International Journal of Information Management Data Insights 2 (2022) 100053

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

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

Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of Covid-19 pandemic

Maryam Mahdikhani

Supply Chain and Information Management Department, College of Charleston, 66 George St, Charleston, SC, United States

ARTICLE INFO

Keywords:
Covid-19
Pandemic
Machine learning
Twitter
Sentiment analysis
Text mining
Big data analytics

ABSTRACT

In this study, public opinion and emotions regarding different stages of the Covid-19 pandemic from the outbreak of the disease to the distribution of vaccines were analyzed to predict the popularity of tweets. More than 1.25 million English tweets were collected, posted from January 20, 2020, to May 29, 2021. Five sets of content features, including topic analysis, topics plus TF-IDF vectorizer, bag of words (BOW) by TF-IDF vectorizer, document embedding, and document embedding plus TF-IDF vectorizer, were extracted and applied to supervised machine learning algorithms to generate a predictive model for the retweetability of posted tweets. The analysis showed that tweets with higher emotional intensity are more popular than tweets containing information on Covid-19 pandemic. This study can help to detect the public emotions during the pandemic and after vaccination and predict the retweetability of posted tweets in different stages of Covid-19 pandemic.

1. Introduction

The coronavirus pandemic, also known as Covid-19, began in December 2019 when several patients from Wuhan Hubei province in China reported severe health symptoms. Since then, Covid-19 has spread across the globe. According to the World Health Organization (WHO) report on July 14th, 2021, there have been 187,519,798 cases of Covid-19, including 4049,372 deaths.[1] In the very early stages of the pandemic, the WHO advocated for isolation and self-quarantine of affected individuals to reduce the number of cases and mortality rates, leading to the largest lockdown in history. Spending time at home and searching for Covid-19-related news became a common preoccupation, and many turned to social media platforms such as Twitter, which became one of the most important means of sharing information and expressing feelings regarding Covid-19 (Mohammed & Ferraris, 2021; Su, Venkat, Yadav, Puglisi, & Fodeh, 2021; Younis et al., 2020).

Twitter users can "retweet" or forward a posted tweet to their network, which speeds up the information sharing process. Thus, retweets can represent Twitter users' interests on a large scale. The popularity of tweets is measured by their content and the volume of retweets. Shahi, Dirkson, & Majchrzak, 2021conducted an exploratory study to examine the sources, spread, and content of misinformation in tweets related to the Covid-19 pandemic. Yousefinaghani, Dara, Mubareka, Papadopoulos, and Sharif (2021) examined the content of four million tweets to learn about public opinion regarding the Covid-19 vaccine. Using Twitter data from several mega-cities worldwide, Yao, Yang, Liu, Keith, & Guan, 2021 employed machine learning techniques to analyze the public's response to the Covid-19 pandemic. To the best of our knowledge, none of the previous studies have investigated the patterns in public responses to the pandemic from its onset to vaccine distribution by analyzing the content of tweets and predicting the popularity of tweets.

This study, address this gap by collecting tweets generated from January 2020 to May 2021 and by analyzing the public opinions and emotions by applying advanced machine learning technique, including the latent Dirichlet allocation (LDA) topic (Blei, Ng, & Jordan, 2003) and CrystalFeel algorithm (Gupta & Yang, 2018). More importantly, the extraction of different categories of content features and the building of a predictive model that assesses the popularity of tweets by using the number of retweets (based on the content of posted tweets) is another gap in the literature that we addressed in this study. The research objectives for this study are as follows: (i) Detecting public emotions in different stages of Covid-19 pandemic using Twitter data. (ii) Exploring the dominant English topics related to Covid-19 on Twitter, and the sentiment associated with them. And (iii) Building a predictive model for retweetability of the posted tweets based on their content. Furthermore,

E-mail address: mahdikhanim@cofc.edu

[1] WHO Coronavirus Disease Dashboard, www.who.int

https://doi.org/10.1016/j.jjimei.2021.100053
Received 6 August 2021; Received in revised form 11 December 2021; Accepted 12 December 2021

the contribution of this study to the literature can be summarized as follows: (i) Analyzing 1251,216 randomly selected tweets from January 20, 2020 to May 29, 2021, which includes tweets from the early stages of the pandemic to tweets related to the distribution of vaccines, can help to understand the public opinions and emotions regarding Covid-19 pandemic at the ongoing pandemic. (ii) This study applied the latent Dirichlet allocation (LDA) topic and CrystalFeel algorithm to detect four basic emotions, fear, anger, joy, and sadness, at different stages of the Covid-19 pandemic. (iii) The proposed approach extracts five different sets of content features from the posted tweets to applies them to three base supervised machine learning algorithms, and an ensemble voting classifier to predict the retweetabilty of the posted tweets. (iv) The experimental results are then compared by four metrics including, accuracy, F1-score, recall, and precision to choose a model with the highest performance. The study further compared the execution time for running each model to choose the most efficient model.

This study is organized by reviewing the literature in section 2, and specifically reviewing the background of the impact of social media and Twitter during the pandemic. The research methodology is introduced in Section 3. Experimental design and analysis along with the models' results are discussed in section 4. The discussion and the implication of the research are presented in section 5. The conclusions and limitations of our work are discussed in Section 6.

2. Literature review

During the Covid-19 pandemic, social media platforms such as Facebook, Instagram, TikTok, and Twitter became even more important as a means to interact and connect with others. Visits to the Twitter increased by 36 percent in 2020 compared with those of the previous year, and users in the United States spent an average of 32.7 min on the platform per day[2]. Access to large datasets on various platforms offer opportunities for scholars to use advanced computational science to gain insights (Kar & Dwivedi, 2020). For instance, Mishra, Urolagin, and Jothi (2019) applied term frequency-inverse document frequency (TF-IDF) and Cosine Similarity on hotels reviews to generate a recommendation system for suggesting proper hotels to the customers. Chintalapudi, Battineni, Canio, Sagaro, & Amenta, 2021 analyzed medical records from digital health systems from 2018 to 2020 by implementing text mining approach to gain insights on improving healthcare quality and assessing patient feedback. Rajendran & Sundarraj, 2021 conducted experiments in two domains including movies and restaurants to gather users browsing history, generate topics by using Latent Dirichlet Allocation (LDA) models, and extract user preferences by enhancing recommendation algorithm. Mishra, Urolagin, and Jothi (2020) also used the reviews data to apply sentiment intensity analyzer and generate a recommendation system for tourist point of interest. This research contributes to two research streams, including the impact of media, and particularly Twitter during pandemics, and retweeting behavior based on the content of tweets.

2.1. Media's impact and particularly Twitter during pandemics

Regarding the first research stream, Odlum and Yoon. (2015) studied the use of Twitter during the Ebola outbreak to monitor information sharing among users and examine the users' behavior and their knowledge of the disease during the pandemic. The result of this study revealed the pattern in the spread of information among the public and highlighted the value of Twitter as a tool for spreading public awareness. Lazard, Scheinfeld, Bernhardt, Wilcox, and Suran (2015) used textual analysis to examine public concerns about the Ebola virus and interest in safety information. The study highlighted the efficiency of using Twitter in public health communication. Jain and Kumar. (2015) examined the use of Twitter in the 2015 H1N1 pandemic (also known as Swine flu) to create an inspection system by analyzing information relevant to Influenza (H1N1) and enhancing public awareness in India. They classified tweets as either relevant or irrelevant to studying public opinion regarding H1N1. Their results highlighted the importance of social media for tracking a disease. Szomszor, Kostkova, and Louis (2011)) analyzed tweets and online media related to the Swine flu pandemic of 2009 to identify the popularity of true information. They found that poorly represented scientific information can still be shared in public and cause harm. Furthermore, several studies have examined Twitter content to analyze how the public expresses their feelings at the onset of pandemics (Baboukardos, Gaia, & She, 2021; Garcia & Berton, 2021; S. Kaur, Kaul, & Zadeh, 2020; Ridhwan & Hargreaves, 2021). By following a quasi-inductive approach, Mittal, Ahmed, Mittal, & Aggarwal, 2021 found that the majority of Twitter users tend to share positive content regarding the lockdown but their opinions could swing over the course of pandemic based on recent developments. Some studies analyzed tweets with a focus on the public's emotions during the Covid-19 pandemic (Gupta et al., 2021; Kabir & Madria, 2021), while others focused on public opinions following the rollout of Covid-19 vaccines (Sv, Tandon, Vikas, & Hinduja, 2021; Yousefinaghani et al., 2021). Kabir and Madria (2021) developed a neural network model to automatically detect a variety of emotions in tweets on Covid-19. They randomly selected ten thousand tweets in English from the United States for their analysis, and their results showed that negative emotions increased during the pandemic. Kaur, Mittal, Khosla, & Mittal, 2021 discussed the use of advanced machine learning tools to predict and analyze the impact of quarantine during Covid-19 pandemic. Rustam et al. (2021) identified sentiments regarding Covid-19 from tweets using a supervised machine learning approach to understand how people made informed decisions on how to handle their circumstances during the pandemic. Mishra, Urolagin, Jothi, Neogi, & Nawaz, 2021 used LDA model on almost 20,000 tweets for tourism sector, sub-domains hospitality and healthcare during Covid-19 pandemic to identify frequent terms and applied state-of-the-art deep learning algorithm to generate a robust sentiment prediction model. This study contributes to this research stream by analyzing 1251,216 Covid-19-related tweets from January 20, 2020, to May 29, 2021 to investigate Twitter users' opinion and feeling about the Covid-19 pandemic during different phases of the pandemic, including the early stage of the disease, during the lockdown, and after the distribution of vaccines.

2.2. Retweeting behavior

Several studies have contributed to this field by proposing methods for predicting the results of important events, such as games, and political elections, using data on the volume of retweet (Abdullah, Nishioka, Tanaka, & Murayama, 2015; Liang et al., 2016). Some studies explored the reasons why users retweet certain information without applying machine learning techniques for prediction. Boyd, Golder, and Lotan (2010) empirically examined several case studies on Twitter to understand and analyze the motivations behind retweeting behavior. Their study highlighted that bias in interpreting tweets caused the spread of false information on Twitter.

Kwak, Lee, Park, and Moon (2010) studied the impact of retweeting on information sharing. To evaluate the popularity of tweets, they ranked users based on their number of followers and followings compared to the volume of retweets. The results of this study showed the volume of retweets based on the tweet's content has a stronger impact than the number of people who follow the Twitter account's user.

Naveed, Gottron, Kunegis, and Alhadi (2011) examined the impact of a tweet's content on its retweet volume. They analyzed two different levels of content-based features in tweets and predicted the retweetability of a given tweet. Guidry, Waters, & Saxton, 2014 analyzed the content of 3415 Twitter updates for 50 nonprofit organizations to examine which type of content is likely to be retweeted and to learn how to engage audi-

[2] Posting less, posting more, and tired of it all: How the pandemic has changed social media, Vox.com, Mar 1st, 2021

ences and facilitate discussions. Marino & lo Presti, 2018 examined the content of tweets of European Commissioners and proposed a retweetability rate to measure citizen engagement based on the content on social media in response to certain events. Chung, Woo, & Lee, 2020 collected the tweets from Women Who Code (WWC) over a one-year period to examine whether certain content and features such as hashtags and photos resulted in differences in retweet volume. Rao, Vemprala, Akello, & Valecha, 2020 studied the alarming vs. reassuring retweet distribution pattern related to Covid-19. To the best of our knowledge, none of the Covid-19-related used an advanced machine learning predictive model to examine the retweetability of tweets based on content. Neogi, Garg, Mishra, & Dwivedi, 2021 generated models to categorize and analyze sentiments based on a collection of tweets pertaining to protests of Indian farmers. We contribute to this research stream by examining the content-based features for predicting the popularity of tweets based on the volume of retweets during the Covid-19 pandemic.

2.3. Topic modeling on tweets related to Covid-19

Recently, several studies adopted topic modeling analysis on tweets to identify public concerns about Covid-19. Abd-Alrazaq, Alhuwail, Househ, Hamdi, & Shah (2020) examined the tweets posted in English related to Covid-19 from February 2020 to March 2020 by adopting Latent Dirichlet Allocation (LDA) for topic analysis. Mackey et al. (2020) explored tweets related to Covid-19 symptoms from March 2020 and applied bi-terms topic model (BTM) to examine the content related to symptoms, testing, and recovery of individuals who had been infected with Covid-19. Stokes, Andy, Guntuku, Ungar, & Merchant (2020) analyzed the public response to Covid-19 based on real-time analysis of 94,467 comments from March 2020 about the pandemic and Covid-19 made in a public forum. They adopted the LDA technique by defining 50 topics and reviewing the top ten words associated with each topic. Lwin et al. (2020) examined worldwide trends of four basic emotions (i.e., fear, anger, sadness, and joy) during the pandemic by analyzing more than 20 million tweets from January 28 to April 9, 2020. They adopted a lexical approach by using the algorithm CrystalFeel and used “wuhan”, “corona”, “nCov”, and “Covid” as search keywords to generate word clouds related to emotions. Cinelli et al. (2020) collected data related to Covid-19 on Twitter, Instagram, YouTube, Reddit, and Gab to examine public engagement on the topic of Covid-19. They extracted all the topics related to Covid-19 by generating word embedding for the text corpus, and then analyzed the topics. This study contributes to the literature by employing the LDA algorithm to identify the most popular topic related to Covid-19 for content feature purposes and applying them into the CrystalFeel algorithm to examine the public’s basic emotions about the Covid-19 pandemic.

3. Research method

In this study, the primary objective was to identify public concerns and basic emotions related to the Covid-19 pandemic at the early stages, during the pandemic and in the post-pandemic phases. Five sets of content features, including topic modeling, topics plus the TF-IDF vectorizer, BOW by the TF-IDF vectorizer, document embedding, and document embedding plus the TF-IDF vectorizer, are then selected. The five sets of features are applied as inputs for the selected classifiers to compare the accuracy of the prediction performance of tweet popularity based on the volume of retweets. Fig. 1 illustrates the system architecture of the research study.

3.1. Tweet collection and preprocessing

Table 1
Example for the tweet’s attributes.

Tweet attributes	Example
User ID	1,245,698,700,736
Text	Virologits weigh in on novel #corornavirus in China’s outbreak
Language	EN
User location	Comunidad de Madrid, Espana
Hashtags	#Coronavirus
User statuse count	805
Retweet count	45

To implement this study, a subset of a dataset of tweets related to Covid-19 were examined which were collected by Chen, Lerman, and Ferrara (2020) from January 20, 2020, to May 29, 2021. In this study, the English tweets for each month are randomly chosen, and narrowed down the dataset to 1251,216 tweet IDs. The tweet IDs further were retrieved to tweets’ complete information by using Hydrator software. A laptop with Quad-Core i7–8750 H processors running at 16X PCI-e lanes was used for analyzing the data.

The following table shows the relevant information about the dataset and an example of one unique record. The data were imported into the Python console by using numpy, nltk, and pandas packages. In Table 1, the user ID represents a unique identifier for the tweet, and EN in our dataset refers to English. Furthermore, the number of tweets that are issued by user ID is shown as the user status count, which describes the user’s activity on Twitter. The number of times that the tweet is shared with the user ID’s network is described as the retweet count.

The raw texts were further cleaned by removing punctuation, usernames, URL links, numbers, pictures, and emojis, and converted text to the lowercase. Furthermore, the stop words such as “the”, “the”, “of”, “in”, “at” were removed. Cleaned tweets were tokenized to be processed from sentences to words for future analysis.

3.2. Retweetability measure

To measure the popularity of tweets based on the volume of the retweets, we considered tweets that had at least one retweet during the period from January 20, 2020, to May 29, 2021. The purpose of this categorization is to describe the process of the binary response variable for future analysis.

3.3. Features extraction

Five different categories of features were chosen for this study: (i) topic modeling, (ii) topic modeling plus TF-IDF vectorizer, (iii) BOW by TF-IDF vectorizer, (iv) document embedding, and (v) document embedding plus TF-IDF vectorizer. The following subsections will cover each set of content features, particularly topic modeling and how basic emotions related to Covid-19 were detected using the CrystalFeel algorithm.

3.3.1. Topics analysis using LDA model

Due to the large volume of tweets and retweets, topic modeling was used to classify text data pertaining to Covid-19 based on the frequency of words in each document. The latent Dirichlet allocation (LDA) model (Blei et al., 2003) was applied to identify the most popular topics in tweets related to Covid-19. LDA model is an unsupervised machine learning algorithm that detects a certain number of topics within documents with a certain probability. Note that each topic is also represented as a probabilistic distribution over words. LDA models a corpus D including M documents, and each document has N_d words to the following generative process (Blei et al., 2003): (i) For each topic $(t \in \{1, \dots, T\}$, chooses a multinomial distribution φ_t from a Dirichlet distribution with parameter β. (ii) For each document $(d \in \{1, \dots, M\}$, chooses a multinomial distribution θ_d from a Dirichlet distribution with parameter α. (iii) For each word w_i in document d chooses: (1) a topic z_i from θ_d and (2) a word w_i from φ_t. In the generative process, the

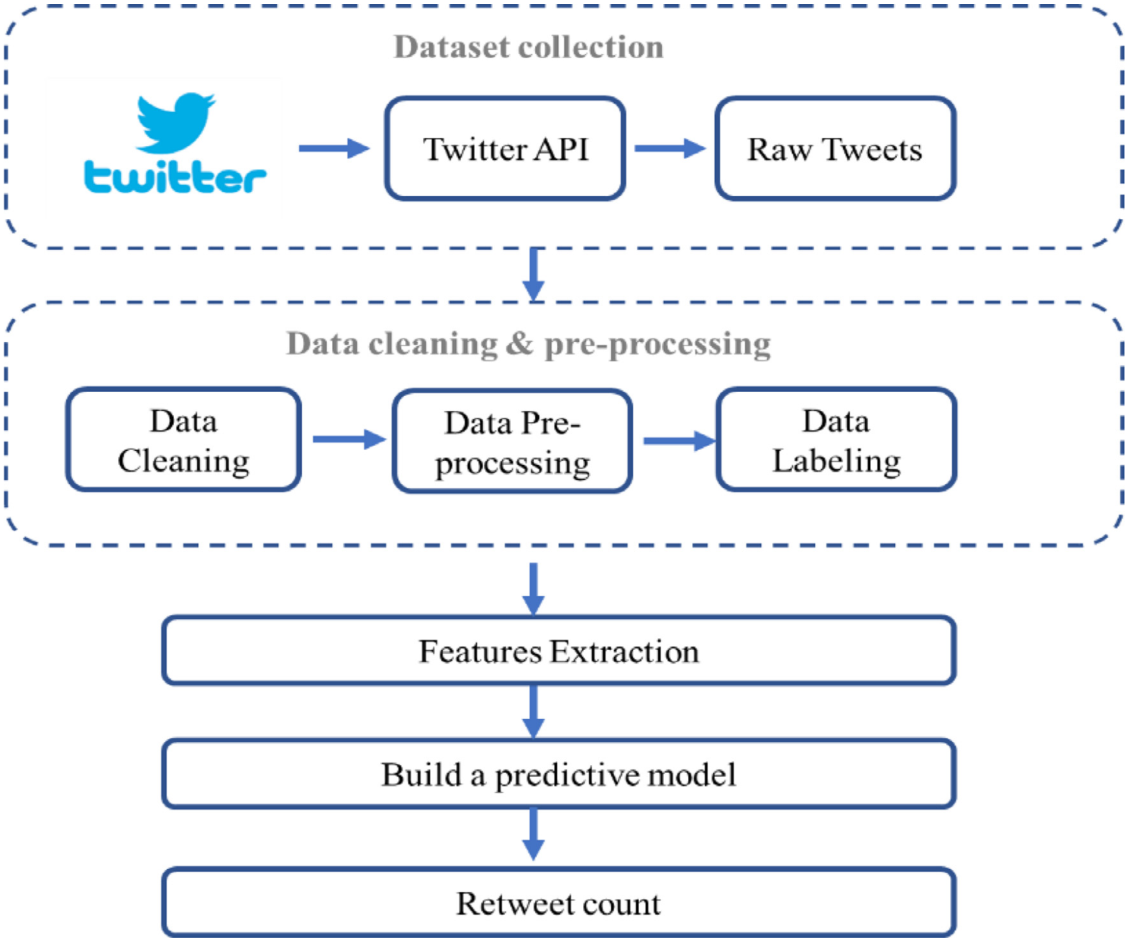


Fig. 1. System architecture of research study.

probability of observed data D is computed as follows:

$$p(D|\alpha,\beta) = \prod_{i=1}^{N_d} \int p(\theta_d|\alpha)\left(\prod_{z_{di}} p(z_{di}|\theta_d)p(w_{di}|z_{di},\beta)\right)d\theta_d \tag{1}$$

In the above equation, z_{di} and w_{di} are word-level variables and θ_d variables are document-level variables. This research aimed to find the optimal number of topics within the documents by calculating the coherence score which is referred as C_v score (Röder, Both, & Hinneburg, 2015) and measures the coherence of the topics by the normalized mutual information (NPMI) metric. NPMI is defined as follows:

$$NPMI\ (w_i, w_j\ ) = \frac{PMI\ (w_i, w_j\ )}{-\ \log\ (p(w_i, w_j\ ))} \tag{2}$$

Where the topic coherence is automatically computed by point wise mutual information (PMI) metric as follows:

$$PMI\ (w_i, w_j\ ) = \ \log \frac{p(w_i, w_j\ )}{p(w_i)\ p(w_j)} \tag{3}$$

And $p(w_i)$ and $p(w_j)$ are probabilities for word w_i and word w_j within the document and $p(w_i, w_j\ )$ is a joint probability of word w_i and word w_j. Given the size of the dataset in this study, applying the LDA model was one of the most effective methodology to extract the features. In this study, Python Scikit-learn's LatentDirichelAllocation function is used with the learning decay of 0.85. Learning decay is a parameter for controlling the learning rate, and its value must be set between 0.5 to 1 to guarantee asymptotic convergence. Fig. 2. shows the optimal number of topics along with the coherence score for the whole dataset. A higher value for the coherence score indicates an optimal number of topics within the documents. The highest coherence value is 0.6088, indicating 38 topics for the whole dataset.

Fig. 3. shows the wordcloud of the most frequent words for all the 38 topics. The LDA algorithm was further applied for tweets related to each phase of the Covid-19 pandemic to identify the most frequent topics and use the CrystalFeel algorithm to detect the four emotions including fear, anger, sadness, and joy.

3.3.1.1. CrystalFeel algorithm. Previous studies analyzed the four emotions in different periods of the pandemic using the CrystalFeel algorithm (Garcia & Berton, 2021; Lwin et al., 2020; Shah, Yan, Qayyum, Naqvi, & Shah, 2021), which has been proven in recent works to be accurate. In this study, the emotional strength scores of the CrystalFeel algorithm (R. K. Gupta & Yang, 2018) were used to label the dominant emotions of fear, anger, sadness, and joy at different phases of the pandemic according to the timeline of WHO tweets and U.S news during the ongoing Covid-19 pandemic. In the CrystalFeel algorithm, topics are labeled based on emotion score (i.e., emotional valence refers to feelings' polarity) in three different categories including: (i) No-specific emotion, (ii) If valence-score is higher than 0.520, then the emotion category is "joy"; (iii) If valence-score is lower than 0.480, then the emotion category is: (1) "anger" if and only if the anger intensity-score is higher than both the fear and sadness intensity-scores, (2) "fear" if and only if the fear intensity-score is higher than both the and sadness intensity-scores, and (3) "sadness" if and only if sadness intensity-score is higher than both the anger and fear intensity-scores (Garcia & Berton, 2021). Fig. 4. illustrates the algorithm 1.

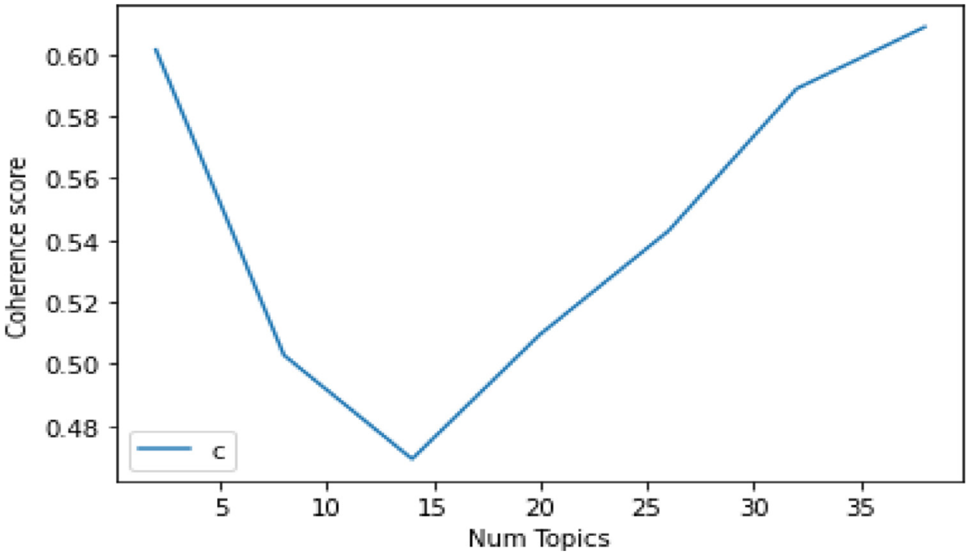


Fig. 2. Coherence measurement for LDA.

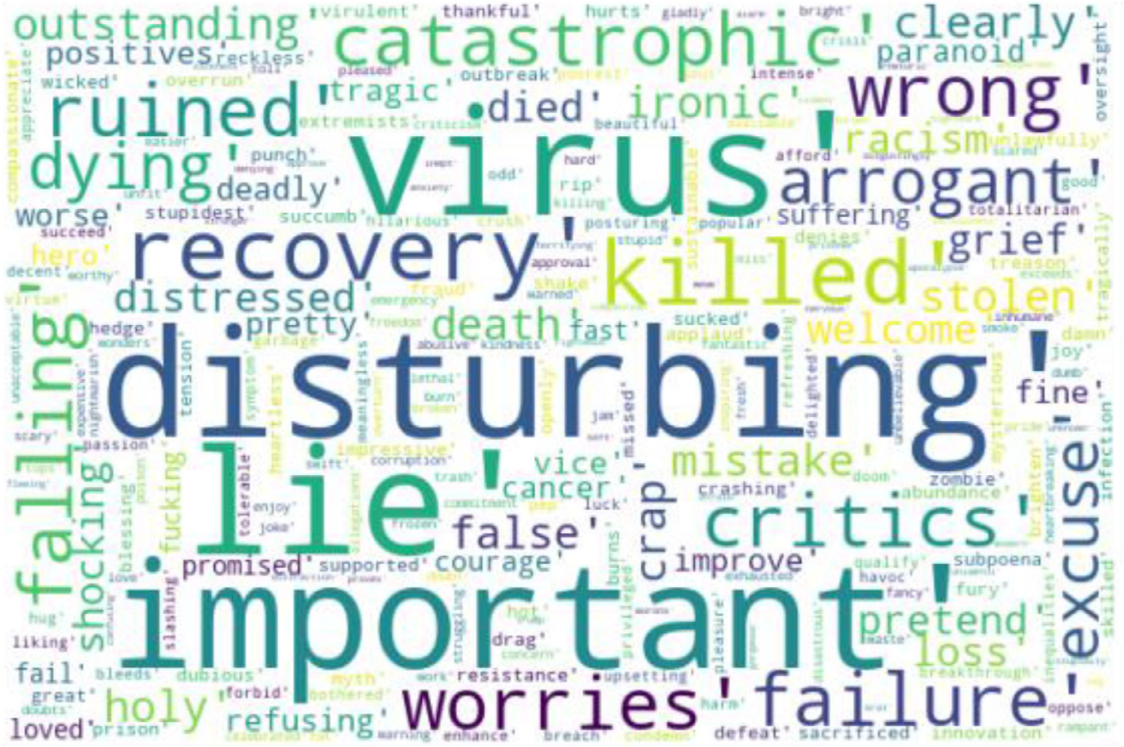


Fig. 3. Wordcloud for all 38 topics with their top thirty frequent words.

Algorithm 1: Emotion score for Label in CrystalFeel

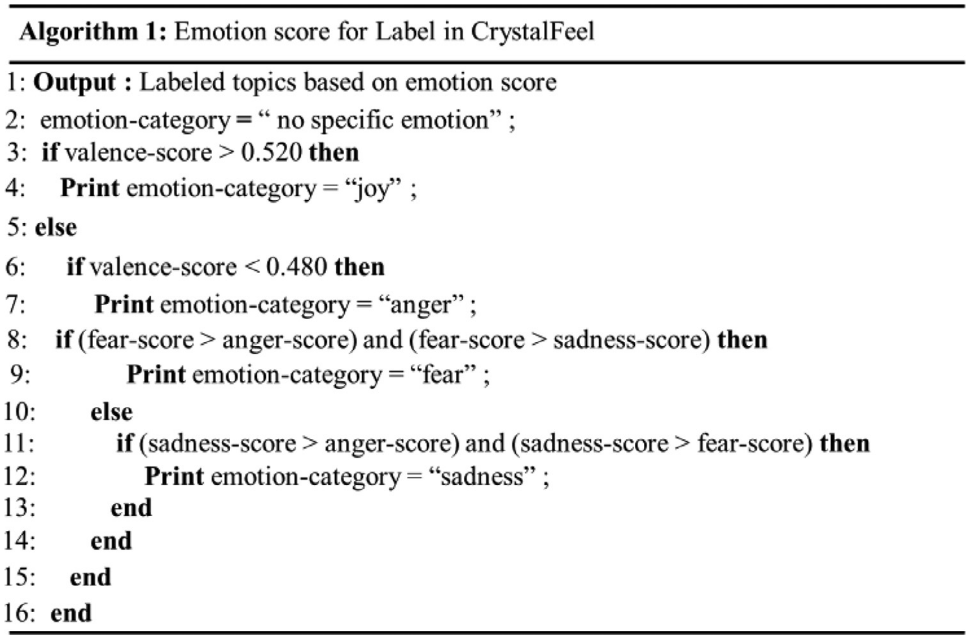
```
Output: Labeled topics based on emotion score
emotion-category = " no specific emotion";
if valence-score > 0.520 then
    Print emotion-category = "joy";
else
    if valence-score < 0.480 then
        Print emotion-category = "anger";
    if (fear-score > anger-score) and (fear-score > sadness-score) then
        Print emotion-category = "fear";
    else
        if (sadness-score > anger-score) and (sadness-score > fear-score) then
            Print emotion-category = "sadness";
        end
    end
  end
end
```

The results of CrystalFeel analysis are shown in Table 2 from January 2020 to May 2021. For each month, the LDA algorithm was applied on the randomly selected tweets, and then the top ten words for each topic were extracted and used as inputs for the CrystalFeel algorithm.

Furthermore, Fig. 5. shows the timeline of the Covid-19 pandemic based on selected WHO tweets and its Covid-19 response[3] and Covid-19 developments in the U.S.[4] for 2020 and 2021. Different stages of the Covid-19 pandemic are shown in Fig. 5. The right side of the figure pertains to WHO tweets and responses to Covid-19, and the left side of the figure shows pandemic-related developments in the U.S. Together, Table 2 and Fig. 5. show that after WHO's announcement of the human-

[3] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline

[4] https://www.ajmc.com/view/a-timeline-of-Covid19-developments



Algorithm 1: Emotion score for Label in CrystalFeel

```
Output : Labeled topics based on emotion score
emotion-category = " no specific emotion" ;
if valence-score > 0.520 then
  Print emotion-category = "joy" ;
else
   if valence-score < 0.480 then
      Print emotion-category = "anger" ;
  if (fear-score > anger-score) and (fear-score > sadness-score) then
         Print emotion-category = "fear" ;
    else
       if (sadness-score > anger-score) and (sadness-score > fear-score) then
          Print emotion-category = "sadness" ;
       end
    end
  end
end
```

Fig. 4. Algorithm 1 Overview.

Table 2
Emotion intensity analysis for tweets.

	Months	Fear Intensity	Anger Intensity	Joy Intensity	Sadness Intensity	Valence Intensity
2020	Jan	0.44	0.441	0.368	0.385	0.468
	Feb	0.465	0.428	0.367	0.381	0.475
	Mar	0.475	0.419	0.369	0.38	0.477
	Apr	0.373	0.367	0.378	0.317	0.543
	May	0.444	0.417	0.367	0.371	0.474
	Jun	0.43	0.442	0.37	0.376	0.463
	Jul	0.445	0.439	0.373	0.47	0.416
	Aug	0.382	0.404	0.336	0.431	0.466
	Sept	0.415	0.447	0.37	0.371	0.469
	Oct	0.471	0.452	0.362	0.38	0.461
	Nov	0.345	0.376	0.378	0.315	0.542
	Dec	0.386	0.384	0.376	0.322	0.522
2021	Jan	0.443	0.426	0.372	0.354	0.475
	Feb	0.304	0.349	0.377	0.311	0.527
	Mar	0.337	0.361	0.381	0.315	0.530
	Apr	0.373	0.392	0.371	0.323	0.521
	May	0.373	0.395	0.372	0.303	0.527

to-human transmission of Covid-19 occurring outside of China, the public response to the Covid-19-ralted news was characterized by anger. By the onset of the pandemic, public response turned to fear. However, there was a sudden increase in joy in April 2020, which marked the beginning of the "lockdown" in the U.S., after the announcement that the unemployment benefits from the U.S. Department of Labor would amount to $600 per week. Countries that experienced the longest stay-at-home orders, saw this joy turned to fear, anger, and sadness in the following months. The announcement that the former president and first lady tested positive for Covid-19 marked when public emotion turned to fear in October 2020. By the end of 2020, public emotions were characterized by joy following the WHO announcement that the Pfizer and Moderna Covid-19 vaccines were effective. However, the beginning of 2021 started with fear which was related to the U.S. 2020 presidential election. The remaining months in 2021 were characterized by joy with the distribution of vaccines in the United States and throughout the world and reopening plans for restaurants and indoor spaces. Fig. 6. shows the line graph of all the four basic emotion from January 2020 to May 202, along with three examples of important events that occurred during the time period.

3.3.2. *Bag of word by using term frequency-inverse document frequency (TF-IDF) vectorizer*

N-gram analysis for extracting features is one of the most reliable, efficient, and fastest techniques for text classification. The process starts by preprocessing language documents by removing unnecessary information, e.g., punctuations, numbers, tags, while keeping necessary terms. N-grams are sequence of words from the documents, and "N" corresponds to the window size of the words in text analysis. In this study, the window size of sequence words for n-gram analysis is one for bag of words, which generates the vocabulary list for all the unique words and their frequencies in the documents. To enhance the performance of classification models, the TF-IDF vectorizer was used to weight the n-gram profiles (Hassan, Gomaa, Khoriba, & Haggag, 2020; Nasser, Karim, el Ouadrhiri, Ali, & Khan, 2021). The highest weight of TF-IDF occurs when a word has high term frequency (TF) in any tweet, and low document frequency (DF) of the word in the entire dataset. In this study, the TF-IDF vectorizer method introduced by Salton and Buckley (1988) was applied to the documents and it is an older method compared to other aforementioned features. The TF-IDF method assumes that the important words in a given document frequently appears in that document

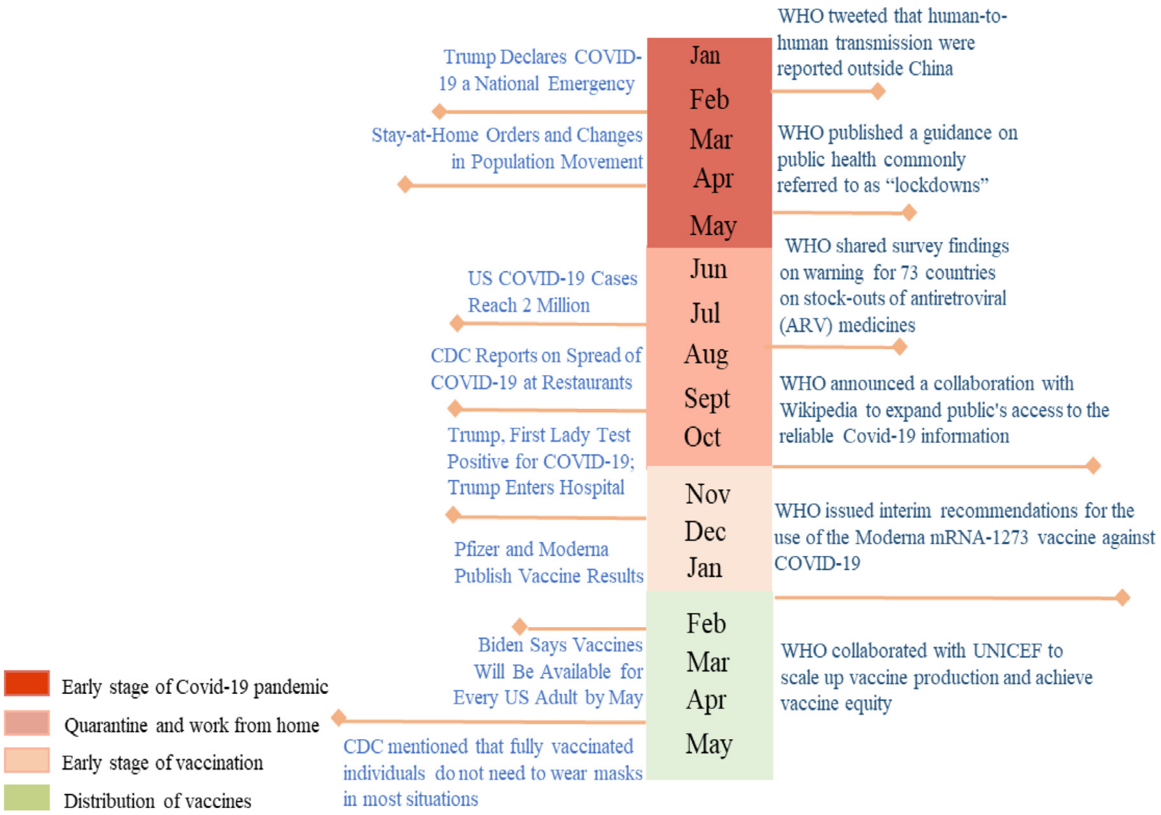

WHO tweeted that human-to-human transmission were reported outside China
Trump Declares COVID-19 a National Emergency
Jan
Feb
Mar
Apr
May
Stay-at-Home Orders and Changes in Population Movement
WHO published a guidance on public health commonly referred to as "lockdowns"
WHO shared survey findings on warning for 73 countries on stock-outs of antiretroviral (ARV) medicines
US COVID-19 Cases Reach 2 Million
Jun
Jul
Aug
Sept
Oct
CDC Reports on Spread of COVID-19 at Restaurants
WHO announced a collaboration with Wikipedia to expand public's access to the reliable Covid-19 information
Trump, First Lady Test Positive for COVID-19; Trump Enters Hospital
Nov
Dec
Jan
WHO issued interim recommendations for the use of the Moderna mRNA-1273 vaccine against COVID-19
Pfizer and Moderna Publish Vaccine Results
Feb
Mar
Apr
May
Biden Says Vaccines Will Be Available for Every US Adult by May
WHO collaborated with UNICEF to scale up vaccine production and achieve vaccine equity
CDC mentioned that fully vaccinated individuals do not need to wear masks in most situations
Early stage of Covid-19 pandemic
Quarantine and work from home
Early stage of vaccination
Distribution of vaccines

Fig. 5. A timeline of Covid-19 developments in USA and WHO's tweets and information related to the recent pandemic in 2020 and early 2021.

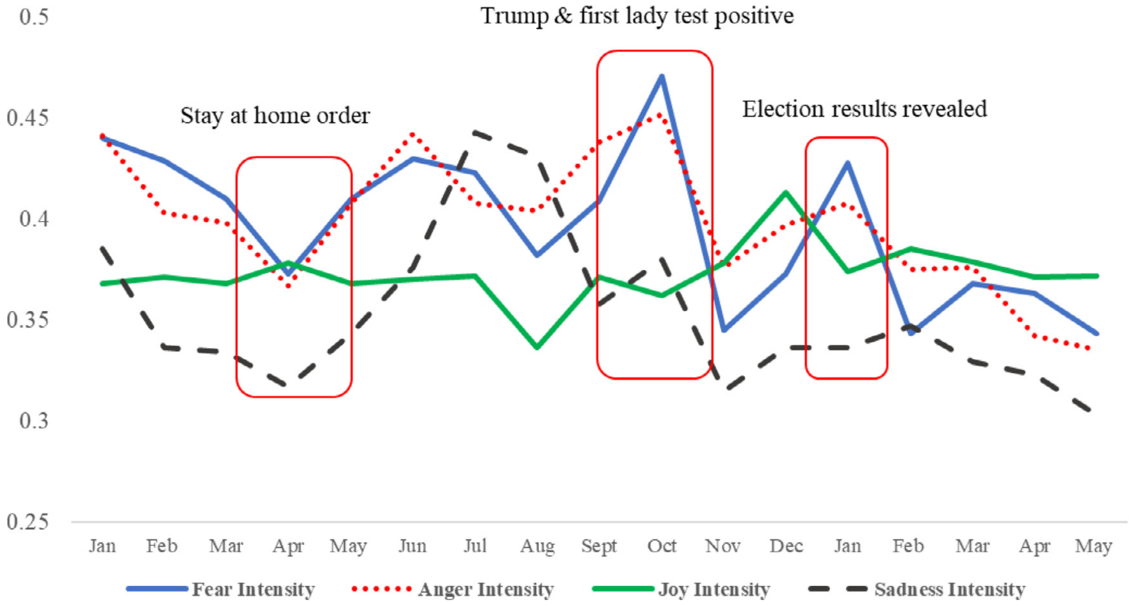


Fig. 6. Emotion intensity score between January 2020 to May 2021.

but rarely appears in other documents which aids in recognizing meaningless terms. Therefore, the frequency of word i within document j is denoted as the parameter tf_{ij} while the frequency of documents with word i is denoted as the parameter df_i. The importance of word i for document j is measured by having a large tf_{ij} and a small df_i and is calculated as follows:

$$TF-IDF_{ij} = tf_{ij} \times \log\left(\frac{N}{df_i + 1}\right) \quad (4)$$

Where "N" is the total number of documents, and $log(\frac{N}{df_i+1})$ represent the inverse document frequency for word i. Table 3 provides examples of tweets along with their top words, TF-IDF values, and detected emotion.

3.3.3. Document embedding

Doc2vector or document embedding is the extension of word embedding for text analysis. Word2Vec can convert tokenize words into a vector that represents the vocabulary of texts within documents. Word2Vec

Table 3
Tweets examples with TF-IDF values, and emotions intensity.

Tweet	Top words in tweet	TF-IDF value		Fear intensity	Anger intensity	Joy intensity	Sadness intensity
Psa facial mask is not effective against wuhan virus	Psa facial effective	0.16513 0.16472 0.16256	0.465		0.398	0.241	0.385
A novel coronavirus is a new strain of the virus that has not been previously identified in humans	Novel new virus	0.12896 0.10394 0.10139	0.528		0.393	0.271	0.414
If coronavirus isn t such a big deal then why are Italian authorities scanning every single passenger landing in the country	Coronavirus authorities scanning	0.13566 0.09949 0.08623	0.558		0.576	0.136	0.451
Why do school kids having been in China need to quarantine themselves for two weeks if Chinese flights bring in 10 000 possible carriers every day let loose to walk among us what about a whole chinese womens soccer team staying in a brisbane hotel they were in Wuhan betrayal	Chinese quarantine day	0.11665 0.09409 0.09111	0.58		0.671	0.146	0.51
More than a 100 people died in China from the novel coronavirus on monday the highest for a day	Novel day people	0.16434 0.16461 0.12312	0.624		0.428	0.264	0.524
As of the end of last year however the number of confirmed cases of vaping related lung illnesses across the country was 2 500 with 54 reported deaths	Reported end number	0.16183 0.15502 0.12341	0.625		0.441	0.212	0.541
We d like to show sweet support to those who have received the covid 19 vaccine starting today bring your vaccine card	Vaccine support like	0.21423 0.17162 0.12543	0.355		0.282	0.413	0.349

enables exploration of the correlation among the words and their contextual information and constructs the network of words. Doc2vec builds a numerical representation of a document where there is a group of words as a unique document to achieve sentence embedding. Thus, when training Word2Vec (Mikolov, Yih, & Zweig, 2013), Doc2vec is also trained. One of the main learning algorithms for Doc2vec that is implemented in this research is distributed bag of word version of paragraph vector (PV-DBOW), which is based on skip-gram. In PV-DBOW, each text is associated with a specific paragraph vector, and each word is associated with a specific word vector in a whole dataset.

The genism package was further imported to Python and created the document-to-vector model to learn the network of documents and to detect similar tweets based on the vector distance.

3.4. *Supervised machine learning algorithms*

Scikit-learn package in Python 3.8 was used to implement three base and effective supervised machine learning algorithms: (i) random forest (RF) (Breiman, 2001) classifier, (ii) stochastic gradient descent (SGD) (Zhang, 2004) classifier, and (iii) logistic regression (LR) (Hosmer, Lemeshow, & Sturdivant, 2013) classifier, and an ensemble voting classifier of the three machine learning algorithms (i.e., RF, SGD, and LR) to enhance accuracy and reduce error rates of classifiers. Each classifier and ensemble approach are explained in detail in the following subsections. Note that, in this study, ensemble voting classifier is referred as EVC for ensemble approach.

3.4.1. *Random forest classifier*

The random forest classifier is a supervised machine learning algorithm. It consists of tree classifiers where each tree is grown with a random vector that is distributed independently and identically, and each tree casts a vote for the most popular class of input vectors (Breiman, 2001). After creation, RF classifier can split into two stages: random forest creation and prediction from the created RF classifier (Biau & Scornet, 2016). The algorithm has the following steps (Neogi, Garg, Mishra, & Dwivedi, 2021) . Step 1 RF randomly selects "k" features from a total of "m" features where $k \ll m$. Step 2 RF calculates the node "d" among the "k" features using the best split point. Step 3 RF uses the optimal split by breaking the node into child nodes. Step 4 RF repeats 1 to 3 steps iteratively until the number of nodes reaches the maximum allocated value. Step 5 RF builds a forest by repeating step 1 to 4 for "n" number time to create "n" number of trees. In this study, RF classifier accuracy was compared with the accuracy of stochastic gradient descent (SGD) and logistic regression (LR), and ensemble voting classifier (EVC).

3.4.2. *Stochastic gradient descent classifier*

The stochastic gradient descent (SGD) classifier is a supervised machine learning algorithm and is a very powerful classifier for building a predictive model (Zhang, 2004). The algorithm has the following steps. Step 1 SGD computes the gradient of the loss function with respect to each feature. Step 2 SGD selects a random initial value for the parameters. Step 3 SGD updates the gradient function by allocating the parameter values. Step 4 SGD calculates the step sizes for each feature with respect to learning rate of algorithm. Step 5 SGD calculates the new parameters. Step 6 SGD repeats step 3 to 5 until the gradient reaches to zero. In SGD classifier, learning rate value has a significant impact on the behavior of gradient descent. Thus, the learning-rate in Python codes is set to "optimal" and the loss function is set to "log" which gives logistic regression, a probabilistic classifier. The log loss function gives the probability of false classifications (Rustam et al., 2021), and can be defines as:

$$logloss = -\frac{1}{N}\sum_{i=1}^{N} y_i . \log\left(p\left(y_i\right)\right) + \left(1 - y_i\right). \log\left(1 - p\left(y_i\right)\right) \tag{5}$$

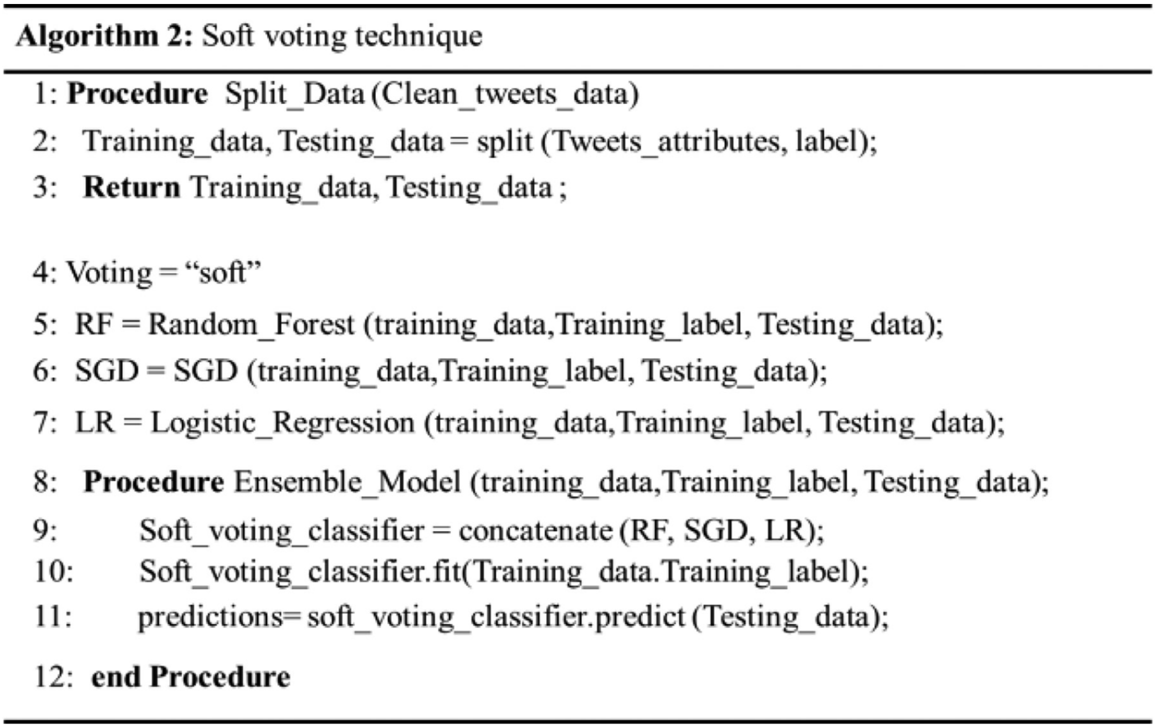
Where N is the number of instances, y_i is the outcome of the i_th instance, and $p(y_i)$ is the probability of the i_th instance for the value y_i.

3.4.3. *Logistic regression classifier*

The logistic regression (LR) classifier is a supervised machine learning algorithm that is used to model the probability of a binary classification problem (Hosmer, Lemeshow, & Sturdivant, 2013). The LR algorithm has the following steps. Step 1 the LR classifier develops the implementation of the sigmoid function. The LR model predicts the binary outcome with sigmoid function as follows:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)}} \tag{6}$$

Where β_0 is the intercept (i.e., the value of bias), β_i, $i \in \{1, 2, \ldots, n\}$ is the inputs coefficient, and X_i, $i \in \{1, 2, \ldots, n\}$ is the input vector. Step 2 the LR classifier determines the cost function. Step 3 the LR classifier



Algorithm 2: Soft voting technique

```
Procedure Split_Data (Clean_tweets_data)
    Training_data, Testing_data = split (Tweets_attributes, label);
    Return Training_data, Testing_data ;

Voting = “soft”
RF = Random_Forest (training_data,Training_label, Testing_data);
SGD = SGD (training_data,Training_label, Testing_data);
LR = Logistic_Regression (training_data,Training_label, Testing_data);
Procedure Ensemble_Model (training_data,Training_label, Testing_data);
        Soft_voting_classifier = concatenate (RF, SGD, LR);
        Soft_voting_classifier.fit(Training_data.Training_label);
        predictions= soft_voting_classifier.predict (Testing_data);
end Procedure
```

Fig. 7. Algorithm 2 overview for soft voting technique.

calculates and updates new coefficients. The value of the coefficient is updated as follows:

$$\beta = \beta + \alpha * (Z - p) * p * (1 - p) * X_i \tag{7}$$

Where α is the learning rate, X_i, $i \in \{1, 2, \ldots, n\}$ is input and Z is the target variable. Step 4 the LR classifier calculates the output with the highest probability. Step 5 the LR classifier repeats steps 1 to 4 and updates the model for each training instance in the dataset. In this classifier, scikit-learn’s LogisticRegression uses liblinear for the solver parameter as a loss function which is the different-different algorithmic style to optimize the loss function, and it supports both L1 and L2 regularization for penalizing the model complexity. Note that, liblinear applies a Newton method for the LR classifier (; Lin, Weng, & Keerthi, 2007).

3.4.4. *Ensemble voting classifier*

An ensemble approach is a combination of classifiers that improves the performance of a classification system (Li, Zong, & Wang, 2007). Classic machine learning methods are trained by using one classification method on the dataset, while ensemble approach is trained by using multiple classifiers. The error rate for ensemble approach is lower than individual classifier’ error rate. To combine the decision of RF, SGD, and LR, this study used soft voting in ensemble approach. The convex combination of the predicted class probabilities was applied for individual classifier. The summation of weights for classifiers was one, and the weighting was chosen based on performance of classifier due to its simplicity and accurate results (Pierola, Epifanio, & Alemany, 2016). In soft voting approach, the predict-proba attribute is used to give the probability of each variable, and shuffles training set, and data points for RF, SGD, and LR classifier. Each classifier computes its prediction with soft voting technique, the majority voting is calculated for the final prediction (Kumari, Kumar, & Mittal, 2021). Fig. 7. illustrates algorithm 2 for the soft voting technique.

Algorithm 2: Soft voting technique

```
Procedure Split_Data (Clean_tweets_data)
    Training_data, Testing_data = split (Tweets_attributes, label)
    Return Training_data, Testing_data
Voting = “soft”
RF = Random_Forest (training_data,Training_label, Testing_data)
SGD = SGD (training_data,Training_label, Testing_data)
LR = Logistic_Regression (training_data,Training_label, Testing_data)
    Procedure Ensemble_Model (training_data,Training_label, Testing_data)
        Soft_voting_classifier = concatenate (RF, SGD, LR);
            Soft_voting_classifier.fit (Training_data.Training_label)
            predictions= soft_voting_classifier.predict (Testing_data)
end Procedure
```

4. Experimental design and analysis

As mentioned in the research method section, tweets related to the Covid-19 pandemic were collected using Twitter APIs (Chen, Lerman, & Ferrara, 2020), and keywords such as: Covid, corona, pandemic, and similar keywords. The study randomly chose 1251,216 tweets written in English that were posted between January 20, 2020 and May 29, 2021. The tweets were labeled as popular and non-popular tweets based on the number of retweets. Each of the classification models used a grid search to find optimal hyper-parameters. The grid search utilized the GridSearchCV object of scikit-learn in Python for all classification models. The results of the models were obtained using five-fold cross-validation with a split ratio of 0.75 to train the classifiers. The optimal hyper-parameters for all the proposed classifiers are summarized in Table 4. Furthermore, to overcome the imbalance data problem, the class weight for each classifier is modified such that higher weight is given to smaller classes to produce optimal results.

4.1. *Model*

The binary response variable in this study was popular versus non-popular tweets based on the volume of retweets, where the tweets with at least one retweet were labeled as popular and tweets with no retweets

Table 4
The hyper-parameters values for all the classifiers.

Classifier	Hyper-parameter	Definition	Optimal value
Random Forest (RF)	n_estimators	The number of trees to be built in the forest	np.arange (5350,25)
	max_depth	The longest path between the root node and the leaf node	np.arange(100,300)
	class-weight	Assigns weights for each class	balanced
	random_state	Sets the random seed given to each estimator at each iteration	25
Stochastic Gradient Descent (SGD)	learning_rate	Controls the over_fitting in the model	optimal
	n_iter	The number of parameter settings that are tried	250
	loss	Measures model(mis) fit given the set of parameters	log
	penalty	Panalized model complexity	l2
	class-weight	Assigns weights for each class	balanced
Logistic Regression (LR)	C-index	The measure of goodness of fit for binary outcome	np.logspace(−4,4.20)
	penalty	Penalized model complexity	['l1′, 'l2′]
	solver	Finds the optimal parameter weights to minimize a cost function	liblinear
	class-weight	Assigns weights for each class	balanced
Ensemble Voting classifier (EVC)	estimators	List of classifiers	('RF', text_classifierRF), ('SGD', text_classifierSGD), ('LG', text_classifierlg)
	voting	Predicts the class label based on argmax of the sums of the predicted probabilities	'soft'
	Flatten_transform	Affects shape of transforms output with the matrix of (n_samples, n_classifiers * n_classes)	TRUE
	weights	weights class probabilities before averaging	[45,35,20]

were labeled as non-popular. Since there were 435,900 non-popular tweets and 815,316 popular tweets, this was an imbalanced dataset. To avoid misleading results due to an imbalanced dataset, an oversampling technique in which the minority class is duplicated was adopted to keep all the relevant information in the training set. Furthermore, three main sets of content features and their combinations were utilized as inputs for three robust and effective machine learning classifiers and an ensemble voting classifier for imbalanced datasets and used to predict the retweetability. To enhance the performance of the classifiers, the feature-extraction function was used from the scikit-learn package in Python 3.8 to extract the lexical features and weight them using a TF-IDF vectorizer. The gensim package was then applied for Doc2vec and LDA, and LatentDirichletAllocation function from Scikit-learn package was used for topics analysis. The parameters for classifiers were also adjusted to prevent poor results. All the classifiers were modified by adding class weights as "balanced" to their cost function where the penalty to the minority class is higher. The scikit-learn Python package provides the class weights for the classifiers. Furthermore, an ensemble voting classifier was applied to enhance the accuracy of prediction and reduce bias, and error rate. This study utilized an ensemble of random forest, stochastic gradient decent, and logistic regression by applying soft voting technique. Furthermore, this research addressed two main components of generating a prediction model. First, tuning the hyper-parameters of each base model, and second, weighting the base models by adopting a soft voting technique to create the prediction model which are explained in the following sections.

4.1.1. *Training time and system configuration*

The classification models were trained for 250 epochs on a system with a RAM of 32 GB. The GPU had a RAM of 8 GB. The unsupervised machine learning algorithms took more than 30 h to train. The supervised machine learning algorithms were efficient and took less time to run and provide outcomes. However, creating an ensemble voting classifier for each set of features took more time for both training and executing the models. By optimizing the hyper-parameters GridSearchCV, and classification models, the performance improved and runtime was more efficient.

4.1.2. *Evaluation metrics*

To evaluate the performance of the selected classifiers, four metrics were chosen: (i) accuracy, (ii) precision, (iii) recall, and (iv) F1-score. The accuracy score is the ratio of correct predictions to total predictions, and range is between zero to 1. The equation for accuracy is as follows:

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

Where TP denotes true positive, FP denotes false positive, TN denotes true negative, and FN denotes false negative. Precision score indicates the proportion of true positive predictions in the list of all the positive predictions. The precision value lies between zero and one, and its equation is as follows:

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

Recall score represents the completeness of a classifier where the number of true positives divided by the total number of true positives and false negatives. The recall value lies between zero to one, and its equation is as follows:

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

F1-score is a harmonic mean of the precision score and recall score, and its value lies between zero to one. The equation for F1-score is as follows:

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{11}$$

The execution time for running the classifiers was utilized to compare and evaluate which classifier consumes a shorter time with more accurate results.

4.2. *Experimental results*

Table 5 summarizes the results of the three supervised machine learning classifiers, and ensemble voting classifier with five sets of features on the Covid-19 tweets. As shown in Table 5, topic modeling has the lowest accuracy of all the classifiers compared to other sets of features. In the first category of features, using topic modeling, EVC has an accuracy of 0.6861, RF has an accuracy of 0.6239, SGD has an accuracy of 0.555, and LR has an accuracy of 0.5506. Adding TF-IDF weighting in topics modeling enhanced the accuracy and F1-score of all the classifiers, particularly the accuracy of the EVC, which raised by 26.43% to the level of 0.9504, and its F1-score increased by 32% to a value of 0.95. Furthermore, with topics plus TF-IDF vectorizer, the RF classifier has an accuracy of 0.9381, and an F1-score of 0.94, the SGD classifier

Table 5
Classifiers results for different sets of features along with the runtime (in second).

	Classifiers	Accuracy	F1-Score	Precision	Recall	Execution time/second	Confusion Matrix [TP FP] [FN TN]
Topic Modeling LDA	Random Forest (RF)	0.6239	0.56	0.48	0.67	110.21	[72,578 78,594] [35,227 116,259]
	Stochastic Gradient Descent (SGD)	0.5550	0.50	0.46	0.57	3.60	[69,411 81,761] [52,798 98,688]
	Logistic Regression (LR)	0.5506	0.63	0.78	0.53	7.34	[118,397 32,775] [103,232 48,254]
	Ensemble Voting classifier (EVC)	0.6861	0.63	0.53	0.76	4140.54	[79,378 70,574] [24,427 128,279]
Topics plus TF-IDF vectorizor	Random Forest (RF)	0.9381	0.94	0.95	0.93	642.61	[143,291 7889] [10,844 140,634]
	Stochastic Gradient Descent (SGD)	0.9304	0.93	0.91	0.95	19.83	[137,874 13,298] [7754 143,759]
	Logistic Regression (LR)	0.9293	0.93	0.91	0.95	39.68	[137,393 13,779] [7613 143,873]
	Ensemble Voting classifier (EVC)	0.9504	0.95	0.94	0.96	12,420.34	[142,490 8682] [6183 145,303]
BOW by TF-IDF vectorizor	Random Forest (RF)	0.9368	0.94	0.94	0.93	879.22	[140,651 8537] [10,578 142,892]
	Stochastic Gradient Descent (SGD)	0.9339	0.94	0.93	0.94	22.34	[143,289 10,098] [9885 139,501]
	Logistic Regression (LR)	0.9308	0.93	0.91	0.95	26.09	[137,476 13,696] [7231 144,255]
	Ensemble Voting classifier (EVC)	0.9437	0.94	0.94	0.95	14,760.07	[142,655 9537] [7498 142,968]
Doc2 vectore vectorizor	Random Forest (RF)	0.8836	0.89	0.91	0.87	593.32	[144,503 13,679] [21,546 122,930]
	Stochastic Gradient Descent (SGD)	0.7829	0.78	0.77	0.79	8.08	[116,916 34,256] [31,443 120,043]
	Logistic Regression (LR)	0.7834	0.78	0.76	0.79	61.31	[115,340 35,832] [29,715 121,771]
	Ensemble Voting classifier (EVC)	0.9206	0.92	0.90	0.93	11,700.15	[135,877 14,288] [9724 142,769]
Doc2 vectore plus TF-IDF vectorizor	Random Forest	0.9214	0.92	0.95	0.88	305.96	[137,268 17,243] [6542 141,605]
	Stochastic Gradient Descent (SGD)	0.9083	0.90	0.93	0.88	12.34	[133,265 18,152] [9593 141,645]
	Logistic Regression (LR)	0.9082	0.90	0.93	0.87	32.46	[131,345 18,875] [8893 143,545]
	Ensemble Voting classifier (EVC)	0.9399	0.94	0.92	0.96	12,240.06	[143,261 12,632] [5534 141,231]

has an accuracy of 0.9304, and an F1-score of 0.93, while the LR classifier has an accuracy of 0.9293, and an F1-score of 0.93. Moreover, BOW by TF-IDF vectorizer for the EVC has a close accuracy to the topics plus TF-IDF vectorizer, with an accuracy of 0.9437 and an F1-score of 0.94, but it also has a longer time runtime of 14,769.07 s. For the RF classifier with BOW by TF-IDF vectorizer, the accuracy is 0.9368 which is slightly lower than the accuracy value for topics plus TF-IDF vectorizer, and F1-score is 0.94 which is as equal as F1-score value for topics plus TF-IDF vectorizer. However, for the SGD classifier with BOW by TF-IDF vectorizer, the accuracy at a value of 0.9339 and F1-score at a value of 0.94 are slightly higher than the accuracy and F1-score value for topics plus TF-IDF vectorizer. For the LR classifier with BOW by TF-IDF vectorizer, the accuracy is 0.9308 which is lower than the accuracy for the SGD classifier. Although, for the fourth set of features, Doc2vector vectorizer, the accuracies for all the classifiers are higher than the accuracies for the classifiers using the first set of features, topic modeling, the performance of classifier is low compared to that of other sets of features. Adding TF-IDF weighting to the Doc2vectore model improved the accuracy of all three classifiers when compared with only applying the Doc2vector feature with the following increases in percentage points: for EVC by 1.93%, for the RF by 3.78%, for the SGD by 12.5%, and for the LR by 12.4%.

In sum, the EVC achieved the highest accuracy compared with the RF classifier, the SGD classifier and the LR classifier for all five sets of features, particularly when using topics plus TF-IDF vectorizer feature with a runtime of 12,420.34 s. Table 5 also shows that although the RF, SGD, and LR classifiers had the shortest runtime of all the models compared with ensemble approach, the accuracy of their models was not as high as the ensemble approach. Fig. 8. shows the F1-score for the four classifiers and all five sets of features. However, with applying ensemble approach and soft voting technique the runtime increased for all five sets of features. The runtime of each model depends on the complexity of the base learners and the size of the dataset. Fig. 9. shows the comparison between the runtime of models by using ensemble approach, and the accuracy of the models. Among all the sets of features, topics plus TF-IDF vectorizer has the highest accuracy, and the runtime is relatively short compared to BOW by TF-IDF vectorizer.

5. Discussion

Inaccurate information related to the ongoing COVID-19 pandemic and the safety of vaccines and their side effects spread quickly through social media, especially via retweets on Twitter. Therefore, it has become more important to address misinformation (Budhwani & Sun, 2020; Forati & Ghose, 2021; Singh et al., 2020). Prior research has explored the essential characteristics of retweet prediction, including retweeting behaviors, emoji and playfulness engagement, and number of followers. However, there is less progress in exploring the content of tweets and in predicting the retweetability over the phases of the pandemic from the initial spread of the virus to the distribution of vaccines. In this study, the content and popularity of tweets and public opinion and emotions were analyzed according to the number of retweets oc-

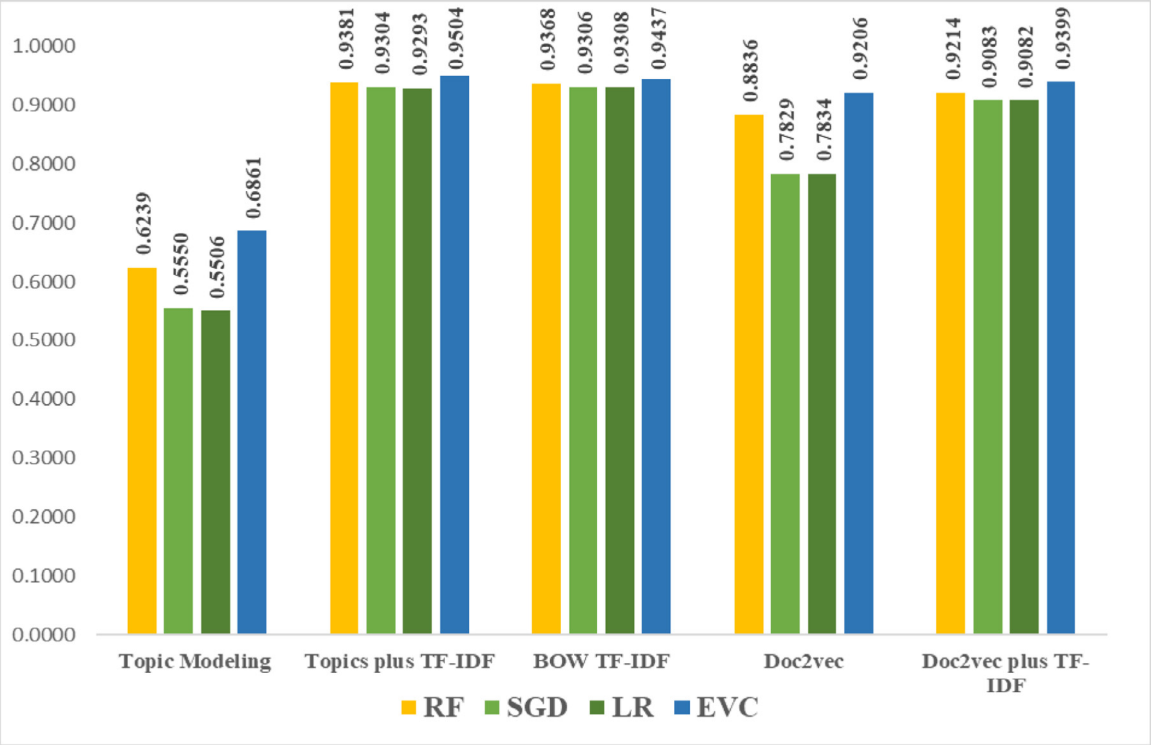


Fig. 8. Accuracy of classifiers for different sets of features.

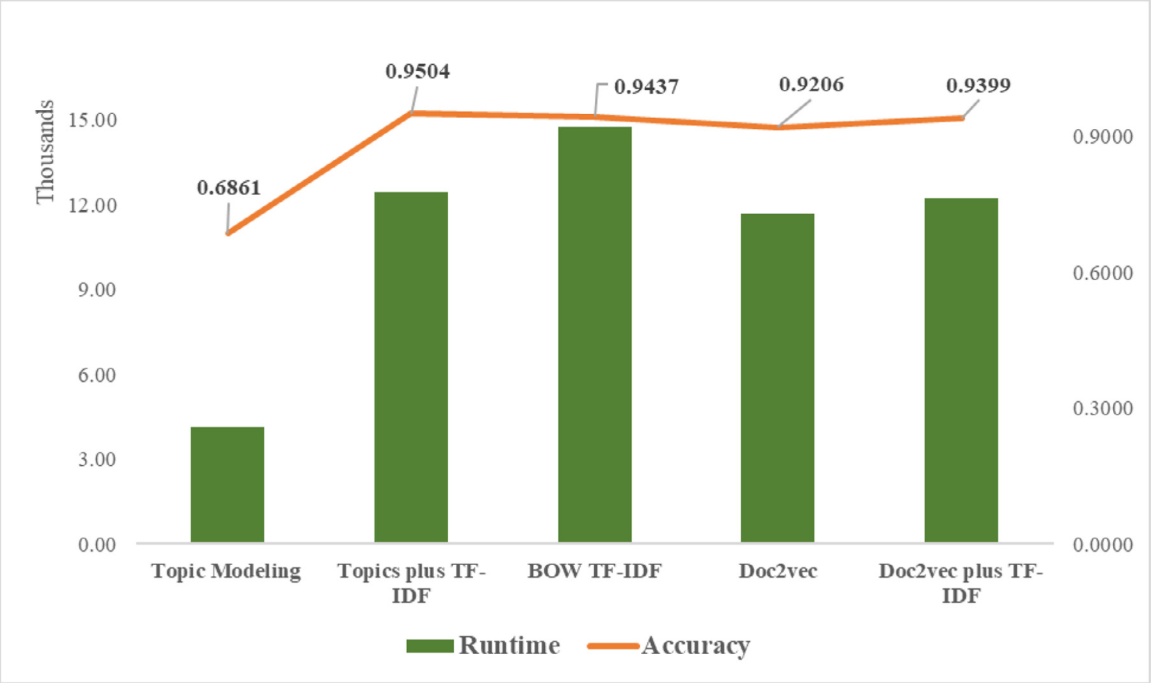


Fig. 9. Comparison of runtime vs accuracy for Ensemble voting classifier.

curring during different phases of the Covid-19 pandemic. Five different sets of content features (i.e., topic modeling, BOW by TF-IDF vectorizer, topics plus TF-IDF vectorizer, Doc2vec, and Doc2vect plus TF-IDF vectorizer) were selected, compared, and then used for three effective and robust classifiers, random forest, stochastic gradient descent, and logistic regression, and an ensemble voting classifier which is a meta classifier to evaluate and compare the outcomes. The results highlighted a strong support for the study's contributions by introducing a novel approach to extract the features from tweets and to predict their retweetability using supervised machine learning algorithms.

The results of this study showed that topics plus TF-IDF vectorizers outperformed other sets of features for all the base classifiers and the ensemble voting classifier. The result of BOW by TF-IDF vectorizers as a content feature set was very close to topics plus TF-IDF vectorizers. One possible explanation is that all tweets pertained to Covid-19, so the performance of the basic text representation was close to that of topic modeling. Moreover, the results of all the experiments in this study confirmed that the EVC has the highest accuracy compared with the state-of-art methods.

5.1. Implications of this study

The results of this study have several theoretical and practical implications. To the best of our knowledge, this is the first study that used the most updated dataset that covers tweets from the onset of the pandemic to the distribution of vaccines. As such, this is the first study that utilized unsupervised machine learning algorithms such as LDA, and document embedding to extract the features and apply them to the supervised machine learning algorithms such as random forest, stochastic gradient descent, and logistic regression, and an optimal ensemble voting model of the selected classifiers to build a predictive model for their retweetability. Furthermore, by applying the LDA algorithm, the most popular topics for each month were identified. The CrystalFeel algorithm was employed to label the public emotions in response to the Covid-19 pandemic, to analyze the patterns in public opinion and emotions, and to extract the most effective features for the predictive model.

In terms of practical implications, the results of this research can be adopted to create a recommendation system for tweets that are relevant to certain events, or as a means of obtaining a higher number of retweets. Identifying patterns in public emotions during the ongoing pandemic can help public health authorities make strategic decisions regarding communication during critical events such as a pandemic. The findings of this study show that although negative emotions, such as anger, fear and sadness were dominant in the early stages of the Covid-19 pandemic, the vaccine rollout and published results on vaccine effectiveness has a positive influence on public emotions. Furthermore, the finding of this study can help to detect and minimize the misleading information related to Covid-19 on Twitter.

6. Conclusion

In this study, the popularity of tweets (based on the number of retweets) was predicted by extracting content features from tweets written in English on the Twitter platform from January 20, 2020, to May 29, 2021. This study shows that the popularity of tweets based on the number of retweets can be drawn from the content of tweets and certain repeated terms during important events such as the Covid-19 pandemic. This section discusses the findings of the study, and its limitations. The results of this study revealed how public opinion changed throughout the stages of the Covid-19 pandemic. The study aimed to select the effective features from the content of the posted tweets by applying unsupervised machine learning algorithms and then to use them as inputs to feed the selected supervised machine learning algorithms for predicting retweetability. Identifying negative and misleading sentiments on popular social media platforms such as Twitter can help to prevent the spread of misinformation. Promoting accurate information and positive sentiments can enhance public awareness regarding certain events such as pandemics. In the proposed approach, the most popular topics at different stages of the pandemic were first identified by using the LDA, and the emotional intensity were detected by employing the CrystalFeel algorithm (Gupta & Yang, 2018) for four emotions: fear, anger, joy and sadness. Second, they were used as one category of content features along with other sets of features to apply them to the selected classifiers. The results showed that topics plus TF-IDF vectorizers feature set had the highest accuracy compared with other sets of content features, and the ensemble voting classifier by ensemble of three machine learning algorithms such as random forest, stochastic gradient decent, and logistic regression had the highest performance when compared with the state-of-art classifiers.

The analysis in this study was limited to tweets written in English and related to Covid-19. Future studies can expand the analysis into different languages. Furthermore, the findings of this study are limited to only users on Twitter platform; future research can explore text content from other social platform to compare the results.

References

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Top Concerns of Tweeters during the COVID-19 Pandemic: Infoveillance study. *Journal of Medical Internet Research, 22*(4), 1–9. 10.2196/19016.

Abdullah, N. A., Nishioka, D., Tanaka, Y., & Murayama, Y. (2015). User's action and decision making of retweet messages towards reducing misinformation spread during disaster. *Journal of Information Processing, 23*(1), 31–40. 10.2197/ipsjjip.23.31.

Baboukardos, D., Gaia, S., & She, C. (2021). Social performance and social media activity in times of pandemic: Evidence from COVID-19-related Twitter activity. *Corporate Governance: The International Journal of Business in Society, ahead-of-print(ahead-of-print), 21*(6), 1271–1289. 10.1108/CG-09-2020-0438.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST, 25*(2), 197–227. 10.1007/s11749-016-0481-7.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022 Jan.

Boyd, D., Golder, S., & Lotan, G. (2010, January). Tweet, Tweet, Retweet: conversational aspects of retweeting on Twitter. *2010 43rd Hawaii International Conference on System Sciences*. 10.1109/HICSS.2010.412.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. 10.1023/A:1010933404324.

Budhwani, H., & Sun, R. (2020). Creating COVID-19 stigma by referencing the novel coronavirus as the "Chinese virus" on twitter: Quantitative analysis of social media data. *Journal of Medical Internet Research, 22*(5), 1–7. 10.2196/19301.

Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance, 6*(2), 1–9. 10.2196/19273.

Chintalapudi, N., Battineni, G., Canio, M. di, Sagaro, G. G., & Amenta, F. (2021). Text mining with sentiment analysis on seafarers' medical documents. *International Journal of Information Management Data Insights, 1*(1), 1–9 100005. 10.1016/J.JJIMEI.2020.100005.

Chung, A., Woo, H., & Lee, K. (2020). Understanding the information diffusion of tweets of a non-profit organization that targets female audiences: An examination of Women Who Code's tweets. *Journal of Communication Management, 25*(1), 68–84. 10.1108/JCOM-05-2020-0036.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., et al. (2020). The COVID-19 social media infodemic. *Scientific Reports, 10*(1), 1–10. 10.1038/s41598-020-73510-5.

Forati, A. M., & Ghose, R. (2021). Geospatial analysis of misinformation in COVID-19 related tweets. *Applied Geography, 133*, 1–10 102473. 10.1016/J.APGEOG.2021.102473.

Garcia, K., & Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing, 101*, 1–15 107057. 10.1016/j.asoc.2020.107057.

Guidry, J. P. D., Waters, R. D., & Saxton, G. D. (2014). Moving social marketing beyond personal change to social change. *Journal of Social Marketing, 4*(3), 240–260. 10.1108/JSOCM-02-2014-0014.

Gupta, R. K., & Yang, Y. (2018). CrystalFeel at SemEval-2018 Task 1: Understanding and detecting emotion intensity using affective lexicons. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. 10.18653/v1/S18-1038.

Gupta, V., Jain, N., Katariya, P., Kumar, A., Mohan, S., Ahmadian, A., et al. (2021). An emotion care model using multimodal textual analysis on COVID-19. *Chaos, Solitons & Fractals, 144*, 1–9 110708. 10.1016/J.CHAOS.2021.110708.

Hosmer, D. W. Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*: 398 (3rd). Hoboken, New Jersey: John Wiley & Sons, Inc..

Jain, V. K., & Kumar, S. (2015). An effective approach to track levels of influenza-A (H1N1) Pandemic in India using Twitter. *Procedia Computer Science, 70*, 801–807. 10.1016/j.procs.2015.10.120.

Hassan, N., Gomaa, W., Khoriba, G., & Haggag, M. (2020). Credibility detection in Twitter using Word N-gram analysis and supervised machine learning techniques. *International Journal of Intelligent Engineering and Systems, 13*(1), 291–300. 10.22266/ijies2020.0229.27.

Kabir, M. Y., & Madria, S. (2021). EMOCOV: Machine learning for emotion detection, analysis and visualization using COVID-19 tweets. *Online Social Networks and Media, 23*, 1–12 100135. 10.1016/J.OSNEM.2021.100135.

Kar, A. K., & Dwivedi, Y. K. (2020). Theory building with big data-driven research – Moving away from the "What" towards the "Why. *International Journal of Information Management, 54*, 1–10. 10.1016/j.ijinfomgt.2020.102205.

Kaur, A., Mittal, N., Khosla, P. K., & Mittal, M. (2021). *Machine Learning Tools to Predict the Impact of Quarantine*, 307–323. 10.1007/978-981-33-4236-1_17.

Kaur, S., Kaul, P., & Zadeh, P. M. (2020). Monitoring the dynamics of emotions during Covid-19 using twitter data. *Procedia Computer Science, 177*, 423–430. 10.1016/j.procs.2020.10.056.

Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering, 2*, 40–46. 10.1016/j.ijcce.2021.01.001.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media?. In *Proceedings of the 19th International Conference on World Wide Web - WWW '10*. 10.1145/1772690.1772751.

Lazard, A. J., Scheinfeld, E., Bernhardt, J. M., Wilcox, G. B., & Suran, M. (2015). Detecting themes of public concern: A text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. *American Journal of Infection Control, 43*(10), 1109–1111. 10.1016/j.ajic.2015.05.025.

Li, S., Zong, C., & Wang, X. (2007). Sentiment Classification through Combining Classifiers with Multiple Feature Sets. In *2007 International* Conference on Natural Language Processing and Knowledge Engineering (pp. 135–140). 10.1109/NLPKE.2007.4368024.

Liang, J., Jiang, B., Yin, R., Wang, C., Tan, J. L., & Bai, S. (2016). RTPMF: Leveraging user and message embeddings for retweeting behavior prediction. *Procedia Computer Science, 80*, 356–365. 10.1016/j.procs.2016.05.351.

Lin, C.-J., Weng, R. C., & Keerthi, S. S. (2007). Trust region Newton methods for large-scale logistic regression. In *Proceedings of the 24th International Conference on Machine Learning - ICML '07* (pp. 561–568). 10.1145/1273496.1273567.

Lwin, M. O., Lu, J., Sheldenkar, A., Schulz, P. J., Shin, W., Gupta, R., et al. (2020). Global sentiments surrounding the COVID-19 Pandemic on Twitter: Analysis of Twitter trends. *JMIR Public Health and Surveillance, 6*(2), 1–4. 10.2196/19447.

Mackey, T., Purushothaman, V., Li, J., Shah, N., Nali, M., Bardier, C., & Cuomo, R. (2020). Machine learning to detect self-reporting of symptoms, testing access, and recovery associated With COVID-19 on Twitter: Retrospective big data infoveillance study. *JMIR Public Health and Surveillance, 6*(2), 1–9. 10.2196/19509.

Marino, V., & lo Presti, L. (2018). From citizens to partners: The role of social media content in fostering citizen engagement. *Transforming Government: People, Process and Policy, 12*(1), 39–60. 10.1108/TG-07-2017-0041.

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *The North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751.

Mishra, R. K., Urolagin, S., & Jothi, J. A. A (2019). A Sentiment analysis-based hotel recommendation using TF-IDF Approach. In *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)* (pp. 811–815). 10.1109/ICCIKE47802.2019.9004385.

Mishra, R. K., Urolagin, S., & Jothi, J. A. A. (2020). Sentiment analysis for POI recommender systems. In *2020 Seventh* International Conference on Information Technology Trends (ITT) (pp. 174–179). 10.1109/ITT51279.2020.9320885.

Mishra, R. K., Urolagin, S., Jothi, J. A. A., Neogi, A. S., & Nawaz, N. (2021). Deep learning-based sentiment analysis and topic modeling on tourism during Covid-19 Pandemic. *Frontiers in Computer Science, 3*, 1–14. 10.3389/fcomp.2021.775368.

Mittal, R., Ahmed, W., Mittal, A., & Aggarwal, I. (2021). Twitter users exhibited coping behaviours during the COVID-19 lockdown: An analysis of tweets using mixed methods. *Information Discovery and Delivery, 49*(3), 193–202. 10.1108/IDD-08-2020-0102.

Mohammed, A., & Ferraris, A. (2021). Factors influencing user participation in social media: Evidence from twitter usage during COVID-19 pandemic in Saudi Arabia. *Technology in Society, 66*, 1–10 101651. 10.1016/J.TECHSOC.2021.101651.

Nasser, N., Karim, L., el Ouadrhiri, A., Ali, A., & Khan, N. (2021). n-Gram based language processing using Twitter dataset to identify COVID-19 patients. *Sustainable Cities and Society, 72*, 1–8. 10.1016/j.scs.2021.103048.

Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Bad news travel fast. In *Proceedings of the 3rd International Web Science Conference on - WebSci '11*. 10.1145/2527031.2527052.

Neogi, A. S., Garg, K. A., Mishra, R. K., & Dwivedi, Y. K. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. *International Journal of Information Management Data Insights*, (2), 1–11. 10.1016/j.jjimei.2021.100019.

Odlum, M., & Yoon, S. (2015). What can we learn about the Ebola outbreak from tweets? *American Journal of Infection Control, 43*(6), 563–571. 10.1016/j.ajic.2015.02.023.

Pierola, A., Epifanio, I., & Alemany, S. (2016). An ensemble of ordered logistic regression and random forest for child garment size matching. *Computers & Industrial Engineering, 101*, 455–465. 10.1016/j.cie.2016.10.013.

Rajendran, D. P. D., & Sundarraj, R. P. (2021). Using topic models with browsing history in hybrid collaborative filtering recommender system: Experiments with user ratings. *International Journal of Information Management Data Insights, 1*(2), 1–12 100027. 10.1016/J.JJIMEI.2021.100027.

Rao, H. R., Vemprala, N., Akello, P., & Valecha, R. (2020). Retweets of officials' alarming vs reassuring messages during the COVID-19 pandemic: Implications for crisis management. *International Journal of Information Management, 55*, 1–6 102187. 10.1016/J.IJINFOMGT.2020.102187.

Röder, M., Both, A., & Hinneburg, A. (2015, February 2). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. 10.1145/2684822.2685324.

Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PloS one, 16*(2), 1–23. 10.1371/journal.pone.0245909.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513–523. 10.1016/0306-4573(88)90021-0.

Shah, A. M., Yan, X., Qayyum, A., Naqvi, R. A., & Shah, S. J. (2021). Mining topic and sentiment dynamics in physician rating websites during the early wave of the COVID-19 pandemic: Machine learning approach. *International Journal of Medical Informatics, 149*(6), 1–16 104434. 10.1016/J.IJMEDINF.2021.104434.

Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2021). An exploratory study of COVID-19 misinformation on Twitter. *Online Social Networks and Media, 22*, 1–16 100104. 10.1016/J.OSNEM.2020.100104.

Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., et al. (2020). A first look at COVID-19 information and misinformation sharing on Twitter. *ArXiv.*, 1–24. http://www.ncbi.nlm.nih.gov/pubmed/32550244.

Stokes, D. C., Andy, A., Guntuku, S. C., Ungar, L. H., & Merchant, R. M. (2020). Public priorities and concerns regarding COVID-19 in an online discussion forum: Longitudinal topic modeling. *Journal of General Internal Medicine, 35*(7), 2244–2247. 10.1007/s11606-020-05889-w.

Su, Y., Venkat, A., Yadav, Y., Puglisi, L. B., & Fodeh, S. J. (2021). Twitter-based analysis reveals differential COVID-19 concerns across areas with socioeconomic disparities. *Computers in Biology and Medicine, 132*, 1–13 104336. 10.1016/J.COMPBIOMED.2021.104336.

Sv, P., Tandon, J., Vikas, & Hinduja, H. (2021). Indian citizen's perspective about side effects of COVID-19 vaccine – A machine learning study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 15*(4), 1–5 102172. 10.1016/J.DSX.2021.06.009.

Szomszor, M., Kostkova, P., & Louis, C. S. (2011, August). Twitter informatics: Tracking and understanding Public Reaction during the 2009 Swine Flu Pandemic. *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. 10.1109/WI-IAT.2011.311.

Yao, Z., Yang, J., Liu, J., Keith, M., & Guan, C. H. (2021). Comparing tweet sentiments in megacities using machine learning techniques: In the midst of COVID-19. *Cities (London, England), 116*, 1–13 103273. 10.1016/J.CITIES.2021.103273.

Younis, J., Freitag, H., Ruthberg, J. S., Romanes, J. P., Nielsen, C., & Mehta, N. (2020). Social media as an early proxy for social distancing indicated by the COVID-19 reproduction number: Observational study. *JMIR Public Health and Surveillance, 6*(4), 1–8. 10.2196/21340.

Yousefinaghani, S., Dara, R., Mubareka, S., Papadopoulos, A., & Sharif, S. (2021). An analysis of COVID-19 vaccine sentiments and opinions on Twitter. *International Journal of Infectious Diseases, 108*, 256–262. 10.1016/J.IJID.2021.05.059.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. *Twenty-First International Conference on Machine Learning - ICML '04*. 10.1145/1015330.1015332.