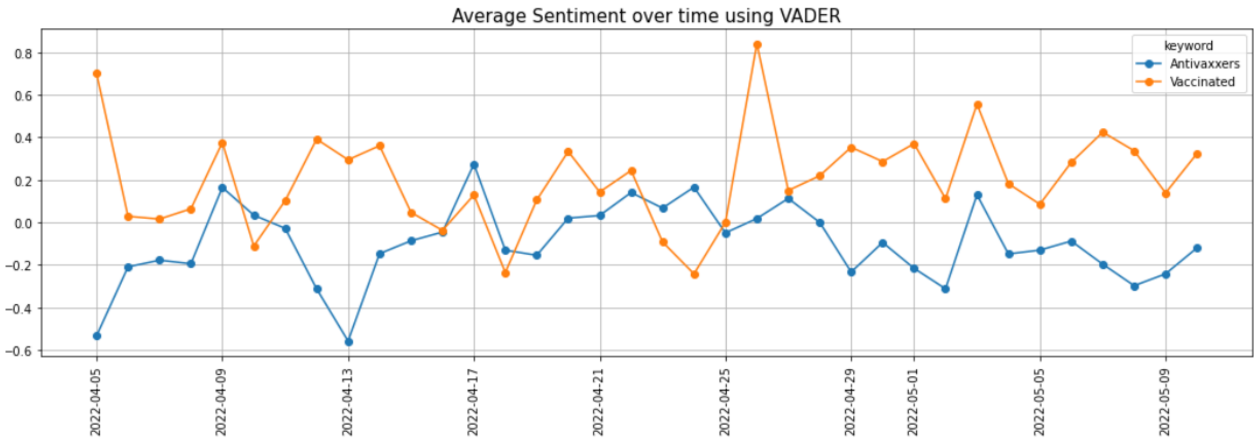
94.96%, while Logistic Regression (LR) remarkably reduced to 83.81% from 91.76% when trained with a Vader annotated dataset. The F1-scores of the models had similar reductions. When the dataset changed from Vader to **Afinn,** all models were **slightly improved**, such as RF accuracy increased from 91.77% to 92.24%.

Experimental results revealed that the models perform better when used with TextBlob annotated data compared to Vader and Afinn. Previous studies [30, 31] show that models perform better when trained on TextBlob labeled data and this study confirms the same.

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

Figure 3: Timeseries of sentiments for **Vaccinated** and **Antivaxxers** using three different Lexicon-Based Approaches





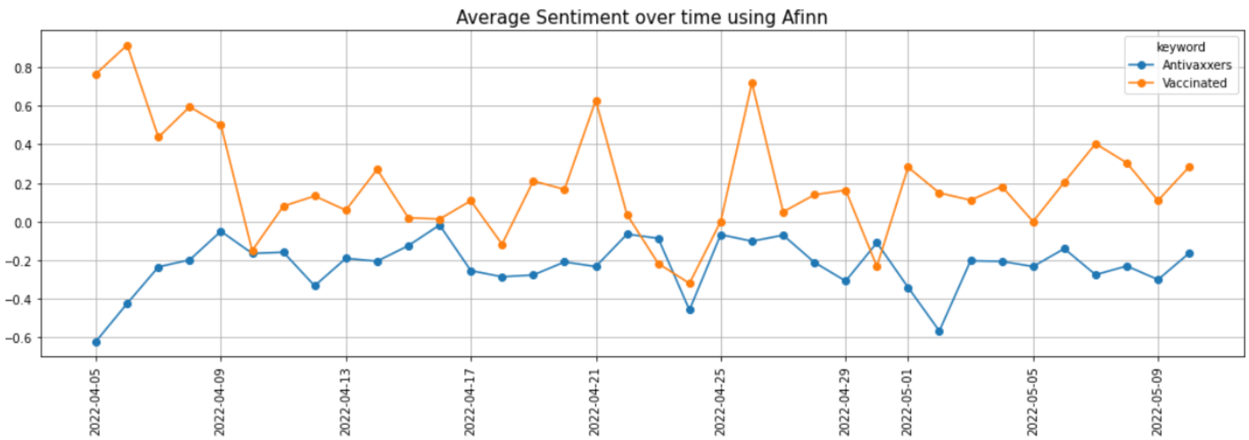
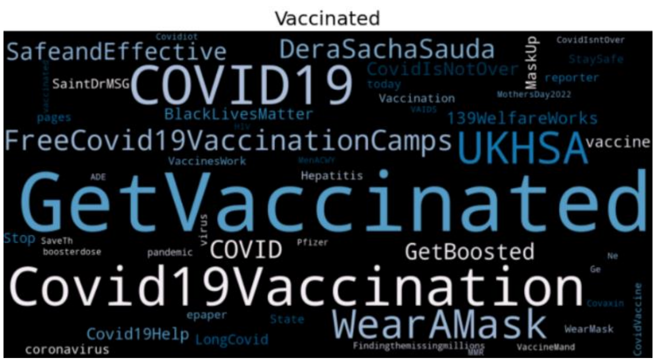


Figure 3 illustrates how the average sentiment score **changes over time** for vaccinated and antivaxxers people using the TextBlob, Vader and Afinn methods. At first glance, we noticed that in all approaches **vaccinated sentiments** are always **more positive than** those of **antivaxxers**. There are great fluctuations in people's feelings in daily tweets within a month. First of all, with **TextBlob** there is one peak in the sentiment score for vaccinated people on 6^{th} of April and only two days with negative tweets on 10^{th} of April and 2^{nd} of May round to -0.1% (a little below zero). The rest of the days the tweets were positive for vaccinated. For antivaxxers, the worst day was on 10^{th} of May with -0.3%, while similarly on other days most tweets had a negative score. In the **Vader** plot, the highest sentiment average was noted on 26^{th} of April reaching almost 1% representing a positive day for vaccinated. Before this day, the sentiment was maintained at high levels, with small drops on 18^{th} and 24^{th} of April. One of the most intense decreases for antivaxxers was noted on 5^{th} and 13^{th} of April, where the sentiment average reached approximately -0.6%. On April 6^{th}, the feelings of the vaccinated became the most positive with the **Afinn** method, something that is also confirmed by TextBlob. Furthermore, there is an increase in the emotions of vaccinated on April 26th with 0.7% lower than the 1% proposed by the Vader approach. For antivaxxers there are several negative sentiments and especially in the tweets that have been created on April 5th similar to Vader and May 2nd. Finally, these graphs clearly show the positive sentiments of the vaccinated, as well as the negative sentiments of the antivaxxers, 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 4: Word clouds with the hashtags that are used by **Antivaxxers** and **Vaccinated**



Following the previous analysis, we created word clouds to find out **which hashtags are most commonly used by vaccinated and antivaxxers**, in addition to the ones we used to collect tweets. The first word cloud in Figure 4 above refers to the vaccinated. The hashtags **«WearAMask»** and **«CovidIsNotOver»** suggest the fear that although Covid-19 is still shrinking, it has not disappeared and the masks must remain in our lives. «SafeandEffective» is a sign that vaccinated people really support vaccines, believe in their productivity and encourage other people to get vaccinated. Also the hashtag «GetBoosted» is another way of expressing someone who has been vaccinated. The second word cloud concerns antivaxxers. The use of the hashtag «Vaccineinjuries» confirms that these people believe that vaccines can cause several side effects. Some antivaxxers 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.

5. CONCLUSIONS

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 are posting their viewpoints regarding the coronavirus. Despite the fact that vaccines are considered as the 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. The aim of this research work was to identify the emotional state of people about coronavirus. A total of 54850 tweets have been retrieved and after data preprocessing, sentiment analysis was performed using three well-known Lexicon-Based techniques. Four Machine Learning models namely Random Forest, Logistic Regression, Decision Tree and XGBoost Classifier were implemented. The classification of emotions using TextBlob contributes to the better performance of the ML models. Best results were achieved with TextBlob and Random Forest. Top accuracy was 96% and top F1-score was 95.98%. 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 in Twitter. It is confirmed that antivaxxers are expressing their displeasure with the COVID-19 vaccine, as the majority of their sentiments are classified as negative regardless of the date.

6. DISCUSSION

By taking into consideration the limitations of the present study, it should be referred that twitter data contains a lot of noise. Although they have handled 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 language and their users were not representative of the general English-speaking public. As a result, their tweets simply reflected the views and feelings of Internet users about vaccination, which could serve as a limitation on the study.

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 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] 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).

[11] 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.

[12] E. Kapoteli, Paraskevas Koukaras and C. Tjortjis, "Social Media Sentiment Analysis Related to COVID-19 Vaccines: Case studies in English and Greek language".

[13] 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

[14] Asderis Georgios-Alexandros, "Sentiment Analysis on Twitter Data: a Detailed Comparison of TextBlob and VADER", 2022

[15] 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.

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

[17] Ahmed, T., Bosu, A., Iqbal, A., & Rahimi, S. (2017, October). SentiCR: a customized sentiment analysis tool for code review interactions. In 2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE) (pp. 106-111). IEEE.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

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

[27] 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.

[28] 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.

[29] Kolchyna, O., Souza, T. T., Treleaven, P., & Aste, T. (2015). Twitter sentiment analysis: Lexicon method, machine learning method and their combination. arXiv preprint arXiv:1507.00955.

[30] 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.

[31] 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.