Monitoring People's Emotions and Symptoms after COVID-19 vaccine

Najwa N. Alshahrani[†] , Sara N. Abduljaleel[†] , Ghidaa A. Alnefaiy[†] and Hanan S. Alshanbari[†]

[†]College of Computer Science and Information System, Umm Al-Qura University, Makkah, Saudi Arabia

Summary
Today, social media has become a vital tool. The world communicates and reaches the news and each other's opinions through social media accounts. Recently, considerable research has been done on analyzing social media due to its rich data content. At the same time, since the beginning of the COVID-19 pandemic, which has afflicted so many around the world, the search for a vaccine has been intense. There have been many studies analyzing people's feelings during a crisis. This study aims to understand people's opinions about available Coronavirus vaccines through a learning model that was developed for this purpose. The dataset was collected using Twitter's streaming Application Programming Interface (API) , then combined with another dataset that had already been collected. The final dataset was cleaned, then analyzed using Python. Polarity and subjectivity functions were used to obtain the results. The results showed that most people had positive opinions toward vaccines in general and toward the Pfizer one. Our study should help governments and decision-makers dispel people's fears and discover new symptoms linked to those listed by the World Health Organization.

Keywords:
COVID-19; Vaccine; Sentiment analysis; social media; Twitter; Tweets

1. Introduction

Social media is a catch-all term for various applications built for the web and implemented through mobile devices to support communication and content sharing. Since social media use is growing, the number of its users has become even more extensive. Companies have begun designing new methods and tools to deal with this increased data load (Batrinca and Treleaven,2015).

For example, almost every large company now has an account on Twitter to get customer feedback about their services or products. In this context, sentiment analysis, also known as opinion mining, is used for classifying specific words into positive or negative ones.

COVID-19 has been declared a pandemic due to the rapid spread of the virus around the world. The pandemic deeply affects all aspects of society, including mental and physical health. Developed countries and large medical companies rushed to start research on a vaccine. Because of the short time available to create it and do clinical trials, there was great anxiety among the general public about whether to take the vaccine.

The rest of this paper is organized as follows: Section 2 discusses related work, Section 3 outlines the methodology, Section 4 contains the results, and Section 5 presents the conclusions.

Manuscript received June 5, 2023
Manuscript revised June 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.6.25

2. Literature Review

Sentiment classification was carried out using three approaches: supervised learning (corpusbased), using the labeled corpus to train the sentiment model(Seyed-Ali and Andreas,2012; Ibrahim et al.,2019; Haq et al.,2018); unsupervised learning (lexiconbased), based on sentiment lexicons (Alex et al.,2017); and a hybrid approach combining the first two methods (Walla et al.,2017). Previously, the widely used

features are n-gram, part-of-of-speech (POS), and syntactic (Oumaima et al.,2020). Many N-gram features generally are superior to character-based features for machine learning classifiers if applied to Arabic text. In classifying emotions, we must understand how to handle lexical negation (Claudia et al.,2019).

In (Mahmud et al.,2017), sentiment analysis is performed to predict the success of the movies by considering the opinions of people and their sentiments on Twitter, which are available in the form of tweets. In this paper, the research has been done on unigram and bigram models. There is no use for a trigram model or an n-gram model, which is a combination of all the models, to improve the accuracy. Moreover, there is no use for hybrid features, such as emoticons, synonyms, and acronyms, at the time of feature extraction to increase the accuracy of the results. The accuracy of this work is 86%.

(Pak and Paroubek,2010) gave a thorough description of Twitter as a data source for performing sentiment analysis along with opinion mining. They distinguished the tweets into positive, negative, and neutral classes, where the tweets were only in English. Each tweet was manually labeled by three different people. Furthermore, the researchers noticed that many of the tweets contained emoticons (i.e., icons expressing the emotions of users toward an entity or a service.), such as ':)', ':(', '=)', '=(', ';)' . Three classification algorithms, namely, Naïve Bayes (NB), Support Vector Machines (SVM), and Conditional Random Fields (CRF) were used, along with features like unigrams, bi-grams, n-grams, etc., to classify the tweets.

Recently, (Giachanou and Crestani , 2006) have conducted a thorough survey of Twitter sentiment analysis (TSA) methods. They identified four different types of (textual) features that have been used so far: semantic, syntactic, stylistic, and Twitter-specific. Semantic features include opinion words, sentiment words, negation, etc., and can be extracted in a manual or semi-automatic manner from opinion and sentiment lexicons. Many researchers have used lexicons that have been developed for other domains, such as SentiWordNet [5Similarly, syntactic features include unigrams, bi-grams, n-grams, term frequencies, and POS. Together with semantic features, they are the most widely used ones.

Whereas some researchers preferred binary weighting scores based on presence/absence, others considered term frequencies. Stylistic features come from non-standard writing styles, such as emoticons, use of slang, and punctuation marks. Lastly, Twitter-specific features include hash-tags, re-tweets, replies, and user names. Many researchers, such as (Hong et al.,2011), have considered the presence/absence or the frequency of these features.

Natural Language Processing (NLP) techniques have also been used in content analysis. One of the simplest techniques determines the presence of a sentiment lexicon (a word expressing a positive or negative sentiment) in an entity, such as tweets. (Asur and Huberman,2010) used tweets to forecast the revenue for movies. They used three million tweets and constructed a linear regression model. Similarly, (Zhou et al.,2013) developed a Tweet sentiment analysis model (TSAM) that could successfully determine society's interest, as well as people's opinions, regarding a social event (in their case, the Australian federal elections). (Sriram et al.,2010) classified tweets using a small set of domain-specific features extracted from the authors' profiles along with the text.

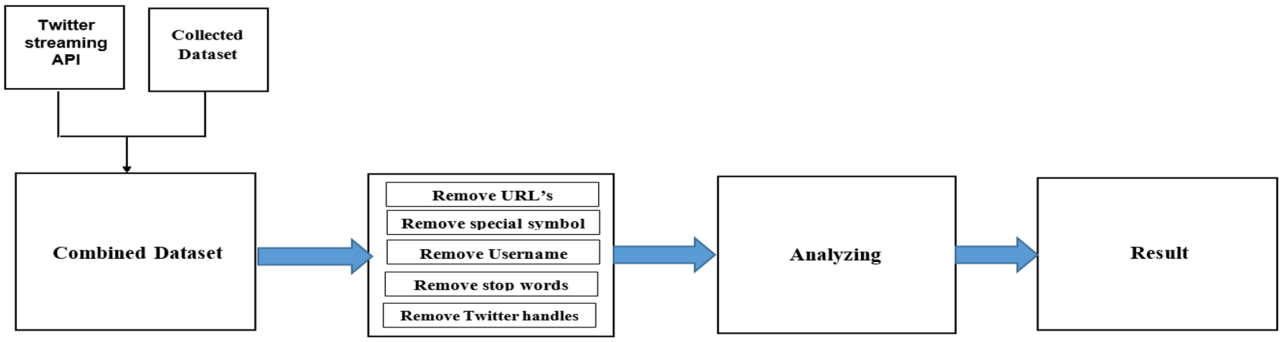

Fig. 1 The structure of our work

3. Methodology

This paper seeks to understand what people think about COVID-19 vaccines. It thus focused on collecting tweets that contained people's opinions. In our project, we concentrated on cleaning and analyzing the tweets. All the tweets were in English. The conceptual architecture of our work is presented in Figure 1.

A list of COVID-19 vaccine symptom keywords was extracted and prepared to be used in a symptom mentions classification. The development of the dataset had four steps: 1) data collection, 2) data preprocessing, 3) opinions annotation, and 4) symptom tweet detection.

A. Data collection

Twitter's streaming API was used to collect the data. Because Twitter's API gave us a limited dataset, another dataset that had already been collected (https://github.com/saranaji66/-COVID-19-vaccine) was combined with it to increase the number of tweets and obtain better results. The total number of tweets was 76,943 and covered the period from Aug 28, 2020, to Feb 17, 2021, which corresponds to when the vaccines were released and started being distributed around the world.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

B. Data preprocessing

After the data from Twitter was collected and combined with the other data, preprocessing techniques were carried out on the final dataset. This involved applying several steps to the entire dataset to reduce the amount of trivial noise and thus clean the data.

The following preprocessing techniques were applied. The first cleaning step was to remove noise; hence, mentions, URL links, emojis, and punctuation were eliminated. Hashtags were also removed, while maintaining their content in tweets by using multiple Python libraries, such as Regular expression(re), which

was used to delete the @ string and the username, and the NLTK library, which contains English stop words. Additional techniques are discussed in the following points.

- Removal of URLs: Twitter data consists of different types of information. If a user posted a link that was of no use for sentiment analysis, the URL was removed from the tweet.
- Removal of special symbols and numbers: There are various types of symbols inserted by users, such as commas (,), full stops (.), etc., which do not express sentiment. Therefore, special symbols were removed from the tweet.
- Remove emoticons: Table. 1 shows the various emoticons that were deleted. Emoticons have become a way for users to express their views and feelings. Emotions play an important role in sentiment analysis.
- Removal of username: Every Twitter user has a unique username; therefore, anything written by a user can be indicated by writing their username proceeded by @. This type of writing is denoted as proper nouns; for example, @username. For the analysis to be effective, this kind of information also has to be removed.
- Remove stop words like "the," "a," "an," "in," etc.
- Remove Twitter handles and #hashtags.

C. Opinion annotation and symptom tweet detection

After all the tweets were cleaned, they were analyzed. Two functions were created. The first one is Subjectivity, which is used to take a tweet and find out if the opinion is a personal text or is supportive, in which case the result is (0), meaning that the opinion is real, while if the result is (+ 1), it is a frequent opinion.

TABLE 1 EXAMPLE OF A TWEET BEFORE AND AFTER PREPROCESSING

Before	👋🏻@JustinTrudeau @cafreeland — no one will be safe from #COVID19 until everyone is safe. Will you invest just 1% of what Canada's spent so far on the response at home to fund global humanitarian efforts and help provide tests, treatments and vaccines for all? #GlobalGoalUnite
After	'safe covid everyone safe will invest canadaCanada spent response home fund global humanitarian efforts help provide tests treatments vaccines globalgoalunite '

The second function is Polarity, which measures how positive or negative a tweet is. If the result is (-1), the tweet is negative; if it is (+1), the tweet is positive. We then created two new columns, one for Subjectivity and one for Polarity, and presented results.

Next, we ran the analysis. If Polarity's output was (0), it was neutral. If it was (-1), it was negative. If it was (+1), it was positive. We then created a new column named Analysis and showed the results in it. Figure 2 shows Polarity and Subjectivity for all the tweets.

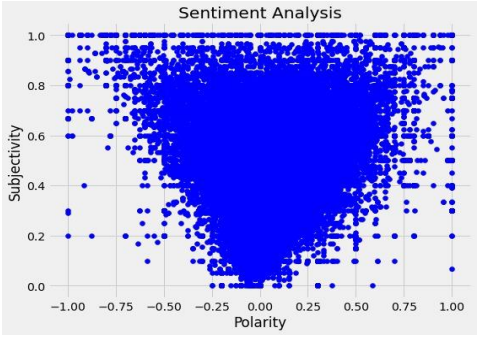

Fig. 2 Polarity and Subjectivity

After preprocessing and analyzing the dataset, we obtained the following results. As shown in Figure 3, the most most popular words people used in their tweets were vaccine, Covid, and people. The majority of users had a positive opinion of the vaccine in general, as shown in Figure 4. The results were then analyzed to understand people's opinions toward each vaccine separately.

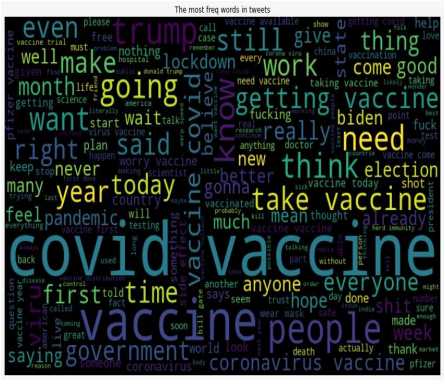

Fig. 3 The most popular words used

Manuscript received June 5, 2023
Manuscript revised June 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.6.25

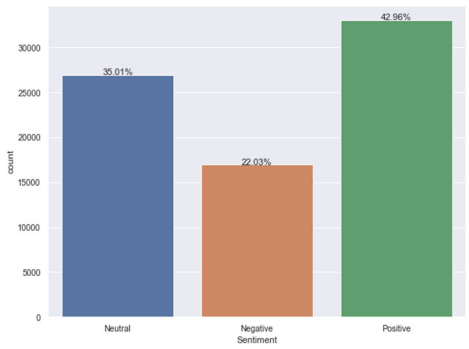

Fig. 4 The Ssentiment analysis towardsof the vaccine

Most people had a positive view of all the vaccines, as shown in Figure 5. The Pfizer vaccine had the highest number of tweets; the Moderna one was second and the Oxford AstraZeneca one third. The difference between the number of positive and neutral tweets for Pfizer was greater than the difference for the other two vaccines.

To find the symptoms that most people mentioned in their tweets about COVID-19 vaccines, the tweets were analyzed using the symptoms' frequency. "Pain" was the most common word, followed by "fever" and "headache," as shown in Figure 6.

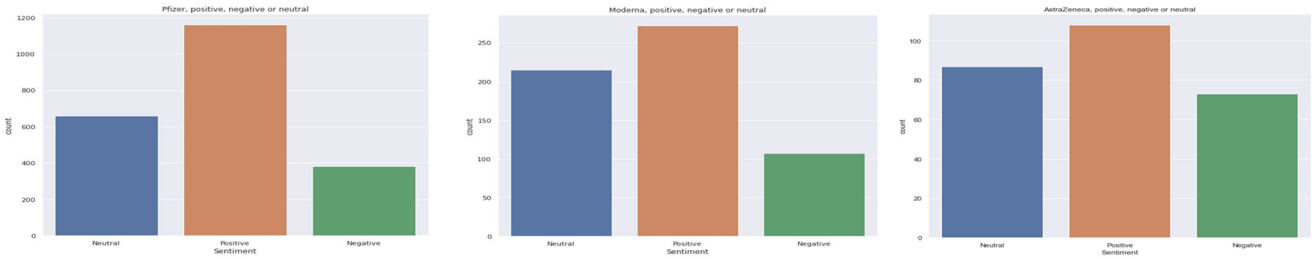

Fig. 5 Sentiment analysis of the most popular vaccine

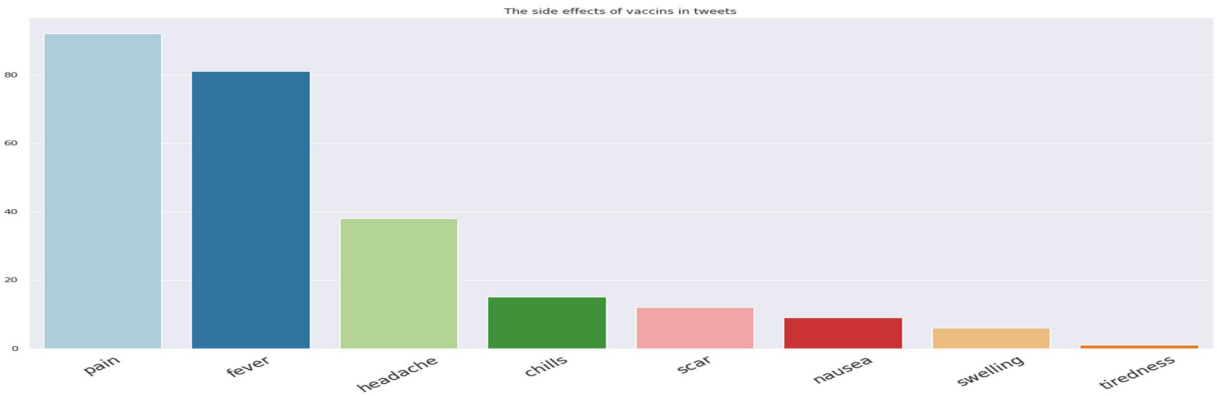

Fig. 6 The symptoms most frequently mntioned.

4. Conclusion

Social media has many advantages. It is used to communicate with the world; it transmits the news and lets people share their opinions. Recently, significant research has been done on social media because it contains considerable amounts of data. With the spread of the COVID-19 virus, researchers analyzed people's opinions about the available vaccines. In our study, the main foci were people's views of the Covid-19 vaccine and what symptoms were mentioned in the tweets. Therefore, the data was extracted using Twitter's streaming API and combined with a previously collected dataset. The final dataset was cleaned and analyzed using Python's programming language. The results were slightly expected. In most countries, the most positive tweets regarded the Pfizer vaccine. Moreover, the symptoms mentioned most often in the tweets were "pain," "fever," and "headache". We look forward to collecting more data on the reasons for these results.

References

[1] Bogdan Batrinca, Philip C. Treleaven. 2015. Social media analytics: A survey of techniques, tools and platforms, AI Soc. 30 (1) 89–116.

[2] Seyed-Ali Bahrainian, Andreas Dengel, Sentiment analysis using sentiment features, in: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) 3, IEEE, 2013,pp. 26–29.

[3] Ibrahim Abu Farha, Walid Magdy, Mazajak .2019. An online arabic sentiment analyser, in:Proceedings of the Fourth Arabic Natural Language Processing Workshop, pp. 192–198.

[4] Haq Nawaz, Tahir Ali, Ali Al-laith, Imran Ahmad, Sridevi Tharanidharan,Shamim Kamal AbdulNazar. 2018.Sentimental analysis of social media to find out customer opinion, in: InternationalConference on Intelligent.

[5] Alex Mircoli, A. Cucchiarelli, C. Diamantini, D. Potena. , 2017. Automatic emotional text annotation using facial expression analysis, in: CAiSE-Forum-DC,pp. 188–196.

[6] Walaa Medhat, Ahmed Hassan, Hoda Korashy.2014. Sentiment analysis algorithms and

[7] applications: A survey, Ain Shams Eng. J. 5 (4) 1093–1113.

[8] Oumaima Oueslati, Erik Cambria, Moez Ben HajHmida, Habib Ounelli. 2020. A review of sentiment analysis research in arabic language, Future Gener.Comput. Syst.

[9] Claudia Diamantini, Alex Mircoli, Domenico Potena, Emanuele Storti, Social information discovery enhanced by sentiment analysis techniques, Future Gener. Comput. Syst. 95 (2019) 816–828.

[10] Quazi Ishtiaque Mahmud, Asif Mohaimen, Md Saiful Islam, Marium-E-Jannat, "A Support Vector Machine mixed with statistical reasoning approach to predict movie success by analyzing public sentiments", IEEE-2017.

[11] Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA), May 2010.

[12] Giachanou and F. Crestani, "Like it or not: A survey of Twitter sentiment analysis methods," ACM Computing Surveys (CSUR), vol. 49,no. 2, pp. 28:1–28:41, Jun. 2016.

[13] Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06), 2006, pp. 417–422.

[14] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in Proceedings of the 20th International Conference Companion on World Wide Web, ser. WWW '11. New York, NY, USA: ACM,2011, pp. 57–58.

[15] S. Asur and B. A. Huberman, "Predicting the future with social media," in Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, ser.WI-IAT '10. Washington, DC, USA: IEEE Computer Society, 2010,pp. 492–499.

[16] X. Zhou, X. Tao, J. Yong, and Z. Yang, "Sentiment analysis on tweets for social events," in Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on, June 2013, pp. 557–562.

[17] Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '10. New York, NY, USA: ACM, 2010, pp. 841–842.