Expert Systems With Applications 212 (2023) 118715

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset

Miftahul Qorib [a], Timothy Oladunni [b], Max Denis [c], Esther Ososanya [d], Paul Cotae [e]

[a] Department of Computer Science and Information Technology, University of the District of Columbia, Washington, DC, United States
[b] Department of Computer Science, Morgan State University, Baltimore, MD, United States
[c] Department of Mechanical and Biomedical Engineering, University of the District of Columbia, Washington, DC, United States
[d] Department of Electrical and Computer Engineering, University of the District of Columbia, Washington, DC, United States
[e] Department of Electrical and Computer Engineering, University of the District of Columbia, Washington, DC, United States

ARTICLE INFO

Keywords:
Covid-19
Vaccine Hesitancy
Sentiment Analysis
Machine Learning
Twitter

ABSTRACT

In 2019 there was an outbreak of coronavirus pandemic also known as COVID-19. Many scientists believe that the pandemic originated from Wuhan, China, before spreading to other parts of the globe. To reduce the spread of the disease, decision makers encouraged measures such as hand washing, face masking, and social distancing. In early 2021, some countries including the United States began administering COVID-19 vaccines. Vaccination brought a relief to the public; it also generated a lot of debates from anti-vaccine and pro-vaccine groups. The controversy and debate surrounding COVID-19 vaccine influenced the decision of several people in either to accept or reject vaccination. Because of data limitations, social media data, collected through live streaming public tweets using an Application Programming Interface (API) search, is considered a viable and reliable resource to study the opinion of the public on Covid-19 vaccine hesitancy. Thus, this study examines 3 sentiment computation methods (Azure Machine Learning, VADER, and TextBlob) to analyze COVID-19 vaccine hesitancy. Five learning algorithms (Random Forest, Logistics Regression, Decision Tree, LinearSVC, and Naïve Bayes) with different combination of three vectorization methods (Doc2Vec, CountVectorizer, and TF-IDF) were deployed. Vocabulary normalization was threefold; potter stemming, lemmatization, and potter stemming with lemmatization. For each vocabulary normalization strategy, we designed, developed, and evaluated 42 models. The study shows that Covid-19 vaccine hesitancy slowly decreases over time; suggesting that the public gradually feels warm and optimistic about COVID-19 vaccination. Moreover, combining potter stemming and lemmatization increased model performances. Finally, the result of our experiment shows that TextBlob + TF-IDF + LinearSVC has the best performance in classifying public sentiment into positive, neutral, or negative with an accuracy, precision, recall and F1 score of 0.96752, 0.96921, 0.92807 and 0.94702 respectively. It means that the best performance was achieved when using TextBlob sentiment score, with TF-IDF vectorization and LinearSVC classification model. We also found out that combining two vectorizations (CountVectorizer and TF-IDF) decreases model accuracy.

1. Background

Based on several studies, Coronavirus Disease 19 (COVID-19), was reported for the first time in Wuhan, China, in 2019 (Kumar et al., 2021), (Mohan & Nambiar, 2020). Face masks, hand sanitation, and social distancing were recommended to reduce the spread and impact of the disease (Kwon et al., 2021). During the peak of the pandemic, educational institutions were closed, large gatherings were outlawed, and international travels were suspended (OECD, 2020). Consequently, developing, underdeveloped, and developed countries were severely impacted by the coronavirus pandemic (Jones, Palumbo, & Brown, 2021). At the end of November 2020, several pharmaceutical companies announced breakthrough in COVIID-19 vaccine research (Pfizer, 2021). Despite the availability of the vaccine, work on COVID-19 is still ongoing research (Jackson et al., 2021).

In early December 2020, the U.S. federal agencies started discussing vaccines' acquisition, storage, and deployment as a solution sought by pharmaceutical companies to combat the spread of the Covid-19

E-mail addresses: miftahul.qorib@udc.edu (M. Qorib), timothy.oladunni@morgan.edu (T. Oladunni), max.denis@udc.edu (M. Denis), eososanya@udc.edu (E. Ososanya), pcotae@udc.edu (P. Cotae).

https://doi.org/10.1016/j.eswa.2022.118715
Received 2 January 2022; Received in revised form 14 July 2022; Accepted 26 August 2022
Available online 5 September 2022

Table 1
Overview of studies with different approaches to analysis.

No	Title	Authors	Datasets	Sentiment Techniques	Classification Method	Classification	Conclusion
1	A Comprehensive Analysis of Approaches for Sentiment Analysis Using Twitter Data on COVID-19 Vaccines	Amrita Mishra, Mohd. Saif Wajid, Upasana Dugal	Daily tweets of vaccines Sputnik V (2000 tweets),Moderna (2000 tweets)Covaxin (2000 tweets)	TextBlob	N/A	N/A	U.S Moderna Vaccine is the most discussed brand among Twitter users, which has the most favorable opinion compared to Covaxin (India) and Sputnik V (Russia) vaccines (Mishra et al., 2021).
2	COVID-19 Vaccine–Related Discussion on Twitter: Topic and Sentiment Analysis Modeling	Joanne Chen Lyu, Ph.D., Eileen Le Han, Ph.D., Garving K Luli, PhD	1,499,421 tweets maintained by Georgia State University's Panacea Lab	Textstem	N/A	N/A	According to this research, the increasing positive sentiments in the tweets related to Covid-19 may imply that Covid Vaccine has higher acceptance rates than previous vaccines (Lyu et al., 2021).
3	Mapping of the Covid-19 Vaccine Uptake Determinants from Mining Twitter Data	Ana Baj-Rogowska	125,495 tweets (Polish language only) were collected using QDA Miner software	Using QDA Miner software to find categories of determinants	The algorithm implemented in the WordStat software	Six groups of determinants (Assurance, Activation, Awareness, Affordability, and Access)	This study states that the Covid-19 vaccine mostly 39.4 % depends on Awareness, and about 27.3 % depends on Access. According to the polls conducted that Covid-19 vaccine hesitancy keeps increasing globally (Baj-Rogowska, 2021).
4	Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis	Siru Liu and Jialin Liu	Collected 2,678,372 tweets using snscrape package in Python to collect tweet IDs and then used the tweepy package to collect the data	VADER	Pruned Exact Linear Time (PELT) algorithm	Positive, negative & neutral and by countries	The study finds that about 42.8 % of tweets are positive, while 30.3 % are negative. Public sentiment increased after Pfizer announced that its vaccine effectiveness reached 90 % (Liu & Liu, 2021).
5	Sentiment Analysis on COVID Tweets: An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models	Ghulam Musa Raza, Zainab Saeed Butt, Seemab Latif, Abdul Wahid	1) 165,116 tweets using Twitter Scraper using keywords Coronavirus2) 41,349 tweets using Twitter Scraper (keyword: work from home, or WFH3) Data from Kaggle	TextBlob	Count Vectorizer and TF-IDF using Logistic Regression, Multinomial Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine and XGBoost	Positive and negative	The highest accuracy is when using Support Vector Machine classifier using vectorization TF-IDF (93.15 %) and Count Vectorizer (93.07 %). In addition, TF-IDF accuracy tends to be 10 % more efficient than Count Vectorizer (Raza et al., 2021).
6	Sentiment analysis on Twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm	Shamrat, F. M. Javed Mehedi Sovon Chakraborty	30,000 tweets using API search based on #Pfizer, #Moderna, and #AstraZeneca	TextBlob	KNN	Positive, negative, and neutral	This study finds that Pfizer has a positive sentiment of 47.29 %, while Moderna is about 46.16 %, and AstraZeneca's rate is 40.08 % (Shamrat et al., 2021).
7	Understanding COVID-19 Vaccine Reaction through Comparative Analysis on Twitter	Yuesheng Luo and Mayank Kejriwal	They have collected 371,940 tweets using API search. After removing non-English and retweets, there are 128,472 tweets left.	VADER	N/A	N/A	Covid-19 vaccine hesitancy remains high even though the government has promoted the benefits of getting vaccinated and making it mandatory for U.S. federal employees (Luo & Kejriwal, 2021).
8	US Based COVID-19 Tweets Sentiment Analysis Using TextBlob and Supervised Machine Learning Algorithms	Rashid Khan, Furqan Rustam, Khadija Kanwal, Arif Mehmood	11, 858 tweets collected by the Department of Computer Science, Abbottabad University	TextBlob	Random Forest, Gradient Boosting Machine, Extra Tree Classifier, Logistic Regression, and	Positive, negative, and neutral	According to this research, TF-IDF features have increased the performance of the Gradient Boosting

(continued on next page)

Table 1 (*continued*)

No	Title	Authors	Datasets	Sentiment Techniques	Classification Method	Classification	Conclusion
		and Gyu Sang Choi	of Science & Technology		Support Vector Machine		Machine to achieve 96 % accuracy (Khan et al., 2021).
9	Use of Two Topic Modeling Methods to Investigate Covid Vaccine Hesitancy	Phillip Ma, Qing Zeng-Treitler and Stuart J. Nelson	Using Tweeter API to collect 3,403,166 tweets	Top2Vec	N/A	N/A	Top2Vec found that topics that influence vaccine hesitancy included safety issues, rushing development and approval, vaccine efficacy, and personal risk judgments (Ma et al., 2021).
10	Using K-Means Clustering Method with Doc2Vec to Understand the Twitter Users' Opinions on COVID-19 Vaccination	Guanjin Wang and Stephen Wai Hang Kwok	Collected 12,134 tweets using Tweeter API search	Spacy	Doc2Vec and K-means clustering	k = 4	Based on this research, there are four essential topics related to the Covid-19 vaccine for Australians: national vaccine rollout, vaccine and death correlation, vaccine approval, and hesitancy (Wang & Kwok, 2021).
11	Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech, and Moderna COVID-19 vaccines	Robert Marcec and Robert Likic	Collected 701,891 tweets using tweeter API search	AFINN lexicon	N/A	N/A	AstraZeneca or Oxford vaccine tends to be negative over time, boosting vaccine hesitancy (Marcec & Likic, 2021).
12	COVID-19 Vaccine Hesitancy on Social Media: Building a	Muric, Yusong Wu, and Emilio Ferrara	Collected 137 million tweets from streaming collections and historical account tweets	Avax (anti-vac score)	N/A	N/A	False claims about COVID-19 vaccines can undercut public trust in the ongoing vaccination campaigns and threaten global public health (Muric et al., 2021).
13	Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic	Erika Bonnevie, Allison Gallegos-Jeffrey, Jaclyn Goldberg, Brian Byrd & Joseph Smyser	Retrieved data from PGP analysts	comparing conversation four months before and after COVID-19 spread	N/A	N/A	Vaccine opponents provoke Covid 19 vaccine opposition and encourage mistrust in health authorities, so exposure to the increase of vaccine opposition might impact the health of a population (Bonnevie, Gallegos-Jeffrey, Goldbarg, Byrd, & Smyser, 2020).
14	Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content	Md Tarique Jamal Ansari, Naseem Ahmad Khan	Collected 820,000 tweets using API search with eight keywords related to Covid 19	TextBlob	N/A	N/A	Most Tweets generally had a negative tone because of a lack of trust. Fear of getting vaccinated may increase because of the prevailing feeling and concern about getting Covid 19 vaccination (Ansari & Khan, 2021).

pandemic (Michaud & Kates, 2020). The CDC (Centers for Disease Control and Prevention), for instance, proposed allocating Covid-19 vaccination through several phases: the healthcare professional workers were the first to be administered COVID-19 Vaccine (phase 1a), followed by essential workers (phase 1b), high-risk medical conditions, and adults 65 years old and older (CDC, 2021). However, anti-vaccine groups constantly challenge the use of vaccines to fight this contagious virus through conspiracy theories and myths. Some anti-vaccine arguments focused on the vaccine development process, and its long-term side effects (Douglas & Joelley, 2014). Thus, researchers and public health experts have been gathering dataset to understand vaccine hesitancy (Loomba, Figueiredo, Piatek, Graaf, & Larson, 2021).

Nowadays, the internet has been the primary source of almost everything, especially vaccination information for most people (McClain, Vogels, Perrin, Sechopoulos, & Rainie, 2021). Popular social media have been a tremendous source of information with its rapid information dissemination, where many individuals express their opinion through social media platforms (McClain et al., 2021), (Wong, Ho, & Olusanya, 2020). While working on social media databases, text mining has been a meaningful tool to collect datasets from some of these large data repositories (Tao, Yang, & Feng, 2020). Text mining on social media seems to be a viable alternative to analyze public opinion on the Covid-19 (Oyebode et al., 2021). According to the World Health Organization (WHO), vaccine hesitancy was one of the significant dangers to worldwide health in 2019 (Piedrahita-Valdés et al., 2021). Therefore, collecting and monitoring data from social media, like Twitter, can provide various viewpoints on vaccines that can benefit public health organizations identify factors contributing to vaccine confidence based

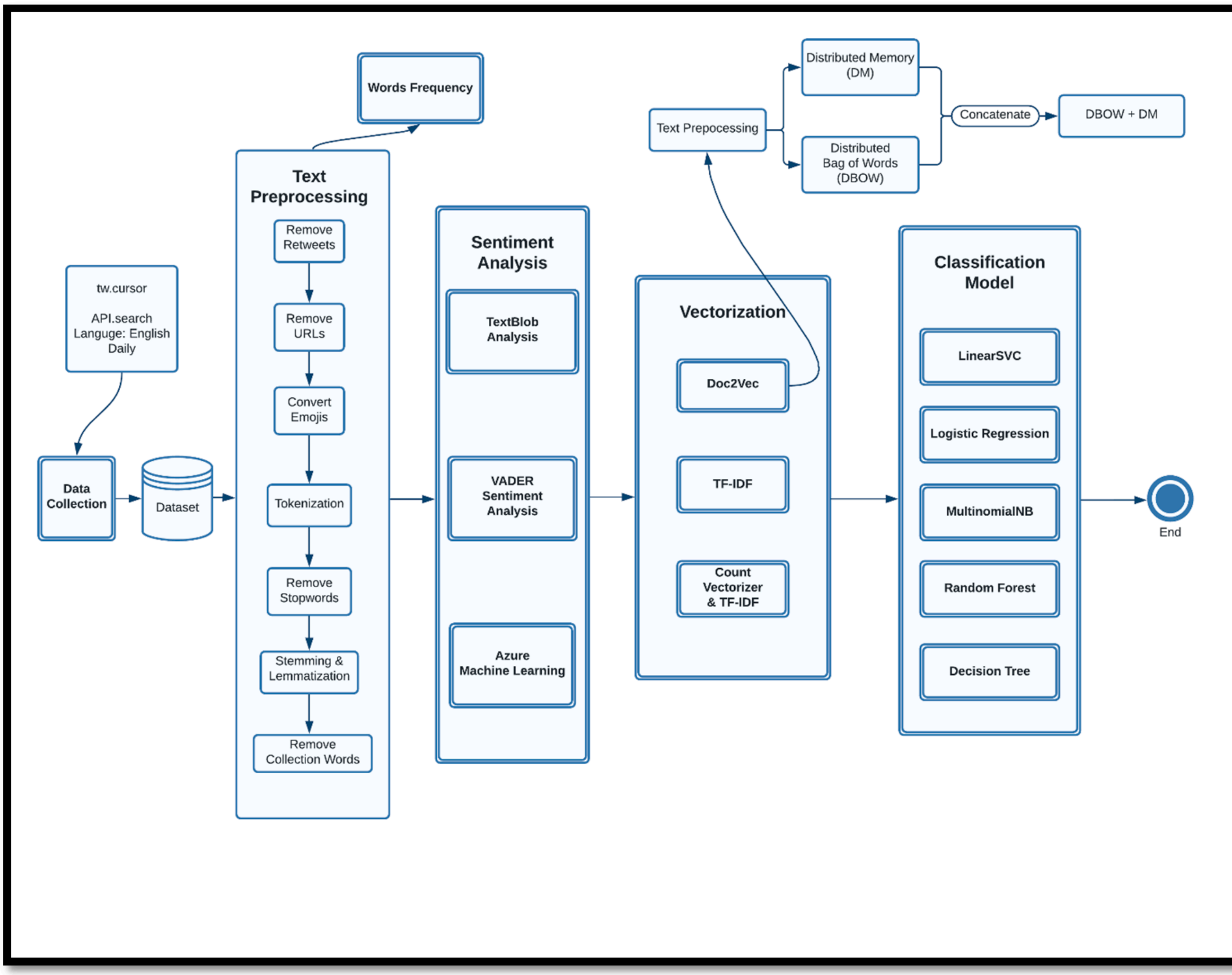

Fig. 1. Experimental Design.

on historical and geographical data (Loomba et al., 2021), (McClain et al., 2021).

Generally, people freely express their feelings, safety, and efficacy of the vaccine through social media (Igoe, 2019), such as Twitter. For this reason, public tweets can be collected using Twitter API authentication tokens (Twittter, 2021). Because almost every-one expresses their senses through social media, anti-vaccination movements used social media to influence people to lower vaccine acceptance rates; frustrating efforts to prevent or reduce the spread of the coronavirus pandemic (Chiou & Tucker, 2018). False claims regarding the Covid-19 vaccine spreading online since the beginning of the pandemics (Muric, Wu, & Ferrara, 2021), (Naeem, Bhatti, & Khan, 2020), can undermine public faith in continuing vaccination campaigns. This poses a threat to global health (Naeem et al., 2020). While rumors and conspiracy theories circulating online can contribute to vaccine hesitancy, we argue that monitoring Covid-19 related data from social media can help track vaccine information in real-time and assist its impact (Islam et al., 2021). Furthermore, understanding public perspectives from behavior change theories can facilitate an effective vaccine promotion strategy (Piltch-Loeb & DiClemente, 2020).

This work is organized as follows: section 2 reviews existing literatures on COVID-19 hesitancy, section 3 discusses the methodology of the study and 4 focuses on the result. Discussions, contributions, and conclusions are discussed in sections 5, 6, and 7 respectively. Future work is itemized in section 8. Finally, we acknowledged our support in section 9.

2. Literature review

2.1. Previous studies

Since the outbreak of COVID-19, there have been several studies on COVID-19 vaccination hesitancy using sentiment analysis. Researchers have used different techniques in analyzing COVID-19 sentiments of the public based on their opinions using the twitter dataset. In most of these works, TextBlob and VADER are two of the most popular COVID-19 twitter sentiment analysis techniques. Mishra, Wajid, and Dugal (2021) used TextBlob sentiment computation on the Twitter dataset to examine different Covid-19 vaccine brands. Raza et al. also analyzed Covid-19 Twitter dataset using TextBlob to understand various classification models on vaccine hesitancy. Khan, Rustam, Kanwal, Mehmood, and Choi (2021) also analyzed datasets from Twitter to examine classification accuracy of various models. Other research studies (Liu & Liu, 2021), (Luo & Kejriwal, 2021) used VADER to examine vaccine hesitancy. Some studies like (Naseem, Khushi, Kim, & Dunn, 2021) discussed sentiment analysis without performing model classifications.

Table 1 shows a comparative study on COVID-19 vaccine sentiment analysis. Comparing and contrasting prior studies will assist the current research in examining different areas that have not been discussed in the previous research. It will also guide new studies offering other contributions to the subject matter. The table is a summary table showing various studies with different approaches that have been done using

Twitter dataset for Covid-19 vaccination hesitancy study.

2.2. Gap in literature

Based on the previous research on Covid-19 vaccine hesitancy as shown in table 1, there seems to be a paucity in the literature that needs to be discussed.

1. Previous works are limited in the scope of sentiment computation techniques – in each of the previous works, researchers only used a single computation method. We believe that comparing several sentiment computation techniques will enhance COVID-19 vaccination hesitancy study.
2. Some previous works were based on downloaded datasets from other institutions. There is a possibility that the current Covid-19 hesitancy situation might have changed. Therefore, addressing the current Covid-19 vaccine hesitancy situation is another critical question to be discussed.
3. Most previous studies on COVID-19 vaccine hesitancy were limited in scope on classification techniques. We believe that applying different classification modeling techniques could enhance the research on Covid-19 vaccine hesitancy. Therefore, addressing which methodology is the most effective classification model for COVID-19 vaccine hesitancy using tweeter dataset is another crucial question to be considered.

In view of the gaps in the previous works on COVID-19 vaccine hesitancy as shown in table 1, this study uses text mining, sentiment analysis and machine learning techniques on COVID-19 Twitter datasets to understand the public's opinions regarding Covid-19 vaccine hesitancy. Achieving this goal, we designed, developed, and evaluated 42 models to determine the best model for classifying twitter COVID-19 dataset into positive, neutral, or negative.

3. Methodology

3.1. Overview

The architectural design of our experiment is shown in Fig. 1. Public tweets extracted from Tweeter via its Application Programming Interface (API) was used for the experiment. As shown in Fig. 1, the first stage of our investigation was the preprocessing stage. Retweets and URLs were removed in the preprocessing step, emojis were converted into words, dataset was cleaned and normalized. We also removed stop words and performed tokenization. Stemming and lemmatization were done as well. We further explored the dataset by computing the Word Frequency. Using *t*-test, we confirmed the positive improvement of the public opinion on Covid-19 vaccination. After the preprocessing stage, we computed sentiment scores of the dataset using Azure Machine Learning, VADER, and TextBlob methods. Vectorization of the dataset was based on Doc2Vec, TF-IDF, and CountVectorizers. Lastly, we built 42 different models to determine the best model for classifying the twitter COVID-19 dataset into positive, neutral, or negative. LinearSVC, Logistic Regression, Multinomial Naïve Bayes, Random Forest, and Decision Tree learning algorithms were trained and evaluated. A detailed explanation of the proposed system is explained in the following sections.

3.2. Data collection

Datasets were collected daily from the public tweets streaming using Twitter API. We utilized: **API.search, q = search_words, lang='en', since=,** and **until =.** Search words used were #vaccine or #covid19 and language was set to English tweets only. One of the challenges we had was that Twitter only allows a researcher to access API search for about 7–9 days, so we could not dig live public tweets longer than nine days. Consequently, data collection was done day by day. we collected live data from Twitter from September 26, 2021, to November 7, 2021. Forty-two thousand seven hundred and ninety-six tweets were collected. Daily live public tweets were saved as CSV (comma-separated value) daily. At the end of data collection, day-to-day public tweets were combined in a big dataset.

3.3. Text preprocessing

Text preprocessing is a vital NLP (Natural Language Processing) procedure. It reconstructs text into a more digestible form for machine learning algorithms. Our text preprocessing steps include removing retweets, URLs, and punctuations, converting emojis to words, tokenization, removing stop words, stemming and lemmatization, and removing collection words.

3.3.1. Retweets

A retweet is when a tweeter user shares a tweet from another user. Sharing of tweets results into duplication. Duplicate tweets might skew word frequency, impede the fitness of the model, and increase space requirement to run the experiment. Therefore, we removed retweets as they contain a content duplication.

3.3.2. URLs and punctuation

The next step is removing URLs. In this step, we removed URLs from the tweets. Cleaning URLs is critical since it does not have meaning and will not affect their sentiment value. Keeping the URLs link will skew word frequency because each tweet has a link. Cleaning up tweet texts includes removing non-useful signs and punctuations, such as #,?, /, \,! using python re package.

3.3.3. Emoji

After cleaning up the tweets from retweets, the next step is to convert emojis into words using emoji.demojize() from the python library. Some people frequently use emojis to express their feeling. Therefore, converting these emojis into phrases might improve tweets' sentiment analysis.

3.3.4. Tokenization and normalization

Tokenization is the process of splitting words in each tweet. In this step, to reduce the length of our code, text tokenization was combined with text normalization. All collection of words from the tweets were converted into lower case as normalized words. Each word from the tweets was then stored in a list collection of words for computation of its polarity.

3.3.5. Stop words

After splitting words from the tweets, the next step is removing stop words. A sklearn package, 'stopwords', was used to remove stop words in English. Removing stop words, such as 'am', 'is', hasn't, aren't, etc., will be helpful to calculate tweet sentiment because these stop words are not helpful for the analysis. These stop words will skew a bucket of phrases from the tweets.

3.3.6. Stemming and lemmatization

Stemming is a process of reducing words into their base, word stem, or root form. For example, words such as looking, looks, or looked are reduced to look. In other words, this is the process of reducing inflection in words to their root form. Occasionally, some of the words might not be valid in the language. 'PorterStemmer' stemming was used for the experiment.

Unlike stemming, lemmatization is a process of reducing the inflected words properly, ensuring that the root word belongs to the language. Lemmatization is based on vocabulary and the form of the words. This process aims to remove inflectional endings and return them to the base or dictionary form. 'WordNetLemmatizer' lemmatization was used for

Table 2
Top 20-Word Frequency.

No	Common Words	Count	No	Common Words	Count
1	covid	6587	11	mand	2163
2	do	4728	12	ncovishield	1941
3	age	3469	13	Pfizer	1806
4	get	3392	14	boost	1689
5	slot	3086	15	shot	1332
6	nag	2757	16	work	1289
7	new	2402	17	gt	1229
8	amp	2361	18	heal	1191
9	people	2257	19	take	1122
10	Pincode	2213	20	lt	1091

the experiment.

3.3.7. Collection of words

Lastly, removing the collection of words (phrases used to search in tweets collection) is critical to the analysis. Collection of words, such as covid19, vaccine, and corona, were removed from the list of the tweets because it would cause frequency skewness. Because each tweet would have at least one of the phrases, thereby making its frequency much higher than other words. Thus, affecting the normal distribution of the dataset.

3.4. Word frequency

After the collection of words were created, word frequency was analyzed to get more information and insight from the Word Frequency Table. As shown in Table 2 (the top 20 words frequency), the word '*covid*' has the highest frequency (6587). It does make sense that the word '*covid*' strongly associated with Covid-19. When we preprocessed the dataset, we did removed collection of words (covid19, vaccine, and corona). Furthermore, the word '*do*' shows up as the second (4728) and the word '*get*' shows up the fourth highest frequency (3392). Both imperative words ('*do*' and '*get*') might explain the urgency for the people to do or get vaccines to combat the spread of the contagious virus. On the other hand, anti-vaccine groups, might also use both words to influence societies not to do or get vaccines.

Another interesting term in the top 20-word list is '*age*'. The phrase '*age*' shows up 3469 times (the third place). There is a possibility that Twitter users discussed the age requirement to get Covid-19 vaccine. It may also point to age regarding the safety of the vaccine for children. Various Covid-19 vaccine providers restrict different minimum ages to get vaccinated. For instance, Pfizer-BioNTech vaccine is available for children aged 5–11 and teens aged 12–17 years old, but Moderna and Johnson & Johnson vaccines are not available for children and teens populations. However, both vaccines (Moderna and Johnson & Johnson) were made available for ages 18 and up. Consequently, there is a possibility that social media platforms were talking about age during that time frame of data collections.

Another interesting word from Table 2 is the word '*mand*'. This word is one of the top 20 words that show up frequently in the tweets. 'mand' is a stem word for mandate or mandatory. It is an indication that people were discussing the government mandatory order to get vaccinated. The previous experience suggested that vaccine mandatory is difficult, but it is not impossible (Hagan, Forman, Mossialos, Ndebele, & Hyder, 2021). Voluntary vaccination might not be optimal in the U.S., so some health systems implemented mandatory vaccines to improve compliance (Reiss & Caplan, 2020). This mandatory policy for people to get vaccinated is still one of the hot topics in COVID-19 vaccination campaign.

In addition, from the top 20-word frequency, the word '*boost*' shows up 1689 times. It might imply that Tweeter users have also discussed vaccine boosters frequently in their tweets. The Covid-19 vaccines might be safe and effective in combating the Coronavirus (Krause et al., 2021), however, booster short is an ongoing discussion. ChAdOx1 nCoV-19 (an AztraZeneca vaccine) for adult and aged-mice is a prime-boost that could enhance immunogenicity in older persons (Silva-Cayetano et al., 2020). Booster discussion in the tweets might relate to the idea of getting vaccine booster as new variants of Coronavirus (Delta and Omicron) have been identified in communities. People, especially pro-vaccine groups, are willing to get vaccine boosters as their efforts to protect their families from the new Covid-19 variants. It probably has to do with requirement of booster shorts.

Table 3
Vaccine Brands in Tweets.

Vaccine Brand	Positive	Neutral	Negative
Pfizer	654	923	204
Moderna	703	81	44
Johnson & Johnson	95	160	22

Table 4
P-Values.

Relation	Azure	VADER	TextBlob
Negative - Positive	9.90872E-45	9.34656E-07	1.80367E-28
Negative - Neutral	0.000285551	9.81157E-21	6.53141E-57
Positive - Neutral	1.13721E-46	8.00554E-14	2.67395E-58

Another interesting word is Pfizer, which shows up 1806 tweets. There is a possibility that the Pfizer-BioNTech brand is frequently discussed in their conversations. On December 11, 2020, the FDA (Food and Drug Administration) issued an Emergency Use Authorization for the Pfizer-BioNTech vaccine to combat Covid-19 (Oliver et al., 2020). In addition, the ACIP (Advisory Committee on Immunization Practices) issued a recommendation for the use Pfizer-BioNTech for individuals aged 16 and more to prevent the Coronavirus. Anaphylaxis, a severe and life-threatening allergic reaction that happens typically within 15 min of vaccination, rarely happens after Pfizer vaccinations (11.1 cases per million doses) (Shimabukuro, 2020). Table 3 shows the vaccine brands in our tweeter dataset and their classification into positive, neutral, or negative.Table 4..

According to Table 3, we notice that the three vaccine brands are mentioned more in positive tweets than negative sentiments. Pfizer-BioNTech is mentioned 654 times in positive tweets but 204 times negatively. We found that Moderna showed up 703 in positive tweets, while it was mentioned 44 times negatively. In addition, Johnson and Johnson's vaccine was found 95 in positive tweets, but it appeared 22 times in negative sentiments. There is an indication that people tend to feel optimistic about the three Covid-19 vaccine brands in the U.S.

3.5. Daily sentiment analysis

Text sentiment analysis is considered as one of the most popular subjects in Natural Language Processing (NLP). It can be used to compute text sentiment scores. If rightly used, text sentiment analysis has the capability of transforming extensive unstructured text data into structured and quantify text opinions. The total of positive, negative, and neutral words in each document is counted to determine the sentiment score of the text documents. For computing text sentiment scores, each positive word counts as $+1$ while the negative and neutral words as -1 and 0 respectively.

Absolute Proportional Difference bounds from 0 to 1.

$$\textbf{\textit{Sentiment}} = \frac{\boldsymbol{P-N}}{\boldsymbol{P+N+O}} \tag{1}$$

Relative Proportional Difference bounds from -1 to 1.

$$\textbf{\textit{Sentiment}} = \frac{\boldsymbol{P-N}}{\boldsymbol{P+N}} \tag{2}$$

Logit scale bounds from -∞ to $+\infty$.

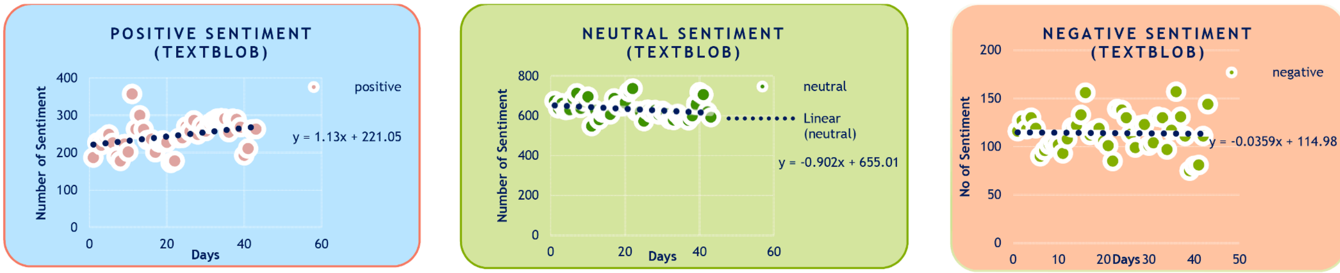

Fig. 2. *TextBlob* Daily Sentiment.

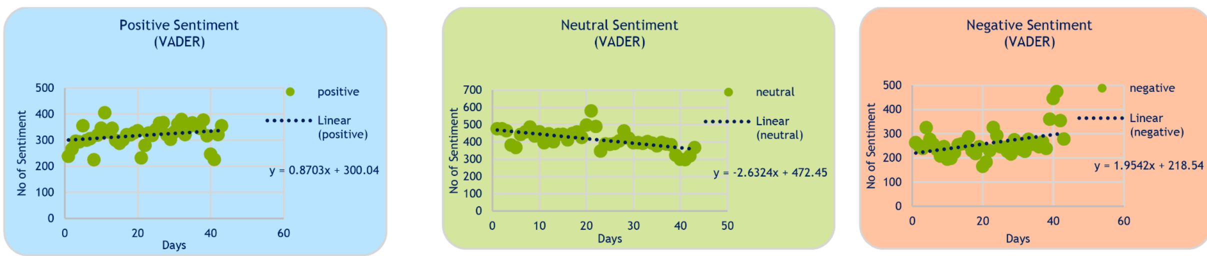

Fig. 3. VADER Daily Sentiment.

$$\textbf{\textit{Sentiment}} = \log(p + 0.5) - \log(N + 0.5) \quad (3)$$

P is a positive word, N is a negative phrase, while O is an emotionless word, not associated with a person's feeling (such as the name of a city, Washington). This sentiment score will be bound from −1 to 1. The Relative Proportional Difference might lead to a possibility that a sentence's score might tend to cluster strongly around endpoints because they may contain primarily or exclusively of either positive or negative contents. This research study will utilize the following sentiment computation methods: TextBlob, VADER and Azure Machine Learning text sentiment computation.

3.5.1. TextBlob

Python package TextBlob() was used to calculate the polarity values of individual tweets on vaccine covid by creating TextBlob objects, which assigned polarity values to the tweets. It identifies the polarity value of the object using the attribute.polarity. Using pandas dataframe, we created tweets polarity values of the tweets. Positive value translated to be a positive emotion, and negative value translated as a negative sentiment response, while polarity close to zero was categorized as neutral sentiment. The result is shown in Fig. 2.

According to the TextBlob Sentiment Analysis (Fig. 2), we notice that daily positive sentiment seems to be increasing about 1.13 per day, while neutral sentiment is decreasing about 0.902 per day and negative sentiment is decreasing by 0.0359 per day. Public positive sentiment increases (42.8 % tweets), while negative tweets are about 30.3 % after Pfizer announced its vaccine effectiveness reach up to 90 % (Liu & Liu, 2021). The figure suggests that there is a positive indicator that people start feeling optimistic about getting vaccinated. It may imply that vaccine hesitancy slightly decreases over time.

3.5.2. Vader

The next method we used was VANDER (Valence Aware Dictionary and Sentiment reasoner). VADER is a lexicon function for analyzing text sentiment, especially attuned to sentiments expressed in social media. VADER is a new package. To utilize the package, first, we installed the required packages (VanderSentiment and twython). After cleaning up the tweets using a command, analyser.polarity_scores, list of tweets was calculated and assigned polarity scores. Positive scores were grouped into positive sentiments, and negative scores were categorized into negative sentiments, while scores closed to zero were listed in neutral sentiments. The result is shown in Fig. 3.

As shown in the VADER Sentiment Analysis (Fig. 3), we notice that positive sentiment slightly increased over time. Positive sentiment increases by 0.8703 each day, while neutral sentiment decreases by 2.6324 overtime. Negative sentiment increases by 1.9542 per day. There is a chance that the number of neutral sentiments slightly switches to the negative sentiment category over time. The number of negative sentiments increases (1.9542) countered by the number of neutral sentiments decreases (2.6324), there is a slight decrease by 0.6782 per day. Based on the VADER sentiment analysis method, there is a positive indicator that people start feeling optimistic about the Covid-19 vaccine. It might imply that Covid-19 vaccine hesitancy slightly decreases over time.

3.5.3. Azure Machine Learning

Results from TextBlob and VADER suggest that positive sentiment

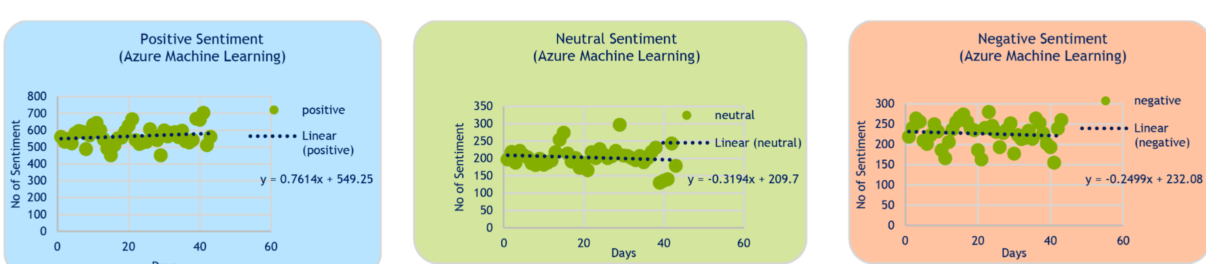

Fig. 4. Azure Machine Learning Daily Sentiment.

about COVID-19 vaccines has increased. We confirmed our result using Azure Machine Learning. Therefore, the third method we used for calculating sentiment scores was Azure Machine Learning. In Microsoft Excel, an add-on Azure Machine Learning from AML Team is available. Using this add-on from Microsoft Excel, we computed the sentiment scores. Polarity scores were grouped into positive, negative, and neutral sentiments. Currently, Azure calculates sentiment scores on social media for Microsoft users that allow them to get more features and adds-on from using its packages (Harfoushi, Hasan, & Obiedat, 2018).

According to the Azure Machine Learning sentiment analysis (Fig. 4), positive sentiment slightly increases. The positive sentiment increases by about 0.7614 over time. On the other hand, both neutral and negative sentiments decrease by 0.3194 and 0.2499, respectively. Like the previous sentiment analysis methods (TextBlob and VADER), Azure Machine Learning also confirms that people have started feeling optimistic about getting the Covid-19 vaccine. This implies that Covid-19 vaccine hesitancy has been decreasing over time.

The daily sentiment analysis result suggests that the public seems to have started feeling optimistic about COVID-19 vaccine. Our inference is in alignment with the work of Kirzinger et al. Per the study, more than half (54 %) of most individuals (92 %) who previously decided to wait and see before getting vaccinated, got a Covid-19 vaccine at least in the beginning of 2021 (Kirzinger, Sparks, & Brodie, 2021). A shift towards positive sentiment vaccination might relate to the effectiveness of the vaccine. A recent study has shown that being unvaccinated is linked with 2.34 times the odds of being infected with the virus compared with fully vaccinated individuals (Cavanaugh, Spicer, Thoroughman, Glick, & Winter, 2021). Buttressing the efficacy and effectiveness of COVID-19 vaccines, scientific research suggests that Covid-19 vaccination helps to reduce the impact of the coronavirus disease (Limaye, Stuar, & Sell, 2021).

As shown in Fig. 4, Azure, VADER and TextBlob seem to agree that there is an increase in positive sentiment to COVID-19 vaccine. To validate this assumption, a one-tailed *t*-test was used to test our hypothesis.

H_0: $B_1 = 0$
H_1: $B_1 > 0$

H_0 is the null hypothesis, H_1 represents the alternative hypothesis, and B_1 describes the slope of the number of positive sentiments over time.

A P-value table (table 5) was created for Azure, VADER, and TextBlob results. The table shows that the p-values between negative and positive sentiment are extremely low (close to zero) for the three sentiment analyzers. Therefore, we reject the null hypothesis. The result of the *t*-test suggests an improvement in public opinion on covid-19. It seems the public have started thinking positively about getting the Covid-19 vaccine. This could be because of perceived benefit of getting the Covid-19 vaccine (public views shifted from being worried to feeling optimistic about having the Covid-19 vaccine). In this case, we deny the belief that there is no improvement in Covid-19 vaccination. The increase in positive sentiments corresponds to the vaccine effectiveness increase, which researchers found to be close to 90 % (Liu & Liu, 2021).

3.6. Vectorizations

Cleaned text was transformed into numerical vectors for an effective pattern recognition and knowledge discovery before learning algorithms were deployed. This is the vectorization phase. TF-IDF, Doc2Vec, and CountVectorizer were the three vectorization techniques used for the experiment.

3.6.1. TF-IDf

TF-IDF is described as Term Frequency-Inverse Document Frequency used to determine what phrases of a corpus might be favorable to use based on each phrase's document frequency (Lilleberg, Zhu, & Zhang, 2015). TF-IDF vectorization combines the weight of two statistics: term frequency and inverse document frequency. The higher TF-IDF values words suggest a more substantial relationship in the text where they appear (Chen & Wang, 2018).

Table 5
Sample representation texts into TF-IDF.

Texts	TF (Term Frequency)	IDF (Inverse Document Frequency)
[He is coughing] *[He is sneezing]* *[He got vaccinated last month]*	[0.33, 0.33, 0.33] [0.33, 0.33, 0.33] [0.20, 0.20, 0.20, 0.20, 0.20]	*'He'* $= \log(\frac{3}{3}) = 0$ *'is'* $= \log(\frac{3}{2}) = 0.1761$ *'vaccinated'* $= \log(\frac{3}{1}) = 0.4771$

$$tf\_idf(t,d) = tf(t,d)*idf(t) \tag{4}$$

where *t* is the number of term appears in the document *d*.

$$tf(t,d) = \frac{number\ of\ time\ term\ (t)\ appears\ in\ document\ (d)}{\textbf{total number of terms in documents}} \tag{5}$$

By designating the total number of documents in a collection as *N*, so we can distinguish t term of the inverse document frequency as:

$$idf(t) = \log\frac{N}{df(t)+1} \tag{6}$$

Table 5 is an example of vectorization using TF-IDF:Table 6..

3.6.2. Doc2Vec

The doc2vec vectorization is an extension of word2vec for training word based on two main algorithms: skip-gram model or Continuous Bag-of-words (CBOW) (Lilleberg et al., 2015). In CBOW, the order of the word in the document does not affect the prediction (Wu & Wang, 2017) because it forecasts the current word according to the situation of the continuous bag-of-words. On the other hand, the skip-gram model's goal is to forecast the context from the target word typically tries to forecast each word from its targeted words (Tafti et al., 2018). The skip-gram model has a training complexity architecture of:

$$Q = Cx(O + O\log_2(V)) \tag{7}$$

Below is an example of transforming texts into vector using Doc2Vec:

Data = ['Her','school','requires','the','students','to','get', 'vaccinated'].

'Her' = [1, 0, 0, 0, 0, 0, 0, 0] *'school'* = [0, 1, 0, 0, 0, 0, 0, 0] *'requires'* = [0, 0, 1, 0, 0, 0, 0, 0].

'the' = [0, 0, 0, 1, 0, 0, 0, 0] *'students'* = [0, 0, 0, 0, 1, 0, 0, 0] *'to'* = [0, 0, 0, 0, 0, 1, 0, 0].

'get' = [0, 0, 0, 0, 0, 0, 1, 0] *'vaccinated'* = [0, 0, 0, 0, 0, 0, 0, 1].

3.6.3. CountVectorizer

CountVectorizer is a technique that transforms a collection of texts to a vector form with the matrix of token counts (Ganesan, 2019). Preprocessing text, such as tokenization is done before transforming texts into vector representations. Each matrix column represents a unique word, while each matrix row represents a datapoint.

Below is an example of how word count is represented in a matrix:

Data = ['I','read','hundreds','of', 'horrendous','stories','of','COVID19′,'vaccine'].

3.7. Classification models

A classification model is a predictive modeling problem where a class label is predicted for a given example of input data. Classification requires a training dataset and calculates best-fit inputs to specific class

Table 6
Visual Representations Words Using CountVectorizer Sparse Matrix.

Doc | Word	I	read	hundreds	of	horrendous	stories	COVID19	vaccine
Data	1	1	1	2	1	1	1	1

labels. The training dataset must sufficiently represent the test and examples for each class label.

In this study, classification models, such as Random Forest, Decision Tree, Logistic Regression, LinearSVC, and Multinomial Naïve Bayes were used as learning algorithms. Labeling was done using the sentiment analysis result. Data was split into 70:30 for training and testing respectively. Labelling was done using the result of the sentiment analysis. Performance evaluation was based on accuracy, precision, recall, and F1 score.

Mathematically, accuracy is defined as a ratio of correctly predicted observations to the total observations. Precision is a ratio of correctly predicted positive observations to the total predicted positive observations. Recall (sensitivity) is a ratio of correctly predicted positive observations in actual class and F1 Score is a weighted average of Precision and Recall.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1\ Score = \frac{2^*(Recall^*Precision)}{Recall + Precision} \tag{11}$$

where TP (True Positives) are the correctly predicted positive values, TN (True Negatives) are the correctly predicted negative values, FP (False Positives) are the incorrectly predicted positive values, and FN (False Negatives) are the incorrectly predicted negative values.

3.7.1. Random Forest

Random Forest (RF) is a tree-based ensemble model which can be used for both regression and classification. It achieves a high accurate prediction by combining several poor learners (decision trees) of the training data and random feature selection (Li, Alam, & Melnokov, 2021), (Svetnik, Liaw, Tong, Culberson, Sheridan, & Feuston, 2003).

Random Forest is described by the following equations:

$$p = mode\{T_1(y), T_2(y), \cdots, T_m(y)\} \tag{12}$$

$$p = mode\{\sum_{m=1}^{m} T_m(y)\} \tag{13}$$

where p is the random forest final prediction, while $T_1(y)$, $T_2(y)$, $T_3(y)$, and $T_m(y)$ are the number of decision trees included in the classification process (Khan et al., 2021). For this experiment, number of estimators (trees) was set to 100 with a maximum depth of 5. All other hyperparameters are the default values in the sklearn library.

3.7.2. Decision Tree

A Decision Tree based its learning on decision that follows a tree like structure (Capozzoli, Cerquitelli, & Piscitelli, 2016). Each line in the tree specifies a test on an attribute, while every branch descending from the node corresponds to a possible value. Each leaf represents class labels associated with the instance (Tan, Steinbach, & Kumar, 2006). The line of code below was set in the sklearn for the minimum sample split. All other hyperparameters values were the default.

'Decision Tree': {'min_samples_split': [1, 2, 5]}.

3.7.3. LinearSVC

Linear SVC (Linear Support Vector Classification) is a support vector classification that supports dense and sparse inputs. This makes it more flexible in choosing penalties and loss functions. Linear SVC performs very well in larger dataset because it tends to be faster to converge with a more significant number of samples (Pedregosa et al., 2011). The hyperparameters' values used in the study were the default in sklearn; LinearSVC().

3.7.4. Logistic Regression

Logistic Regression implements a linear equation with explanatory variables to predict the response variable (Banerjee, 2019). Nonlinearity is introduced using a logistical function or logistical curve (Khan et al., 2021).

$$l = \log_b \frac{p}{1 - p} = \beta_{0+}\beta_1 x_1 + \beta_2 x_2 \cdots \beta_n x_n \tag{14}$$

where p is the estimated probability, β_0 is y-intercept, β_1 is the coefficient of variable x_1, β_1 is the coefficient of variable $x_{1,}$ β_2 is the coefficient of variable x_2 and β_n is the coefficient of variable $x_{n.}$ Default hyperparameters' value in sklearn was used for the study.

LogisticRegression(random_state = 0).

3.7.5. Multinomial Naïve Bayes

Naïve Bayes is a classification model based on Bayes' Theorem, assuming all predictors are independent. It assumes that the presence of a particular class feature is not related to others. Naive Bayes classifier is the simplest Bayesian network model, which is a simple probabilistic classifier based on applying Bayes' theorem with naïve (strong) independence assumptions among the features (Li et al., 2021). In this Naïve Bayes classifier, default hyperparameters' values in sklearn were used for the experiment; MultinomialNB().

4. Result

We examined the impact of stemming, lemmatization, and a combination of stem and lemmatization on the models' performances. The results of each method are presented in the following sections.

4.1. Stemming model performance

Table 7 represents classification models' performance using stemming. For this study *porter stemmer* was used for the experiment. Previous studies have shown the effectiveness of *porter stemmer* as compared to other stemmers (Lovins, Paice, Dowson, Krovetz, and Xerox Stemmer). It has been shown also that it has less error rate (Jivani, 2011). The major short coming of the porter stemmer lies in its inability to produce actual words. It is also very time-consuming. As shown in table 7, 12 out of 14 (85.71 %) TextBlob sentiment computation has the highest model accuracies (models 1 to 6).

As shown in the table, the best classification model performance is TextBlob + TF-IDF + LinearSVC with a model accuracy of 0.94860 and precision of 0.94572. The second-best model accuracy is 0.90724 with a precision of 0.91290 when we use TextBlob + CountVectorizer + TF-IDF + LinearSVC model classifier. On the other hand, the table shows the lowest model performance is Azure + Doc2Vec + Decision Tree classification with an accuracy of 0.41428 and precision of 0.38715.

The result of stemming is quite promising, with its highest accuracy of 0.94860 with a precision of 0.94572. However, we continued our

Table 7
Classification Model Sorted by Accuracy – Stemming Only.

No	Model Classifier	Accuracy	Precision	Recall	F1 Score
1	TextBlob + TF-IDF + LinearSVC	0.94860	0.94572	0.91426	0.92851
2	TextBlob + CountVect + TF-IDF + LinearSVC	0.90724	0.91290	0.86076	0.88257
3	TextBlob + TF-IDF + Logistic Regression	0.89942	0.91293	0.82187	0.85457
4	TextBlob + TF-IDF + Decision Tree	0.87967	0.84332	0.82302	0.83217
5	TextBlob + CountVect + TF-IDF + Logistic Regression	0.86658	0.89754	0.77938	0.81687
6	TextBlob + CountVect + TF-IDF + Decision Tree	0.86510	0.84021	0.81912	0.82871
7	VADER + TF-IDF + LinearSVC	0.85771	0.85701	0.84628	0.85064
8	VADER + CountVect + TF-IDF + LinearSVC	0.83433	0.83385	0.82236	0.82695
9	VADER + TF-IDF + Logistic Regression	0.82185	0.82538	0.80648	0.81331
10	VADER + CountVect + TF-IDF + Logistic Regression	0.81829	0.82543	0.80119	0.80935
11	TextBlob + TF-IDF + Naive Bayes	0.80783	0.86128	0.63985	0.63803
12	TextBlob + CountVect + TF-IDF + Naive Bayes	0.80006	0.82759	0.68320	0.70646
13	Azure + TF-IDF + LinearSVC	0.79416	0.73953	0.71010	0.71069
14	Azure + TF-IDF + Logistic Regression	0.79112	0.74902	0.69424	0.70158
15	VADER + CountVect + TF-IDF + MultinomialNB	0.74235	0.74274	0.73933	0.73746
16	VADER + TF-IDF + Naive Bayes	0.73400	0.74331	0.72418	0.72437
17	TextBlob + Doc2Vec + Logistic Regression	0.70831	0.66623	0.59899	0.61586
18	VADER + CountVect + TF-IDF + Decision Tree	0.70091	0.69891	0.68212	0.68738
19	Azure + TF-IDF + Naive Bayes	0.69439	0.69109	0.54774	0.54707
20	Azure + CountVect + TF-IDF + Logistic Regression	0.69219	0.65686	0.56371	0.57906
21	TextBlob + Doc2Vec + LinearSVC	0.68798	0.63012	0.62769	0.62855
22	Azure + TF-IDF + Decision Tree	0.67430	0.60593	0.59470	0.59970
23	Azure + CountVect + TF-IDF + LinearSVC	0.67349	0.60665	0.57705	0.58598
24	Azure + CountVect + TF-IDF + Naive Bayes	0.66983	0.64740	0.53447	0.54062
25	VADER + TF-IDF + Decision Tree	0.66636	0.66481	0.64463	0.65025
26	VADER + Doc2Vec + LinearSVC	0.62248	0.61704	0.60678	0.60861
27	VADER + Doc2Vec + Logistic Regression	0.61609	0.60898	0.60559	0.60613
28	Azure + CountVect + TF-IDF + Decision Tree	0.60986	0.53648	0.52407	0.52895
29	Azure + Doc2Vec + Logistic Regression	0.59179	0.50196	0.49374	0.49659
30	Azure + Doc2Vec + LinearSVC	0.58462	0.50421	0.49768	0.50053
31	TextBlob + Doc2Vec + Random Forest	0.57497	0.38007	0.40963	0.37739
32	Azure + CountVect + TF-IDF + Random Forest	0.57014	0.19005	0.33333	0.24208
33	Azure + TF-IDF + Random Forest	0.57009	0.19003	0.33333	0.24206
34	Azure + Doc2Vec + Random Forest	0.56983	0.46977	0.33562	0.24646
35	TextBlob + CountVect + TF-IDF + Random Forest	0.54171	0.51035	0.35524	0.27234
36	TextBlob + TF-IDF + Random Forest	0.52079	0.17360	0.33333	0.22830
37	TextBlob + Doc2Vec + Decision Tree	0.44092	0.39177	0.39412	0.38791
38	VADER + CountVect + TF-IDF + Random Forest	0.43750	0.47264	0.36124	0.25049
39	VADER + TF-IDF + Random Forest	0.43411	0.47563	0.34919	0.22971
40	VADER + Doc2Vec + Random Forest	0.42636	0.57728	0.35538	0.24341
41	VADER + Doc2Vec + Decision Tree	0.42238	0.41423	0.41439	0.41395
42	Azure + Doc2Vec + Decision Tree	0.41428	0.38715	0.39132	0.37913

investigation by exploring the lemmatization method.

4.2. Lemmatization model performance

Table 8 represents the model performance of lemmatization. We noticed that TextBlob sentiment computations have all the highest accuracies (models 1 to 6). In addition, when the first 14 models are considered, TextBlob occupy 10 places in performance; this is 71.43 %.

As shown in Table 8, we noticed that TextBlob sentiment with TF-IDF vectorization and LinearSVC model classifier (TextBlob + TF-IDF + LinearSVC) has the best model performance (0.96472 accuracies and 0.96529 precision). Follow closely is TextBlob sentiment with TF-IDF vectorization and Decision Tree model classifier (accuracy of 0.95701 and precision of 0.93806). In addition, the top 6 models used TextBlob text sentiment calculations.

We observed that the lemmatization methodology performance is higher than the stemming method (table 7). Switching from stemming to lemmatization improved the previous highest model classification (TextBlob + TF-IDF + LinearSVC) from accuracy of 0.94860 to 0.96472 and its precision of 0.94572 to 0.96529. The superiority of lemmatization over stemming agrees with (Balakrishnan & Lloyd-Yemoh, 2014).

Table 8 (lemmatization) is an improvement over table 7 (stemming). We believe that exploring lemmatization after potter stemming could produce an interesting result. Therefore, we continued our investigation by combining both stemming and lemmatization methodologies to produce *Stemming/Lemmatization* model. The result of our experiment shows the effectiveness of our strategy.

4.3. Stemming with lemmatization model performance

Table 9 represents the model performance of potter stemming before lemmatization. We noticed that TextBlob Sentiment Analysis performed better than the other two methods (VADER and Azure). We also observed that 10 out of 14 TextBlob precision and F1scores (about 71.43 %) have the higher scores compared with the other two methods. On the other hand, 4 out of 14 (about 28.57 %) VADER methods are higher than TextBlob.

Comparing recall scores, we noticed that 8 out of 14 (about 57.14 %) VADER methods have the highest scores, while 7 out of 14 (about 42.86 %) TextBlob methods have the highest scores. Consequently, TextBlob sentiment computations perform better than VADER or Azure sentiment computation. On the other hand, Azure Machine Learning has the lowest accuracy compared with other methods. The low accuracy could be because of an imbalanced dataset (Jobs, 2017) since the classification model assumes comparable data among classes. Furthermore, Azure Machine Learning requires users to operate manually, including data preprocessing, exploration, choosing methods, and model validation (Harfoushi et al., 2018). The accuracy of the logistic regression model is significantly less (Price, Masood, & Aroraa, 2021), while the multiclassification Random Forest achieves much higher accuracy.

According to Table 9, Accuracy Classification Model by combining stem and lemmatization, we notice that the highest accuracy is 0.96752

Table 8
Classification Model Sorted by Accuracy – Lemmatization Only.

No	Model Classifier	Accuracy	Precision	Recall	F1 Score
1	TextBlob + TF-IDF + LinearSVC	0.96472	0.96529	0.91967	0.94033
2	TextBlob + TF-IDF + Decision Tree	0.95701	0.93806	0.90851	0.92217
3	TextBlob + TF-IDF + Logistic Regression	0.91963	0.93903	0.82141	0.86574
4	TextBlob + CountVect + TF-IDF + LinearSVC	0.91939	0.92588	0.84231	0.87739
5	TextBlob + CountVect + TF-IDF + Decision Tree	0.89477	0.85294	0.84115	0.84682
6	TextBlob + CountVect + TF-IDF + Logistic Regression	0.88527	0.91888	0.76034	0.81441
7	VADER + TF-IDF + LinearSVC	0.86028	0.86027	0.85026	0.85432
8	TextBlob + TF-IDF + Naive Bayes	0.83832	0.82734	0.70023	0.72176
9	VADER + CountVect + TF-IDF + LinearSVC	0.83464	0.83457	0.82270	0.82742
10	VADER + TF-IDF + Logistic Regression	0.83283	0.83627	0.81997	0.82591
11	TextBlob + CountVect + TF-IDF + Naive Bayes	0.81953	0.85299	0.64357	0.68705
12	VADER + CountVect + TF-IDF + Logistic Regression	0.81837	0.82536	0.80133	0.80944
13	Azure + TF-IDF + LinearSVC	0.78727	0.72203	0.70326	0.70040
14	Azure + TF-IDF + Logistic Regression	0.78586	0.73274	0.68993	0.69308
15	VADER + TF-IDF + Naive Bayes	0.75479	0.76576	0.74469	0.74780
16	VADER + CountVect + TF-IDF + MultinomialNB	0.74359	0.74362	0.74066	0.73870
17	TextBlob + Doc2Vec + Logistic Regression	0.73160	0.63511	0.62151	0.62739
18	TextBlob + Doc2Vec + LinearSVC	0.72093	0.61981	0.62480	0.62094
19	VADER + CountVect + TF-IDF + Decision Tree	0.71571	0.71331	0.69836	0.70334
20	Azure + TF-IDF + Naive Bayes	0.69720	0.69272	0.55073	0.54801
21	Azure + CountVect + TF-IDF + Logistic Regression	0.69258	0.65917	0.56402	0.57927
22	Azure + CountVect + TF-IDF + LinearSVC	0.67569	0.60125	0.57755	0.58798
23	Azure + CountVect + TF-IDF + Naive Bayes	0.67038	0.65240	0.53717	0.54073
24	Azure + TF-IDF + Decision Tree	0.67033	0.60545	0.60138	0.60325
25	VADER + TF-IDF + Decision Tree	0.66028	0.65912	0.64025	0.64499
26	TextBlob + CountVect + TF-IDF + Random Forest	0.64779	0.54876	0.33926	0.27337
27	TextBlob + TF-IDF + Random Forest	0.64334	0.21445	0.33333	0.26099
28	TextBlob + Doc2Vec + Random Forest	0.64148	0.47299	0.33554	0.26484
29	VADER + Doc2Vec + LinearSVC	0.63564	0.62974	0.62185	0.62364
30	VADER + Doc2Vec + Logistic Regression	0.63556	0.63070	0.62037	0.62256
31	Azure + CountVect + TF-IDF + Decision Tree	0.60565	0.52988	0.51873	0.52310
32	Azure + Doc2Vec + Logistic Regression	0.58400	0.50405	0.50034	0.50205
33	Azure + Doc2Vec + LinearSVC	0.58377	0.50394	0.49726	0.50011
34	Azure + TF-IDF + Random Forest	0.57664	0.19221	0.33333	0.24383
35	Azure + CountVect + TF-IDF + Random Forest	0.57122	0.19275	0.33672	0.24402
36	Azure + Doc2Vec + Random Forest	0.56835	0.52277	0.33344	0.24180
37	TextBlob + Doc2Vec + Decision Tree	0.47286	0.39484	0.40676	0.38863
38	VADER + Doc2Vec + Random Forest	0.46810	0.45498	0.43408	0.41749
39	VADER + CountVect + TF-IDF + Random Forest	0.43820	0.80494	0.36185	0.25174
40	Azure + Doc2Vec + Decision Tree	0.41506	0.38244	0.38549	0.37552
41	VADER + Doc2Vec + Decision Tree	0.41280	0.40209	0.40374	0.40213
42	VADER + TF-IDF + Random Forest	0.41005	0.13668	0.33333	0.19387

(TF-IDF + LinearSVC, using TextBlob) with a precision of 0.96921. The next higher accuracy is 0.95771 (TextBlob + TF-IDF + Decision Tree) with an accuracy of 0.94101. On the other hand, the lowest model performance is Azure with Doc2Vec vectorizer and Decision Tree classification model (an accuracy of 0.41062 and precision of 0.37792). In addition, the top 6 highest model performances are when we are using TextBlob sentiment computation.

Combining stem and lemmatization has slightly increased the model performance. Its highest accuracy (TextBlob + TF-IDF + LinearSVC) has improved from 0.96472 to 0.96752, while its precision increases from 0.96529 to 0.96921. Stemmer has the advantage to work well on larger documents and is capable of removing the prefixes (Jivani, 2011), while lemmatization uses vocabularies and morphological analysis of words in addition to try to remove inflectional endings which help return the words to their dictionary forms (Balakrishnan & Lloyd-Yemoh, 2014). There is a possibility that combining both advantages help model classifications improve their performances.

5. Discussion

According to the three simulations (Stemming, lemmatization, and combining stemming with lemmatization), we noticed that the highest model performance is TextBlob with TF-IDF vectorization and LinearSVC classification model. In addition, lemmatization produces higher model performance than stemming text preprocessing. Next, model performances slightly improved when we combined both stemming and lemmatization. Furthermore, we observed that the top 6 highest model performances used TextBlob sentiment computation. Moreover, combining stem and lemmatization consistently indicates that TextBlob has better model classifier performances compared with VADER and azure sentiment computations.

The highest model performance in our experiment is 0.96752 with a precision of 0.96921 (TextBlob + TF-IDF + LinearSVC). This result is slightly higher than the previous research (0.9600) (Khan et al., 2021). The previous research used TextBlob sentiment computation Gradient Boosting Machine with TF-IDF vectorization. Using the LinearSVC classification model, the model's accuracy increased by 0.752 %. There is a possibility that changing the classification learning algorithm from Gradient Boosting Machine to LinearSVC (Linear Support Vector Classification) slightly improved the model's performance. In addition, using three different sentiment analyzers, our experiment shows that the top six highest classification model accuracies are based on TextBlob sentiment analysis. It suggests that TextBlob sentiment analyser performed better than the other two text sentiment computation methods.

Looking at the stemming with lemmatization model table (Table 9), seven out of the ten top highest classification model accuracies are TextBlob (Rank 1, 2, 3, 4, 5, 6, and 8), while VADER has three out of ten the top highest model accuracy (Rank 7, 9, and 10). None of Azure sentiment is among the top ten highest performance. The table also shows that four out of the top ten classification models are LinearSVC (Rank 1, 3, 7, and 9). In addition, three out of ten Logistic Regression classifications are in the top ten best performance (Rank 4, 6, and 10).

Table 9
Classification Model Sorted by Accuracy – Stemming with Lemmatization.

No	Sentiment Analysis + Classification Models	Accuracy	Precision	Recall	F1 Score
1	TextBlob + TF-IDF + LinearSVC	0.96752	0.96921	0.92807	0.94702
2	TextBlob + TF-IDF + Decision Tree	0.95771	0.94101	0.91367	0.92645
3	TextBlob + (CountVec + TF-IDF) + LinearSVC	0.92203	0.92852	0.84690	0.88146
4	TextBlob + TF-IDF + Logistic Regression	0.91857	0.93783	0.81502	0.86030
5	TextBlob + (CountVec + TF-IDF) + Decision Tree	0.89820	0.85951	0.84597	0.85241
6	TextBlob + (CountVec + TF-IDF) + Logistic Regression	0.88652	0.91933	0.76103	0.81541
7	VADER + TF-IDF + LinearSVC	0.86577	0.86464	0.85532	0.85913
8	TextBlob + TF-IDF + Naïve Bayes	0.84287	0.83771	0.69200	0.71643
9	VADER + (CountVec + TF-IDF) + LinearSVC	0.83916	0.83983	0.82849	0.83309
10	VADER + TF-IDF + Logistic Regression	0.83061	0.83398	0.81574	0.82211
11	VADER + (CountVec + TF-IDF) + Logistic Regression	0.82234	0.83018	0.80588	0.81416
12	TextBlob + (CountVec + TF-IDF) + Naïve Bayes	0.82094	0.85339	0.64189	0.68564
13	Azure + TF-IDF + LinearSVC	0.79416	0.73953	0.71010	0.71069
14	Azure + TF-IDF + Logistic Regression	0.79112	0.74902	0.69424	0.70158
15	VADER + (CountVec + TF-IDF) + Decision Tree	0.78386	0.78037	0.77047	0.77434
16	VADER + TF-IDF + Naïve Bayes	0.75339	0.75657	0.74880	0.74753
17	VADER + (CountVec + TF-IDF) + Naïve Bayes	0.75278	0.75287	0.75056	0.74870
18	VADER + TF-IDF + Decision Tree	0.74638	0.74446	0.73001	0.73478
19	TextBlob + Doc2Vec + Logistic Regression	0.74320	0.65226	0.61604	0.63075
20	TextBlob + Doc2Vec + LinearSVC	0.72459	0.62300	0.62512	0.62260
21	Azure + TF-IDF + Naïve Bayes	0.69439	0.69109	0.54774	0.54707
22	Azure + (CountVec + TF-IDF) + Logistic Regression	0.69242	0.65825	0.56329	0.57861
23	Azure + TF-IDF + Decision Tree	0.67430	0.60593	0.59470	0.59970
24	Azure + (CountVec + TF-IDF) + LinearSVC	0.67349	0.60665	0.57705	0.58598
25	Azure + (CountVec + TF-IDF) + Naïve Bayes	0.66983	0.64740	0.53447	0.54062
26	TextBlob + (CountVec + TF-IDF) + Random Forest	0.65013	0.54954	0.33926	0.27393
27	TextBlob + TF-IDF + Random Forest	0.64428	0.21476	0.33333	0.26122
28	TextBlob + Doc2Vec + Random Forest	0.64273	0.45865	0.33449	0.26311
29	VADER + Doc2Vec + Logistic Regression	0.63642	0.63759	0.60798	0.61244
30	VADER + Doc2Vec + LinearSVC	0.63330	0.62823	0.61528	0.61781
31	Azure + (CountVec + TF-IDF) + Decision Tree	0.60597	0.53074	0.52005	0.52421
32	Azure + Doc2Vec + Logistic Regression	0.58899	0.50428	0.49589	0.49923
33	Azure + Doc2Vec + LinearSVC	0.58439	0.50055	0.49383	0.49673
34	Azure + (CountVec + TF-IDF) + Random Forest	0.57014	0.19005	0.33333	0.24208
35	Azure + TF-IDF + Random Forest	0.57009	0.19003	0.33333	0.24206
36	Azure + Doc2Vec + Random Forest	0.56835	0.35612	0.33344	0.24182
37	VADER + Doc2Vec + Random Forest	0.47831	0.45927	0.44226	0.43626
38	TextBlob + Doc2Vec + Decision Tree	0.46055	0.39063	0.40317	0.38169
39	VADER + (CountVec + TF-IDF) + Random Forest	0.44069	0.47452	0.36439	0.25630
40	VADER + TF-IDF + Random Forest	0.42862	0.47316	0.35301	0.23417
41	VADER + Doc2Vec + Decision Tree	0.42846	0.41947	0.42036	0.41929
42	Azure + Doc2Vec + Decision Tree	0.41062	0.37792	0.38044	0.37077

Furthermore, two out of ten Decision Tree accuracies (Rank 2 and 5) are in the top ten highest performance. Three out of ten Logistic Regression accuracies (Rank 4, 6, and 10). Lastly, there is only one Naïve Bayes (Rank 8) in the top ten highest performance. LinearSVC contributes 40 % among the top ten classification models, while Logistic Regression model classifier adds 30 % in the top ten performance. In the top ten model classifier performance, about 20 % of Decision Tree model classifiers are in the top ten model performance, and only 10 % of Naïve Bayes are in the top ten classification models. We also observed that LinearSVC with TF-IDF vectorization (No. 1) achieves an accuracy of 0.96752 with a 0.96921 precision score. When we combine CountVectorizer and TF-IDF for LinearSVC classification (No. 3), the accuracy goes down to 0.92203 with a 0.92852 precision score.

Furthermore, the Decision Tree model classifier vectorized with TF-IDF has an accuracy of 0.95771 with a precision score of 0.94101 (No. 2). However, when TF-IDF was combined with CountVectorizer on the Decision Tree classifier, the accuracy goes down to 0.89820 with a precision score of 0.85951 (No. 5). In addition, Logistic Regression vectorized with TF-IDF has an accuracy of 0.91857 (No. 4) with a precision score of 0.93783. Nevertheless, when we combine TF-IDF and CountVectorizer vectorizations on Logistic Regression classification, the accuracy goes to 0.82234 with a precision score of 0.83018. According to the classification model accuracies, we can conclude that combining both CountVectorizer + TF-IDF causes a lower accuracy compared with TF-IDF vectorization only. There is a possibility that combining both vectorizations could increase the inputs' complexity, which influences the model's ability to learn the datasets.

6. Contributions

We would like to highlight the following contributions of this study to COVID-19 vaccine hesitancy using twitter dataset.

- We designed, developed, and evaluated 42 classification models. Vocabulary normalization was based on potter stemming, lemmatization, and potter stemming with lemmatization. For each vocabulary normalization strategy, we built 42 different classifiers. The result of our experiment shows an improvement over previous works. For instance, Raza, Butt, Latif, and Wahid (2021) used CountVectorizer and TF-IDF with Logistic Regression, Multinomial Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine, and XGBoost using two classes (positive and negative). The best accuracy of their work was 93.17 %. Wang and Kwok (2021) built Doc2Vec and K-means clustering without presenting accuracy. Shamrat et al. (2021) used KNN with three classes (positive, neutral, and negative) without model accuracy. Secondly, some previous studies only performed sentiment analysis without applying classification modeling. As shown in Table 1, Mishra et al. (2021), Lyu, Han, and Luli (2021), Luo and Kejriwal (2021), Ma, Zeng-Treitler, and Nelson (2021), Muric et al. (2021), and Ansari and Khan (2021) did not apply classification modeling in their research studies; these studies only analyzed vaccine hesitancy by computing text sentiment scores of the tweets.

- Researchers, such as Naseem et al. (2021), performed classification modeling (accuracy 0.82) but did not compute sentiment scores. Raza et al. (2021) removed neutral tweets from the classification model; used only positive and negative tweets in their discussions. To the best of our knowledge, as shown in the summary table (Table 1) the only previous research (2021) closest to the result of our experiment achieved the highest classification model accuracy of 0.9600. Gradient Boosting Machine was used as a learning algorithm in the study. In this study, we experimented with three different sentiment analytical methods (VADER, TextBlob and Azure), three different vocabulary normalization strategies (potter stemming, lemmatization and potter stemming with lemmatization), three different vectorization methodologies (CountVec, TF-IDF and Doc2Vec). Our experiment achieved the highest model accuracy of 0.96752 with a precision score of 0.96921.
- We experimented by feeding the vectorized text of CountVectorizer as input into TF-IDF to undergo another vectorization. This methodology reduced models' performance. A drop in the accuracy might relate to an increased complexity. This may be because using both vectorization methodologies cause text representations to become more complex, eventually resulting in lower classification performances. Increased complexity might have had a significant influence on the capability of the model to learn the data effectively. Consequently, using standalone TF-IDF vectorization with model classifiers tends to have a better performance than the combination of vectorizations.
- Our analysis suggests that the daily positive sentiments has slightly increased over time. There is a possibility that the public have started feeling positive and optimistic about Covid-19 vaccine. Our inference agrees with other studies, such as Kirzinger et al. (2021). Researchers have shown that individuals who are hesitant to get vaccinated will likely change position after realizing the effectiveness of the vaccine (Cavanaugh et al., 2021) and (Limaye et al., 2021).

7. Conclusion

The worldwide Covid-19 pandemic, first reported in Wuhan, China, in 2019, has caused tremendous stress to the global community. Schools were closed or switched to online to prevent the spread of the contagious virus. Due to mask mandate, many businesses have been shut down, and other companies could not operate normally. Since the introduction of COVID-19 vaccines, some people hesitated to get vaccinated because of controversies, conspiracies, myths, or safety issues, while other groups supported vaccination as an effective mitigation strategy. Researchers have been examining the level of Covid-19 vaccine hesitancy as part of the effort to combat the infectious virus. Pros and cons about getting vaccinated have become an essential topic on social media platforms. Consequently, social media platforms have been a primary data source about individual opinions on the Covid-19 vaccine hesitancy.

The result of our study shows that Covid-19 vaccine hesitancy is gradually decreasing over time, suggesting that societies' positive opinions on getting vaccinated have gradually increased. This suggests that there is a positive feeling about COVID-19 vaccination. Daily graphs on positive sentiments continue to increase, while neutral and negative sentiment graphs decrease. Some individuals voluntarily opt to get vaccine boosters because of positive feelings about the Covid-19 vaccination.

Finally, comparing various classification models, our experiment shows that LinearSVC using TextBlob sentiment computation with TF-IDF vectorization achieved the highest model accuracy of 0.96752, which is higher than the previous research study Gradient Boosting Machine model classifier. However, combining two vectorizations (CountVectorizer and LinearSVC) causes classification model accuracy to decrease this may be due to increased complexity.

8. Future work

To improve sentiment analysis of the public tweets, we need to increase the number of public tweets data. So, we keep recording live streaming public tweets daily using similar keywords. In addition, we will use neural network modeling in the future for a better classification model. We will test the dataset using CNN, LSTM, and BERT models.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported in part by the National Science Foundation grant number 2032345.

References

Ansari, M., & Khan , N. (2021). Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content. *ELECTRON J GEN MED, 18*(6). doi:10.29333/ejgm/11316.

Baj-Rogowska, A. (2021). Mapping of the Covid-19 Vaccine Uptake Determinants From Mining Twitter Data. *IEEE Xplore*. https://doi.org/10.1109/access.2021.3115554

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances.

Banerjee, P. (2019). Logistic Regression Classifier Tutorial. Retrieved from https://www.kaggle.com/prashant111/logistic-regression-classifier-tutorial.

Bonnevie, E., Gallegos-Jeffrey, A., Goldbarg, J., Byrd, B., & Smyser, J. (2020). Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. *Journal of Communication in Healthcare*. https://doi.org/10.1080/17538068.2020.1858222

Capozzoli, A., Cerquitelli, T., & Piscitelli, M. (2016). Chapter 11 - Enhancing energy efficiency in buildings through innovative data analytics technologies. ScienceDirect. doi:10.1016/b978-0-12-803663-1.00011-5.

Cavanaugh, A. M., Spicer, K. B., Thoroughman, D., Glick, C., & Winter, K. (2021). *Reduced Risk of Reinfection with SARS-CoV-2 After COVID-19 Vaccination*. Kentucky: CDC.

CDC. (2021). *Interim List of Categories of Essential Workers Mapped to Standardized Industry Codes and Titles.* CDC. Retrieved from https://www.cdc.gov/vaccines/covid-19/categories-essential-workers.html.

Chen, D., & Wang, H. (2018). Research on Short Text Classification Algorithm Based on Neural Network. *IEEE Xplore*. https://doi.org/10.1109/imcec.2018.8469376

Chiou, L., & Tucker, C. (2018). Fake News and Advertising on Social Media: A Study of the Anti-Vaccination Movement. *SSRN*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3209929.

Douglas, K., & Joelley, D. (2014). The Effects of Anti-Vaccine Conspiracy Theories on Vaccination Intentions. *Plos One*. https://doi.org/10.1371/journal.pone.0089177

Ganesan, K. (2019). 10+ Examples for Using CountVectorizer. Retrieved from https://kavita-ganesan.com/how-to-use-countvectorizer/#.YbvN7WjMI2x.

Hagan, K., Forman, R., Mossialos, E., Ndebele, P., & Hyder, A. (2021). COVID-19 vaccine mandate for healthcare workers in the United States: A social justice policy. *Taylor & Francis Online*. https://doi.org/10.1080/14760584.2022.1999811

Harfoushi, O., Hasan, D., & Obiedat, R. (2018). Sentiment Analysis Algorithms through Azure Machine Learning: Analysis and Comparison. *Modern Applied Science, 12*(7). https://doi.org/10.5539/mas.v12n7p49

Igoe, K. J. (2019). Establishing the Truth: Vaccines, Social Media, and the Spread of Misinformation. Retrieved from *Executive and Continuing Professional Education* https://www.hsph.harvard.edu/ecpe/vaccines-social-media-spread-misinformation/.

Islam, M., Kamal, A.-H., Kabir, A., Southern, D., Khan, S., Hasan, S., … Seale, H. (2021). COVID-19 vaccine rumors and conspiracy theories: The need for cognitive inoculation against misinformation to improve vaccine adherence. *Plos One*. https://doi.org/10.1371/journal.pone.0251605

Jackson, J., Weiss, M., Schwarzenberg, A., Nelson, R., Sutter, K., & Sutherland, M. (2021). Global Economic Effects of COVID-19. Retrieved from *Congressional Research Service* https://sgp.fas.org/crs/row/R46270.pdf.

Jivani, A. G. (2011). A Comparative Study of Stemming Algorithms. *International Journal Comp. Tech. Application*.

Jobs, A. (2017). Imbalanced Data : How to handle Imbalanced Classification Problems. Retrieved from https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/.

Jones, L., Palumbo, D., & Brown, D. (2021). Coronavirus: How the pandemic has changed the world economy. Retrieved from *BBC News* https://www.bbc.com/news/business-51706225.

Khan, R., Rustam, F., Kanwal, K., Mehmood, A., & Choi, G. (2021). US Based COVID-19 Tweets Sentiment Analysis Using TextBlob and Supervised Machine Learning Algorithms. *IEEE*. https://doi.org/10.1109/ICAI52203.2021.9445207
Kirzinger, A., Sparks, G., & Brodie, M. (2021). KFF COVID-19 Vaccine Monitor. *Their Own Words, Six Months Later*. KFF.
Krause, P., Fleming, P., Peto, P., Longini, P., Figueroa, P., & Sterne, P. (2021). Considerations in boosting COVID-19 vaccine immune responses. *The Lancet*. https://doi.org/10.1016/S0140-6736(21)02046-8
Kumar, A., Singh, R., Kaur, J., Pandey, S., Sharma, V., Thakur, L., … Kumar, N. (2021). Wuhan to World: The COVID-19 Pandemic. *Frontiers*.
Kwon, S., Joshi, A., Lo, C.-H., Drew, D., Nguyen, L., Guo, C.-G., … Chan, A. (2021). Association of social distancing and face mask use with risk of COVID-19. *Nature Communications*. https://doi.org/10.1038/s41467-021-24115-7
Li, J., Alam, H., & Melnokov, O. (2021). An Evaluation of Tweet Sentiment Classification Methods. *IEEE Xplore*. https://doi.org/10.1109/csci51800.2020.00057
Lilleberg, J., Zhu, Y., & Zhang, Y. (2015). Support vector machines and Word2vec for text classification with semantic features. *IEEE Xplore*. https://doi.org/10.1109/icci-cc.2015.7259377
Limaye, R., Stuar, E., & Sell, T. K. (2021). *How Can I Talk to My Friends and Family About Getting Vaccinated for COVID-19?* Johns Hopkins Bloomberg School of Public Health.
Liu, S., & Liu, J. (2021). Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis. *Elsevier, 5499–5505*. https://doi.org/10.1016/j.vaccine.2021.08.058
Loomba, S., Figueiredo, A.d., Piatek, S. J., Graaf, K. D., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Retrieved from *Nature Human Behaviour* https://www.nature.com/articles/s41562-021-01056-1.
Luo, Y., & Kejriwal, M. (2021). Understanding COVID-19 Vaccine Reaction through Comparative Analysis on Twitter. *arxiv*. Retrieved from https://arxiv.org/abs/2111.05823.
Lyu, J., Han, E., & Luli, G. (2021). COVID-19 Vaccine-Related Discussion on Twitter: Topic Modeling and Sentiment Analysis. *Journal of Medical Internet Research, 23*(6). https://doi.org/10.2196/24435
Ma, P., Zeng-Treitler, Q., & Nelson, S. (2021). USE OF TWO TOPIC MODELING METHODS TO INVESTIGATE COVID VACCINE HESITANCY. Retrieved from *International Conferences ICT, Society, and Human Beings*. https://www.ict-conf.org/wp-content/uploads/2021/07/04_202106C030_Ma.pdf.
Marcec, R., & Likic, R. (2021). Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. *Postgraduate Medical Journal*. https://doi.org/10.1136/postgradmedj-2021-140685
McClain, C., Vogels, E., Perrin, A., Sechopoulos, S., & Rainie, L. (2021). The Internet and the Pandemic. *Paw Research Center*.
Michaud, J., & Kates, J. (2020). Distributing a COVID-19 Vaccine Across the U.S. - A Look at Key Issues. *KTF*.
Mishra, A., Wajid, M. S., & Dugal, U. (2021). A Comprehensive Analysis of Approaches for Sentiment Analysis Using Twitter Data on COVID-19 Vaccine. *Journal of Informatics Electrical and Electronics Engineering (JIEEE)*. https://doi.org/10.54060/jieee/002.02.009
Mohan, B.S., & Nambiar, V. (2020). COVID-19: An Insight into SARS-CoV-2 Pandemic Originated at Wuhan City in Hubei Province of China. ISSN: 2474-3658.
Muric, G., Wu, Y., & Ferrara, E. (2021). COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Dataset of Anti-vaccine Content, Vaccine Misinformation and Conspiracies. *arxiv*. Retrieved from https://arxiv.org/abs/2105.05134.
Naeem, S., Bhatti, R., & Khan, A. (2020). An exploration of how fake news is taking over social media and putting public health at risk. *Health Information & Libraries Journal*. https://doi.org/10.1111/hir.12320
Naseem, U., Khushi, M., Kim, J., & Dunn, A. (2021). Classifying vaccine sentiment tweets by modelling domain-specific representation and commonsense knowledge into context-aware attentive GRU. *arxiv*. Retrieved from https://arxiv.org/abs/2106.09589.
OECD. (2020). The impact of COVID-19 on student equity and inclusion: Supporting vulnerable students during school closures and school re-openings. *Organisation for Economic Co-operation and Development*. Retrieved from https://www.oecd.org/coronavirus/policy-responses/the-impact-of-covid-19-on-student-equity-and-inclusion-supporting-vulnerable-students-during-school-closures-and-school-re-openings-d593b5c8/.
Oliver, S., Gargano, J., Marin, M., Wallace, M., Curran, K., Chamberland, M., … Dooling, K. (2020). The Advisory Committee on Immunization Practices' Interim Recommendation. *MMWR.*. https://doi.org/10.15585/mmwr.mm6950e2
Oyebode, O., Ndulue, C., Adib, A., Mulchandani, D., Suruliraj, B., Orji, F., … Orji, R. (2021). Health, Psychosocial, and Social Issues Emanating From the COVID-19 Pandemic Based on Social Media Comments: Text Mining and Thematic Analysis Approach. Retrieved from *JMIR Medical Informatics, 6*(4) https://medinform.jmir.org/2021/4/e22734.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Retrieved from *Journal of Machine Learning Research, 12*(85), 2825–2830 https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.
Pfizer (2021). Real-World Evidence Confirms High Effectiveness of Pfizer-BioNTech COVID-19 Vaccine and Profound Public Health Impact of Vaccination One Year After Pandemic Declared. Pfizer. Retrieved from https://www.pfizer.com/news/press-release/press-release-detail/real-world-evidence-confirms-high-effectiveness-pfizer.
Piedrahita-Valdés, H., Piedrahita-Castillo, D., Bermejo-Higuera, J., Guillem-Saiz, P., Bermejo-Higuera, J., Guillem-Saiz, J., & Machío-Regidor, F. (2021). Vaccine Hesitancy on Social Media: Sentiment Analysis from June 2011 to April 2019. *MDPI*. https://doi.org/10.3390/vaccines9010028
Piltch-Loeb, R., & DiClemente, R. (2020). The Vaccine Uptake Continuum: Applying Social Science Theory to Shift Vaccine Hesitancy. *MDPI*. https://doi.org/10.3390/vaccines8010076
Price, A., Masood, A., & Aroraa, G. (2021). Azure Machine Learning. In: Hands-on Azure Cognitive Services. doi:10.1007/978-1-4842-7249-7_10.
Raza, G., Butt, Z., Latif, S., & Wahid, A. (2021). Sentiment Analysis on COVID Tweets: An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models. *IEEE Xplore*. https://doi.org/10.1109/ICoDT252288.2021.9441508
Reiss, D. R., & Caplan, A. L. (2020). Considerations in mandating a new Covid-19 vaccine in the USA for children and adults. *Journal of Law and the Biosciences, 8*(2). https://doi.org/10.1093/jlb/lsaa025
Shamrat, F. M., Chakraborty, S., Imran, M. M., Muna, J. N., Billah, M. M., Das, P., & Rahman, M. O. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. *Indonesian Journal of Electrical Engineering and Computer Science, 23*(1), 463–470.
Shimabukuro, T. (2020). Allergic Reactions Including Anaphylaxis After Receipt of the First Dose of. *MMWR.*. https://doi.org/10.15585/mmwr.mm7002e1
Silva-Cayetano, A., Foster, W., Innocentin, S., Gilbert, S., Lambe, T., & Linterman, M. (2020). A booster dose enhances immunogenicity of the COVID-19 vaccine candidate ChAdOx1 nCoV-19 in aged mice. *Clinical and Translational Artcle*. https://doi.org/10.1016/j.medj.2020.12.006
Svetnik, V., Liaw, A., Tong, C., Culberson, J., Sheridan, R., & Feuston, B. (2003). *Journal of Chemical Information and Computer Sciences*, 1947 - 1958. doi:10.1021/ci034160g.
Tafti, A., Behravesh, E., Assefi, M., LaRose, E., Badger, J., Mayer, J., … Peissig, P. (2018). bigNN: An open-source big data toolkit focused on biomedical sentence classification. *IEEE Xplore*. https://doi.org/10.1109/BigData.2017.8258394
Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Retrieved from. Pearson Addison-Wesley https://www-users.cse.umn.edu/~kumar001/dmbook/sol.pdf.
Tao, D., Yang, P., & Feng, H. (2020). Utilization of text mining as a big data analysis tool for food science and nutrition. *Comprehensive Reviews in Food Science and Food Safety, 19*(2), 875–894. https://doi.org/10.1111/1541-4337.12540
Twittter (2021). Overview of the different authentication methods. Retrieved December 2, 2021, from https://developer.twitter.com/en/docs/tutorials/authenticating-with-twitter-api-for-enterprise/authentication-method-overview.
Wang, G., & Kwok, S. (2021). Using K-Means Clustering Method with Doc2Vec to Understand the Twitter Users' Opinions on COVID-19 Vaccination. *IEEE Xplore*. https://doi.org/10.1109/BHI50953.2021.9508578
Wong, A., Ho, S., & Olusanya, O. (2020). The use of social media and online communications in times of pandemic COVID-19. *Journal of the Intensive Care Society*. https://doi.org/10.1177/1751143720966280
Wu, C., & Wang, B. (2017). Extracting Topics Based on Word2Vec and Improved Jaccard Similarity Coefficient. *IEEE Xplore*. https://doi.org/10.1109/DSC.2017.70