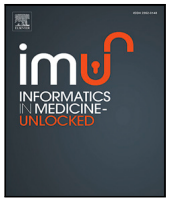
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Informatics in Medicine Unlocked 31 (2022) 100969

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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



COVID-19 analytics: Towards the effect of vaccine brands through analyzing public sentiment of tweets

Khandaker Tayef Shahriar [a], Muhammad Nazrul Islam [b], Md. Musfique Anwar [c], Iqbal H. Sarker [a,*]

[a] Department of Computer Science and Engineering, Chittagong University of Engineering & Technology, Chittagong 4349, Bangladesh
[b] Department of Computer Science and Engineering, Military Institute of Science and Technology, Dhaka 1216, Bangladesh
[c] Jahangirnagar University, Dhaka, Bangladesh

ARTICLE INFO

Keywords:
Data analytics
Covid-19 vaccine
Tweet
Sentiment analysis
Deep learning

ABSTRACT

The COVID-19 outbreak has created effects on everyday life worldwide. Many research teams at major pharmaceutical companies and research institutes in various countries have been producing vaccines since the beginning of the outbreak. There is an impact of gender on vaccine responses, acceptance, and outcomes. Worldwide promotion of the COVID-19 vaccine additionally generates a huge amount of discussions on social media platforms about diverse factors of vaccines including protection and efficacy. Twitter is considered one of the most well-known social media platforms which have been widely used to share a public opinion on vaccine-related problems in the COVID-19 pandemic. However, there is a lack of research work to analyze the public perception of COVID-19 vaccines systematically from a gender perspective. In this paper, we perform an in-depth analysis of the coronavirus vaccine-related tweets to understand the people's sentiment towards various vaccine brands corresponding to the gender level. The proposed method focuses on the effect of COVID-19 vaccines on gender by taking into account *descriptive, diagnostic, predictive, and prescriptive* analytics on the Twitter dataset. We also conduct experiments with deep learning models to determine the sentiment polarities of tweets, which are positive, neutral, and negative. The results reveal that LSTM performs better compared to other models with an accuracy rate of 85.7%.

1. Introduction

The outbreak of the novel Coronavirus Disease (COVID-19) has hit the earth drastically and billions of people are affected worldwide. According to a recent report by the World Health Organization (WHO), more than 260 million people have been infected with the virus, and over 5 million deaths are caused by the COVID-19.[1] COVID-19 has affected public life extensively and compared to past epidemics such as the Spanish Flu of 1918 and the Black Death of the 13th century, it is considered to be the most serious epidemic of this century. [1]. Now it has become a common goal for all countries to eliminate COVID-19 and return to normal life activities. Many countries officially have followed a specific series of measures to lessen the spread of COVID-19 including closing borders, reducing work in public locations (e.g., gyms, restaurants, shopping malls), learning from homes, limiting travel, and wearing masks, maintaining social distance, and personal hygiene. These measures have had a positive effect on controlling the spread of the virus. However, the latest COVID-19 Omicron variant put the countries around the world again in concern [2]. The presence of COVID-19 may be durable in the future and vaccination is the most effective long time way to handle the COVID-19 pandemic situation. Thus, the countries and pharmaceutical companies around the world have begun the development of vaccines and clinical trials from the starting of the pandemic outbreak.

As of December 01, 2021, there are 160 vaccine candidates and 24 of them have been approved by various countries around the world.[2] At first, Pfizer/BioNTech vaccine received emergency validation by WHO on December 31, 2020, and then AstraZeneca, Covishield, Janssen, Moderna.[3] Each country provides approval for multiple vaccinations and evolves precise guidelines to inspire all citizens to be vaccinated. In most countries, current vaccination rates have no longer yet reached the minimal standards for restricting the expansion of the pandemic. People's lack of knowledge and belief in vaccines, and skeptical attitudes towards vaccines are the possible reasons for low vaccination

* Corresponding author.
E-mail address: iqbal@cuet.ac.bd (I.H. Sarker).

[1] https://covid19.who.int/.
[2] https://covid19.trackvaccines.org/.
[3] https://www.who.int/news/item/31-12-2020-who-issues-its-first-emergency-use-validation-for-a-covid-19-vaccine-and-emphasizes-need-for-equitable-global-access.

https://doi.org/10.1016/j.imu.2022.100969
Received 18 February 2022; Received in revised form 11 May 2022; Accepted 12 May 2022
Available online 20 May 2022

rates in many countries. They fear that if the vaccine is not adequately tested, it could lead to chronic disease. The spread of incorrect information about COVID-19 on social media platforms is another practical reason, which can discourage skeptics. So far, the COVID-19 pandemic has been very gendered. The current COVID-19 pandemic has primary and secondary effects related to gender differences. The primary effects include differences between men and women in case fatality, while secondary effects involve differences in social and economic outcomes as a result of the pandemic, including the risk of domestic violence [3], economic and employment insecurity, and increased workload [4]. More men than women die from COVID-19 and the behaviors associated with COVID-19 are different in all genders [4]. Women are more compliant with public health rules set in during the pandemic [5] but are less concerned about COVID-19 if they are doing economic work that puts them at a higher risk of infection [6]. Thus, gender-based analysis of various vaccine brands is important to consider the social impact, side effects, and gender-specific suitability of vaccination. In this paper, we examine gender differences in attitudes and expected behavior regarding COVID-19 vaccines and gender perceptions in various vaccines brands approved by different countries by mining Twitter data in terms of sentiment analysis.

Social media platforms (e.g., Facebook, Twitter, Instagram) and online forums (e.g., Kaggle, StackOverflow, Yahoo) offer researchers with a simple and reliable repository of information [7–10]. People can post freely, comment on a specific post or interact with others on these platforms [11,12]. Twitter is the world's fastest-growing and one of the largest social media site [13]. Therefore, the COVID-19 vaccine-related discussion on Twitter helps us to detect people's concerns about the impact of vaccination. This paper investigates the COVID-19 vaccine-associated tweets on Twitter and detects sentiment polarities in tweets. Tweets represent the direction of public opinion towards popular vaccine brands. Moreover, identifying attitudes of people about various vaccines and conducting sentiment analysis can help policymakers and government officials to understand people's concerns and take appropriate action according to the necessity of proper vaccination, which could also be a concern for developing today's smart cities as well [14].

Twitter content has contributed to the creation of a huge volume of unstructured data [15,16]. The most common applications for Twitter content are topic discovery, data analysis, opinion mining, and sentiment analysis. Deep learning is one of the most common techniques to analyze Twitter data [17]. However, effective strategies and analytical methods for handling the ever-growing data generated by the Twitter platform are in great demand [18]. This paper introduces a method for analyzing and predicting sentiments from the perspective of various COVID-19 vaccine brands to evaluate the effects of vaccines on gender. This method uses data analysis techniques and conducts four kinds of data analysis, which include descriptive, diagnostic, predictive, and prescriptive analysis [19]. Useful information on Twitter datasets can be extracted by implementing the proposed framework to perform an in-depth analysis of the COVID-19 pandemic. The main contributions of this research paper are as follows:

- We have analyzed gender differences in attitudes and behaviors regarding the COVID-19 vaccine by investigating Twitter data.
- We have introduced a new framework based on data analytics to evaluate different types of vaccine brands from the gender perspective and predict sentiments of COVID-19 tweets.
- Descriptive, diagnostic, predictive, and prescriptive data analytics strategies are explored to generate COVID-19 vaccine-related insightful information for providing decision support to physicians, experts, and policymakers.
- The prescriptive analysis is provided for both males and females, using the consequences of predictive evaluation.

The work mentioned in this paper is organized as follows: We review the related works in Section 2. Section 3 provides a summary of the datasets used in this work. A data analytics-based framework is presented to predict sentiment polarities in Section 3. Details about the experimental results are described in Section 4. A general discussion of the method discussed above is given in Section 5. Conclusion and future works are provided in Section 6.

2. Related work

Structured data analysis of social media platforms can be broadly used in emerging public health-related issues [20,21]. The COVID-19 outbreak enables a lot of discussion on social media platforms every minute. This type of massive quantity of discussions in form of posts and comments provides an important source of data to get the attention of the researchers [22–25]. There has been some research into the COVID-19 pandemic tweets for analysis and mining of useful information hidden within the post.

Some research work has analyzed sentiment predictions as a tool to investigate the human response to the pandemic through their participation in social media. Li et al. [26] analyzed American–Chinese posts on Twitter and Weibo to evaluate emotions and presented sharp variations in different cultures in how humans perceive the COVID-19 situation. Stella et al. [27] explored the psychological and social consequences in Italy due to the lockdown. They found the presence of complex sentiments where there was unity, faith, and hope with fear and anger. To analyze how people from different regions reacted to the pandemic situation, Dubey [28] analyzed the sentiments in COVID-19 tweets from 12 countries. The results showed the existence of grief, fear, and disgust among people around the world with confidence to deal with the pandemic environment. Zhou et al. [29] analyzed sentiments during the pandemic by investigating COVID-19-related tweets in New South Wales (NSW) in Australia. They observed a dynamic change in mood over time by classifying tweets by local government areas (LGA). Yin et al. [30] introduced a new framework to perform topic and sentiment analysis in COVID-19 tweets. They found a higher rate of positive tweets than negative tweets, which was compatible with other similar works. Imran et al. [31] proposed an LSTM based model with multiple layers for distinguishing sentiment variations. Satu et al. [32] introduced TClustVID which scans COVID-19 public tweets for positive, neutral, and negative sentiment analysis purposes.

In addition to the growing popularity of the vaccine, many researchers have worked on social media content about the COVID-19 vaccine. Kwok et al. [33] extracted topics and sentiments related to COVID-19 vaccines from Australian Twitter users. They categorized each tweet for positive and negative sentiments along with eight emotions such as fear, anger, trust, anticipation, surprise, joy, sadness, and disgust. They found the sentiment polarities where two-thirds of all tweets were positive and one-third were negative. Lyu et al. [34] used similar methods like [33] for topic identification and sentiment analysis of the COVID-19 vaccine-related public discussions on social media to better understand community concerns, perceptions, and feelings that may help to achieve the goals of immunization. Protests against vaccination in the United States during the COVID-19 pandemic led to an increase in Twitter conversations that were measured by Bonnevie et al. [35]. At first, they collected such types of tweets and then categorized them into themes. They monitored for four months and then there was a significant increase in vaccine resistance on Twitter. This increased number of exposures to the vaccine could confuse the public with an anti-vaccine attitude, which may cause a long-term negative impact on human health. Thus it is important to respond to the words used by vaccine opponents to ensure widespread support for the COVID-19 vaccine. Thelwall et al. [36] tried to find out what kind of skepticism information was shared on Twitter, which could be helpful to combat misguided attitudes. The primary topics mentioned were conspiracy, vaccine development speed, and the safety of vaccines.

Table 1
Various vaccine brands in the dataset-1.

Vaccine brand	Reference tag
Pfizer	Pfizer, pfizer, Pfizer–BioNTech, pfizer–bioNtech, BioNTech, biontech
Covaxin (Bharat Biotech)	covax, covaxin, Covax, Covaxin, Bharat Biotech, bharat biotech, BharatBiotec, bharatbiotech
Sputnik V	russia, sputnik, Sputnik, V
AstraZeneca (Covishield)	sii, SII, adar poonawalla, Covishield, covishield, astra, zenca, Oxford–AstraZeneca, astrazeneca, oxford–astrazeneca, serum institute
Moderna	moderna, Moderna, mRNA-1273, Spikevax

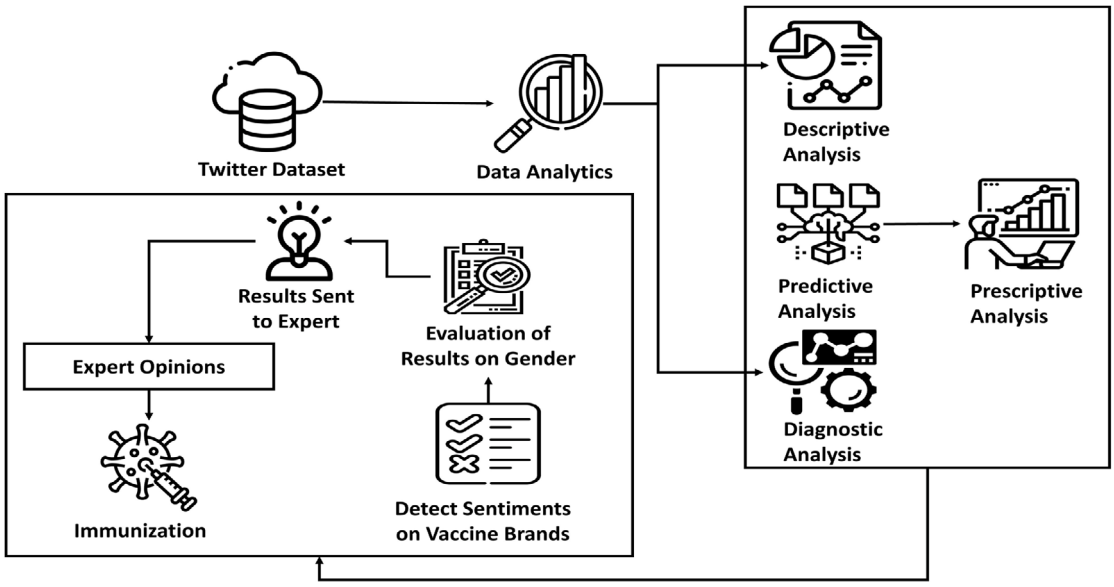


Fig. 1. Proposed data analytics based framework.

Right-wing views, conspiracy theories, or fears of a critical situation were the possible reasons for the majority (79%) of those who declined on Twitter to get vaccinated. Especially in an apolitical way, other issues were written on Twitter by a significant proportion (18%) of those who refused to be vaccinated. Sakr et al. [37] presented a framework based on data analysis in intelligent health care services. Cloud computing, IoT, sensor technology, and big data analytics are the areas of research where they focused primarily. They discussed the diverse challenges of adopting new technologies and developed a structured framework to facilitate a smart and intelligent healthcare system. Shahbaz et al. [38] found recognition of key data analysis strategies in healthcare systems. The authors used questionnaire research and in AMOS v21, they evaluated 224 practical answers. Ragini et al. [39] used data analytics techniques to visualize and analyze the sentiments about the various basic needs of people affected by the disaster. For COVID-19 analysis and prediction, Ahmed et al. [40] created a framework for health monitoring. By implementing the data analytics approach, they performed descriptive, diagnostic, predictive, and prescriptive analyses.

The researchers illustrated various sentiment analysis and data analytics methods using tweets in the literature reviews mentioned above. Some research works used text mining techniques to determine public opinion and sentiments about various aspects of the COVID-19 pandemic such as vaccines, cultures, etc. Few researchers also implemented data analytics strategies to promote applications in health care organizations. Encouraged by the above works, in this paper, we present a data analytics-based framework for evaluating the effects of various COVID-19 vaccine brands on a gender perspective using Twitter data that can help physicians, government officials, and policymakers to vaccinate or immunize people in an organized and effective way.

3. Methodology

In this work, we focus on analyzing the sentiments of the tweets related to the COVID-19 vaccine on Twitter to assess the effect of COVID-19 vaccines on gender. For experimental purposes, we use the Twitter datasets related to the COVID-19 vaccine. To perform step by step data analysis, we have explored descriptive, diagnostic, predictive, and prescriptive data analytics strategies. Fig. 1 presents the overall framework to analyze and predict sentiments. From Fig. 1, we can observe that there are four different data analysis modules in the proposed framework. We collect two datasets to extract the required features. Then we perform descriptive analysis to present the general information about the required features. The diagnostic analysis uses data visualization methods to show the relationship between different features.

Predictive analysis is applied for sentiment detection using various deep learning models. In this work, we use the Long Short Term Memory (LSTM) model with word2vec (skip-gram version) [41] word embedding model to predict sentiment polarities in tweets and create a data analytics-dependent framework that determines the effects of various COVID-19 vaccine brands on gender. In the predictive analysis, we present and observe the prediction and evaluation results of various deep learning models to choose the best model for implementation. The prescriptive analysis is performed by combining expert opinions. We analyze tweets using a variety of data analytics strategies [19]. In the following sections, we provide details of datasets and analysis methods.

3.1. Preparing dataset

In this work, we collect the publicly available dataset of the COVID-19 vaccine-related tweets from Kaggle as dataset-1.[4] The dataset-1 contains 220844 tweets about the COVID-19 vaccines used on large scale and tweets are collected in the time interval of 12 December 2020 to 24 November 2021. We find the gender that is male/female from user_name attribute of the dataset-1 as shown in Fig. 2, because of the absence of the sex attribute in the dataset. We use nameparser api to

[4] https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets.

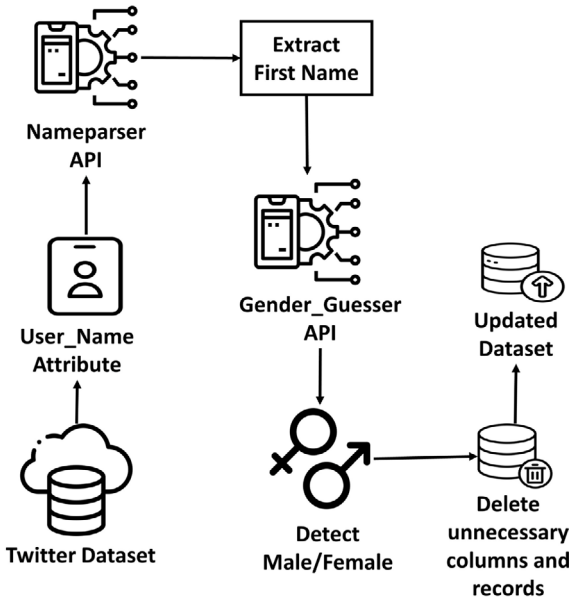
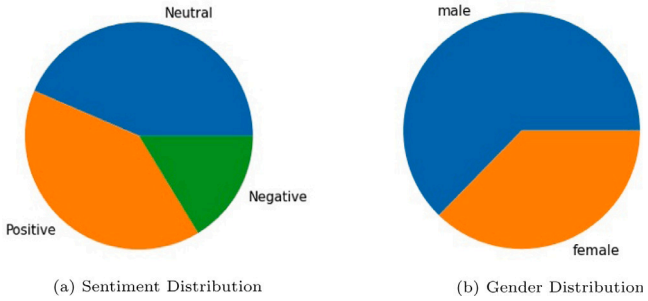


Fig. 2. Gender extraction from user name.



(a) Sentiment Distribution (b) Gender Distribution

Fig. 3. Distribution of attributes.

Table 2
The COVID-19 vaccine brands in the dataset.

Vaccine brands	Description
Pfizer	Approved in 112 countries, 46 trials in 21 countries
Covaxin (Bharat Biotech)	Approved in 12 countries, 7 trials in 1 country
Sputnik V	Approved in 74 countries, 22 trials in 7 countries
AstraZeneca (Covishield)	Approved in 47 countries, 2 trials in 1 country
Moderna	Approved in 79 countries, 33 trials in 8 countries

collect first name from the username and then apply gender_guesser API to detect gender from the username. We delete the tweets without male/female gender information and get 59,750 tweets. We extract five popular vaccine brands from the tweets of dataset-1 by creating reference tags for 5 vaccines: Pfizer, Covaxin(Bharat Biotech), Sputnik, AstraZeneca(Covishield), Moderna as shown in Table 1. For example, in Table 1 we use the keyword "Russia" to extract the tweets related to the vaccine brand "Sputnik V" because it was manufactured by the Gamaleya Research Institute of Epidemiology and Microbiology in Russia [42]. Similarly, we use other significant reference tags as shown in Table 1 to extract proper tweets related to respective vaccine brands. Since dataset-1 does not contain sentiment polarity as the target label. Therefore, to train the deep learning models in a supervised manner we find another labeled dataset-2[5] of COVID-19 tweets that contains sentiment polarities as target label. We train the models by using the tweets of dataset-2, which contains about 45,000 tweets labeled as either negative, neutral or positive sentiment, and tweets are collected in the time interval of 04 January 2020 to 04 December 2020. We consider dataset-2 to train the model because it contains manually tagged target labels with sentiment polarities of tweets related to COVID-19. If we try to train a transformer such as BERT, RoBERTA, and ALBERT from scratch, it will need a very large dataset to obtain better performance. Therefore, in this paper, we have experimented with classical deep learning classifiers like RNN, CNN, GRU, BiLSTM, and LSTM instead of highly successful transformers used in many NLP tasks [43].

We implement the trained model generated by utilizing dataset-2 to get the sentiment polarities of tweets of dataset-1. Handling noisy, unstructured and ill-formatted Twitter data is one of the most essential jobs for us. We normalize the Twitter dataset by performing a series of preprocessing steps including the functions of replacing mentions, hyperlinks, and hashtags with an empty string, converting words into lowercase, dealing with contractions, restoring punctuation with space, eliminating words less than two characters, removing space from words, deleting stop words, managing Unicode and non-English words. We prepare updated dataset-1 for further analysis by dropping unnecessary columns except for gender, original text, preprocessed text, vaccine brands, and sentiment polarity from the dataset as shown in Fig. 2.

3.2. Descriptive analysis

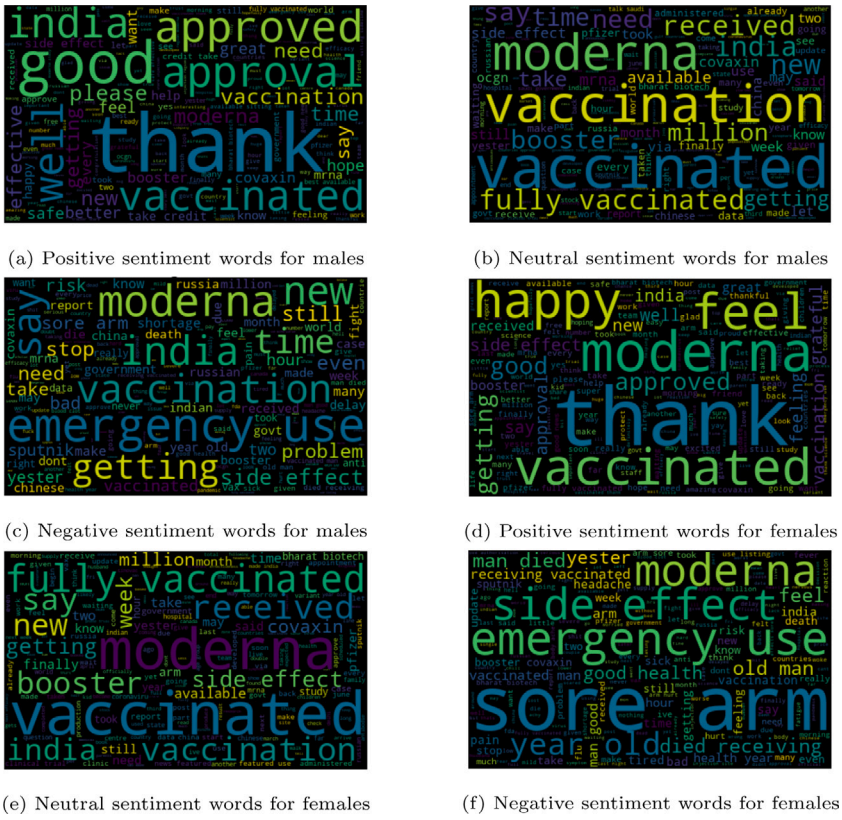
The descriptive analysis enables us to get comprehensive information about various features available in the updated dataset. People can easily interpret visual graphs, summaries, or detailed descriptions of features in the descriptive analysis. Descriptive analysis can be considered a simple form of analysis to know the properties of the dataset. Analyzing the vaccine-related sentiments of tweets from the perspective of the gender of this research work simply reveals the general statistics of the data. Such as country-based statistics of vaccine brands, the total number of sentiment polarities of tweets, and the number of males and females present in the dataset. Various tools are used for descriptive analysis such as column or bar charts, pie charts, tables, or written descriptions [19]. Fig. 3 and Table 2 provide examples of descriptive analysis. The distribution of different attributes can be seen in the updated dataset using the pie graph. Fig. 3(a) shows that very few sentiments of tweets are negative compared to positive and neutral in the dataset. Similarly, Fig. 3(b) presents information on gender distribution. From Fig. 3(b), we can analyze that most people are male. It is visible in Table 2 that the tweets in the dataset include five popular vaccine brands[6] with a number of approved and trials cases in different countries. Visualizing these graphs and table is useful for us to see the insights of the tweets.

3.3. Diagnostic analysis

One of the advanced forms of data analysis techniques refers to diagnostic analysis. The insights of data in diagnostic analysis can be understood by answering the question "Why did it happen?" [19]. It considers various features in the dataset and includes information to determine relationships amongst those features. Diagnostic analysis refers to data acquisition, data mining, and correlation methods. It extensively analyzes data and explains the causes of behavior and events. In analyzing the sentiments of tweets, a diagnostic analysis explores the insights by using the information of different attributes. For example, it

[5] https://www.kaggle.com/datatattle/covid-19-nlp-text-classification.

[6] https://covid19.trackvaccines.org/vaccines/approved/#vaccine-list.



(a) Positive sentiment words for males

(b) Neutral sentiment words for males

(c) Negative sentiment words for males

(d) Positive sentiment words for females

(e) Neutral sentiment words for females

(f) Negative sentiment words for females

Fig. 4. Prevalent words in tweets in the dataset.

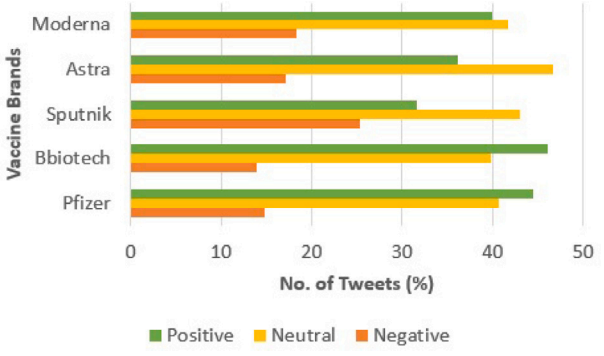


Fig. 5. Vaccine brands & sentiment polarities.

may help to determine all the gender's perceptions of various vaccine brands. The causes of varying the sentiment polarities associated with vaccine brands from the perspective of gender are explored in the diagnostic analysis. It can be seen that in tweets, high-frequency words and various prevalent words related to vaccines are collected to explore the general understanding of human attitudes on different vaccines at the gender level. Examples of diagnostic analysis are shown in Fig. 4, we extract common words from tweets of males and females and apply the word cloud for visualization. Tweets classified as positive sentiment frequently contain words like thank, good, happy, approval, well, safe, etc. for both males and females as shown in Figs. 4(a) and 4(d). These words about the vaccines indicate a positive sentiment that helps us to rely on the vaccines to protect against infection. Tweets with negative sentiment contain common words like emergency, side effect, die, sick, risk, etc. for both genders as shown in Figs. 4(c) and 4(f). However, high-frequent words in Fig. 4(f) seem to offer more insight, like sore, arm, and headache give further details about what is frustrating female candidates more. Neutral sentiments are often more descriptive. Although they do not express certain emotions, they shed light on relevant themes for users. In this case, keywords like vaccinated, booster, moderna, received, etc indicate how people use the platform as shown in Figs. 4(b) and 4(e). Learning what people like and dislike about vaccine brands can help to get information that goes beyond the frequency of words, and can even lead to informed decision-making.

An example of diagnostic analysis is shown in Fig. 5. In Fig. 5, we analyze the sentiment polarities of tweets in terms of percentage against different vaccine brands; it can be seen that the positive attitudes of the people are higher than negative attitudes towards vaccine brands. For Covaxin (Biotech), we get the highest difference in the percentage of positive and negative sentiments, where the percentage of positive sentiments of tweets is 46.15% and negative sentiments of tweets is 13.96%, which indicates the positive attitude of people towards Covaxin is higher than other vaccine brands. In Fig. 6, the relationship between gender and the polarity of positive, neutral, and negative sentiments has been explored in the case of vaccine brands. From Figs. 6(a) and 6(d), it is seen that most attitudes towards Pfizer and AstraZeneca vaccines are more positive for females than males. In Figs. 6(b), 6(c) and 6(e), it can be seen that male attitudes are positive towards vaccine brands Covaxin, Sputnik V, and Moderna compared to female attitudes. From visualization results, by analyzing the tweets it can be seen that females are favorable to Pfizer and AstraZeneca vaccines and males are favorable to Covaxin, Sputnik V, and Moderna vaccines, which can be helpful to determine proper immunization mechanisms for physicians, government officials, and policymakers.

3.4. Predictive analysis

The predictive analysis supports predicting what the future may hold using existing data. It estimates data insights and delivers potential advice to institutions with possible details. In addition, it also infers

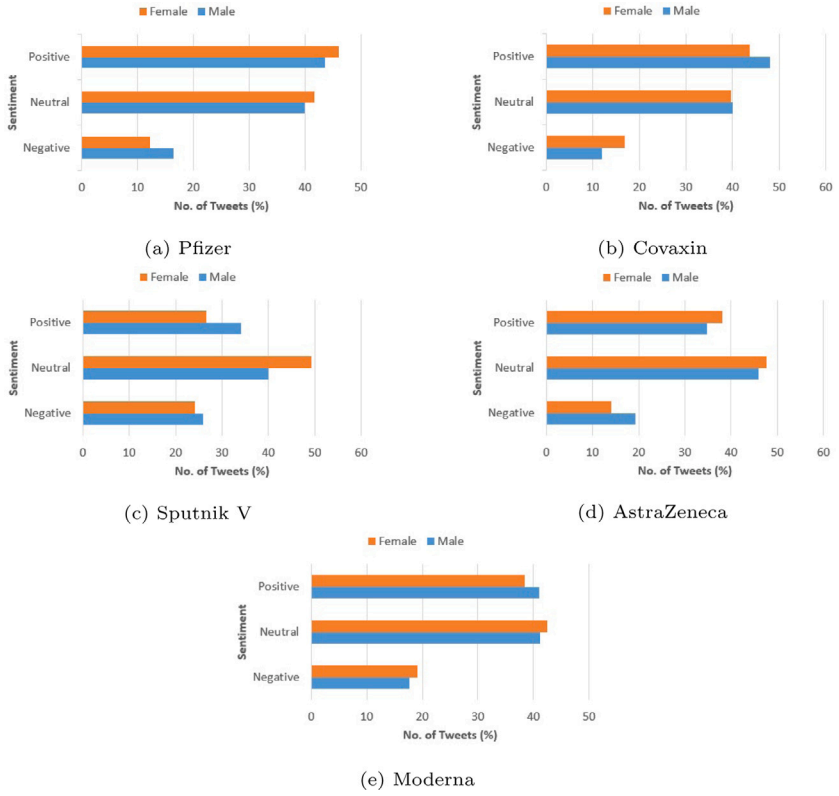


Fig. 6. Sentiment polarities of tweets for vaccine brands from gender perspective.

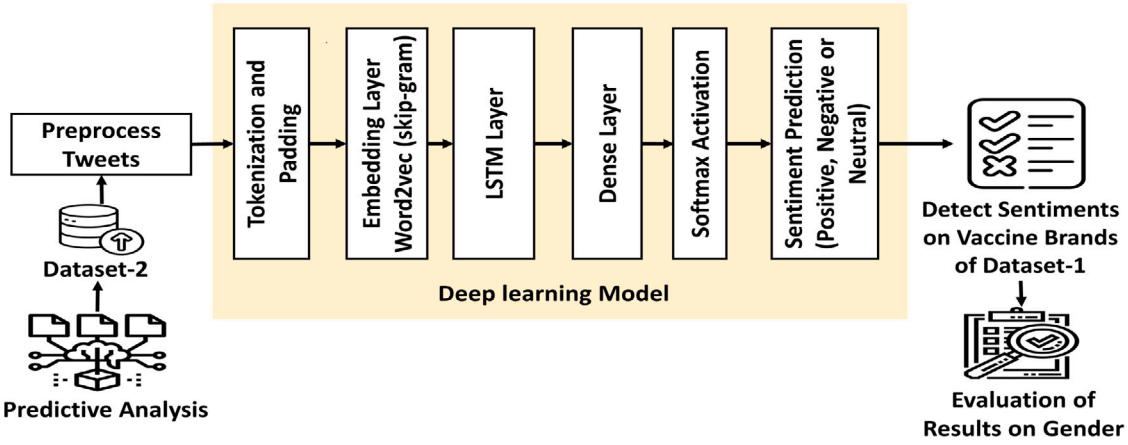


Fig. 7. Predictive analysis of tweets using deep learning model.

the potential possibilities for future outcomes. It is required to train a machine learning or deep learning model with recorded data that considers key hidden patterns and trends of data. The trained model is then used in current data to predict results. In determining people's attitudes towards a specific event, predictive analyses are used to detect sentiment polarities of people's conversations. The diagram is shown in Fig. 7 for predictive analysis showing that the preprocessed tweets are selected from dataset-2. We divide the Dataset-2 into train and test sets. The deep learning model is trained using the training set and feature knowledge. The test set is used to test the effectiveness of the trained model and to predict the sentiments. Then we detect sentiments of tweets on vaccine brands of dataset-1. Different evaluation parameters are used to analyze the results at the gender level as shown in Fig. 7 and refer to experts for feedback.

In this work, we use LSTM [44], which is a deep learning-based model for predictive analysis. In Fig. 7, the overall architecture of the LSTM with the word2vec (skip-gram version) word embedding model implemented in this research work is depicted. In the proposed approach, we use Word2vec (Skip-gram) model as an embedded layer that takes the tokenized tweet as input. The output of this layer is fed into the LSTM layer. In the skip-gram version, the model uses the current word to predict the surrounding window of context words. Skip-gram is better for infrequent words than Continuous Bag of Words (CBOW) in Word2vec [45]. We use the LSTM model to extract hidden features by capturing raw text details due to the long length of the sentence in tweets [46]. The LSTM is used to capture multiple features by monodirectional mapping because of the multiple sentiment labels in tweets. We use 256 hidden neurons in the LSTM layer. LSTM is popular in natural language processing tasks and is considered one of the most successful versions of RNN [44]. The outputs of the LSTM layer are passed through a fully connected dense layer with a softmax activation function for multiclass sentiment classification i.e. positive,

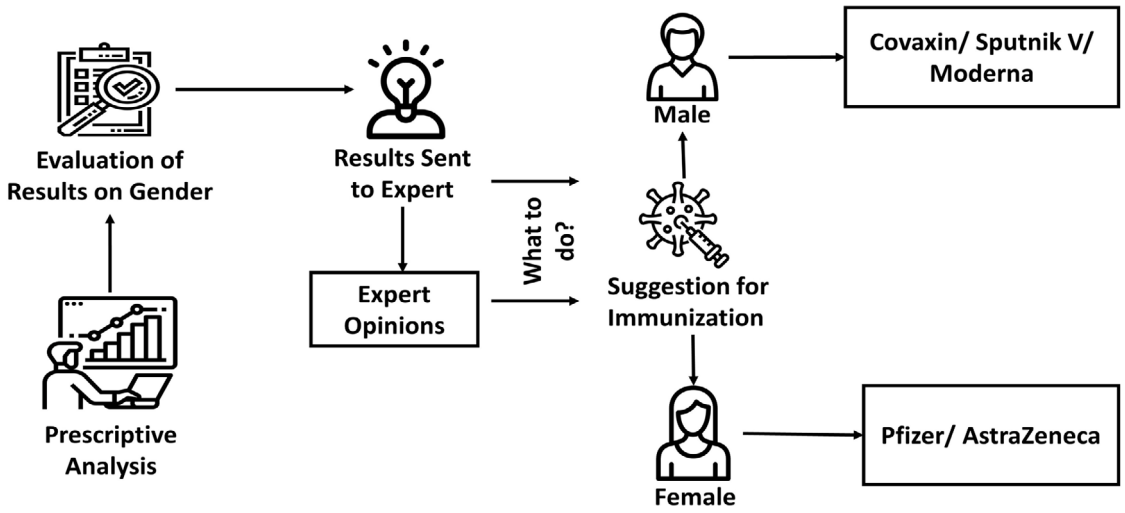


Fig. 8. Prescriptive analysis for further suggestions and expert opinions.

negative, and neutral as shown in Fig. 7 by compiling the model with RMSProp optimizer [47].

3.5. Prescriptive analysis

The prescriptive analysis uses the benefit of predictive analysis and helps users to decide on various actions that can direct them to a solution. It evaluates the outcome of future predictions and advises on possible consequences before making decisions. It provides an insight into what will happen and an explanation of why it will happen, thus offering advice on how to benefit from predicted results. It recommends a variety of practical activities and outlines the possible effects of each event. In the context of determining the effects of vaccine brands on the sentiment analysis of Twitter data, it plays an important role in directing responsible people or guiding bodies to establish appropriate vaccinations to achieve positive physical and mental health outcomes for both males and females by reducing the side effects caused by the impact of the vaccination [48]. Fig. 8 shows the transformation of the results of the predictive analysis to the experts to obtain an important suggestion. In our case study, we send the results of the LSTM model to the experts for further analysis to make wise decisions. Based on the gender implications of sentiment analysis of various vaccine brands, the expert suggests to people need to be expected and correct vaccination. It can help reduce the side effects or unwanted health effects due to immunization.

4. Evaluation & experimental results

In this section, we first discuss the predictions and outcomes of the trained LSTM model. We use different graphs to visualize data. Secondly, different evaluation metrics are used to test the overall performance of the LSTM model. In addition, we also consider comparison results for LSTM and other deep learning models. In Figs. 9 and 10, we analyze the evaluation of the effectiveness of the trained model. Fig. 9 shows the accuracy of the training and validation of the trained LSTM model. We use the size of epoch 10 to train the model.

As can be seen in Fig. 9, we have found the highest validation accuracy of the LSTM model which is about 86% in epoch number 3. Similarly, in Fig. 10, loss of the training and validation is seen as the validation loss to the model increases with training. To analyze the model, we calculate accuracy, precision, recall, and f1-measure using different metrics, such as true positive, true negative, false positive, and false negative values.

Fig. 11 presents the confusion matrix of the model. It is clear from the matrix that few data are classified incorrectly by the model. For example, 54 instances among 619 of the Neutral class are predicted as Positive. In the Negative class, 67 data out of 1633 are mistakenly classified as Neutral. Fig. 11 depicts that the Positive class achieves the highest rate of correct classification (87%) while the Neutral acquired the lowest (80%). The possible reason for high incorrect predictions in the Neutral class may be the presence of fewer tweets of the neutral class in dataset-2. We implement the trained LSTM model to enumerate the sentiment polarities of tweets of dataset-1. To check the validity

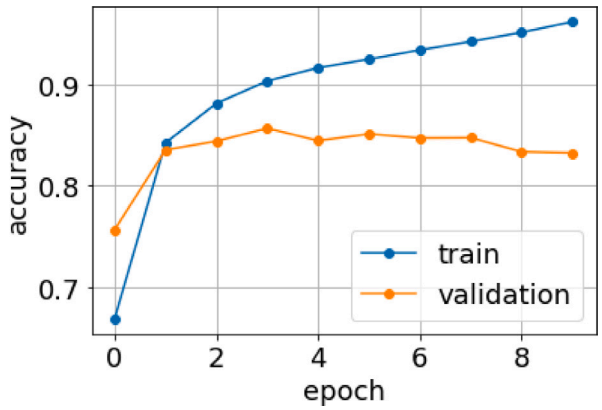


Fig. 9. Accuracy graph of the LSTM model.

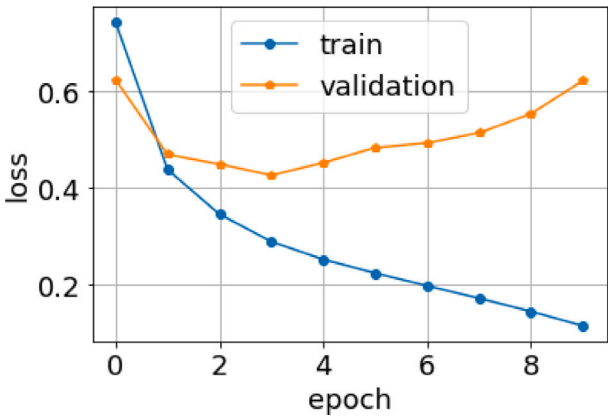


Fig. 10. Loss graph of the LSTM model.

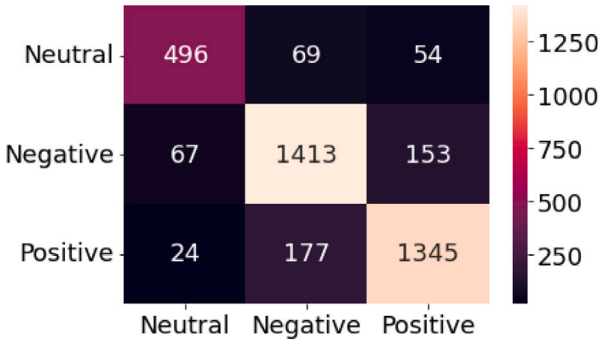


Fig. 11. Confusion matrix of the LSTM model.

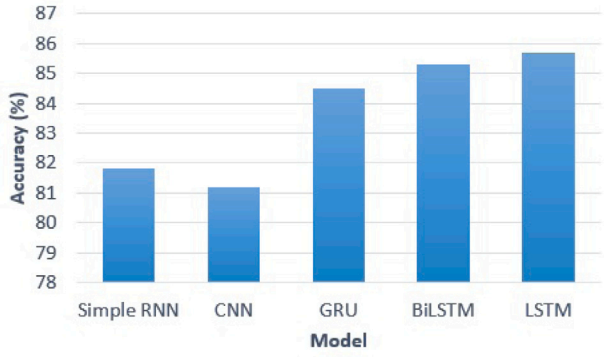


Fig. 12. Performance comparison of different models.

Table 3
Evaluation of metrics of different models.

Model	Precision	Recall	f1-measure
Simple RNN	0.82	0.82	0.82
CNN	0.82	0.81	0.81
GRU	0.85	0.85	0.84
BiLSTM	0.85	0.85	0.85
LSTM	0.86	0.86	0.86

of the trained LSTM model in dataset-1, we randomly select 100 explainable tweets from dataset-1 and then manually label them with mentioned sentiment polarities. We find 85 manually labeled instances matched with the target polarities generated by the trained LSTM model.

In addition, by implementing the word2vec (skip-gram) version, we also train different deep learning models on dataset-2 and compare their results with the LSTM model. From Fig. 12, we can observe that all the deep learning models work well and generate good results with a level of accuracy of more than 80%. The results of the comparison show that LSTM surpasses other models and achieves maximum accuracy of about 86%. The results of the Table 3 show that the performance of the LSTM model in terms of accuracy, recall, and f1-measure for weighted averages of positive, neutral, and negative sentiments is higher than that of other deep learning models.

5. Discussion

Overall, our proposed data analytics-based framework evaluates the effect of the COVID-19 vaccine brands on gender level from social media texts. According to our knowledge, this is the first approach that implements data analytics techniques such as descriptive, diagnostic, predictive, and prescriptive analysis [19] to analyze the sentiments of tweets to determine public perceptions about different COVID-19 vaccine brands from the gender level. The proposed approach is very effective to provide helpful suggestions for experts, government officials, and policymakers to apply appropriate vaccinations. The number of key observations of our approach is highlighted as follows.

To explore the required features, in our proposed framework we have performed the descriptive analysis of the updated dataset-1. Here, we have observed that the number of male candidates is higher than the number of female candidates and most people provide positive sentiments instead of negative ones in vaccine-related tweets as shown in Fig. 3. The diagnostic analysis provides important findings by examining the tweets in terms of high-frequency and prevalent words for both males and females as shown in Fig. 4. We have found that in positive tweets, people express their gratitude for being vaccinated. They think that, by grabbing proper vaccination, the impacts of the pandemic can be minimized as early as possible and can be returned to normal life. People mostly complain about their concerns after being vaccinated in negative tweets, such as emergency, sore arm, etc. We have also found the relationship between different sentiment polarities concerning vaccine brands from the perspective of gender level in the diagnostic analysis as shown in Fig. 6, which concludes that Pfizer or AstraZeneca is suitable for females while Covaxin, Sputnik V, or Moderna is suitable for males based on their expressed sentiments on tweets. The predictive analysis helps to determine the sentiment polarities of vaccine-related tweets. The experiment results in the predictive analysis present that the LSTM performs better compared to other basic deep learning models achieving the highest accuracy of 85.7%. Followed by the predictive analysis, the prescriptive analysis helps the experts to provide useful suggestions by categorizing vaccine brands between males and females as shown in Fig. 8.

The observations from the outcomes of the different data analytics strategies present that the dominant positive sentiments towards specific vaccine brands of gender may be helpful for experts to suggest an appropriate immunization process. However, there are some limitations to this paper. We concentrate our work only on analyzing five popular vaccine brands. We could not consider the location and age of people while analyzing the tweets about vaccine brands. Thus context-aware technology [49] and AI techniques [50] could be useful. To detect gender from user name, we follow the machine-dependent approach, while manually labeled datasets with sex or gender may be produced more effective results. In addition to analysis of the effect of COVID-19 vaccine brands, our proposed data analytics-based framework can also be considered in other domains of application, such as education, agriculture, cyber-security, etc. by analyzing social media texts on various events where human current interests are involved.

6. Conclusion

The number of textual data in social media is increasing day by day. People express their needs, emotions, and opinions on social media platforms like Twitter on the outbreak of the COVID-19 pandemic. Thus in this research work, we present a systematic framework by taking the advantages of data analytics and sentiment analysis that might be beneficial to provide useful suggestions for appropriate vaccination by physicians, government officials, and policymakers. Our proposed approach might be helpful in the visualization of vaccine-related pandemic information, effectiveness assessment of various vaccine brands, prevention of side effects, and providing mental and physical satisfaction for both males and females. By implementing a data analytics strategy, we are enabled to perform descriptive, diagnostic, predictive, and prescriptive analyses. Through insightful data analysis, we conclude by providing helpful suggestions to experts to take necessary steps for the proper distribution of various vaccine brands amongst males and females. In the future, we will consider people's age, location and other contextual information to assist the people for taking exact action in advance.

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.

References

[1] Jianlong Zhou, Yang Shuiqiao, Xiao Chun, Chen Fang. Examination of community sentiment dynamics due to COVID-19 pandemic: a case study from a state in Australia. SN Comput Sci 2021;2(3):1–11.

[2] Ewen Callaway, Ledford Heidi. How bad is omicron? What scientists know so far. Nature 2021;600(7888):197–9.

[3] Women's Aid, UK. The Impact of COVID-19 on Women and Children Experiencing Domestic Abuse, and the Life-Saving Services that Support Them. Women's Aid UK; 2020.

[4] Clare Wenham, Smith Julia, Morgan Rosemary. COVID-19: the gendered impacts of the outbreak. Lancet 2020;395(10227):846–8.

[5] Vincenzo Galasso, Pons Vincent, Profeta Paola, Becher Michael, Brouard Sylvain, Foucault Martial. Gender differences in COVID-19 attitudes and behavior: Panel evidence from eight countries. Proc Natl Acad Sci 2020;117(44):27285–91.

[6] Mathieu Boniol, McIsaac Michelle, Xu Lihui, Wuliji Tana, Diallo Khassoum, Campbell Jim. Gender equity in the health workforce: analysis of 104 countries. No. WHO/HIS/HWF/Gender/WP1/2019.1.. World Health Organization; 2019.

[7] Noha Alduaiji, Datta Amitava, Li Jianxin. Influence propagation model for clique-based community detection in social networks. IEEE Trans Comput Soc Syst 2018;5(2):563–75.

[8] Jianxin Li, Cai Taotao, Deng Ke, Wang Xinjue, Sellis Timos, Xia Feng. Community-diversified influence maximization in social networks. Inf Syst 92(2020):101522.

[9] Hui Yin, Yang Shuiqiao, Song Xiangyu, Liu Wei, Li Jianxin. Deep fusion of multimodal features for social media retweet time prediction. World Wide Web 2021;24(4):1027–44.

[10] Shuiqiao Yang, Huang Guangyan, Xiang Yang, Zhou Xiangmin, Chi Chi-Hung. Modeling user preferences on spatiotemporal topics for point-of-interest recommendation. In: 2017 IEEE International Conference on Services Computing. IEEE; 2017, p. 204–11.

[11] Shuiqiao Yang, Huang Guangyan, Cai Borui. Discovering topic representative terms for short text clustering. IEEE Access 2019;7:92037–47.

[12] Haixin Jiang, Zhou Rui, Zhang Limeng, Wang Hua, Zhang Yanchun. Sentence level topic models for associated topics extraction. World Wide Web 2019;22(6):2545–60.

[13] Lima Ana CES, de Castro Leandro N. Automatic sentiment analysis of Twitter messages. In: 2012 Fourth international conference on computational aspects of social networks. IEEE; 2012, p. 52–7.

[14] Sarker Iqbal H. Smart city data science: Towards data-driven smart cities with open research issues. Internet of Things 2022;100528.

[15] Kwon Ohbyung, Lee Namyeon, Shin Bongsik. Data quality management data usage experience and acquisition intention of big data analytics. Int J Inf Manage 2014;34(3):387–94.

[16] Kigon Lyu, Kim Hyeoncheol. Sentiment analysis using word polarity of social media. Wirel Pers Commun 2016;89(3):941–58.

[17] Patrick Jansson, Liu Shuhua. Distributed representation, LDA topic modelling and deep learning for emerging named entity recognition from social media. In: Proceedings of the 3rd workshop on noisy user-generated text. 2017, p. 154–9.

[18] Gama João, Zliobaitė Indrė, Bifet Albert, Pechenizkiy Mykola, Bouchachia Abdelhamid. A survey on concept drift adaptation. ACM Comput Surv (CSUR) 2014;46(4):1–37.

[19] Sarker Iqbal H. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. SN Comput Sci 2021;2(5):1–22.

[20] Jiahua Du, Michalska Sandra, Subramani Sudha, Wang Hua, Zhang Yanchun. Neural attention with character embeddings for hay fever detection from twitter. Health Inf Sci Syst 2019;7(1):1–7.

[21] Rubina Sarki, Ahmed Khandakar, Wang Hua, Zhang Yanchun. Automated detection of mild and multi-class diabetic eye diseases using deep learning. Health Inf Sci Syst 2020;8(1):1–9.

[22] Wang Xinjue, Deng Ke, Li Jianxin, Yu Jeffery Xu, Jensen Christian S, Yang Xiaochun. Efficient targeted influence minimization in big social networks. World Wide Web 2020;23(4):2323–40.

[23] Qiao Tian, Li Jianxin, Chen Lu, Deng Ke, Li Rong-hua, Reynolds Mark, et al. Evidence-driven dubious decision making in online shopping. World Wide Web 2019;22(6):2883–99.

[24] Jiao Yin, Tang MingJian, Cao Jinli, Wang Hua, You Mingshan, Lin Yongzheng. Vulnerability exploitation time prediction: an integrated framework for dynamic imbalanced learning. World Wide Web 2022;25(1):401–23.

[25] Fuyong Zhang, Wang Yi, Liu Shigang, Wang Hua. Decision-based evasion attacks on tree ensemble classifiers. World Wide Web 2020;23(5):2957–77.

[26] Xiaoya Li, Zhou Mingxin, Wu Jiawei, Yuan Arianna, Wu Fei, Li Jiwei. Analyzing COVID-19 on online social media: Trends, sentiments and emotions. 2020, arXiv preprint arXiv:2005.14464.

[27] Massimo Stella, Restocchi Valerio, Deyne Simon De. # Lockdown: Network-enhanced emotional profiling in the time of Covid-19. Big Data Cogn Comput 2020;4(2):14.

[28] Dubey Akash Dutt. Twitter sentiment analysis during COVID-19 outbreak. 2020, Available at SSRN 3572023.

[29] Hui Yin, Yang Shuiqiao, Li Jianxin. Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media. In: International conference on advanced data mining and applications. Cham: Springer; 2020, p. 610–23.

[30] Shuiqiao Yang, Jiang Jiaojiao, Pal Arindam, Yu Kun, Chen Fang, Yu Shui. Analysis and insights for myths circulating on Twitter during the COVID-19 pandemic. IEEE Open J Comput Soc 1(2020):209–19.

[31] Imran Ali Shariq, Daudpota Sher Muhammad, Kastrati Zenun, Batra Rakhi. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. Ieee Access 8(2020):181074-181090.

[32] Satu Md Shahriare, Khan Md Imran, Mahmud Mufti, Uddin Shahadat, Summers Matthew A, Quinn Julian MW, et al. TClustVID: A novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets. Knowl-Based Syst 226(2021):107126.

[33] Kwok Stephen Wai Hang, Vadde Sai Kumar, Wang Guanjin. Tweet topics and sentiments relating to COVID-19 vaccination among Australian Twitter users: Machine learning analysis. J Med Internet Res 2021;23(5):e26953.

[34] Lyu Joanne Chen, Han Eileen Le, Luli Garving K. COVID-19 vaccine–related discussion on Twitter: topic modeling and sentiment analysis. J Med Internet Res 2021;23(6):e24435.

[35] Erika Bonnevie, Gallegos-Jeffrey Allison, Goldbarg Jaclyn, Byrd Brian, Smyser Joseph. Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. J Commun Healthcare 2021;14(1):12–9.

[36] Mike Thelwall, Kousha Kayvan, Thelwall Saheeda. Covid-19 vaccine hesitancy on English-language Twitter. Profla Inf (EPI) 30(2):2021.

[37] Sherif Sakr, Elgammal Amal. Towards a comprehensive data analytics framework for smart healthcare services. Big Data Res 2016;4:44–58.

[38] Muhammad Shahbaz, Gao Changyuan, Zhai LiLi, Shahzad Fakhar, Hu Yanling. Investigating the adoption of big data analytics in healthcare: the moderating role of resistance to change. J Big Data 2019;6(1):1–20.

[39] Ragini J Rexiline, Anand PM Rubesh, Bhaskar Vidhyacharan. Big data analytics for disaster response and recovery through sentiment analysis. Int J Inf Manage 42(2018):13–24.

[40] Imran Ahmed, Ahmad Misbah, Jeon Gwanggil, Piccialli Francesco. A framework for pandemic prediction using big data analytics. Big Data Res 2021;25:100190.

[41] Long Ma, Zhang Yanqing. Using Word2Vec to process big text data. In: 2015 IEEE International conference on big data (Big Data). IEEE; 2015, p. 2895–7.

[42] Ian Jones, Roy Polly. Sputnik V COVID-19 vaccine candidate appears safe and effective. Lancet 2021;397(10275):642–3.

[43] Aysu Ezen-Can. A comparison of LSTM and BERT for small corpus. 2020, arXiv preprint arXiv:2009.05451.

[44] Sarker Iqbal H. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Comput Sci 2021;2(6):1–20.

[45] Tomas Mikolov, Chen Kai, Corrado Greg, Dean Jeffrey. Efficient estimation of word representations in vector space. 2013, arXiv preprint arXiv:1301.3781.

[46] Al-Smadi Mohammad, Talafha Bashar, Al-Ayyoub Mahmoud, Jararweh Yaser. Using long short-term memory deep neural networks for aspect-based sentiment analysis of arabic reviews. Int J Mach Learn Cybern 2019;10(8):2163–75.

[47] Taqi Arwa Mohammed, Awad Ahmed, Al-Azzo Fadwa, Milanova Mariofanna. The impact of multi-optimizers and data augmentation on TensorFlow convolutional neural network performance. In: 2018 IEEE Conference on multimedia information processing and retrieval. IEEE; 2018, p. 140–5.

[48] Ahamad Md Martuza, Aktar Sakifa, Uddin Md Jamal, Rashed-Al-Mahfuz Md, Azad AKM, Uddin Shahadat, et al. Adverse effects of COVID-19 vaccination: machine learning and statistical approach to identify and classify incidences of morbidity and post-vaccination reactogenicity. 2021, Medrxiv.

[49] Sarker Iqbal H, Han Alan Colman Jun, Watters Paul A. Context-aware machine learning and mobile data analytics: automated rule-based services with intelligent decision-making. Springer Nature; 2021.

[50] Sarker Iqbal H. Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. SN Comput Sci 2022;3(2):1–20.