Covid-19 and Vaccine Tweet Analysis

Eren Alp, Bedirhan Gergin, Yiğit Ahmet Eraslan, Mert Can Çakmak, and Reda Alhajj

Abstract Social networks are the most effective instruments for gathering information about people's opinions and perceptions on a variety of subjects and concerns. People spend hours a day on social media to express their ideas, viewpoints, and answers with others. In this chapter, Covid-19 and Vaccine tweets that are taken from two different time manners were analyzed. Python was used to perform experiments on a variety of tweets. After collecting and preprocessing the data, various visualization techniques were used to show the results for most occurred words and sentiment analysis for positivity and negativity of tweets.

Keywords COVID-19 · Vaccine · Tweets · Opinion · Sentiment analysis

1 Introduction

Data that people poured into the internet like reactions and comments on the topics have the potential to reveal valuable insights on human emotions. Thus, the analysis of people's ideas and comments can play a crucial role to understand people's behavior and response in various ways. With the increasing number of microblogs and social media, people have begun to express their opinions on a wide variety of topics on Twitter and other similar platforms. As they are growing and spreading rapidly these tools became more useful to understand and model various events.

In this chapter, a dataset formed of collected tweets from Twitter was used. Twitter contains a large number of short messages created by the users of this microblogging platform. The contents of the messages vary from personal thoughts to public statements.

E. Alp · B. Gergin · Y. A. Eraslan · M. C. Çakmak · R. Alhajj (✉)
Department of Computer Science, University of Calgary, Calgary, AB, Canada

Department of Computer Engineering, Istanbul Medipol University, Istanbul, Turkey

Department of Heath Informatics, University of Southern Denmark, Odense, Denmark
e-mail: yaeraslan@st.medipol.edu.tr; bgergin@albany.edu; ralhajj@ucalgary.ca

T. Özyer (eds.), *Social Media Analysis for Event Detection*, Lecture Notes in Social Networks, https://doi.org/10.1007/978-3-031-08242-9_9

As a microblogging and social networking website, Twitter has become very popular and has grown rapidly. An increasing number of people are willing to post their opinions on Twitter, which is now considered a valuable online source for opinions. As a result, Twitter sentiment analysis provides a quick and efficient tool to evaluate public opinion for business marketing or social research. In this project sentiment analysis is done about Covid-19 and Vaccine tweets. First word occurrences and some visualizations were used and sentiment analysis was done.

Sentiment is an attitude, thought, or judgement prompted by feeling. Sentiment analysis is the process of determining and measuring the tone, attitude, opinion, and emotional state of responses. More precisely, it is the concept of deciding whether a specific conversation is positive, negative, or neutral. In our study just negativity and positivity of tweets were categorized.

The rest of this chapter is organized as follows. Section 2 covers the related work. Section 3 describes the methodology. Section 4 presents the results. Section 5 is the conclusions.

2 Literature Review

There are works about sentimental analysis, measuring the of the user, and topic modeling. In the Sentiment Analysis and Influence Tracking using Twitter paper [1], the authors mention that how Twitter data is used as a corpus for analysis by the application of sentiment analysis and a study of different algorithms and methods that help to track the influence and impact of a particular user/brand active on the social network. They used Twitter API, Twitter Streaming API, and Twitter Search API for data collection. For analysis preprocessing, techniques such as tokenization, normalization, and part of speech (POS) tagging are used. To determine the influence of the user PeopleRank and TwitterRank algorithms are used. Using these data collection APIs data can be collected from Twitter easily and ranking algorithms can help to calculate the influence of the user.

In the Detecting Real-World Influence Through Twitter paper [2] the authors investigated the issue of detecting the real-life influence of people based on their Twitter account. For the dataset CLEF RepLab, 2014 dataset is used. Social Network Analysis (SNA), Principal Component Analysis (PCA), bag of words, POS, linear classifiers which are Support Vector Machine (SVM) and libLinear, logistic regression, logic boost, multinominal Naïve Bayes are used for determining real-world influence. Since bots are not real influence in the real world this is helpful to detect someone's real influence value. In the Topic Modeling of Twitter Conversations paper [3], the authors presented a way to analyze large amounts of textual data from Twitter conversations efficiently and effectively. Specifically, it was explained how to capture the narratives that people share on Twitter about social events, reduce their complexity, and provide plausible explanations. For this Latent Dirichlet Allocation (LDA) method is used. By using this method, the topics from contexts can be extracted efficiently and effectively.

In the Extracting health-related causality from Twitter messages using natural language processing paper [4], the authors evaluated an approach to extracting causalities from tweets using natural language processing (NLP) techniques. Twitter Streaming API is used for dataset collection. To extract causality, lexicon syntactic relations and NLP pipeline operations which are lemmatizing, POS and dependency parsing are used. Since a good causality relationship sentence results in the good influence of a person when a reader reads that sentence so that this can be used for determining the influence of the user. However, because there are so many distinct methods to express cause and effect relationships in a phrase, it's difficult to keep track of them all.

In the Investigating the Relationship between Trust and Sentiment Agreement in Arab Twitter Users paper [5] the authors proposed a research methodology framework for investigating the relationship between trust and sentiment agreement on Twitter and explain the framework by applying it to a use case from Saudi Arabia. For this, the adaptation of the EigenTrust Algorithm which is the MarkovTrust algorithm is used. Also, surface analysis, deep analysis, and shallow analysis algorithms are used to determine the relationship between trust and sentiment agreement. Since the context and sentiment have been taken into consideration, determining the trust of the user will be more accurate.

In the Influence Analysis of Emotional Behavior and User Relationships Based on Twitter Data paper [6], the authors analyzed the influence of emotional behavior on user relationships based on Twitter data using two dictionaries of emotional words. For the collection of data random sampling, for calculation emotion score Keyword Matching, and the testing Brunner-Munzel test is used. By looking at emotional behaviors the influence of the user can be determined.

To sum up, the related work is summarized in Table 1.

3 Methodology

3.1 Data Collection

Implementing the sentiment algorithm and using it for further steps in the project, as well as a data collection technique. Collecting the data from a social media website was done through a scraper. A scraper is a type of software used to copy content from a website. In this project Snscrape was used for this purpose. Snscrape is a scraper for social networking services (SNS). It scrapes things like user profiles, hashtags, or searches and returns the discovered items, e.g., the relevant posts.

Shown in Fig. 1 is an example data collection that were taken from Twitter and transformed into csv file.

Table 1 Related works

Author	Period	Title	Method/remarks
Pramila M. Chawan	2012	Sentiment analysis and influence tracking using Twitter	"Paper makes a sentiment analysis on an entity level; mine people's idea on specific entities instead of whole tweets, scrutinize each tweet. Use three main features for scoring: followers, mentions and retweets, and lists, these are used for ratio to users" [1]
Peiyao Li, Weiliang Zhao, Jian Yang, and Jia Wu	2019	CoTrRank: trust evaluation of users and tweets	"Develop a trust ranking approach named CoTrRank. It mainly uses a coupled dual network. By evaluating the coupling affect in users and tweets. Values are derived with their original meanings in different trust spaces. The results of experiment show that the CoTrRank provides better evaluations of the trustworthiness of users and tweets when it is compared with other methods." [7]
Jean-Valère Cossu, Nicolas Dugué, Vincent Labatut	2015	Detecting real-world influence through Twitter	"Project analyze Twitter-based features with comparing and allowing to measure the offline effects and influence of users. Look for specific characteristics in twitter that can explain people known to be influential in their real-life." [2]
Son Doan, Elly W. Yang, Sameer S. Tilak, Peter W. Li, Daniel S. Zisook and Manabu Torii	2018	Extracting health-related causality from twitter messages using natural language processing	"Causality extraction is done by outputs that are dependency parser of Lexico-syntactic patterns. These techniques were used to help and improve the preciseness of information extraction. Paper shows that dependency parser with lexicon-syntactic relations yields high precision, which is an important feature for big data set mining." [4]
Kiichi Tago and Qun Jin	2018	Influence analysis of emotional behaviors and user relationships based on Twitter data	"Paper conduct three different experiments: calculate the average emotion score of a user, calculate the average emotion score using emotional tweets, and calculate the average emotion score using emotional tweets, with not including users of few emotional tweets. Then analyze by Brunner–Munzel test for the influence of emotional behaviors to user relationships. From the result it is understand that a positive user is more active than a negative user for building a user relationship in a specific situation." [6]

(continued)

Table 1 (continued)

Author	Period	Title	Method/remarks
Areeb Alowisheq and Sarah O Al-Humoud	2017	Investigating the relationship between trust and sentiment agreement in Arab Twitter users	"It conducts a research method for identifying the relationship between trust and sentiment for Arab Twitter users." [5]
Younggue Bae and Hongchul Lee	2012	Sentiment analysis of Twitter audiences: measuring the positive or negative influence of popular twitterers	"Paper identify between the positive and negative audiences of popular tweet users. Then, find that the audience are influenced by the sentiments used in the tweets by popular users. Thirdly, from these two findings it develops a positive-negative measurement for influence. Finally, by a Granger causality analysis, it is understood that sentiment change of the audience was related to the real- world sentiment landscape of popular users." [8]
Cano Basave, A. E.; Mazumdar, S. and Ciravegna, F.	2011	Social influence analysis in microblogging platforms a topic sensitive based approach	"Paper suggests the use of lexical profiles forming dominant users depending upon the retweet Twitter graph. Establishes a different version of the PageRank algorithm for examining user's relevance of a retweet connection." [9]
Juyup Sung, Seunghyeon Moon, and Jae-Gil Lee	2013	The influence in Twitter: are they really influenced?	"Paper tenders a development of PageRank algorithm, which is InterRank. It regards both relationship and topical similarity among users. It suggests that topical similarity act upon dominance." [10]
Eliana Sanandres, Camilo Madariaga, Raimundo Abello	2018	Topic modeling of Twitter conversations	"Paper suggests a technique for topic modeling on Twitter chatting which is Latent Dirichlet Allocation to decide the topics that are talked." [3]

(continued)

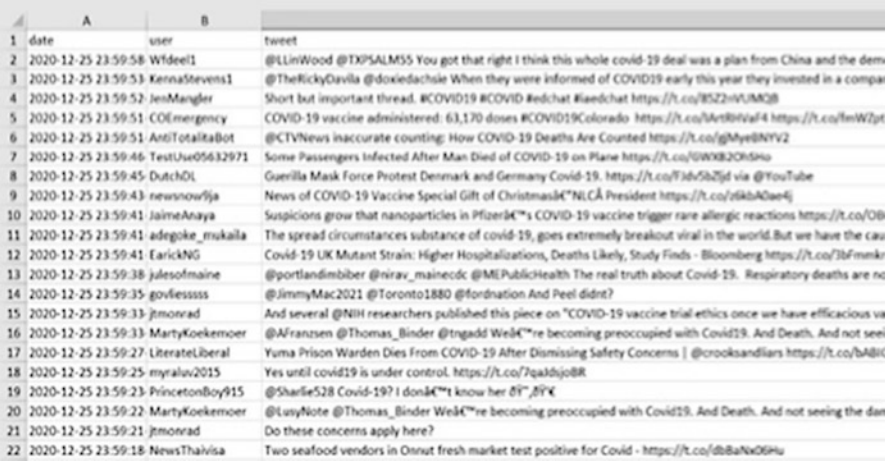
Table 1 (continued)

Author	Period	Title	Method/remarks
Liangjie Hong and Brian D. Davison	2010 Empirical study of topic modeling in Twitter "Paper suggests a solution for normal topic model algorithms that have been used on social media. It proposes that training a topic model with clustered text, it can be achieved better accuracy and preferable performance." [11]		
Christan Grant, Clint P. George, Chris Jenneisch, and Joseph N. Wilson	2011	Online topic modeling for real-time Twitter search	"Paper aims to get the attractive and topical social media entries from the dataset. It uses topic modeling algorithm for examination in the dataset." [12]
Ełłas Jonsson, Jake Stolee	2016	An evaluation of topic modelling techniques for Twitter	"Paper assesses of different topic modelling algorithms and analyze them by looking their performance on Twitter texts." [13]
Yefeng Ruana, Arjan Durresia, Lina Alfantoukha	2018	Using Twitter Trust Network for Stock Market Analysis	"Paper suggests that using the trust between users on microblogs, this can improve the mutual affinity with financial data in the stock market." [14]

3.2 Preprocessing

The preprocessing steps are:

1. Lower Tweets: Text are converted to lowercase.
2. Remove the URLs: Links starting with "http" or "https" or "www" are replaced by empty string.
3. Remove mentions, retweet and hashtags: Words starting with "", "#", "RT" are removed.



	A	B	
1	date	user	tweet
2	2020-12-25 23:59:58	Wfdeel1	@LLinWood @TXPSALMS5 You got that right I think this whole covid-19 deal was a plan from China and the dem
3	2020-12-25 23:59:53	KennaStevens1	@TheRickyDavila @doxiedachsie When they were informed of COVID19 early this year they invested in a compa
4	2020-12-25 23:59:52	JenMangler	Short but important thread. #COVID19 #COVID #edchat #iaedchat https://t.co/85Z2nVUMQB
5	2020-12-25 23:59:51	COEmergency	COVID-19 vaccine administered: 63,170 doses #COVID19Colorado https://t.co/lArtRHVaF4 https://t.co/fmWZpt
6	2020-12-25 23:59:51	AntiTotalitaBot	@CTVNews inaccurate counting: How COVID-19 Deaths Are Counted https://t.co/gjMyeBNYV2
7	2020-12-25 23:59:46	TestUse05632971	Some Passengers Infected After Man Died of COVID-19 on Plane https://t.co/GWXB2OhSHo
8	2020-12-25 23:59:45	DutchOL	Guerilla Mask Force Protest Denmark and Germany Covid-19. https://t.co/f3dv5bZljd via @YouTube
9	2020-12-25 23:59:43	newsnow9ja	News of COVID-19 Vaccine Special Gift of Christmasâ€"NLCÂ President https://t.co/z6kbA0ae4j
10	2020-12-25 23:59:41	JaimeAnaya	Suspicions grow that nanoparticles in Pfizerâ€™s COVID-19 vaccine trigger rare allergic reactions https://t.co/OB
11	2020-12-25 23:59:41	adegoke_mukaila	The spread circumstances substance of covid-19, goes extremely breakout viral in the world.But we have the cau
12	2020-12-25 23:59:41	EarickNG	Covid-19 UK Mutant Strain: Higher Hospitalizations, Deaths Likely, Study Finds - Bloomberg https://t.co/3bFmmkr
13	2020-12-25 23:59:38	julesofmaine	@portlandimbiber @nirav_mainecdc @MEPublicHealth The real truth about Covid-19. Respiratory deaths are no
14	2020-12-25 23:59:35	govliesssss	@JimmyMac2021 @Toronto1880 @fordnation And Peel didnt?
15	2020-12-25 23:59:33	jtmonrad	And several @NIH researchers published this piece on "COVID-19 vaccine trial ethics once we have efficacious va
16	2020-12-25 23:59:33	MartyKoekemoer	@AFranzsen @Thomas_Binder @tngadd Weâ€™re becoming preoccupied with Covid19. And Death. And not seei
17	2020-12-25 23:59:27	LiterateLiberal	Yuma Prison Warden Dies From COVID-19 After Dismissing Safety Concerns | @crooksandliars https://t.co/bABIC
18	2020-12-25 23:59:25	myraluv2015	Yes until covid19 is under control. https://t.co/7qaJdsjoBR
19	2020-12-25 23:59:23	PrincetonBoy915	@Sharlie528 Covid-19? I donâ€™t know her ðŸ˜‚ðŸ€
20	2020-12-25 23:59:22	MartyKoekemoer	@LusyNote @Thomas_Binder Weâ€™re becoming preoccupied with Covid19. And Death. And not seeing the dan
21	2020-12-25 23:59:21	jtmonrad	Do these concerns apply here?
22	2020-12-25 23:59:18	NewsThaivisa	Two seafood vendors in Onnut fresh market test positive for Covid - https://t.co/dbBaNxO6Hu

Fig. 1 Example Covid-19 tweet data from Snscrape

4. Remove symbols: Emoticons, symbols and pictographs, transport and map symbols, flags, other language characters and dingbats are removed.
5. Remove non alphabet characters: Replacing characters except Digits and Alphabets with a space.
6. Remove consecutive letters three or more: 3 or more consecutive letters are replaced by 2 letters. (eg: "Cooool" to "Cool")
7. Remove punctuations: Punctuations are removed from the sentence since it is not affecting the meaning of the sentence.
8. Remove stopwords: The stopwords are not add much meaning to a sentence.

Shown in Tables 2 and 3 are examples of data and results before and after preprocessing.

Table 2 Tweet examples

Covid test tweets
@TheRickyDavila @doxiedachsie When they were informed of COVID19 early this year they invested in a company that makes body bags. Why would they care about the nations virus death toll? These two are despicable and must be voted out! Let's go, Georgia!
Short but important thread. #COVID19 #COVID #edchat #iaedchat https://t.co/85Z2nVUMQB
COVID-19 vaccine administered: 63,170 doses #COVID19Colorado https://t.co/IArtRHVaF4 https://t.co/fmWZptXtYA
@CTVNews inaccurate counting: How COVID-19 Deaths Are Counted https://t.co/gjMyeBNYV2
Some Passengers Infected After Man Died of COVID-19 on Plane https://t.co/GWXB2OhSHo

Table 3 Preprocessed tweet examples

Preprocessed tweets
Informed covid19 early year invested company makes body bags would care nations virus death toll two despicable must vote let go Georgia
Short important thread covid19 covid edchat iaedchat
Covid 19 vaccine administered 63 170 doses covid19colorado
Inaccurate counting covid 19 deaths counted
Passengers infected man died covid 19 plane

3.3 Vectorization

In this part every single word occurrence was counted to fill the word occurrence matrix with words and their number of occurrences. This can be counted as n-grams. An n-gram is a contiguous sequence of n items from a given sample of text. In our case n is equal to 1, which means single word was counted not group of words. After vectorization, we obtained one word occurrence matrix for each csv file.

3.4 Sentiment Analysis

There are different types of sentiment analysis types, some of them are; polarity and subjectivity analysis, positivity and negativity analysis, emotion detection. Our project includes positivity and negativity analysis meaning that the result for every tweet is positive or negative. While implementing this, the Naive Bayes Classifier method from TextBlob library in Python was used. The Naive Bayes Classifier is wrapping the same named method from NLTK library in Python and this method classifies movies using a pre-trained model, or the coder can manually train the model with related data. We choose the second approach and trained the model with our labeled tweets dataset, then tested and accuracy was found. Finally, the unlabeled data was given to model and obtained their positivity and negativity values.

3.5 Visaulization

The results were all numbers, but they are more meaningful when visualization is good. So, the Matplotlib library of Python was used to draw bar charts, plots, and pie charts. Wordcloud method from TextBlob library was also used for more colorful results for word occurrences.

4 Result and Discussion

In this study, four different Dataset were analyzed. Two datasets from December 2020 about Vaccine (380,000 tweet) and Covid-19(318,000 tweet) and two dataset from January 2021 about Vaccine (500,000 tweet) and Covid-19(212,000 tweet). Accuracy of the sentiment analysis algorithm after training is determined as “0.6”.

In this section, the results of the visualization process and criticization of the results are included. The bar charts and word clouds are the result of vectorization. The table shows us the sentiment analysis result for each dataset.

By considering the datasets collected in December, 2020, occurrences of the most common words related to “Vaccine” in the analyzed tweets are shown in Fig. 2. Occurrences of the most common words about COVID are displayed in Fig. 3. The same two results for the data collected in January 2021 are shown in Figs. 4 and 5, respectively. Comparing Figs. 2 and 3 with Figs. 4 and 5, respectively, it is obvious

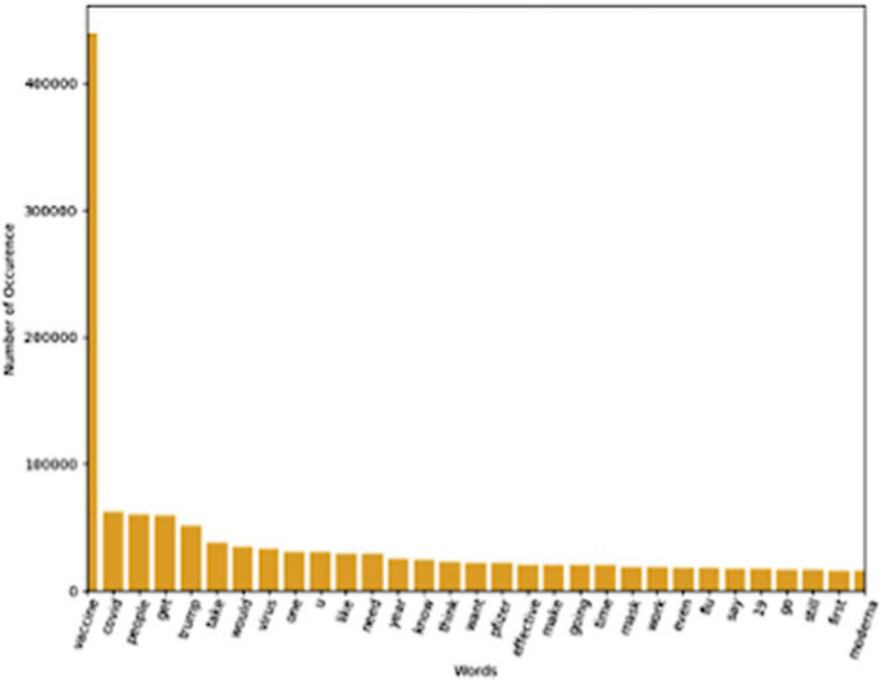


Fig. 2 Most occurred words in tweets about vaccine in December, 2020

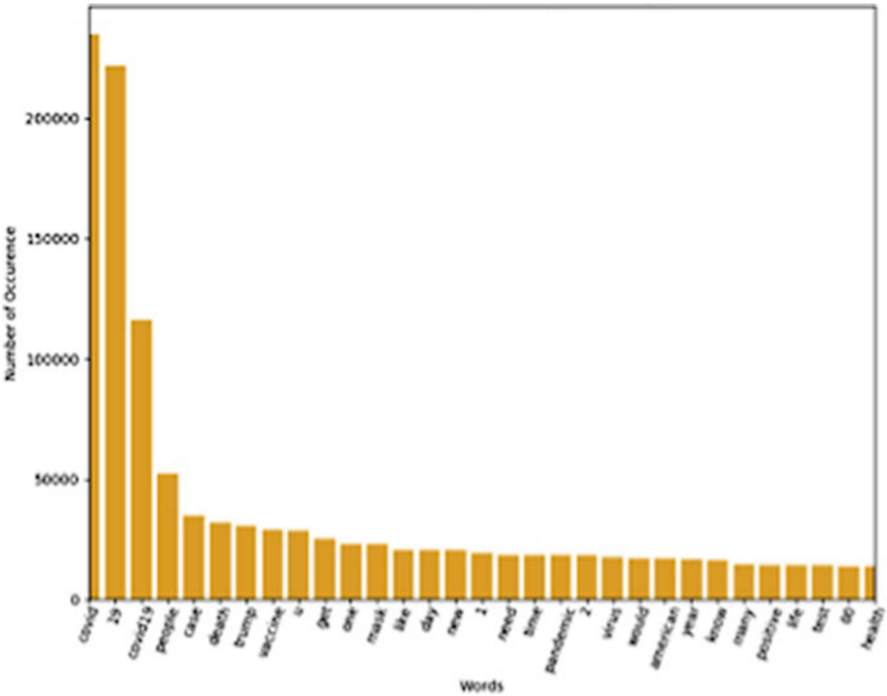


Fig. 3 Most occurred words in tweets about Covid in December, 2020

that the number of occurrences for the common words decreased from December 2020 to January 2021. This may be attributed to various factors, including the following. December is mostly characterized as a vital month with holidays season where people organize a lot of indoor and outdoor activities, travels, etc. On the other hand, January is considered a calm month where people recover from the activities and travel they completed in December. Thus, the drop in the interest in the covid and vaccine can be seen as normal. Further, in January, people are more uninterested in discussing the pandemic after one year of suffering from its health, societal and economic consequences. People tend to be more interested in returning back to normal life style. The most important words discussed during these two periods for "Vaccine" and "Covid" related tweets are reflected in the

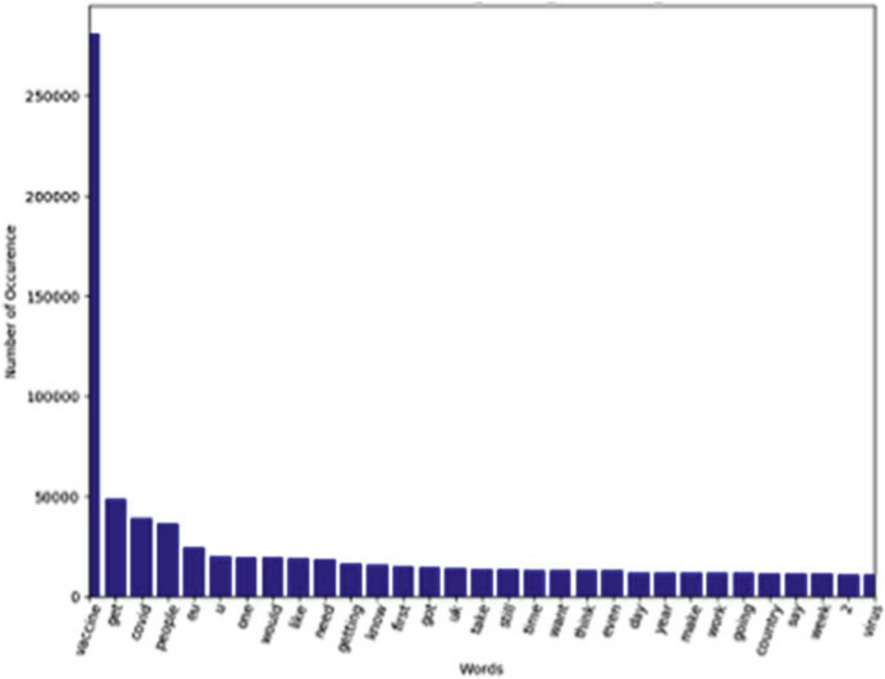


Fig. 4 Most occurred words in tweets about vaccine in January, 2021

word clouds shown in Figs. 6, 7, 8 and 9. The related to sentiments for these two periods (December 2020 and January 2021) concerning "Vaccine" and "Covid" related tweets are shown in Figs. 10, 11, 12 and 13.

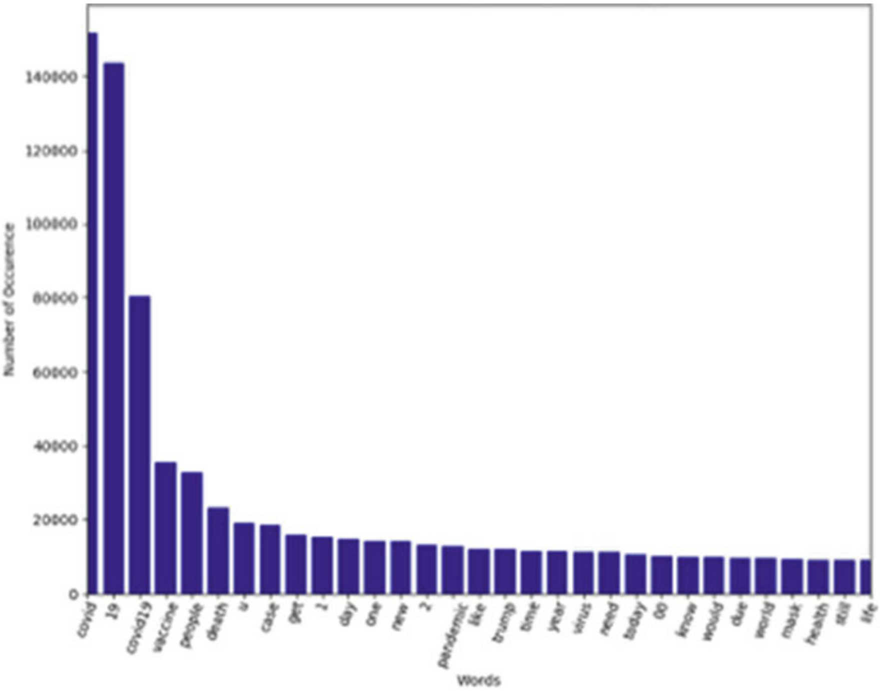


Fig. 5 Most occurred words in tweets about Covid in January, 2021

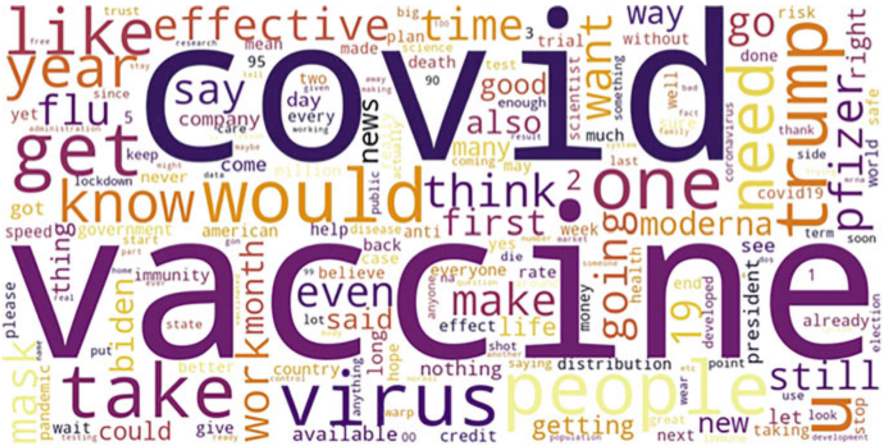


Fig. 6 Wordcloud of tweets about vaccine in December, 2020

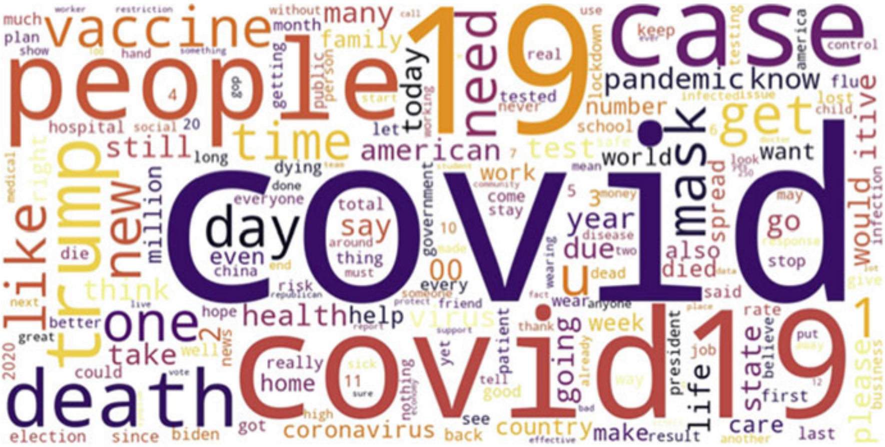


Fig. 7 Wordcloud of tweets about Covid in December, 2020

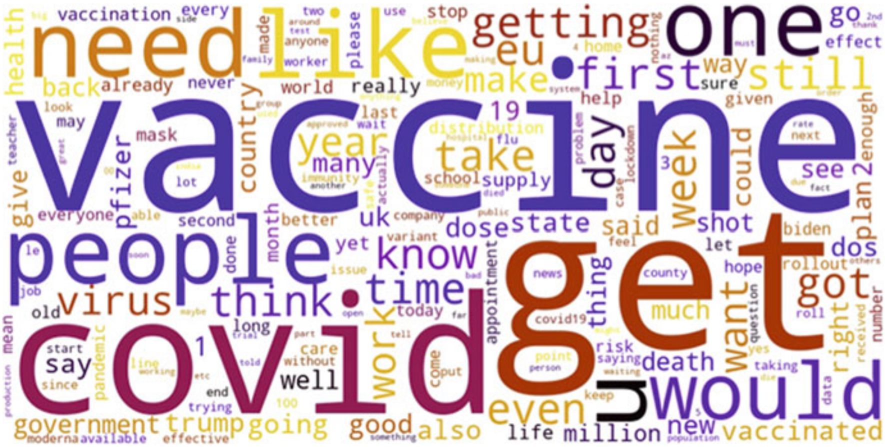


Fig. 8 Wordcloud of tweets about vaccine in January, 2021

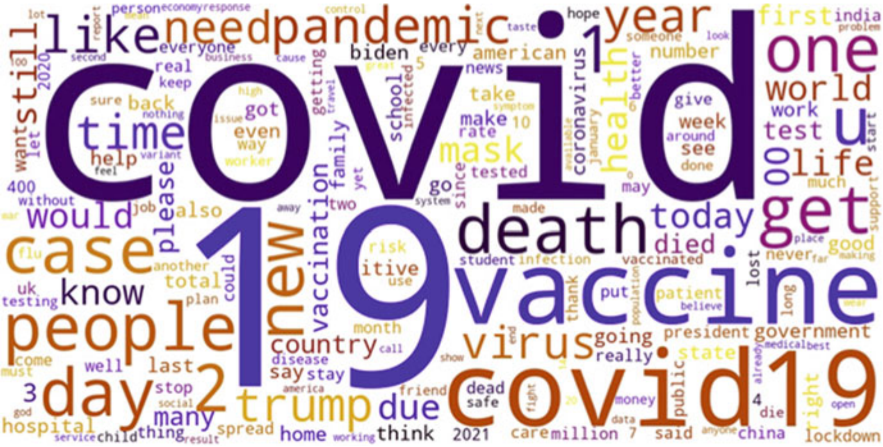


Fig. 9 Wordcloud of tweets about Covid in January, 2021

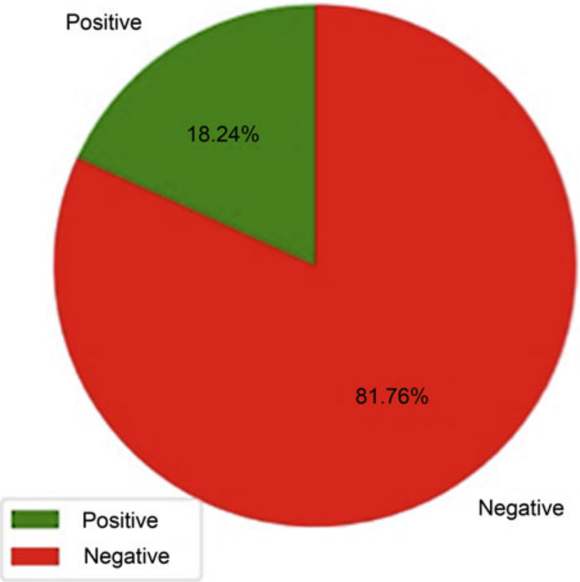


Fig. 10 Sentiment of tweets about Covid in December, 2020

Fig. 11 Sentiment of tweets about vaccine in December 2020

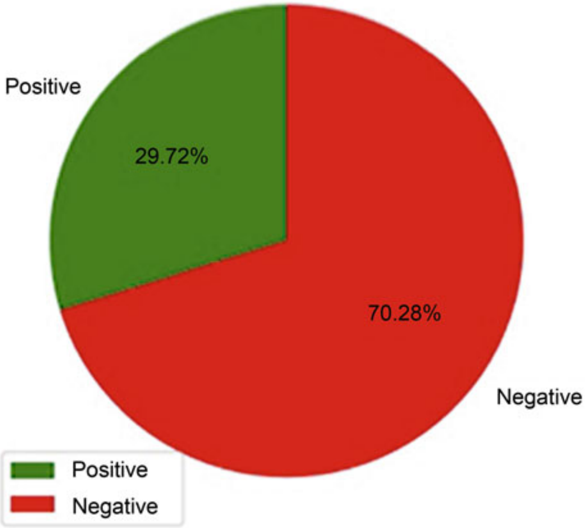
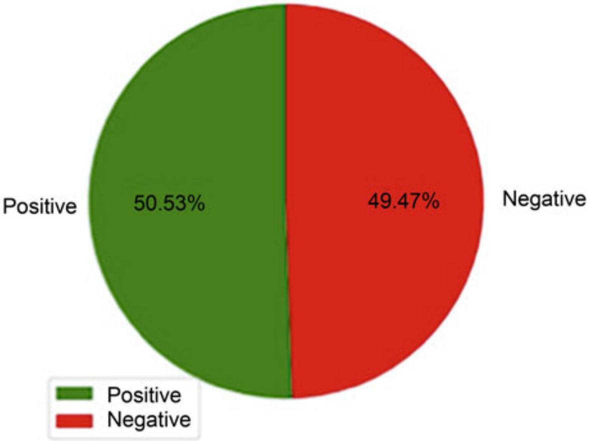


Fig. 12 Sentiment of tweets about Covid in January 2021



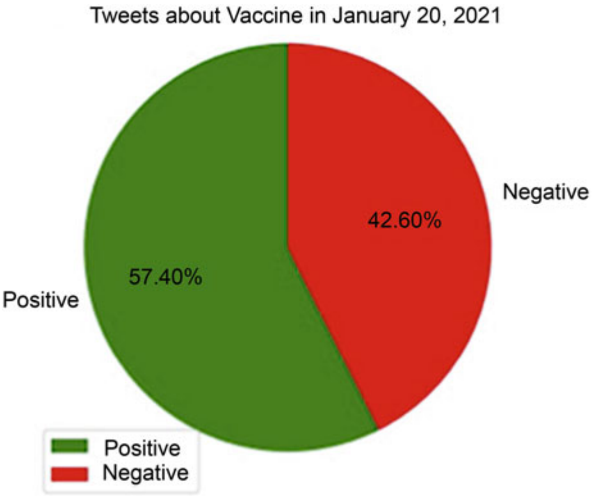


Fig. 13 Sentiment of tweets about vaccine in January 2021

5 Conclusion

As a result of this study, several conclusions could be derived. First of all, for the sentiment analysis algorithm, 0.6 accuracy was determined. This accuracy can be developed with further methods of preprocessing or with a better and much more efficient training algorithm. Also, the algorithm include just positive and negative evaluation. This can be expanded thorough more complex and a better algorithm with adding the neutrality. Even further, some evaluation techniques can be used with different degrees. All these evaluations are effective in our results. We can see the most occurred words in the tables and changes through the months in that trend. Also we see that negativity is seen more in the Covid tweets, whereas positivity is seen more in Vaccine tweets. But this result can be doubted since accuracy is 0.6 and also algorithm omits the neutral tweets. These results should be considered for further developments and works.

References

1. Chawan P (2012) Sentiment analysis and influence tracking using Twitter. Int J Adv Res Comput Sci Elect Eng, 1
2. Cossu J, Dugué N, Labatut V (2015) Detecting real-world influence through Twitter. In: 2015 second European network intelligence conference, Karlskrona, pp 83–90
3. Sanandres E, Llanos R, Camilo MO (2018) Topic modeling of Twitter conversations
4. Doan S, Yang EW, Tilak SS et al (2019) Extracting health-related causality from twitter messages using natural language processing. BMC Med Inform Decis Mak 19, 79

5. Alowisheq A, Alrajebah N, Alrumikhani A, Al-Shamrani G, Shaabi M, Al-Nufaisi M, Alnasser A, Al-Humoud S (2017) Investigating the relationship between trust and sentiment agreement in Arab Twitter users, pp 236–245
6. Tago K, Jin Q (2018) Influence analysis of emotional behaviors and user relationships based on Twitter data. Tsinghua Sci Technol 23(1):104–113. https://doi.org/10.26599/TST.2018.9010012
7. Li P, Zhao W, Yang J, Wu J (2019) CoTrRank: trust evaluation of users and tweets. In: Proceedings of the twenty-eighth international joint conference on artificial intelligence. Twenty-eighth international joint conference on artificial intelligence IJCAI-19
8. Bae Y, Lee H (2012) Sentiment analysis of twitter audiences: measuring the positive or negative influence of popular twitterers. J Am Soc Inf Sci Technol 63(12):2521–2535
9. Cano Basave AE, Mazumdar S, Ciravegna F (2014) Social influence analysis in microblogging platforms – a topic-sensitive based approach. Semantic Web 5(5):357–403
10. Sung J, Moon S, Lee J-G (2013) The influence in Twitter: are they really influenced? In: Behavior and social computing. Springer International Publishing, New York City, pp. 95–105
11. Hong L, Davison BD (2010) Empirical study of topic modeling in Twitter. In: Proceedings of the first workshop on social media analytics – SOMA '10. The First Workshop
12. Grant C, George C, Jenneisch C, Wilson J (2011) Online topic modeling for real-time Twitter search, NIST Special Publication: SP 500-296, The Twentieth Text REtrieval Conference (TREC 2011) Proceedings. https://trec.nist.gov/pubs/trec20/t20.proceedings.html (accessed July 1, 2022)
13. Jonsson E (2016) An evaluation of topic modelling techniques for Twitter. http://www.cs.toronto.edu/~jstolee/projects/topic.pdf (accessed July 1, 2022)
14. Ruan Y, Durresi A, Alfantoukh L (2018) Using Twitter trust network for stock market analysis. Knowl Based Syst 145:207–218