


How Do People View COVID-19 Vaccines: Analyses on Tweets About COVID-19 Vaccines Using Natural Language Processing and Sentiment Analysis

Victor Chang, Aston University, UK*

 <https://orcid.org/0000-0002-8012-5852>

Chun Yu Ng, Teesside University, UK

Qianwen Ariel Xu, Teesside University, UK

Mohsen Guizani, Mohamed Bin Zayed University of Artificial Intelligence, UAE

M. A. Hossain, Vice President Office, Cambodia University of Technology and Science, Cambodia

ABSTRACT

The COVID-19 pandemic has been the most devastating public health crisis in the recent decade and vaccination is anticipated as the means to terminate the pandemic. People's views and feelings over COVID-19 vaccines determine the success of vaccination. This study was set to investigate sentiments and common topics about COVID-19 vaccines by machine learning sentiment and topic analyses with natural language processing on massive tweets data. Findings revealed that concern on COVID-19 vaccine grew alongside the introduction and start of vaccination programs. Overall positive sentiments and emotions were greater than negative ones. Common topics include vaccine development for progression, effectiveness, safety, availability, sharing of vaccines received, and updates on pandemics and government policies. Outcomes suggested the current atmosphere and its focus over the COVID-19 vaccine issue for the public health sector and policymakers for better decision-making. Evaluations on analytical methods were performed additionally.

KEYWORDS

Clustering, COVID-19, Machine Learning, Sentiment Analysis, Topic Analysis, Twitter, Vaccine

I. INTRODUCTION

Began in December 2019, an outbreak of a novel coronavirus disease (COVID-19) was first found in China (World Health Organization, 2020). COVID-19 was officially defined as a pandemic by the World Health Organisation in March 2020 following a mass global spread. By the beginning of the current study in Feb 2021, the cumulative confirmed infection cases exceeded 110 million and the death toll over 2.5 million worldwide (World Health Organization, 2021), making it the worst public health crisis in the recent decade. Government, medical professionals and pharmaceutical companies had endeavored to develop vaccines that produce immunity to COVID-19. The goal of vaccination is to achieve herd immunity that hopefully ends the pandemic. Acceptance of vaccines is important

DOI: 10.4018/JGIM.300817

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

as the success of herd immunity depends on the scale of the population vaccinated (Fontanet and Cauchemez, 2020). However, acceptance of the COVID-19 vaccine was claimed varied globally and unpromising in certain areas of the world (Sallam, 2021; Malik et al., 2020). Thus, it is necessary to monitor and understand the sentiments and opinions of the general public to build confidence for vaccination and identify skeptics that lead to a reduction in public confidence (de Figueiredo et al., 2020; Lazarus et al., 2021; Bloom et al., 2020). Investigation on social sharing in the general population is also essential as social interactions, especially dissemination of information, would induce to influence on public perceptions over topics like epidemics (Funk et al., 2009).

Social media allows the population to share their daily happenings, feelings, and thoughts over events within their communities, providing massive textual data for potential sentiment analyses. With a publicly available application programming interface (API) enabling convenient data gathering, Twitter is one of the most widely used and representative social media platforms commonly employed as a data source for text mining and analysis. Due to the social distancing measures to control the spread of the disease, social media usage became even more prevalent, playing a critical role in keeping people connected and informed during the COVID-19 pandemic (Nabity-Grover et al., 2020; Mehla et al., 2021). This results in immense textual data for various text mining or analysis. Traditional surveys typically have small sample sizes, closed questions and limited spatiotemporal granularity. Compared to the traditional surveys, analysis results on social media data grant an overview of the sentiments and opinions of larger communities and changes over time. In order to deal with the drawback of social media data being mostly unstructured, natural language and machine learning algorithms can be employed to mine sentiments and topics from texts.

The main objective of this study was to discover sentiments and identify public opinions related to the COVID-19 vaccine from social media data. Therefore, tweets from Twitter were acquired and analyzed using natural language processing and machine learning algorithms, including sentiment analysis and topic analysis. Temporal changes were also examined to understand the people's view on COVID-19 vaccines throughout pandemics and vaccine development. Evaluation of different sentiment and topic analytical methods applied was the secondary objective. This study was set to contribute to the literature by

- 1) expanding the understanding of public sentiments and emotions over a crucial sub-topic of COVID-19, vaccination,
- 2) identifying topics worth public health professionals' and stakeholders' notices for critical decision making, and
- 3) demonstrating and comparing multiple sentiments and topic analysis methodologies on vast social media textual data.

II. LITERATURE REVIEW

Social media data has been widely used for analyses in many previous works of literature in the recent decade. Applying machine learning algorithms, researchers had analyzed and extracted opinions and sentiments from Twitter data on many different fields, such as business (Jagdale et al., 2019; Uma and Thirupathi, 2018), politics (Adarsh and Ravikumar, 2015), digital technology (Maindola et al., 2018), and journalism (Donnell and Hutchinson, 2015). For the public health sector, early in 2009, Chew and Eysenbach (Chew and Eysenbach, 2010) began analyzing tweets for the H1N1 influenza pandemic, demonstrating social media was a source of opinions worth stakeholders' looking to design proper responses to public health concerns. Since then, a couple of works of literature have shed more light on Twitter data usage in evaluation or surveillance on different epidemics and pandemics, including seasonal flu (Comito et al., 2018), Zika virus (Ahmed et al., 2020), and Ebola (Kim et al., 2016). Recently, various researchers have published tweets analyses on the COVID-19 pandemic. Most of them aimed to investigate general sentiments towards COVID-19 using different means of sentiment

analyses, e.g., Rajput et al. (2004), Bhat et al. (2020), Dubey (2020) and Lwin et al. (2020). Some researchers employed further techniques to extract popular topics from COVID-19 related tweets (Abd-Alrazaq et al., 2020; Xue et al., 2020; Melo and Figueiredo, 2021). Apart from general topics, specific themes about COVID-19 were studied, for instance, the use of controversial terms and racial discrimination (Chen et al., 2004), facemask usage (Sanders et al., 2021; Yeung et al., 2020), and world leaders' responses to COVID-19 (Rufai and Bunce, 2020). Researches related to the COVID-19 vaccine were relatively few and geographically limited (Dubey, 2021; Hussain et al., 2021).

Sentiment analysis is defined as a classification process that identifies sentiments and classifies their polarity or category (Medhat et al., 2014). It usually involves two major procedures, feature selection and sentiment classification. Feature selection refers to extracting and selecting text features from target documents, including terms presence and frequency, parts of speech, opinion words and phrases, and negations (Aggarwal and Zhai, 2012). These are usually incorporated as the pre-processing steps for textual data. Bag of Words (BoW) is thus created for later procedures. For Sentiment classification, three commonly used approaches are the machine learning approach, lexicon-based approach and hybrid approach (Maynard and Funk, 2011). While the machine learning approach applies supervised or unsupervised machine learning algorithms to define words, sentences, or documents into different sentiments, the lexicon-based approach relies on an opinion lexicon or dictionary. The lexicon can be produced manually or by growing a seed list of words by searching in the well-known corpora for their synonyms and antonyms (Kim and Hovy, 2004) or used along with a set of linguistic constraints to identify additional opinion words and their orientations (Hatzivassiloglou and McKeown, 1997). For practical implementations, lexicon-based sentiment analysis tools were usually applied, along with a hybrid approach incorporating unsupervised clustering in certain cases (Rajput et al., 2004; Bhat et al., 2020; Dubey, 2020; Lwin et al., 2020; Sanders et al., 2021).

Topics in tweets can be acquired simply by clustering. Clustering dividing text documents into groups by similarity allows differentiation of topics within the set of documents (Kolini and Janczewski, 2017; Lu et al., 2013; Zhang et al., 2018). Besides clustering, previous research works had adopted other topic modeling approaches for data extraction, including cosine similarity using term frequency-inverse document frequency (TF-IDF) vectors, latent semantic analysis (LSA), and probabilistic topic modeling (Boyack et al., 2011; Ding and Chen, 2014; Suominen and Toivanen, 2016). In general, topic modeling is different from traditional clustering in that clustering groups texts into a certain number of clusters as output. In contrast, topic modeling builds clusters of words rather than clusters of texts. One text, therefore, can contain a mixture of topics generated by topic modeling. LSA is a simple topic extraction method that illustrates topics by plotting together the similar words, which are usually words that frequently appear together and interpreting dimensions depicted in the plot (Karl et al., 2015). The more sophisticated probabilistic topic modeling, like Latent Dirichlet Allocation (LDA), assumes each document or text is a mixture of a set of topic probabilities. In contrast, each topic is a probability distribution over words (Blei et al., 2003; Blei, 2012). The objective of LDA is set to determine a set of model parameters, topic proportions and topic-word distributions (Bagheri et al., 2014). LDA is a well-developed topic modeling method that is highly generalizable to new documents (Blei et al., 2003). Thus, it is now one of the most popular topic modeling methods that researchers in different fields utilize for textual mining or analyses, e.g., Jacobi et al. (2016) and Tong and Zhang (2016).

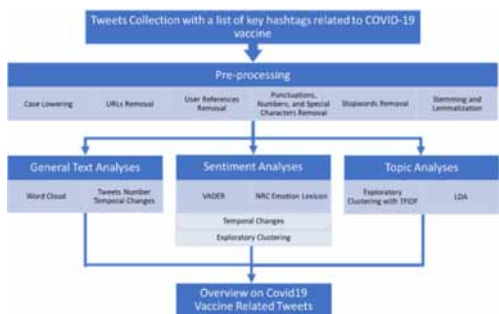
III. METHODS

A. Overview

Tweets were collected by purposive sampling, with only COVID-19 vaccine-related tweets retrieved. Data have first undergone a series of pre-processing before entering analyses. Sentiment analyses and topic analyses were then carried out on the processed data. For sentiment analyses, TextBlob and NRC Emotion Lexicon were utilized to investigate polarity, subjectivity and emotions in the data.

Sentiment and emotion scores were employed in unsupervised exploratory clustering to explore the possible grouping of posts by sentiments and emotions. Topic analyses included another exploratory clustering using frequencies of terms' appearance in the data and Latent Dirichlet Allocation (LDA) topic modeling. Separate topic analyses were performed for data of different polarities defined by sentiment analyses as well. With the results of these analyses, a general understanding of major sentiments and topics towards COVID-19 vaccines was anticipated. The overall workflow was summarized in Figure 1. Python (Version 3.9) was used to collect, process and analyze the data.

Figure 1. The overall workflow for the current study.



B. Data Source

Tweets about the COVID-19 vaccine were searched using a list of COVID-19 vaccine-related key hashtags (see Table 1) and fetched through Twitter API. A total of 1,149,231 tweets containing one or more key hashtags were collected with posting date from Apr 30th, 2020, when the first COVID-19 vaccine development was announced by AstraZeneca (2020) to Feb 28th, 2021. Retweets were filtered out in the search process; thus, all retrieved tweets were original tweets. Tweets containing media such as pictures or videos were included, but only the text components were collected. Each piece of data contained (1) ID of the tweet, (2) full text of the tweet, (3) posting date, (4) user's Twitter ID, and (5) geolocation if available. All the tweets retrieved were written in English. However, since most data did not provide geolocation, locations where most of the tweets were posted, were undetermined. Many were posted in the UK, USA, or India for the limited number of tweets that provided geolocation. For ethical concerns, only tweets that had been set to be open to the public by the users were collected. Full texts of the tweets were analyzed aggregately without study on any individual tweet. Personal information was not contained in any of the data. All the data would be properly discarded following the end of the study.

Table 1. Key hashtags for tweet searching

Key hashtags	
#covidvaccine	(#vaccine #covid19)
#covid19vaccine	(#vaccines #covid19)
#covidvaccines	(#vaccine #covid)
#covid19vaccines	(#vaccines #covid)
#covidvaccination	(#corona #vaccine)
#covid19vaccination	(#coronavirus #vaccine)
#coronavaccine	(#corona #vaccines)
#coronavaccines	(#coronavirus #vaccines)
#coronavaccination	(#vaccination #covid19)
#coronavirusvaccine	(#vaccination #covid)
#coronavirusvaccines	(#corona #vaccination)
#coronavirusvaccination	(#coronavirus #vaccination)
#covidjab	(#jab #covid19)
#covid19jab	(#jab #covid)
#coronajab	(#corona #jab)
#coronavirusjab	(#coronavirus #jab)

Before entering further analyses, as standard procedures for natural language processing, noise within the raw data was cleaned by following pre-processing procedures to allow precise and efficient language analyses.

- 1) *Case Lowering*: All letters were converted to lowercase. Identical words were therefore properly recognized and counted. It reduced the dimensions of the data to provide more accurate and effective analyses.
- 2) *URLs Removal*: User-provided URLs for external websites and system-generated URLs for certain tweets without meaning for analyses were removed.
- 3) *User References and Key Hashtags Removal*: Twitter users can use the symbol “@” followed by a username to tag another user in their tweets. These usernames and the key hashtags used in data collection did not help understand the messages, so they were replaced by empty space. Other hashtags, however, were retained as they often contain the subject of the tweet or useful information related to the topic of the tweet.
- 4) *Punctuations, Numbers, and Special Characters Removal*: Analyses were entirely focused on English words. Thus, characters other than English letters were cleared.
- 5) *Stop words Removal*: Stop words such as pronouns, prepositions and articles frequently appeared in the text but served no function for meaningful evaluations. Stop words from a list provided by Natural Language Toolkit (NLTK) were removed from the current data. NLTK is a library that gives easy access to massive lexical resources and allows users to perform various natural language processing tasks such as categorization and classification (Natural Language Toolkit, 2021; Patel et al., 2021).
- 6) *Lemmatization and Stemming*: Lemmatization is a process of converting the inflectional forms of words to the dictionary form of the words. Stemming also aims at reducing inflectional forms

but using a process of cutting the ends of words. These processes were applied to turn multiple inflectional forms of words with the same or similar meanings into the same word to reduce the dimensionality of the text. Utilizing WordNet Lemmatizer (2021), lemmatization was done by stemming with PorterStemmer based on Porter (1980)'s algorithm. Both the Lemmatizer and PorterStemmer were embedded in NLTK. A version of data without lemmatization and stemming was retained for sentiment analyses for more accurate results.

- 7) *Tokenization*: Every single word in the text was turned into to token. Words with fewer than three letters were discarded as they were mostly abbreviations or meaningless. BoW for each tweet was therefore created.

Empty data after the cleaning processes were discarded and 1,103,584 tweets remained in the data into analyses.

C. Sentiment Analyses

The pre-processed data first underwent sentiment analyses utilizing TextBlob and NRC Emotion Lexicon. Sentiment and emotion scores were generated for each tweet to illustrate the sentiment polarity and emotions expressed in the tweet (Karn et al., 2018). Overall polarity and emotions scores indicated people's general feelings expressed on the COVID-19 vaccine. Results were also obtained and compared monthly to identify any temporal change throughout the year. Unsupervised exploratory clustering was performed using sentiment and emotion scores as attributes to extract possible patterns of tweets with different emotional expressions. Details of the three main components of sentiment analyses were explained as follow:

- 1) *TextBlob*: TextBlob (2020) is a Python library that provides a simple API for natural language processing, including part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, tokenization, etc. TextBlob actively used NLTK to support its tasks. Despite being a small and simple library, TextBlob supports a variety of complex analyses and operations on textual data, making it one of the commonly used natural language processing libraries. In the current study, TextBlob was employed to discover sentiments within the collected tweets, specifically subjectivity and polarity.

By using a lexicon-based approach, sentiment in textual data is defined by its semantic orientation and the intensity of each word in the BoW extracted from data. A pre-trained or pre-defined dictionary is needed to classify positive and negative words and determine the level of positivity and negativity of each word. Entering the sentiment analysis and a bag of words represented each pre-processed tokenized tweet. By referencing the pre-trained dictionary, any word in the tweet with a defined sentiment score was assigned with the respective score. Eventually, the sentiment of the whole tweet was calculated by taking a mean of all the words' sentiment scores. The dictionary used in TextBlob was trained from a massive Sentiment Polarity Dataset supplied by NLTK (NLTK Data, 2005).

Two scores, polarity and subjectivity, were returned by TextBlob sentiment analysis for each tweet. Polarity score ranged from -1 to +1, where -1 defined a negative sentiment and 1 defined a positive sentiment. Negation words like "not" reversed the polarity. Subjectivity scores ranged between 0 and 1. It deduced how much a tweet presented personal opinions and factual information. Tweets that expressed more personal opinions rather than factual information would score high in subjectivity. TextBlob also considered a parameter called "intensity" while estimating subjectivity. Intensity determines if a word modifies the next word, such as adverbs "very", "so", etc. With the polarity scores computed for every tweets, tweets were defined in three groups, "positive" (polarity score > 0), "negative" (polarity score < 0) and "neutral" (polarity score = 0). Separate topic analyses were conducted with these groups of data.

- 2) *NRC Emotion Lexicon*: Stick to the lexicon-based approach, the second part of sentiment analysis employed NRC Emotion Lexicon as the dictionary for sentiment deduction. NRC stands for National Research Council Canada, where this lexicon was developed. The NRC Emotion Lexicon consisted of a large list of English words and their associations with eight basic emotions, including “Fear”, “Anger”, “Anticipation”, “Trust”, “Surprise”, “Sadness”, “Disgust”, and “Joy”, and two sentiments, negative and positive (Mohammad, 2021). The emotion and sentiment annotations were manually done by crowdsourcing through popular online crowdsourcing platform Amazon’s Mechanical Turk (Mohammad and Turney, 2013). Applying the same mechanism as that of TextBlob analysis, NRC Emotion Lexicon was served as the pre-trained dictionary to assign each word in tweet valence scores for emotions and sentiments. Every tweet, therefore, attained average scores for each of the eight emotions and two sentiments.

Python library NRCLex (2019) was utilized in the current study to measure emotional affect in the collected tweets. The library expanded the original NRC Emotion Lexicon dictionary of approximately 14,000 words to 27,000, borrowing the WordNet synonyms database from NLTK. This allowed improved accuracy on emotion and sentiment detection than simply using the original NRC Emotion Lexicon. NRCLex scanned for words contained in NRC Emotion Lexicon and counted the word frequency for each emotion and sentiment for every tweet. A percentage score was calculated for each emotion and sentiment as the frequency of that emotion or sentiment over the total number of words marked with any emotion or sentiment. Thus, scores ranging from 0 to 1 were assigned for every emotion and sentiment for each tweet.

- 3) *Unsupervised Clustering*: The aim of clustering was to deduce any possible groupings or patterns within the data (Manning et al., 2008). Tweets might cluster together by their sentiment and emotion levels, i.e., there might be certain combinations of subjectivity, polarity and emotions commonly seen in tweets, and these tweets might be grouped to form different types of tweets. As there were no pre-determined types of labels for the tweets available in the data, exploratory clustering was the appropriate method to discover the potential categorization of the tweets. This unsupervised machine learning method divides the data into clusters with their similarity based on a set of attributes.

The well-known K-means clustering method was applied to perform exploratory clustering in current data. With a pre-set number of k, every item in the dataset is classified as belonging to one of the k groups with the nearest mean, or shortest distance, to the centroid of the respective group. The centroid describes the cluster and is the central point in the multidimensional space surrounded by all members of the cluster. The distance from every data point to its cluster centroid was defined by the Euclidean distance in the multidimensional space. The Euclidean distance between data point X and Y in m dimension space can be expressed as:

$$d(X, Y) = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \dots + (X_m - Y_m)^2} \quad (1)$$

K-means clustering strives for minimal within-cluster dispersion by iterative re-calculations of cluster centroids and reallocating data points to centroids. The K-means clustering process starts with the initial assignment of k centroids based on the pre-set number of k. Each data instance’s distance to the centroids is then calculated and thus, data instances are assigned to the closest centroid. It is followed by re-estimation of the centroids using means of every data point within the cluster. Data points are assigned to centroid and centroids re-calculation steps repeat until the convergence of minimal distortion within each cluster for all clusters is obtained. This is the state where no further re-assignment of data points is needed. The distortion is defined by the criterion function as follow:

$$E = \sum_{i=1}^x \sum_{x \in C_i} (x - x_i)^2 \quad (2)$$

where the target object is x , x_i indicates the average of cluster C_i , and E is the sum of the squared error of all objects in the database.

K-means was chosen as the clustering methodology for the current study over its clustering method counterpart, hierarchical clustering, mainly due to its efficiency. Compared to a hierarchical clustering algorithm, K-means is less resource expensive and faster, especially for large datasets and a high number of attributes, since it involves only calculations and storage of distance between centroids and data points while the distance between every data point has to be computed and stored for hierarchical clustering (Jain, 2010). The downside of K-means clustering is that a pre-determined k is needed before the process, which could be difficult to decide (Kanungo et al., 2002). Moreover, different assignments of initial cluster centroids can result in different outcome clusters. With the random assignment of initial centroids, resulting clusters could be varied each run of the algorithm, creating confusing outcomes. Efficiency and desired number of iterations would also be affected by initial centroid selection (Abdul Nazeer and Sebastian, 2009).

In this study, ten sentiment and emotion attributes were entered into the clustering process, including subjectivity and polarity scores from TextBlob analysis and scores of eight emotions from NRC Emotion Lexicon. The two sentiment scores from NRC Emotion Lexicon were excluded as they overlapped with the polarity score from TextBlob, which also defined positivity and negativity. This helped to reduce the number of variables and enhanced the efficiency of the analysis. Since all attributes had a similar range, no standardization or normalization was applied before clustering.

In order to determine better k numbers before actual clustering, the elbow method and gap statistics investigation were employed to estimate the performance of different setups. Elbow method simply ran K-means clustering with a couple of k number settings and, for each setting, computed the average within-cluster distortion, or sum of squares, across all clusters. By plotting the average distortion scores with k numbers and locating the k corresponding to a “knee” in the plot, the k that the drop of average distortion starts diminishing, the best k is suggested (Syakur et al., 2018). Gap statistics is another method to estimate the best number of k in K-means clustering invented by researchers at Stanford University (Tibshirani et al., 2001). A similar approach is applied for gap statistics, with the statistics plotted across different k numbers, the optimal k is which after it, the gap statistics increase rate slows down (Tibshirani et al., 2001).

The K-means clustering tool, KMeans, from a well-developed machine learning library in Python, Scikit-Learn (2021), was utilized for the actual clustering procedure. As mentioned before, different initial centroids would possibly result in different outcome clusters. This problem can be minimized by comparing multiple runs with different initial centroids to find the best results. Although it would be time-consuming, K-means clustering results should eventually converge with enough iterations and trials (Bottou and Bengio, 1995). KMeans from Scikit-Learn offered “ n_init ” function allowing n runs of the algorithm with different initial centroid assignments to be compared to obtain the final results. KMeans also applied the “ k -means++” initialization scheme, which initializes the centroids to be distant from each other (Vassilvitskii and Arthur, 2006), to pursue better results than normal random centroid initialization. The functional command for KMeans and parameters set for the current study were illustrated below:

`Kmeans(n_clusters=k, init='k-means++', n_init=10, max_iter=300, random_state=10)`

where “ $n_clusters$ ” was k number,

“ $init$ ” was the method for centroid initialization,

“ n_init ” was the number of time algorithm will be run with different centroid seeds,

“ max_iter ” was the maximum number of iterations for a single run, and

“*random_state*” was parameter controls random number generation for centroid initialization, allowing replicable result generation if necessary.

Clustering outcomes were evaluated by internal validation, visual inspection and hands-on evaluation over statistics among attributes for each cluster. External validation, which required “true label” for the grouping, was unachievable for current data. For internal validation, two clustering validity statistics, Calinski-Harabasz Index and Silhouette coefficient, were adopted. Invented by Calinski and Harabasz in 1974 (Calinski and Harabasz, 1974), Calinski-Harabasz Index, which is also known as the Variance ratio criterion, is a measure of similarity of data points to their own cluster (cohesion within clusters) compared to the separation of centroid with other cluster centroids (dispersion between clusters). The higher the score of the Calinski-Harabasz Index, the better the performance of clusters, which means clusters are dense internally and well separated from each other. Silhouette coefficient took a similar approach but a different calculation method. It measures intra-cluster cohesion by mean distance between a data point and all other points in the same cluster and inter-cluster separation by mean distance between a data point and all other points in the next nearest cluster (Rousseeuw, 1987). Same as Calinski-Harabasz Index, a higher Silhouette coefficient indicates better cluster distribution. The silhouette coefficient for each single data point also demonstrates whether the data point is correctly placed in a cluster (Starczewski and Krzyżak, 2015). While a negative score suggests incorrect placement and a positive score means good assignment, a score tends to zero indicates overlapping between clusters.

Apart from validity statistics, clusters were investigated by visual inspection and reviewing attributes’ scores within clusters. The inter-cluster distance was depicted by the inter-cluster distance map reducing dimensions into two principal components. Boxplots for each emotion attribute were generated and reviewed to determine whether different clusters actually behaved differently.

D. Topic Analyses

Topic analyses are aimed at uncovering hidden textual structures and identifying common topics mentioned in collected tweets. Results were obtained by using two unsupervised machine learning methods, exploratory clustering and Latent Dirichlet Allocation (LDA) topic modeling. The two analyses were applied on overall data and separately on data divided by polarity from sentiment analyses, i.e., to reveal common topics for positive, negative, and neutral tweets. For each method, models with different types of topics were generated and compared by validity statistics and manual evaluations on topic themes.

Frequencies for every single term that appeared in the tweets were vectorized and used as attributes in both analyses. In particular, Term Frequency - Inverse Document Frequency (TF-IDF) was calculated and used to encode the data to obtain vectors for the analyses. TF-IDF for each word was produced by multiplying the term frequency of the word in a document with the inverse document frequency of the word across all documents, or how common the word is in the entire set of documents (Salton, 1991). Equation (1) was the mathematical expression for TF-IDF score is as follows: w is a word in document d , and D refers to the whole set of documents.

$$tf\ idf(w, d, D) = tf(w, d) \times idf(w, D) \quad (3)$$

Where:

$$\begin{aligned} tf(w, d) &= \log(1 + \text{freq}(w, d)) \\ idf(w, D) &= \log\left(\frac{|D|}{|\{d \in D : w \in d\}|}\right) \end{aligned}$$

TF-IDF aimed to highlight more interesting or important words to a document, which frequently appeared in a document but were not very commonly found across documents (Aizawa, 2003). TF-IDF value for each word increases by the appearance in a document but is gradually down-scaled by the appearance in other documents in the set. The TF-IDF vectorizer from Scikit-Learn machine learning toolbox was employed to learn and compute the TF-IDF value for each term and encode all tweets into vectors with normalization suitable for clustering and LDA.

1) *Exploratory Clustering*: This analysis was set to discover the potential grouping of the tweets by the number of the appearance of terms or words. Similar to analysis on sentiment, being an exploratory analysis, clustering was applied to divide the tweets into different groups. The overall clustering approach was akin to sentiment analysis, except TF-IDF vectors for each tweet were used as attributes. K-means was employed again as the clustering method, executed by the same KMeans clustering tool from Scikit-Learn machine learning library in Python. Distance measurement between the data point and cluster centroids were computed by Euclidean distance, although the computational time was multiplied with much more variables. The Elbow method was also appointed to estimate a better number of k prior to the clustering. Average Silhouette coefficient and Silhouette coefficient plot were examined to evaluate the internal validity of the resulting model.

Aiming to reveal common topics within the collected tweets, the most frequently used words in each cluster were listed and examined for potential summarization of topics.

2) *LDA*: The second attempt of topic analysis implemented a topic modeling method, Latent Dirichlet Allocation (LDA). To achieve a model of topics, LDA first assumes a number of the topic (k) within the whole sample of documents and distributes the k topic across the documents by giving a topic to each word. Then, for each word (w) in a document (d), assume it has been assigned a wrong topic and probabilistically re-assign a topic for it, based on the current topics in d and the number of times the topic was assigned to w across the whole set of documents. The algorithm repeats this process for each document to adjust the topic assigned to every word until a stable model is generated (Blei et al., 2003).

In the current study, LDA topic modeling was implemented using the LDA modeling tool, LdaMulticore, in an open-source Python library, Gensim. Gensim (2021) is designed to process raw, unstructured digital texts using unsupervised machine learning algorithms. LdaMulticore is a version of the LDA modeling tool with parallel multiprocessing ability, allowing multiple CPU cores to work together to speed up the process. Before entering the modeling process, a corpus such as TF-IDF vectors and a dictionary with all the words in the dataset was generated. The functional command for LdaMulticore and parameters set for the current study were illustrated below:

```
ldamulticore(corpus=corpus_tfidf, num_topics=n, id2word=dictionary, passes=10, workers=3)
```

where “corpus” was corpus,

“num_topics” was the number of topics in the model generated,

“id2word” was dictionary mapping from word IDs to words,

“passes” was the number of passes through the corpus during training, and

“workers” was the number of workers processes to be used for parallelized processing.

Besides manual assessment on keywords for each result topic, topic models generated were also evaluated through the measure of topic coherence. Coherence within a topic refers to the level at which each statement of fact in the topic supports one another. There are different measures of topic coherence, but they all aim to score a topic by measuring the degree of semantic similarity between high-scoring words in the topic. These measurements also help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference. UMass

coherence developed by Mimno et al. (2011) was hired as a coherence measure in the current study as it corresponds well to human coherence judgments and can recognize particular semantic issues in a topic model without human evaluation or an external reference corpus. It is based on document co-occurrence counts, a one-preceding segmentation and a logarithmic conditional probability.

$$C_{UMass} = \frac{2}{N \cdot (N-1)} \sum_{i=2}^N \sum_{j=1}^{i-1} \log \frac{P(w_i, w_j) + \epsilon}{P(w_j)} \quad (4)$$

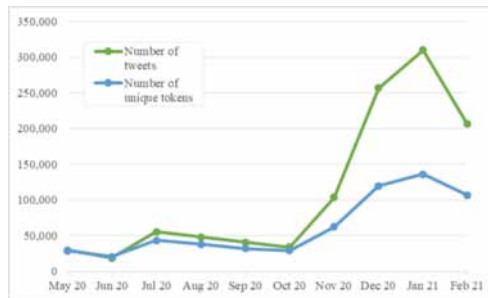
Since the UMass coherence score is calculated over a log of probabilities, it is negative in nature. Scores close to zero or less negative represent good coherence. The coherence scores in this study were computed by topic coherence pipeline, CoherenceModel, available in Gensim.

IV. RESULTS

A. Descriptive Results

There were 315,041 unique tokens or words in total found across all 1,103,584 processed tweets. A monthly number of tweets related to the COVID-19 vaccine and unique tokens remained similar from April to October 2020 (Figure 2). There was a drastic increase in both numbers recorded since November 2020 and reached the highest in January 2021, of over 300,000 tweets and about 140,000 unique tokens. This indicated a huge increase in people's concern over COVID-19 vaccines from November 2020.

Figure 2. The number of tweets and unique tokens across months (Tweets posted on Apr 30th, 2020 were included in May 2020).

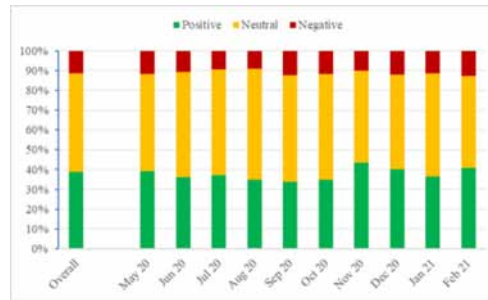


The top-50 words were shown on a word cloud (Figure 3). “Peopl”, which was the stemmed form of “people”, was the most found word in the data, with over 73,000 counts. “Pfizer” was the most mentioned brand making COVID-19 vaccine, having more than twice the counts of the second brand, “Moderna”. There were not any obvious sentiment-related words seen in the top-10 words, and within a top-50 word, only “thank” and “good” were found positive in tone.

B. Sentiment Analyses

The mean subjectivity and polarity scores by TextBlob analysis for all data were 0.275 (SD = 0.303) and 0.086 (SD = 0.225) respectively. Low subjectivity

Figure 5. Tweets ratio with different polarity across months



Analyses with NRC Emotion Lexicon yielded similar results to TextBlob analyses that positive tweets were dominant over negative ones. The mean score for positive sentiment across entire dataset was 0.378 (SD = 0.350), which was much higher than that for negative sentiment (mean = 0.082, SD = 0.146). The temporal change was unseen as the monthly means had patterns very much alike (Figure 6).

Figure 6. Monthly means of positive and negative sentiment scores from NRC Emotion Lexicon.



Regarding the eight emotions considered in NRC Emotion Lexicon, “Trust” was the most prominent emotion with a mean score of 0.089 for overall data. The second highest score was 0.051 for “Fear”, followed by 0.035 for both “Joy” and “Sadness” (Figure 7). “Anticipation” was not detected in the entire dataset resulting in a zero score.¹ Change over time was minimal, although a gradual rise of “Trust” and “Anger” along with dropping of “Fear” and “Surprise” from May 2020 to recent could still be observed (Figure 8). About 20% of the total tweets scored 0 in every NRC sentiment and emotion, which meant one in five collected tweets was neutral or emotionless recognized by NRC Emotion Lexicon. Above all, the NRC Emotion Lexicon analyses went along with the TextBlob analyses results that positive expressions were relatively prevalent compared to negative ones in the current sample.

Figure 7. Mean scores for the eight emotions from NRC Emotion Lexicon.

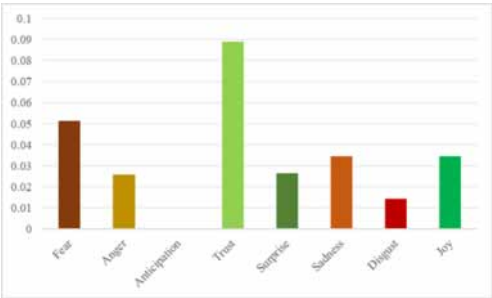
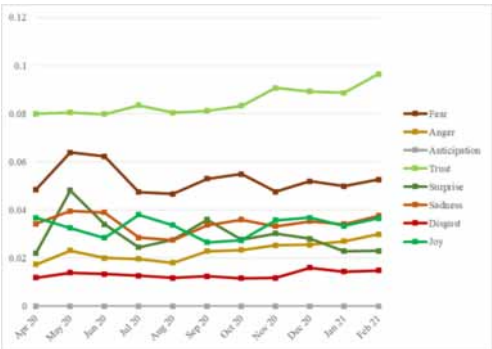


Figure 8. Monthly Mean scores for the eight emotions from NRC Emotion Lexicon.



Before undergoing exploratory clustering by sentiment and emotion scores, the elbow method and Gap Statistics were utilized to determine a better number of clusters. However, both methods showed unclear clues for best cluster number (see WCSS chart in Figure 9 and Gap Statistic chart in Figure 10). Thus, k-means clustering with k ranging from 2 to 6 was performed. Results were then manually checked and compared between models to decide which model fits and explains the data well.

5-Cluster model was selected as the best model to represent the data. It was the model with the highest number of dimensions while having none of the clusters overlapped with others in the principal component chart, indicating a decent separation between clusters (Figure 11). Calinski-Harabasz Index (50,127) and average Silhouette (0.387) coefficient of this model were acceptably high and comparable to that of the fewer cluster number models, implying satisfactory inter-cluster distance and intra-cluster cohesion (Figure 12). Silhouette coefficient analysis revealed that adequate cluster distribution with only a small number of samples is mis-clustered (Figure 13). Each cluster's characteristics could be observed in cluster-wised statistics on each of the ten attributes (Figure 14), indicating the feasibility of explaining the data by these clusters. The biggest cluster, Cluster 0, consisted of neutral or emotionless tweets, with low means scores in all attributes. In contrast, Cluster 3 and Cluster 4 contained polarised tweets. Cluster 3 stored mostly negative tweets with high subjectivity scores, negative polarity scores and relatively high scores in negative emotions like "Fear", "Anger", "Sadness" and "Disgust". Cluster 4 showed more positive sentiment, scoring high in subjectivity, positive polarity, and positive emotions including "Joy", "Trust" and "Surprise". While mixed feelings are seen in Cluster 1 tweets, data in Cluster 2 interestingly displayed minimal subjectivity, polarity and other emotions, but the highest scores of "Trust".

Figure 9. WCSS vs. a number of clusters (k)

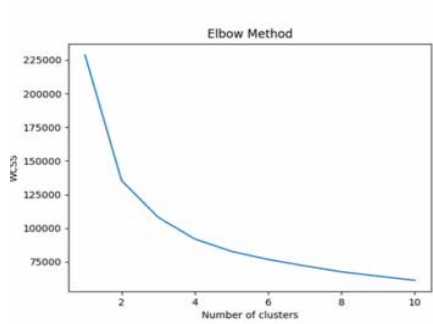


Figure 11. Intercluster distance map for 5-cluster model

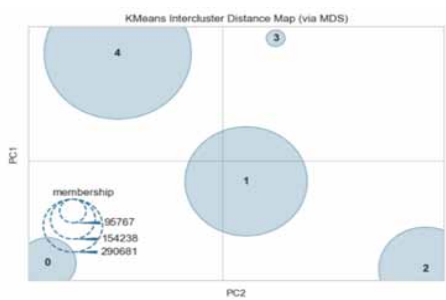


Figure 12. Average Silhouette coefficients for models with a different number of clusters (k)

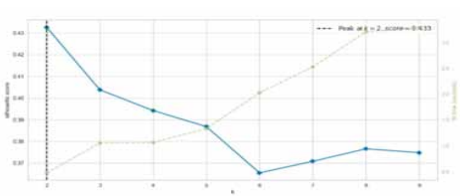
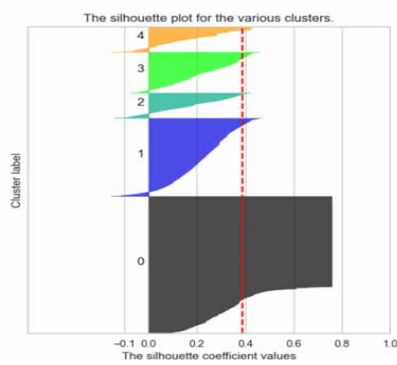


Figure 13. Silhouette plot showing Silhouette scores for each data point.



D. Topic Analyses

For the exploratory clustering using TF-IDF vectors as attributes, the elbow method suggested 5 as the number of k for K-means clustering (Figure 15). However, for further evaluation of the final result by comparison with different setups, clustering with k ranging from 2 to 7 was performed.

Figure 14. Boxplots showing statistics on every attribute for each cluster.

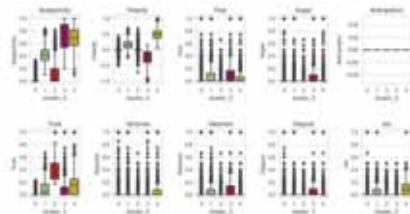


Figure 15. Boxplots showing statistics on every attribute for each cluster.

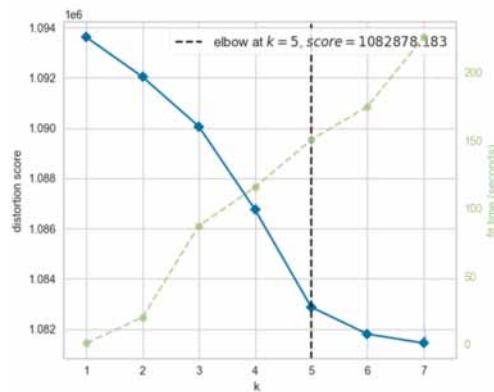


Figure 16. Intercluster distance map for 5-cluster model of TF-IDF clustering.

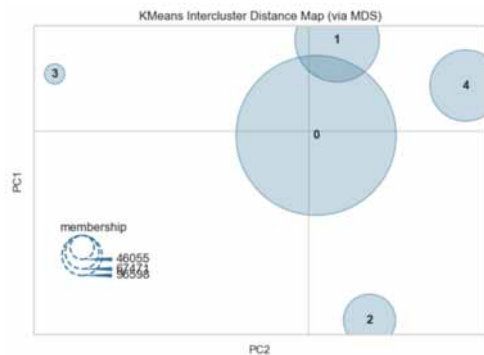
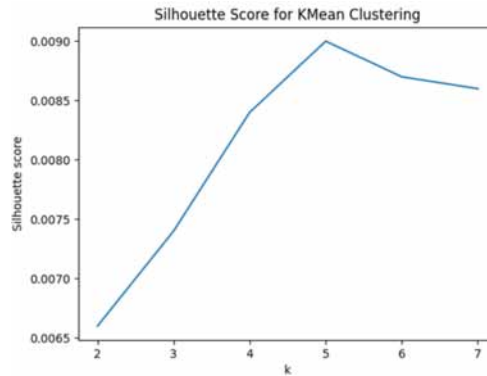


Figure 17. Average Silhouette scores for models of TF-IDF clustering with a different number of clusters (k).



The top-10 frequently appeared words and a total number of tweets for each cluster in the final 5-cluster model were listed in Table 2. Visual inspection with PCA plot revealed potential overlapping between the two biggest clusters, cluster 0 and cluster 1 (Figure 16). Otherwise, the overall distance between clusters was acceptable. Nevertheless, the low average Silhouette coefficient and rather unbalanced distributions indicated the unstable nature of this model, although it had comparatively better performance than models with other numbers of k (Figure 17).

Interpretation of topics or characteristics based on top-10 words of the clusters was performed manually. Interpretation for each cluster was as follow:

- 1) *Cluster 0*: Mainly consisted of updates of everyday news and vaccine development. Particularly, “thank” was one of the top-10 words showing a proportion of tweets on COVID-19 vaccine expressed gratitude.
- 2) *Cluster 1*: Healthcare related, medical updates and issues under the pandemic
- 3) *Cluster 2*: Vaccine trials, development progression and data for different brands of COVID-19 vaccine.
- 4) *Cluster 3*: Updates and information about vaccines developed in India.
- 5) *Cluster 4*: Updates in Russia mainly, and possibly in another part of the world.

Table 2. Top-10 frequently appeared words (lemmatized and stemmed) and a number of tweets in each cluster of the 5-cluster model

Cluster	Top-10 words	Number of tweets
0	update news work health today trial dose make thank year	884,575
1	people health plead nh today medicalupd doctor insure policies medicine pharmaceutical	66,203
2	Pfizer BioNTech Moderna effect approved dose data say end FDA	46,932
3	India update covax in India fight covishield drive news serum trial institute	32,375
4	Russia Russian patent Putin world country sputnik suspend giant pass	18,490

Separate clustering for the three polarity groups of data yielded different sets of keywords as expected (Table 3). These models were not stable reading the low average Silhouette coefficients. More perplexing combinations of top words were found compared to the model for the whole dataset. It was hard to conclude for each cluster, but for negative tweets, common topics seemed to include late vaccine rolling out, complaints to pharmaceutical business exploitations, or difficulties for getting vaccinated, etc. For positive tweets, people expressed belief in the safety and effectiveness of the vaccines and hope for an end of the pandemic. Messages were even more diverse for neutral tweets, but they seemed about news updates around different places in the world.

Table 3. Top-10 frequently appeared words and silhouette coefficient in each cluster of the three cluster models by polarity.

Polarity	Cluster	Top-10 words
Positive (Average Silhouette coefficient = 0.007)	0	new Pfizer today health pandemic safe India dose effective available
	1	patents giants suspend passes driving profits vote governments restricting control
	2	people vaccinated million need getting dose health new received safe
	3	year-end everywhere pool the fastest property intellectual planet knowledge end benefit
	4	news good great Pfizer India latest oxford positive Moderna breaking
Negative (Average Silhouette coefficient = 0.003)	0	India johnson dry run health Vardhan harsh retweet today biotech
	1	Pfizer Moderna BioNTech doses dose late AstraZeneca mRNA cold effect
	2	pandemic long health like trump getting news virus needs time
	3	people vaccinated need vulnerable want like million know black sick
	4	pharmaceutical business pharma business model clash giant move industry strategy make
Neutral (Average Silhouette coefficient = 0.037)	0	vaccine AstraZeneca Oxford vacuna unite fight France pour China vaccines
	1	pandemic updates people health Russia lockdown covaxin Moderna today update
	2	news India breaking update breaking news pandemic daily news viral modi health
	3	India updates India fights pandemic covaxin covishield drive pmmodi update Russia
	4	Pfizer BioNTech Moderna Pfizer BioNTech FDA use emergency AstraZeneca approve dose Canada

LDA topic modeling trials were performed with a number of topics ranging from 1 to 10 and other parameters remained constant. Models generated from these trials were compared with each other by coherence scores and heuristic judgment on the perplexity of topics. The peak of coherence scores was found at the 4-topic model (Figure 18), with the value of -3.314 indicating a satisfactory topic coherence within the model. Manual comparison of keywords and meanings from different models suggested that starting from 5-topics, overlapping between topics gradually increases. Thus, the 4-topic model has claimed the best model for the current data. Keywords and their weights in each topic for the 4-topic model were presented in Table 4. The four topics were interpreted as follow:

- 1) *Topic 1:* Personal sharing about receiving vaccination, such as availability or receiving first dose of COVID-19 vaccines.
- 2) *Topic 2:* Information about vaccine development progress of different brands around the world.
- 3) *Topic 3:* Political talks about policies for pandemic and public figures related, mainly from the United States.

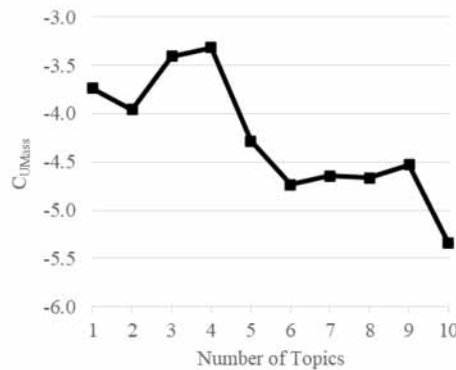
- 4) *Topic 4:* Keywords consisted of a couple of verbs that were hard to interpret. It was possibly related to general circumstances and what people did or needed to do under the pandemic.

Table 4. Keywords and weighs and coherence score (CUMass) in each topic of the topic models

Tweets	Topic	Keywords and weighs
All (CUMass = -3.314)	1	0.007* ^{***} first ^{***} + 0.007* ^{***} get ^{***} + 0.006* ^{***} receiv ^{***} + 0.006* ^{***} health ^{***} + 0.006* ^{***} do ^{***} + 0.005* ^{***} vaccin ^{***} + 0.005* ^{***} dose ^{***} + 0.005* ^{***} peopl ^{***} + 0.005* ^{***} avail ^{***} + 0.005* ^{***} state ^{***}
	2	0.010* ^{***} india ^{***} + 0.010* ^{***} trial ^{***} + 0.008* ^{***} pfizer ^{***} + 0.007* ^{***} moderna ^{***} + 0.007* ^{***} develop ^{***} + 0.007* ^{***} russia ^{***} + 0.006* ^{***} patent ^{***} + 0.006* ^{***} china ^{***} + 0.006* ^{***} world ^{***} + 0.006* ^{***} approv ^{***}
	3	0.016* ^{***} trump ^{***} + 0.011* ^{***} lockdown ^{***} + 0.010* ^{***} biden ^{***} + 0.008* ^{***} pandem ^{***} + 0.007* ^{***} billgat ^{***} + 0.006* ^{***} usa ^{***} + 0.006* ^{***} sar ^{***} + 0.005* ^{***} cov ^{***} + 0.005* ^{***} miami ^{***} + 0.005* ^{***} iot ^{***}
	4	0.006* ^{***} get ^{***} + 0.005* ^{***} peopl ^{***} + 0.005* ^{***} take ^{***} + 0.005* ^{***} pandem ^{***} + 0.005* ^{***} need ^{***} + 0.004* ^{***} work ^{***} + 0.004* ^{***} viru ^{***} + 0.004* ^{***} know ^{***} + 0.004* ^{***} make ^{***} + 0.004* ^{***} let ^{***}
Positive (CUMass = -2.476)	1	0.007* ^{***} pandemic ^{***} + 0.006* ^{***} trump ^{***} + 0.005* ^{***} safe ^{***} + 0.005* ^{***} end ^{***} + 0.005* ^{***} way ^{***} + 0.005* ^{***} get ^{***} + 0.005* ^{***} everyone ^{***} + 0.005* ^{***} tell ^{***} + 0.005* ^{***} make ^{***} + 0.004* ^{***} good ^{***}
	2	0.006* ^{***} pfizer ^{***} + 0.006* ^{***} effective ^{***} + 0.006* ^{***} new ^{***} + 0.005* ^{***} first ^{***} + 0.005* ^{***} moderna ^{***} + 0.005* ^{***} trials ^{***} + 0.005* ^{***} russia ^{***} + 0.005* ^{***} trial ^{***} + 0.005* ^{***} news ^{***} + 0.004* ^{***} china ^{***}
	3	0.011* ^{***} first ^{***} + 0.006* ^{***} today ^{***} + 0.006* ^{***} get ^{***} + 0.005* ^{***} workers ^{***} + 0.005* ^{***} dose ^{***} + 0.004* ^{***} year ^{***} + 0.004* ^{***} great ^{***} + 0.004* ^{***} vaccinated ^{***} + 0.004* ^{***} thank ^{***} + 0.004* ^{***} news ^{***}
	4	0.011* ^{***} patents ^{***} + 0.006* ^{***} plan ^{***} + 0.006* ^{***} countries ^{***} + 0.006* ^{***} join ^{***} + 0.006* ^{***} suspend ^{***} + 0.006* ^{***} governments ^{***} + 0.005* ^{***} profits ^{***} + 0.005* ^{***} supply ^{***} + 0.005* ^{***} vote ^{***} + 0.005* ^{***} rich ^{***}
Negative (CUMass = -3.625)	1	0.009* ^{***} trump ^{***} + 0.005* ^{***} fake ^{***} + 0.004* ^{***} biden ^{***} + 0.004* ^{***} pandemic ^{***} + 0.004* ^{***} lockdown ^{***} + 0.004* ^{***} billgates ^{***} + 0.003* ^{***} false ^{***} + 0.003* ^{***} people ^{***} + 0.003* ^{***} news ^{***} + 0.003* ^{***} propaganda ^{***}
	2	0.007* ^{***} get ^{***} + 0.006* ^{***} people ^{***} + 0.004* ^{***} like ^{***} + 0.004* ^{***} long ^{***} + 0.004* ^{***} getting ^{***} + 0.004* ^{***} one ^{***} + 0.004* ^{***} virus ^{***} + 0.004* ^{***} take ^{***} + 0.003* ^{***} know ^{***} + 0.003* ^{***} pandemic ^{***}
	3	0.006* ^{***} trials ^{***} + 0.006* ^{***} trial ^{***} + 0.006* ^{***} pfizer ^{***} + 0.005* ^{***} astrazeneca ^{***} + 0.005* ^{***} johnson ^{***} + 0.005* ^{***} china ^{***} + 0.004* ^{***} moderna ^{***} + 0.004* ^{***} countries ^{***} + 0.004* ^{***} late ^{***} + 0.004* ^{***} oxford ^{***}
	4	0.007* ^{***} health ^{***} + 0.005* ^{***} vulnerable ^{***} + 0.004* ^{***} people ^{***} + 0.004* ^{***} doses ^{***} + 0.004* ^{***} run ^{***} + 0.004* ^{***} workers ^{***} + 0.003* ^{***} dry ^{***} + 0.003* ^{***} care ^{***} + 0.003* ^{***} today ^{***} + 0.003* ^{***} india ^{***}
Neutral (CUMass = -3.889)	1	0.015* ^{***} lockdown ^{***} + 0.014* ^{***} russia ^{***} + 0.008* ^{***} billgates ^{***} + 0.008* ^{***} trump ^{***} + 0.006* ^{***} russian ^{***} + 0.005* ^{***} usa ^{***} + 0.005* ^{***} pandemic ^{***} + 0.005* ^{***} america ^{***} + 0.004* ^{***} israel ^{***} + 0.004* ^{***} gates ^{***}
	2	0.006* ^{***} say ^{***} + 0.005* ^{***} effect ^{***} + 0.005* ^{***} people ^{***} + 0.005* ^{***} make ^{***} + 0.005* ^{***} world ^{***} + 0.004* ^{***} country ^{***} + 0.004* ^{***} trial ^{***} + 0.004* ^{***} virus ^{***} + 0.004* ^{***} develop ^{***} + 0.004* ^{***} pfizer ^{***}
	3	0.006* ^{***} health ^{***} + 0.006* ^{***} get ^{***} + 0.005* ^{***} people ^{***} + 0.004* ^{***} pandemic ^{***} + 0.004* ^{***} via ^{***} + 0.004* ^{***} work ^{***} + 0.003* ^{***} need ^{***} + 0.003* ^{***} take ^{***} + 0.003* ^{***} vaccinated ^{***} + 0.003* ^{***} science ^{***}
	4	0.023* ^{***} india ^{***} + 0.017* ^{***} pfizer ^{***} + 0.011* ^{***} astrazeneca ^{***} + 0.010* ^{***} updates ^{***} + 0.010* ^{***} moderna ^{***} + 0.009* ^{***} oxford ^{***} + 0.009* ^{***} pandemic ^{***} + 0.009* ^{***} china ^{***} + 0.008* ^{***} update ^{***} + 0.008* ^{***} news ^{***}

LDA topic models for separate polarities were also generated. Keywords weigh and coherence scores were also listed in Table 4. For positive tweets, topics expressed hopefulness on the end of the pandemic, gratefulness on the introduction of effective vaccines, gratitude for getting vaccinated, and agreeing on some policies. For negative tweets, topics focused on politics in the USA, long and unending pandemic, disbelief on vaccine trials, effects or safety, and issues about care for vulnerable people and health workers. Topics for neutral tweets also considered policies, vaccine developments and people's circumstances under the pandemic.

Figure 18. Coherence scores (C_{UMass}) for models with a different number of topics.



V. DISCUSSION

The main objective of this study was to uncover public perceptions and concerns over the COVID-19 vaccine by machine learning and evaluate the effectiveness of the applied language analysis method. Tweets number on its base reflected how prevalent people are concerned and talk about an issue. The number of tweets collected in this study showed that the COVID-19 vaccine had not been a popular topic in daily social media sharing until late 2020. Specifically, in November 2020, the news was out about rolling out COVID-19 vaccines in December 2020 both in the United Kingdom and the United States, which could be why the rapid rising of tweets calling COVID-19 vaccines. With the first jab approved and started using in December 2020 in UK and USA, COVID-19 vaccination-related sharing continued to rise. With over 300,000 tweets gathered in January, which meant around 10,000 tweets per day, it reflected the huge popularity of the topic early this year.

Sentiment analyses were performed to capture sentiment and emotional expressions in the sample. COVID-19 vaccine-related tweets in the current sample showed low emotional engagement overall. A relatively large proportion of tweets were marked neutral. Previous research using TextBlob on general COVID-19 sentiments showed a similar pattern (Manguri et al., 2020). Upon manual checking on the part of the data, it was believed that these neutral or emotionless tweets were predominantly facts or news sharing without adding personal thoughts or feelings since news sharing had become one of the biggest social media usage (Kümpel et al., 2015). Regardless of the unpolarized tweets, positive tweets were found much more prevalent than negative ones in the current sample, showing that people viewed the COVID-19 vaccine positively in general. This finding went along with the outcome of a previous study on sentiment towards COVID-19 in the USA in early 2020 which concluded positive sentiment outweighed negative sentiment (Hung et al., 2020). Change across months was found minimal, except for a small spike of positive tweets in November 2020. This

was aligned with previous research results showing excitement on the announcement and start of COVID-19 vaccination (Hussain et al., 2021).

Regarding specific emotions, current results showed “Trust” was the most expressed emotion among the eight considered. The score for “Fear” was also high, although not as near “Trust”. These results were different from a previous study also utilizing NRC Emotion Lexicon on a Twitter discussion about the COVID-19 pandemic, which presented higher “Fear” than “Trust” (Xue et al., 2020). Another sentiment analysis study on Indians’ views over two brands of COVID-19 vaccines demonstrated similar results that people tended to trust than fear about the vaccines (Dubey, 2021). There seemed to be a considerable proportion of people who were hopeful for COVID-19 vaccines to be trustworthy and competent to defeat the virus. At the same time, certain views might have asserted fear over the lack of effectiveness of the vaccines and the inability to stop the pandemic (Soares et al., 2021). For temporal changes, the peaking of “Fear” and “Surprise” in May 2020 indicated diverse views over the first announcement of vaccine development. People were either excited about the coming of vaccines or skeptical about it. With more news and knowledge about the vaccines exposed throughout the year, both two emotions gradually diminished across months. As increasing evidence arises to support the vaccines, “Trust” level kept enhancing. At the same time, “Anger” was noted to slightly increase over time, probably due to people’s urge to get the vaccine to end the pandemic while politics or other issues hindered the implementation of vaccination schemes. Also, the effect of vaccination became questionable as the COVID case number remained high despite many people having applied the vaccine. These brought up the “Anger” level and reached the peak in February 2021.

Both two sentiment analysis methods employed in this study contribute to valuable results. TextBlob provided convenient and user-friendly implementation of sentiment analyses supplying accurate and meaningful results. Lightweight and short processing time also support its suitability for analyzing massive textual data like tweet analyses in the current study. Polarity and subjectivity were simple figures yet offered precious insight on the level of sentiment in the tweets. NRC Emotion Lexicon is one of the few lexicons involved in assessing various emotions, and it provides good emotion recognizing ability (Tabak and Evrim, 2016). In this study, we produced useful emotion scores that enhanced the dimensionality of the results, which brought in additional insights. With a workable library for easy implementation and continuous development, the lexicon could become one of the best sentiment and emotion lexicons for users.

Unsupervised exploratory clustering successfully found out possible patterns of tweets by sentiment and emotions. This machine learning method is suitable for data mining on unlabeled data like the current sample, discovering potential groupings within the data. After executing and comparing several trials, a 5-cluster model was decided to explain the data. The 5 clusters represented five types of tweets: (1)*Neutral or emotionless*, (2)*Positive*, (3)*Negative*, (4)*Mixed feelings*, and (5)*Merely trust*, respectively. These groups were found stable and valid with decent distributions and clustering validity statistics. As mentioned above, neutral or emotionless tweets composed the largest group consisting mostly of news and facts. For the interesting “*merely trust*” group, with some manual reviewing of the data, tweets were not mainly composed by news sharing contrasting to the “*neutral*” group, but several advertisement-like posts were found. These tweets did not use strongly or noticeably emotional wordings but were more affirmative, possibly aiming to support or promote COVID-19 vaccination. It could be possible that the increase in “Trust” level over time was related to the growing number of tweets. In general, exploratory clustering enabled differentiation of various types of tweets in the current sample based on sentiment and emotions, further facilitating understanding of the data.

Topic analyses were done by exploratory clustering and LDA topic modeling. A 5-cluster model was elected to represent the current data despite unbalanced distributions and unsatisfactory stability of the clusters. Clustering enabled the grouping of tweets using term frequencies as attributes differentiating types of tweets based on their commonly used words. Frequently seen words revealed different types of tweets covering news updates, healthcare issues, vaccine development progresses.

Particularly, there were two groups of tweets focused on updates in India and Russia, perhaps mainly shared by the Indian and Russian populations, respectively.

LDA topic modeling appointed a more sophisticated approach to defined topics on words level, assigning a topic for each word. A 4-topic model was chosen to explain the data. The topics included personal sharing of vaccination availability and receiving periods, politics, and general circumstances under pandemics. One topic aligned with the clustering result was vaccine development progress, suggesting it as the people's major concern regarding COVID-19 vaccine issues.

To summarize, the most common topic people shared in their tweets was vaccine development progress. Other prevalent topics included vaccine effectiveness, safety and availability, personal sharing of vaccination experience, updated situations and issues of the pandemic, and politics.

Results of topic analyses on polarity separated tweets denoted both pro and against views persisted over topics like vaccine effectiveness and safety, time to end the pandemic, policies or politics, and profit and business of pharmaceuticals. Since positive tweets number was much higher than negative ones, people were generally happy on such issues. Nonetheless, skepticism, such as disbelief in vaccines efficacy, perception of vaccination as political propaganda or exploitation from pharmaceutical corporates, etc., were crucial for public health practitioners and policymakers to designate suitable measures to alleviate any adverse events concerns and rebuild confidence.

By considering the two topic analytical methods, LDA topic modeling was recommended as better than exploratory clustering (Kelaiaia and Merouani, 2016). While observing the results in this study, exploratory clustering attempted to separate tweets having distinctive wordings but did not necessarily differentiate topics. Although there were clear differences in top words found in clusters, the groupings differentiated data in other dimensions like geographic (countries) instead of defining topics. There was also unbalanced distribution resulting in a large cluster harder to interpret. Sub-clustering on that cluster would potentially decipher more meaningful results (Sanders et al., 2021). Altogether, clustering was claimed not a decent method for topic analyses. LDA, on the other hand, adequately extracted topics on word level. By assuming multiple topics for every tweet, there were no grouping issues. The only con for LDA was since no grouping and labeling of tweets, manual checking on raw data for in-depth reviewing of the results was inhibited. Some researchers also suggested combining LDA and clustering for text mining (Alhawarat and Hegazi, 2018; Bui et al., 2017).

VI. CONCLUSION

Understanding the general population's perception of health issues like pandemics and vaccination help crucially promote public health. Social media platforms offer convenient resources for semantic data analyses to enable recognition of people's views by their daily expression and sharing. This study demonstrated using machine learning natural language analytical methods, particularly sentiment and topic analyses, to analyze extensive tweets data related to COVID-19 vaccination. Findings revealed that attitude towards COVID-19 vaccines was positive in general. Common topics identified by topic analyses facilitate understanding of specific concerns on COVID-19 vaccine issues. Overall, the current study suggested that analyses of social media data employing machine learning allow effective gathering of knowledge and insights on population perceptions. Another contribution of this research is that it provides decision-makers with a streamlined solution (Figure 1) to identify the general population's perception of health issues effectively.

Several limitations should be addressed in this study. First and foremost, although multiple different wording combinations and various expressions of COVID-19 and vaccination were used, the key hashtags for tweets searching might not cover all related tweets. Hashtags keep evolving and being newly invented. Some users might also use related hashtags, such as some researchers suggested combining LDA and clustering for text mining brands' names instead of actually mentioning COVID-19 or vaccination. Secondly, due to limitations in the Twitter content redistribution policy, besides limited geolocation gathered, most demographic information of the users like age, gender,

social statuses, etc., were not provided. Therefore, analyses on demographics like geographical difference, age difference and gender difference were restricted. Besides, due to limited user information provided, fake or robot users could not be recognized. Their tweets could be included in the data, which might induce bias. Additionally, tweets data could not represent the entire population despite Twitter being one of the major social media platforms. Not to mention people who do not or cannot use the Internet, not all online users use Twitter as their main sharing platform. Thus, one must be cautious that results in this study might not be generalizable to other social media platforms or the general public. Regarding language processing, many special forms of wording like a combination of words (e.g., “keep social distance”) and alternative short forms usually found on the Internet (e.g., “luv” for love) could be missed by lemmatization stemming or dictionaries for sentiment analyses. Finally, since the current data was unlabelled, external validation was unable to apply to the results of all the unsupervised analyses. Internal validation relies on validity statistics and manual interpretations could be subjective and less reliable.

Researchers may enlarge the extent of data collection for future studies by including a wider range of search words and streaming data across different social media platforms. Language processing algorithms need continuous evolution and development to adapt to constantly changing semantic expressions on the Internet to perform more accurate evaluations. Pre-analysis manual labeling of the data by blind participants would allow external validation of analysis results or construction of a more sophisticated classification tool to perform real-time sentiment and topic recognition.

In future work, we will incorporate further improved language analytical algorithms, sophisticated visualizations, and advanced real-time analytical tools. The instant language mining can be anticipated for public health practitioners, healthcare providers and policymakers, providing valuable information for rapid judgments and decision-making for an ongoing or upcoming public health crisis.

FUNDING AGENCY

Publisher has waived the Open Access publishing fee.

ACKNOWLEDGMENT

This work is partly supported by VC Research (VCR 0000156) for Prof Chang.

REFERENCES

- Abd-Alrazaq, D., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top concerns of tweeters during the COVID-19 pandemic: Infoveillance study. *Journal of Medical Internet Research*, 22(4), e19016. doi:10.2196/19016 PMID:32287039
- Adarsh, M. J., & Ravikumar, P. (2015). Survey: Twitter data analysis using opinion mining. *International Journal of Computers and Applications*, 128(5).
- Aggarwal, C. C., & Zhai, C. (2012). *Mining Text Data*. Springer Science & Business Media. doi:10.1007/978-1-4614-3223-4
- Ahmed, W., Bath, P. A., Sbaffi, L., & Demartini, G. (2020). Zika outbreak of 2016: insights from Twitter. In *Proceedings of Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing* (pp. 447–458). Springer. doi:10.1007/978-3-030-49576-3_32
- Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 39(1), 45–65. doi:10.1016/S0306-4573(02)00021-3
- Alhawarat, M., & Hegazi, M. (2018). Revisiting k-means and topic modeling, a comparison study to cluster arabic documents. *IEEE Access: Practical Innovations, Open Solutions*, 6, 42740–42749. doi:10.1109/ACCESS.2018.2852648
- AstraZeneca. (2020). *AstraZeneca and Oxford University announce landmark agreement for COVID-19 vaccine*. <https://www.astrazeneca.com/media-centre/press-releases/2020/astrazeneca-and-oxford-university-announce-landmark-agreement-for-covid-19-vaccine.html>
- Bagheri, M. S., & de Jong, F. (2014). ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences. *Journal of Information Science*, 40(5), 621–636. doi:10.1177/0165551514538744
- Baid, P., Gupta, A., & Chaplot, N. (2018). Sentiment Analysis of Movie Review using Machine Learning Techniques. *International Journal of Engineering & Technology*, 7(2), 676.
- Bhat, M., Qadri, M., Beg, N.-A., Kundroo, M., Ahanger, N., & Agarwal, B. (2020). Sentiment analysis of social media response on the Covid19 outbreak. *Brain, Behavior, and Immunity*, 87, 136–137. doi:10.1016/j.bbi.2020.05.006 PMID:32418721
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Blei, M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. doi:10.1145/2133806.2133826
- Bloom, B. R., Nowak, G. J., & Orenstein, W. (2020). "When Will We Have a Vaccine?"—Understanding Questions and Answers about Covid-19 Vaccination. *The New England Journal of Medicine*, 383(23), 2202–2204. doi:10.1056/NEJMp2025331 PMID:32897660
- Bottou, L., & Bengio, Y. (1995). Convergence properties of the k-means algorithms. *Advances in Neural Information Processing Systems*, 585–592.
- Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., Schijvenaars, B., Skupin, A., Ma, N., & Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS One*, 6(3), 1–11. doi:10.1371/journal.pone.0018029 PMID:21437291
- Bui, Q. V., Sayadi, K., Amor, S. B., & Bui, M. (2017). Combining Latent Dirichlet Allocation and K-means for documents clustering: effect of probabilistic based distance measures. In *Proceedings of Asian Conference on Intelligent Information and Database Systems*. Springer. doi:10.1007/978-3-319-54472-4_24
- Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1), 1–27.
- Cascini, F., Pantovic, A., Al-Ajlouni, Y., Failla, G., & Ricciardi, W. (2021). Factors Associated with COVID-19 Vaccine Hesitancy. *Vaccines*, 9(3), 300. doi:10.3390/vaccines9030300 PMID:33810131

- Chen, L., Lyu, H., Yang, T., Wang, Y., & Luo, J. (2020). *In the eyes of the beholder: Sentiment and topic analyses on social media use of neutral and controversial terms for covid-19*. arXiv preprint arXiv:2004.10225.
- Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of Tweets during the 2009 H1N1 outbreak. *PLoS One*, 5(11), e14118. doi:10.1371/journal.pone.0014118 PMID:21124761
- Comito, C., Forestiero, A., & Pizzuti, C. (2018). Twitter-based influenza surveillance: an analysis of the 2016-2017 and 2017-2018 seasons in Italy. *IDEAS 2018: Proceedings of the 22nd International Database Engineering & Applications Symposium*, 175-183. doi:10.1145/3216122.3216128
- DubeyD. (2021). Public Sentiment Analysis of COVID-19 Vaccination Drive in India. Available at SSRN 3772401. 10.2139/ssrn.3772401
- de Figueiredo, C., Simas, C., Karafillakis, E., Paterson, P., & Larson, H. J. (2020). Mapping global trends in vaccine confidence and investigating barriers to vaccine uptake: A large-scale retrospective temporal modelling study. *Lancet*, 396(10255), 898–908. doi:10.1016/S0140-6736(20)31558-0 PMID:32919524
- de Melo, T., & Figueiredo, C. M. (2021). Comparing News Articles and Tweets About COVID-19 in Brazil: Sentiment Analysis and Topic Modeling Approach. *JMIR Public Health and Surveillance*, 7(2), e24585. doi:10.2196/24585 PMID:33480853
- Ding, W., & Chen, C. (2014). Dynamic topic detection and tracking: A comparison of HDP, C-word, and cocitation methods. *Journal of the Association for Information Science and Technology*, 65(10), 2084–2097. doi:10.1002/asi.23134
- DubeyA. D. (2020). Twitter sentiment analysis during COVID19 outbreak. Available at SSRN 3572023.
- Fontanet, A., & Cauchemez, S. (2020). COVID-19 herd immunity: Where are we? *Nature Reviews. Immunology*, 20(10), 583–584. doi:10.1038/s41577-020-00451-5 PMID:32908300
- Funk, S., Gilad, E., Watkins, C., & Jansen, V. A. (2009). The spread of awareness and its impact on epidemic outbreaks. *Proceedings of the National Academy of Sciences of the United States of America*, 106(16), 6872–6877. doi:10.1073/pnas.0810762106 PMID:19332788
- Gensim 4.0.0. (2021). Available: <https://radimrehurek.com/gensim/>
- Hatzivassiloglou, V., & McKeown, K. (1997). Predicting the semantic orientation of adjectives. In *35th Annual Meeting of The Association for Computational Linguistics and 8th Conference of The European Chapter of The Association for Computational Linguistics*. Association for Computational Linguistics.
- Hung, M., Lauren, E., Hon, E. S., Birmingham, W. C., Xu, J., Su, S., Hon, S. D., Park, J., Dang, P., & Lipsky, M. S. (2020). Social network analysis of COVID-19 Sentiments: Application of artificial intelligence. *Journal of Medical Internet Research*, 22(8), e22590. doi:10.2196/22590 PMID:32750001
- Hussain, A., Tahir, A., Hussain, Z., Sheikh, Z., Gogate, M., Dashtipour, K., Ali, A., & Sheikh, A. (2021). Artificial Intelligence–Enabled Analysis of Public Attitudes on Facebook and Twitter Toward COVID-19 Vaccines in the United Kingdom and the United States: Observational Study. *Journal of Medical Internet Research*, 23(4), e26627. doi:10.2196/26627 PMID:33724919
- Jacobi, W., van Atteveldt, W., & Welbers, K. (2016). van Atteveldt, and K. Welbers, “Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4(1), 89–106. doi:10.1080/21670811.2015.1093271
- Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2019). Sentiment analysis on product reviews using machine learning techniques. In *Cognitive Informatics and Soft Computing* (pp. 639–647). Springer. doi:10.1007/978-981-13-0617-4_61
- Jain, K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. doi:10.1016/j.patrec.2009.09.011
- Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 881–892. doi:10.1109/TPAMI.2002.1017616

- Karl, A., Wisnowski, J., & Rushing, W. H. (2015). A practical guide to text mining with topic extraction. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(5), 326–340. doi:10.1002/wics.1361
- Karn, A. L., Qiang, Y. E., Karna, R. K., & Wang, X. (2018). News Sentiment Incorporation in Real-Time Trading: Alpha Testing the Event Trading Strategy in HFT. *Journal of Global Information Management*, 26(4), 18–35. doi:10.4018/JGIM.2018100102
- Kelaiaia, A., & Merouani, H. F. (2016). Clustering with probabilistic topic models on arabic texts: A comparative study of LDA and K-means. *The International Arab Journal of Information Technology*, 13(2), 332–338.
- Kim, E. H. J., Jeong, Y. K., Kim, Y., Kang, K. Y., & Song, M. (2016). Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. *Journal of Information Science*, 42(6), 763–781. doi:10.1177/0165551515608733
- Kim, S. M., & Hovy, E. (2004). Determining the sentiment of opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. COLING. doi:10.3115/1220355.1220555
- Kolini, F., & Janczewski, L. (2017). Clustering and topic modelling: A new approach for analysis of national cyber security strategies. In *Pacific Asia Conference on Information Systems (PACIS)*. Association for Information Systems.
- Kümpel, A. S., Karnowski, V., & Keyling, T. (2015). News sharing in social media: A review of current research on news sharing users, content, and networks. *Social Media + Society*, 1(2), 1–14. doi:10.1177/2056305115610141
- Lazarus, J. V., Ratzan, S. C., Palayew, A., Gostin, L. O., Larson, H. J., Rabin, K., Kimball, S., & El-Mohandes, A. (2021). A global survey of potential acceptance of a COVID-19 vaccine. *Nature Medicine*, 27(2), 225–228. doi:10.1038/s41591-020-1124-9 PMID:33082575
- Lu, Y., Zhang, P., Liu, J., Li, J., & Deng, S. (2013). Health-related hot topic detection in online communities using text clustering. *PLoS One*, 8(2), e56221. doi:10.1371/journal.pone.0056221 PMID:23457530
- Lwin, M. O., Lu, J., Sheldenkar, A., Schulz, P. J., Shin, W., Gupta, R., & Yang, Y. (2020). Global sentiments surrounding the COVID-19 pandemic on Twitter: Analysis of Twitter trends. *JMIR Public Health and Surveillance*, 6(2), e19447. doi:10.2196/19447 PMID:32412418
- Maindola, P., Singhal, N., & Dubey, A. D. (2018). Sentiment Analysis of Digital Wallets and UPI Systems in India Post Demonetization Using IBM Watson. In *2018 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE. doi:10.1109/ICCCI.2018.8441441
- Malik, A. A., McFadden, S. M., Elharake, J., & Omer, S. B. (2020). Determinants of COVID-19 vaccine acceptance in the US. *EClinicalMedicine*, 26, 100495.
- Manguri, K. H., Ramadhan, R. N., & Amin, P. R. M. (2020). Twitter sentiment analysis on worldwide COVID-19 outbreaks. *Kurdistan Journal of Applied Research*, 5(3), 54–65. doi:10.24017/covid.8
- Maynard, D., & Funk, A. (2011). Automatic detection of political opinions in tweets. In *Extended Semantic Web Conference*. Springer.
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113. doi:10.1016/j.asej.2014.04.011
- Mehla, L., Sheorey, P. A., Tiwari, A. K., & Behl, A. (2021). Paradigm Shift in the Education Sector Amidst COVID-19 to Improve Online Engagement: Opportunities and Challenges. *Journal of Global Information Management*, 30(5), 1–21. doi:10.4018/JGIM.290366
- Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*. Association for Computational Linguistics.
- Mohammad, S. M. (n.d.). *NRC Word-Emotion Association Lexicon (aka EmoLex)*. <https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>
- Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3), 436–465. doi:10.1111/j.1467-8640.2012.00460.x

Nabity-Grover, T., Cheung, C. M. K., & Thatcher, J. B. (2020). Inside out and outside in: How the COVID-19 pandemic affects self-disclosure on social media. *International Journal of Information Management*, 55, 102188. doi:10.1016/j.ijinfomgt.2020.102188 PMID:32836645

Natural Language Toolkit 3.6.2. (2021). Available: <https://www.nltk.org/>

Nazeer, K. A., & Sebastian, M. P. (2009). Improving the accuracy and efficiency of the k-means clustering algorithm. In *Proceeding of the World Congress on Engineering*. International Association of Engineering.

NLTK Data. (2005). *Sentiment Polarity Dataset Version 2.0*. NLTK. <http://www.cs.cornell.edu/people/pabo/movie-review-data/>

NRCLEx. (2019). Available: <https://pypi.org/project/NRCLEx/>

O'Donnell, P., & Hutchinson, J. (2015). Pushback journalism: Twitter, user engagement and journalism students' responses to 'The Australian'. *Australian Journalism Review*, 37(1), 105–123.

Patel, S., Chiu, Y. T., Khan, M. S., Bernard, J. G., & Ekandjo, T. A. (2021). Conversational Agents in Organisations: Strategic Applications and Implementation Considerations. *Journal of Global Information Management*, 29(6), 1–25. doi:10.4018/JGIM.20211101.0a53

Porter, M. F. (1980). *An algorithm for suffix stripping*. Program. doi:10.1108/eb046814

Rajput, N. K., Grover, B. A., & Rath, V. K. (2020). *Word frequency and sentiment analysis of twitter messages during coronavirus pandemic*. arXiv preprint, arXiv:2004.03925

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.

Rufai, S., & Bunce, C. (2020). World leaders' usage of Twitter in response to the COVID-19 pandemic: A content analysis. *Journal of Public Health (Oxford, England)*, 42(3), 510–516. doi:10.1093/pubmed/fdaa049 PMID:32309854

Sallam, M. (2021). COVID-19 vaccine hesitancy worldwide: A concise systematic review of vaccine acceptance rates. *Vaccines*, 9(2), 160. doi:10.3390/vaccines9020160 PMID:33669441

Salton, G. (1991). Developments in automatic text retrieval. *Science*, 253(5023), 974–979. doi:10.1126/science.253.5023.974 PMID:17775340

Sanders, A. C., White, R. C., & Severson, L. S. (2021). *Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse*. medRxiv, 2021: 2020.08. 28.20183863.

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge University Press.

Scikit-Learn 0.24. (2021). Available: <https://scikit-learn.org/stable/>

Starczewski, A., & Krzyżak, A. (2015). Performance evaluation of the silhouette index. In *International Conference on Artificial Intelligence and Soft Computing*. Springer. doi:10.1007/978-3-319-19369-4_5

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, 67(10), 2464–2476. doi:10.1002/asi.23596

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series. Materials Science and Engineering*, 336(1), 012017. doi:10.1088/1757-899X/336/1/012017

Tabak, F. S., & Evrim, V. (2016). Comparison of emotion lexicons. In *In 2016 HONET-ICT* (pp. 154–158). IEEE. doi:10.1109/HONET.2016.7753440

TextBlob v.0.16.0. (2020). Available: <https://textblob.readthedocs.io/en/dev/>

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 63(2), 411–423. doi:10.1111/1467-9868.00293

Tong, Z., & Zhang, H. (2016). A text mining research based on LDA topic modelling. In *International Conference on Computer Science, Engineering and Information Technology*. CCSEIT. doi:10.5121/csit.2016.60616

Vassilvitskii, S., & Arthur, D. (2006). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics.

WordNet Lemmatizer. (2021). Available: https://www.nltk.org/_modules/nltk/stem/wordnet.html

World Health Organization. (2020). *Coronavirus disease (COVID-19)*. <https://www.who.int/publications/m/item/weekly-epidemiological-update---12-october-2020>

World Health Organization. (2021). *Weekly Operational Update on COVID-19*. <https://www.who.int/publications/m/item/weekly-operational-update-on-covid-19---1-march-2021>

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., & Zhu, T. (2020). Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *Journal of Medical Internet Research*, 22(11), e20550. doi:10.2196/20550 PMID:33119535

Yeung, N., Lai, J., & Luo, J. (2020). *Face Off: Polarized Public Opinions on Personal Face Mask Usage during the COVID-19 Pandemic*. 10.1109/BigData50022.2020.9378114

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., & Zhang, G. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics*, 12(4), 1099–1117. doi:10.1016/j.joi.2018.09.004

ENDNOTE

- ¹ This result was subsequently found due to a bug in the NRCLex tool that skipped the frequency counting for all the “Anticipation” words.

Victor Chang is a Professor of Business Analytics, Aston University, UK, since mid-May 2022. He was previously a Professor of Data Science and IS at Teesside University, UK between September 2019 and mid-May 2022. He was a Senior Associate Professor, Xi'an Jiaotong-Liverpool University between June 2016 and Aug 2019. He was a Senior Lecturer at Leeds Beckett University, UK between Sep 2012 and May 2016. Within 4 years, he completed Ph.D. (CS, Southampton) and PGCert (HE, Fellow, Greenwich) while working for several projects. Before becoming an academic, he achieved 97% on average in 27 IT certifications. He won an IEEE Outstanding Service Award in 2015, best papers in 2012, 2015 & 2018, 2016 European award: Best Project in Research, 2017 Outstanding Young Scientist and numerous awards since 2012. He is widely regarded as a leading expert on Big Data/Cloud/IoT/security. He is a visiting scholar/PhD examiner at several universities, an Editor-in-Chief of IJOCL & OJBD, former Editor of FGCS, Associate Editor of TII & Info Fusion, founding chair of international workshops and founding Conference Chair of IoTBDs, COMPLEXIS, FEMIB & IoTBDSC. He was involved in projects worth more than £14 million in Europe and Asia. He published 3 books and edited 2 books. He gave 30 keynotes internationally as a top researcher.

Chun Yu Ng graduated with MSc in Data Science from Teesside University, UK. He was Prof Chang's and Prof Hossain's student.

Qianwen Ariel Xu completed MSc in Business Analytics from XJTLU and the University of Liverpool with Distinctions. She has worked as a Research Assistant. She is working with PhD under Prof Chang's supervision.

Mohsen Guizani (S'85–M'89–SM'99–F'09) received the BS (with distinction), MS and PhD degrees in Electrical and Computer engineering from Syracuse University, Syracuse, NY, USA, in 1984, 1986, and 1990, respectively. He is currently a Professor and Associate Provost at MBZUAI, Abu Dhabi, UAE. Previously, he worked in different institutions in the USA: University of Idaho, Western Michigan University, University of West Florida, University of Missouri, University of Colorado-Boulder, and Syracuse University. His research interests include wireless communications and mobile computing, applied machine learning, cloud computing, security and its application to healthcare systems. He was elevated to the IEEE Fellow in 2009. He was listed as a Clarivate Analytics Highly Cited Researcher in Computer Science in 2019, 2020 and 2021. Dr. Guizani has won several research awards including the "2015 IEEE Communications Society Best Survey Paper Award" as well as 4 Best Paper Awards from ICC and Globecom Conferences. He is the author of nine books and more than 800 publications. He is also the recipient of the 2017 IEEE Communications Society Wireless Technical Committee (WTC) Recognition Award, the 2018 AdHoc Technical Committee Recognition Award, and the 2019 IEEE Communications and Information Security Technical Recognition (CISTC) Award. He served as the Editor-in-Chief of IEEE Network and is currently serving on the Editorial Boards of many IEEE Transactions and Magazines. He was the Chair of the IEEE Communications Society Wireless Technical Committee and the Chair of the TAOS Technical Committee. He served as the IEEE Computer Society Distinguished Speaker and is currently the IEEE ComSoc Distinguished Lecturer.

Alamgir Hossain received the DPhil degree from the Department of Automatic Control and Systems Engineering, University of Sheffield, UK. He is currently serving as the Vice President and Professor of Artificial Intelligence of the Cambodia University of Technology and Science. Prior to this, he also served in the Teesside University (Research Lead of the Centre for Digital Innovation, Head of the National Horizon Centre), Anglia Ruskin University at Cambridge (Director of IT Research Institute), University of Northumbria (Research Lead of the Computational Intelligence Group), University of Bradford, University of Sheffield, Sheffield Hallam University and University of Dhaka (Chairman, Department of Computer Science and Engineering). He has extensive research experience in artificial intelligence, image processing, predictive modelling, cybersecurity, intelligent decision support systems, educational technology, robotics and adaptive control systems. He has led many funded projects as an international lead investigator, worth over £16 million. With a publication in Nature, he has published over 350 refereed research articles, contributed to 12 books, received the "F C Williams 1996" award (UK) and Lifetime achievement award (Channel S, London). He has received over 12,000 citations and won five best paper awards at international conferences. He had 17 successful PhD completions under his direct supervision. He also served as a guest editor of many high-quality journals and chair/co-chairs of many conferences. See further details of articles and citations in his Google Scholar account.