Twitter Sentiment Analysis Based on COVID-19 Vaccination

Fatemeh Haghighi - University of Guelph
Mahmoud Hazari - University of Guelph
School of Computer Science- December 2021

***Abstract*—Anti-vaccination attitudes have been an issue since the development of the first COVID-19 vaccines. The increasing use of social media as a source of health information may contribute to vaccine hesitancy due to anti-vaccination content widely available on social media, including Twitter. So having sentiment analysis over vaccination's content on Twitter is of great importance. We aimed to implement a sentiment analysis on Twitter topics related to the COVID-19 vaccination. We labeled the available data points and classified the tweets into two positive and negative subgroups. After predicting the overall attitude of the labeled tweets with LSTM, we attempted to categorize the negative tweets and identify the reasons for the hesitancy to receive the COVID-19 vaccine. Based on our findings, most people have a positive attitude towards vaccination. Additionally, we found that LSTM can predict positive or negative sentiment based on the labeled tweets with 88% accuracy.**

I. Introduction

In the past two years, CoronaVirus, also called COVID-19, has become one of the biggest problems we have faced. The spread of the COVID-19 had a huge impact on every person's life in recent years and it changed the way we used to live. Nowadays, we have to wear a facemask everywhere to protect ourselves, our families, and our society; however, we know that this is not enough and the best protection against this virus is vaccination. COVID-19 vaccines are effective and can decrease the risk of spreading this virus. COVID-19 vaccines also protect individuals from getting seriously ill even if they get the virus. Despite all of the mentioned benefits of the vaccination, some people are still hesitant to receive their vaccines due to the reasons such as misinformation, vaccines' side effects, political views, and conspiracy theories. These people consistently share their views about the COVID-19 vaccination with others, and on some levels, they can affect people and lead them to not get the vaccine.
Social media platforms are one of the most important places to find negative and positive comments and views about the vaccination. These platforms are one of the places where people with different levels of knowledge and understanding can share their emotions and thoughts. For instance, Twitter is one of the most popular platforms in which we can find countless numbers of tweets in favor or against COVID-19 vaccinations.

Some studies have been done related to the sentiment analysis about the COVID-19 vaccination. An experiment showed that many individuals in Asian countries have positive sentiments towards the vaccine and are taking the first dose of the COVID-19 vaccine [1]. Vaccines have triggered hope for protection against COVID-19 and provoked anti-vaccine campaigns, so we need to analyze public views. In one research, the sentiment of Bangladesh people towards vaccination has been analyzed[2]. In another study, researchers conclude and analyze the sentiment and manifestation of the user of Twitter based on the main COVID-19 related trends. They have shown that social media platforms' overall positive manifestation and presence have strengthened over time.[3] Moreover, anti-vaccination attitudes have been an issue since the development of the first vaccines. The increasing use of social media as one of the health information resources may help vaccine hesitancy due to anti-vaccine content widely available on social media, including Twitter. In a study, researchers evaluated the performance of different natural language processing models to identify anti-vaccination tweets published during the COVID-19 pandemic[4]. Social media platforms such as Twitter are an inevitable part of our daily lives. These platforms are effective tools for spreading news, photos, or any other type of information. Besides the positive side of these platforms, they are often used for propagating malicious data or information. The spread of misinformation related to the COVID-19 pandemic and its threat could be a severe challenge for each country; thus, researchers tend to have misinformation detection regarding the COVID-19 trends in another study. [5]
As mentioned, it is important to find out the grounds of the negative comments on social media related to COVID-19 vaccination. The purpose of this study is to classify tweets of Twitter into positive, negative, and at last try to find the main reason behind the negative tweets and categorize them using unsupervised algorithms. The work conducted here will enable governments and other organizations to take appropriate measures against those people who are hesitant to receive vaccines.
The paper is organized as follows; section 2 provides a general description of our related work and background. Preliminaries are presented in section 3 of the report, while material and method are discussed in section 4. In section 5, we describe our final results and discuss this paper, and in section 7, we explain the limitations we encountered during this research.

II. Related Work and Background

The purpose of this review is to provide an overview of the issues relating to sentiment analysis on social media platforms. In a study over the Twitter platform, users' sentiments and manifestations related to the main CoronaVirus trends were analyzed, using Natural Language Processing and sentiment classification methods, including Recurrent

Neural Networks. This study showed that the trained model works much more accurately, with a smaller margin of error, in determining emotional polarity in today's 'modern' often with ambiguous tweets[3]. Moreover, in another study, researchers tend to have misinformation detection regarding the COVID-19 trends. To reach that aim, they considered two other approaches alongside Recurrent Neural Networks(long short-term memory); a multichannel convolutional neural network (MC-CNN) and k-nearest neighbors (KNN). [5]

By discovering effective vaccines for COVID-19, a new trend showed in COVID-related topics on Twitter. Consequently, the need for analyzing sentiments towards vaccination became important. In most of the works that have been done on sentiment analysis of vaccination, recurrent neural networks family has been used as one of the effective approaches. For example, in one study, researchers proposed two models to know the sentiment of Asian region tweets. They used a machine learning-based model, Naive Bayes, and a deep learning-based model, LSTM, to classify positive, negative, and neutral tweets. This experiment showed that several people in Asian countries have positive sentiments towards the vaccine and are taking the first dose of the COVID-19 vaccine.[1]

In another study, Traditional machine learning methods, recurrent neural network (RNN) algorithm, several different types of recurrent neural networks, including simple RNNs, Gated Recurrent Units (GRUs), LSTMs, and small BERT models, were used to analyze the sentiment of Bangladesh people towards the vaccination[2]. On the other hand, some researchers were just studying the negative tweets since identifying anti-vaccination tweets could provide useful information for formulating strategies to decrease anti-vaccination sentiments. For instance, in a study, the performance of different natural language processing models to identify anti-vaccination tweets published during the COVID-19 pandemic was evaluated. For this aim, the performance of bidirectional encoder representations from Transformers(BERT) and the bidirectional long short-term memory networks with pre-trained Glove embeddings with classic machine learning methods including support vector machine and Naive Bayes were compared.[4]

Although various aspects of sentiment analysis have been studied so far, we aimed to have sentiment analysis of the Twitter data-set about COVID-19 vaccination through a semi-supervised method, through which we first labeled and classified the tweets. Then we predict the overall sentiment of Twitter's users using the LSTM network. Moreover, we aim to analyze the negative tweets and cluster them to find the reason behind the vaccine hesitancy.

III. Preliminaries

A. Embedding

In natural language processing (NLP), word embedding is a term used for the representation of words for text analysis, typically in the form of a real-valued vector that encodes the meaning of the word such that the words that are closer in the vector space are expected to be similar in meaning. Word embeddings can be obtained using a set of language modeling and feature learning techniques where words or phrases from the vocabulary are mapped to vectors of real numbers. Conceptually it involves the mathematical embedding from space with many dimensions per word to a continuous vector space with a much lower dimension.
The initial embedding techniques dealt with only words. You would generate an embedding for each word in the set given a set of words. The simplest method was to one-hot encode the sequence of words provided so that each word was represented by 1 and other words by 0. While this effectively represented words and other simple text-processing tasks, it didn't really work on the more complex ones, such as finding similar words. Basically, a word embedding converts the word and identifies the semantics and syntaxes of the word to build a vector representation of this information. Some popular word embedding techniques include Word2Vec, GloVe, ELMo, FastText, etc.

B. Word2Vec

The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence. As the name implies, word2vec represents each distinct word with a particular list of numbers called a vector. The vectors are chosen carefully such that a simple mathematical function (the cosine similarity between the vectors) indicates the level of semantic similarity between the words represented by those vectors.

C. Sentence Embedding

In the case of large text, using only words would be very tedious and the information would limit what we can extract from the word embeddings. Sentence embedding techniques represent entire sentences and their semantic information as vectors. This helps the machine in understanding the context, intention, and other nuances in the entire text.Just like Word Embedding, Sentence Embedding is also a very popular research area with very interesting techniques that break the barrier in helping the machine understand our language.

- Doc2Vec
- SentenceBERT
- InferSent
- Universal Sentence Encoder

D. Doc2Vec

An extension of Word2Vec, the Doc2Vec embedding is one of the most popular techniques out there. It is an unsupervised algorithm and adds to the Word2Vec model by introducing

another paragraph vector. Also, there are two ways to add the paragraph vector to the model.

1) PVDM: Memory version of Paragraph Vector: We assign a paragraph vector sentence while sharing word vectors among all sentences. Then we either average or concatenate the (paragraph vector and words vector) to get the final sentence representation. If you notice, it is an extension of the Continuous Bag-of-Word type of Word2Vec where we predict the next word given a set of words. It is just that in PVDM, we predict the next sentence given a set of sentences.

2) PVDOBW: Distributed Bag of Words version of Paragraph Vector: Just lime PVDM, PVDOBW is another extension, this time of the Skip-gram type. Here, we just sample random words from the sentence and make the model predict which sentence it came from.

E. Dimension Reduction

Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension. Working in high-dimensional spaces can be undesirable for many reasons; raw data is often sparse due to the curse of dimensionality, and analyzing the data is usually computationally intractable. Dimensionality reduction is common in fields that deal with large numbers of observations and/or large numbers of variables, such as signal processing, speech recognition, neuroinformatics, and bioinformatics. Methods are commonly divided into linear and nonlinear approaches. Approaches can also be divided into feature selection and feature extraction. Dimensionality reduction can be used for noise reduction, data visualization, cluster analysis, or as an intermediate step to facilitate other analyses.

Feature selection approaches try to find a subset of the input variables (also called features or attributes). The three strategies are: the filter strategy, the wrapper strategy, and the embedded strategy. Data analysis such as regression or classification can be done in the reduced space more accurately than in the original space.

Feature projection (also called feature extraction) transforms the data from the high-dimensional space to a space of fewer dimensions. As in principal component analysis (PCA), the data transformation may be linear, but many nonlinear dimensionality reduction techniques also exist.For multidimensional data, tensor representation can be used in dimensionality reduction through multilinear subspace learning.

F. Principal component analysis

The main linear technique for dimensionality reduction, principal component analysis, performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. In practice, the covariance matrix of the data is constructed and the eigenvectors on this matrix are computed. The eigenvectors that correspond to the largest eigenvalues (the principal components) can now be used to reconstruct a large fraction of the variance of the original data. Moreover, the first few eigenvectors can often be interpreted in terms of the large-scale physical behavior of the system, because they often contribute the vast majority of the system's energy, especially in low-dimensional systems. Still, this must be proven on a case-by-case basis as not all systems exhibit this behavior. The original space (with the dimension of the number of points) has been reduced (with data loss, but hopefully retaining the most important variance) to the space spanned by a few eigenvectors.

G. Clustering

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."

It does it by finding some similar patterns in the unlabelled dataset, such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns. It is an unsupervised learning method; hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset. After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML systems can use this id to simplify the processing of large and complex datasets.

H. K-means

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. We'll define a target number k, which refers to the number of centroids we need in the dataset. A centroid is the imaginary or real location representing the center of the cluster. Every data point is allocated to each of the clusters by reducing the in-cluster sum of squares. In other words, the K-means algorithm identifies k number of centroids and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The 'means' in the K-means refers to averaging of the data; that is, finding the centroid. To process the learning data, the K-means algorithm in data mining starts with the first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved

I. Mini batch k-means

Mini Batch K-means algorithm's main idea is to use small random batches of data of a fixed size, so they can be stored in memory. Each iteration a new random sample from the dataset is obtained and used to update the clusters and this is repeated

until convergence. Each mini batch updates the clusters using a convex combination of the values of the prototypes and the data, applying a learning rate that decreases with the number of iterations. This learning rate is the inverse of the number of data assigned to a cluster during the process. As the number of iterations increases, the effect of new data is reduced, so convergence can be detected when no changes in the clusters occur in several consecutive iterations.
The empirical results suggest that it can obtain a substantial saving of computational time at the expense of some loss of cluster quality, but not extensive study of the algorithm has been done to measure how the characteristics of the datasets, such as the number of clusters or its size, affect the partition quality.

J. Density-based spatial clustering of applications

DBSCAN stands for density-based spatial clustering of applications with noise. It is able to find arbitrary shaped clusters and clusters with noise. The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster.There are two key parameters of DBSCAN:

- EPS: The distance that specifies the neighborhoods. Two points are considered to be neighbors if the distance between them are less than or equal to eps
- minPts: Minimum number of data points to define a cluster.

Based on these two parameters, points are classified as core point, border point, or outlier:

- Core Point: A point is a core point if there are at least minPts number of points (including the point itself) in its surrounding area with radius eps.
- Border point: A point is a border point if it is reachable from a core point and there are less than minPts number of points within its surrounding area.
- Outlier: A point is an outlier if it is not a core point and not reachable from any core points.

A starting point is selected at random at it's neighborhood area and is determined using radius eps. If there are at least minPts number of points in the neighborhood, the point is marked as core point and a cluster formation starts. If not, the point is marked as noise. Once a cluster formation starts, all the points within the neighborhood of the initial point become a part of cluster. If these new points are also core points, the points that are in the neighborhood of them are also added to cluster. Next step is to randomly choose another point among the points that have not been visited in the previous steps. Then the same procedure applies. This process is finished when all points are visited. By applying these steps, DBSCAN algorithm is able to find high density regions and separate them from low density regions. A cluster includes core points that are neighbors and all the border points of these core points. The required condition to form a cluster is to have at least one core point.

K. Recurrent Neural Networks

A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data or time series data. These deep learning algorithms are commonly used for ordinal or temporal problems, such as language translation, natural language processing (NLP), speech recognition, and image captioning. Like feedforward and convolutional neural networks (CNNs), recurrent neural networks utilize training data to learn. They are distinguished by their "memory" as they take information from prior inputs to influence the current input and output. While traditional deep neural networks assume that inputs and outputs are independent of each other, the output of recurrent neural networks depends on the prior elements within the sequence. While future events would also be helpful in determining the output of a given sequence, unidirectional recurrent neural networks cannot account for these events in their predictions.

L. LSTM

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture. LSTM has feedback connections. It can process not only single data points, but also entire sequences of data. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM networks are well-suited for classifying, processing, and making predictions based on time series data. There can be lags of unknown duration between important events in a time series.

IV. Material and Method

A. Dataset

The dataset used in this project is the information of tweets posted from December 2020 to October 2021, containing 212981 data rows. Each data point contains information regarding a tweet, including the username of the tweet's owner, user location, user description, the creation date of the account, user followers, user friends, user favorites, whether the user is verified, the number of retweets, the source of tweet, hashtags, the date tweet has posted, and the text of the tweet.

B. Data Processing

The first step in data processing was detecting the data points that were noisy or not valuable. So, we gathered the NAN values in each column and analyzed them. Having NAN values in username and source could be a sign of noise or invaluable data since each user in Twitter has a username and it is impossible to have an account without any username. Similarly, based on the settings that exist for each tweet, every tweet is along with its source that shows from which device or operating system the tweet has posted. As a result, the data points with NAN username or source were removed from the dataset.
Another approach for detecting the validity of data points was checking the creation date of the account and the date that the

tweet was posted; if the published date of the tweet is earlier than the creation date of the account, it means the owner of the account is not real, and it could be a bot. This statement checked over each data point and based on that, there was no tweet posted from a bot. After validating the datapoint, we processed and cleaned the text of tweets. Text cleaning is the process of preparing raw text for natural language processing so that machines can understand human language. There are some steps that should be considered to reach the clean text for NLP works:

- Removing URLs, hashtags, etc
- Removing Stop Words
- Stemming and Lemmatization
- Removing Punctuations

1) Removing URLs, hashtags: In this part, we removed URLs, hashtags, single quotes, emails, new lines, mentioned, and converted all the characters to lowercase.

2) Removing Stop Words: The words which are generally filtered out before processing a natural language are called stop words. These are actually the most common words in any language and do not add much information to the text. Examples of a few stop words in English are “the”, “a”, “an”, “so”, “what”. Stop words are available in abundance in any human language. By removing these words, we removed the low-level information from our text in order to give more focus to the important information. The removal of such words does not show any negative consequences on the model we trained for our task. Removal of stop words definitely reduced the dataset size and thus reduced the training time due to the fewer number of tokens involved in the training.

3) Stemming and Lemmatization: Stemming and Lemmatization is Text Normalization techniques in the field of natural language processing that are used to prepare text, words, and documents for further processing. Languages we speak and write are made up of several words often derived from one another. When a language contains words derived from another word as their use in the speech changes, it is called Inflected Language. In grammar, inflection is the modification of a word to express different grammatical categories such as tense, case, voice, aspect, person, number, gender, and mood. An inflection expresses one or more grammatical categories with a prefix, suffix, or infix, or another internal modification such as a vowel change. The degree of inflection may be higher or lower in a language. Inflected words will have a common root form. Stemming and Lemmatization are widely used in tagging systems, indexing, SEOs, Web search results, and information retrieval. Stemming is the process of reducing inflection in words to their root forms, such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language. Stem (root) is the part of the word to which we add inflectionally (changing/deriving) affixes such as (-ed,-ize, -s,-de,mis). So stemming a word or sentence may result in words that are not actual words. Stems are created by removing the suffixes or prefixes used with a word.

4) Removing punctuations: the grammatical rules and signs which exist in the text or sentences could be counted as noise and do not carry any semantic information. So we first tokenized each text and then removed the words representing punctuation from the raw text.

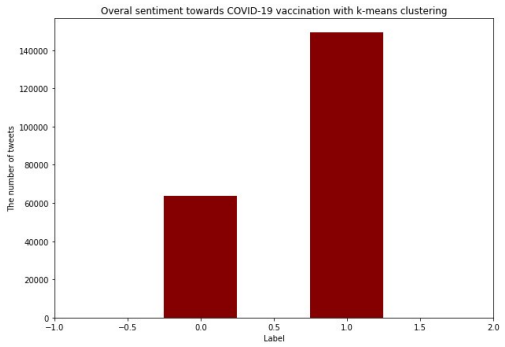

Fig. 1. Overall Sentiment towards COVID-19 Vaccination with K-means Clustering

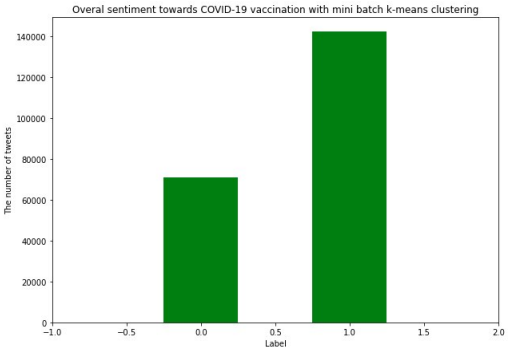

Fig. 2. Overall Sentiment towards COVID-19 Vaccination with Mini Batch K-means Clustering

V. Results and Discussion

The Clustering algorithms used for labeling the tweets, K-Means and Mini Batch K-Means, were not different in terms of the distribution of labels, shown in Figures 1, 2. However, when we manually observed the tweets and attempted to assign a label (positive or negative) to each, we understood the clustering output from the Mini Batch K-Means algorithm is more accurate and the labels were more separated than the K-Means labels. As a result of clustering algorithms and manually labeling, the number zero represents the negative tweets while the one represents the positive tweets.

As mentioned in the method section, the LSTM network trained over two sets of target data, K-Means targets and Mini Batch K-Means targets. The comparison of loss values and accuracy values in each epoch between train and validation data of these models are available in figures 3, 4, 5, 6. In both models, the trend exists in loss values and accuracy values are as expected and there is no over-fitting in them. As shown in table1, the accuracy and loss values of LSTM’s prediction over K-Means labels is better compared to the target generated by Mini Batch K-Means. Although the manual labeling with Mini Batch K-Means seemed more accurate, the selection of

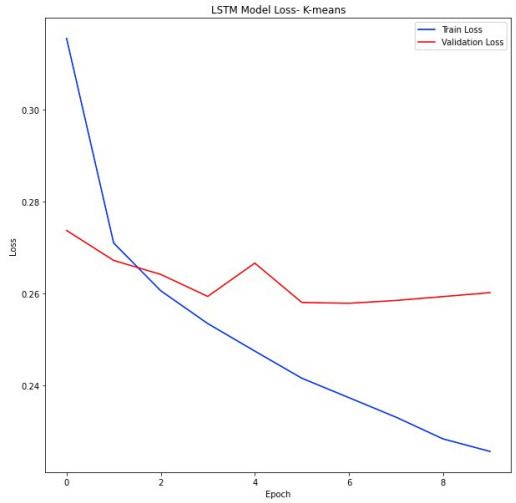

Fig. 3. LSTM Model Loss values for K-means

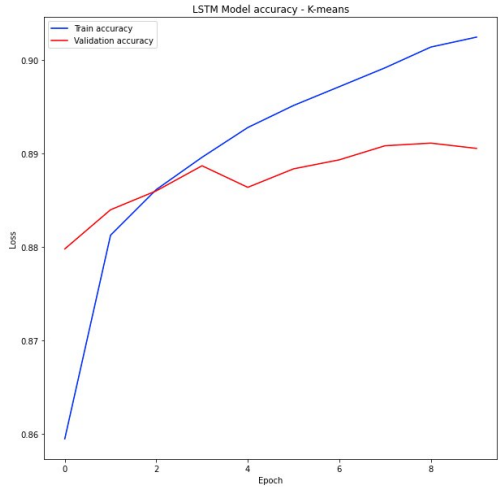

Fig. 4. LSTM Model Accuracy values For K-means

K-Means is a better choice for clustering. We could say that the LSTM network can predict the sentiment of tweets regarding the COVID-19 vaccination with an accuracy of 88%.

For analysis over negative tweets, two algorithms were assessed. The first one was DBSCAN, which detected 24 clusters out of negative tweets and 61465 noise points. We expected to obtain three clusters but the available dataset is not balanced in terms of positive and negative tweets regarding COVID-19 vaccination. Furthermore, Almost all of the negative tweets have been removed by Twitter since December 2020, and the small majority of them are at hand for analysis. Generally, the COVID-19 vaccination datasets are filtered and there is no exact information from the negative thought regarding this issue. As a result, we could not accurately analyze the negative tweets and label them. By looking at tweets, we understood most of them are positive and the existing negative tweets do not have any reasons or descriptions.

TABLE I
TABLE 1: THE COMPARISON OF LOSS AND ACCURACY VALUES OVER LSTM WITH K-MEANS TARGETS AND LSTM WITH MINI BATCH K-MEANS TARGETS

	Accuracy	**Loss**
LSTM with K-Means Targets	89%	26%
LSTM with Mini Batch K-means Targets	89%	27%

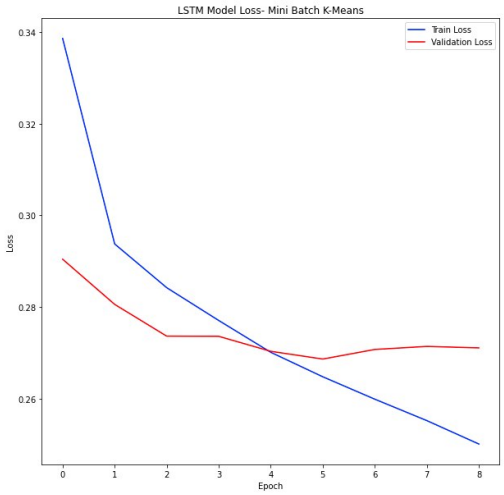

Fig. 5. LSTM Model Loss values For Mini Batch K-means

Considering the limitation mentioned, to cluster the negative tweets into misinformation, conspiracy theory, and political view categories the K-Means algorithm was used and divided the negative tweets into three categories. The distribution of the negative tweets in these clusters is shown in Figure 7. But since the tweets do not carry any other information and are not of big numbers, we could not manually label them.

Overall we conducted a project for sentiment prediction over an unlabeled dataset using the Twitter platform. We labeled the dataset through an unsupervised method and implemented an LSTM network that predicted the overall sentiment with

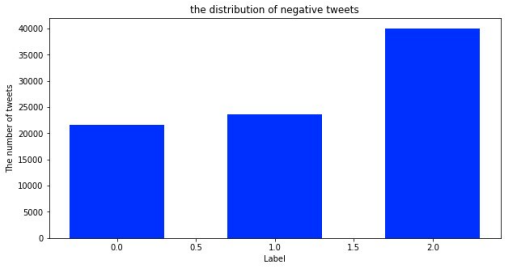

Fig. 6.
Distribution of Negative Tweets

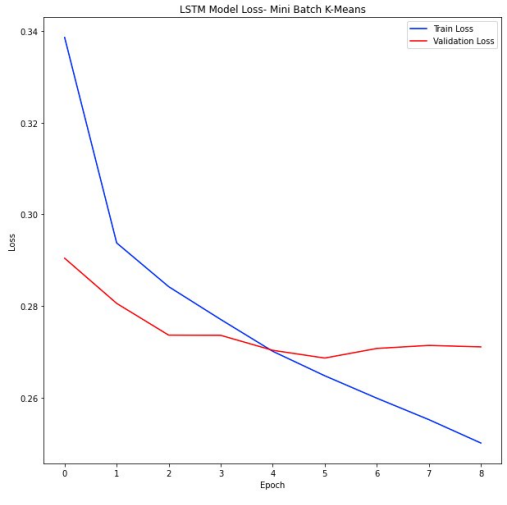

Fig. 7.
LSTM Model Loss - Mini Batch K-means

an accuracy of 88%. Moreover, we analyzed the negative tweets and understood that although there is a lack of data and information regarding the negative tweets, they could be clustered into 24 clusters.

VI. Limitation

This project had the primary goal of identifying the common reasons behind negative tweets about COVID-19 vaccines and categorized them into areas such as conspiracy theories, misinformation, and political views. However, we encountered some limitations during this project that led to the conclusion that we are not able to fully analyze negative tweets at this time. The first and foremost reason is the nature of sentiment and text analysis. For instance, there might not be enough context for reliable sentiment analysis on each tweet. The Twitter policy on negative tweets was also a hindrance to our progress. In response to the worldwide vaccination against COVID-19, Twitter announced they might ask customers to remove the negative tweets they have posted about the case of COVID-19 vaccination. Also, Twitter announced that it would directly remove tweets that contained inaccurate or misleading information. In addition, all of the datasets available to us were unable to fully satisfy our approach of identifying negative tweets as a result of the reasons mentioned earlier, such as the Twitter filtering policy. Also, the dataset had a problem in that if the tweet was long or if it included a picture, it would turn into a link and we couldn't use it for sentiment analysis. Among the sources we could use for our study, this dataset was the only available and most appropriate that we were able to find.

References

[1] Ranjan Raj Aryal and Ankit Bhattarai, *Sentiment Analysis on COVID-19 Vaccination Tweets using Naive Bayes and LSTM*, Advances in Engineering and Technology,An International Journal, 2021

[2] Anjir Ahmed Chowdhury, Argho Das, Suben Kumer Saha, Mahfujur Rahman, and Khandaker Tabin Hasan , *Sentiment Analysis of COVID-19 Vaccination from Survey Responses in Bangladesh*, Research Square, 2021

[3] László Nemes and Attila Kiss, *Social Media Sentiment Analysis Based on COVID-19*, Journal of Information and Telecommunication, 2020.

[4] QuyenG.To1* ,KienG.To2 ,Van-AnhN.Huynh2, NhungT.Q.Nguyen3, DiepT.N.Ngo2, Stephanie J.Alley, AnhN.Q.Tran, AnhN.P.Tran , NganT.T.Pham, ThanhX.Bui and Corneel Vandelanotte, *Applying Machine Learning to Identify Anti-Vaccination Tweets During the COVID-19 Pandemic*, International Journal of Environmental Research and Public Health, 2021

[5] Mohammed N. Alenezi, Zainab M. Alqenaei, *Machine Learning in Detecting COVID-19 Misinformation on Twitter*, Future Internet, 2021

[6] Purva Huilgol, *Top 4 Sentence Embedding Techniques using Python* August 25, 2020

[7] Wikipedia, Word2vec From *Wikipedia, the free encyclopedia*

[8] Benefits of Getting a COVID-19 Vaccine, *Centers for Disease Control and Prevention*

[9] Michael J. Garbade, *Understanding K-means Clustering in Machine Learning*, Towards Data Science, September 12, 2018