JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 07 Issue 01
2024

Research Article
https://doi.org/10.56979/701/2024

Navigating the Infodemic: The Impact of Social Media Rumors on Public Response to the COVID-19 Pandemic in Pakistan

Ramesha Rehman[1*], Syeda Mariyum Nizami[2], Rabia Younas[2], and Khalid Masood[1]

[1]Department of Computer Science, Faculty of Science, Lahore Garrison University, Pakistan.
[2]Department of Information Technology, Faculty of Science, Lahore Garrison University, Pakistan.
[*]Corresponding Author: Ramesha Rehman. Email: ramesharehman@lgu.edu.pk

Received: January 07, 2024 **Accepted:** May 01, 2024 **Published:** June 01, 2024

Abstract: The overall spread of the Coronavirus disease 2019 (COVID-19) irresistible sickness came about with a pandemic that has compromised a large number of lives. Infodemics is a well-known problem of interest. Nowadays, social media platforms are excellently representing the public sentiments and opinions about current events. Twitter is one of the most popular social media network that has captured the attention of researchers for studying public sentiments. Pandemic and Infodemics prediction on the basis of public sentiments expressed on Twitter has been an intriguing field of research. In this study we are focusing on people who have highest number of followers in Pakistan with most tweets related to Covid-19. The manually annotated dataset contains 2000 tweets of 1000 users for training and 380 tweets for test data from June to July 2020. For data processing we have manually labeled and added features to the dataset with the help of Senti Word Net. In the proposed model, KORONV is collecting data of tweets which will show the hashtags of COVID-19. Multiple machine learning algorithms are applied and the Long Short Term Memory (LSTM) gives the best accuracy of 98%. These techniques will be used to recognize patterns on the basis of existing algorithms, data sets and to develop adequate solution concepts that will be used for identification and classification between positive negative and neutral sentiment classification.

Keywords: Machine learning; COVID19; Data mining; LSTM; Deep learning.

1. Introduction

As indicated by Statista [1], an expected 2.95 billion individuals in 2019 adopted internet based life around the world. The number is anticipated to increase to 3.43 billion by 2023. One of the significant reason is the rapid evolution and access of the internet. Various reviews from Twitter have been collected to capture the online Sentiments of new users around the globe, and after analyzing the pattern of new users it can be observed that internet has made a huge impact on individual's life. Moreover, government and significant habitats for sickness control, including the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC), are depending on interpersonal networking as a source for coping up with the developing pandemic by routinely dispersing direction. The clouded side of online life was shown in a tidal wave of fake and unpredictable tweets that ran from offering forged fixes to utilizing internet-based life as a stage to dispatch cyberattacks on basic data systems. This drove the United Nations to caution against an expansion of bogus data about the infection and the rise of the COVID-19 Infodemic [2].

Infodemics is a well-known problem of interest. In this era of technology, Twitter has gain a lot of attention for expressing views and opinions on any current issue. Stock market prediction on the basis of public sentiments expressed on Twitter has been an intriguing field of research. In this research we have collected the data set from the tweets which shows the hashtags of COVID-19. In the proposed methodology we have collected the data of 1000 users. The data is collected to rank the topics related to COVID-19. Analysis of our simulations depends upon given three parameters

1.1. Granger

This widgets performs a series of statistical tests to determine the series that cause other series so we can use the former to forecast the latter**.**

1.2. Environment Monitoring Management System

Environmental monitoring management system (EMMS) will play more important role. This study integrated the big data technology into EMMS. First, it describes the traditional EMMS which has basic functions. Next it introduced big data technologies and analyzed EMMS system structure integrating with big data and gives an actual case study.

1.3. Correlation

It is used to study the closeness of the relationship between two or more variables i.e. the degree to which the variables are associated with each other.

2. Literature Review

The two most well-known small scale blogging social networking sites are Facebook and Twitter. There are some applied and used contrasts between among Twitter and Facebook, One of the progressively vital and significant difference is how Twitter clients have more open profiles rather than Facebook where users limit essentially precious data as notifications from users that are not a part of their framework. Truth be told, Twitter accounts are as a matter of course open so new users are naturally submitted on the open Twitter site as mentioned in Figure 1. [1].

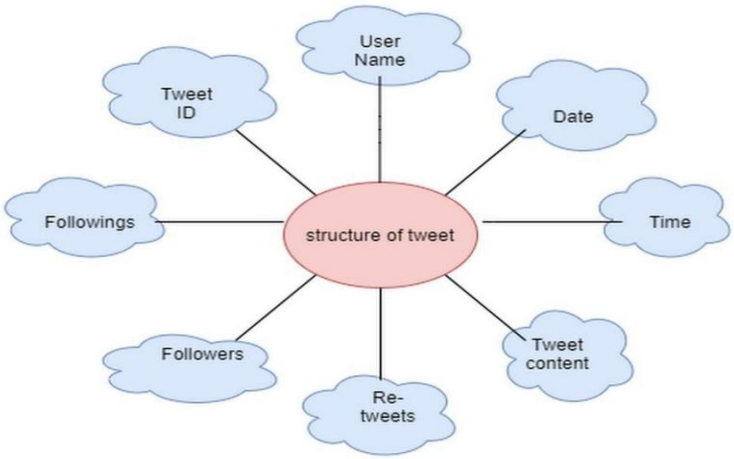

Figure 1. Structure of a tweet

2.1. Covid-19 & Twitter

Infodemic is of great concern with the recent evolution of Covid-19 pandemic in the world. Twitter serves as a source of information about Covid-19 related news and tweets. A study examined 4 million tweets of the user with hashtag of Covid-19 from March to April 2020. The tweets were categorized with five different discussion topics regarding Covid-19 and Latent Dirichlet Allocation (LDA) model is used to identify the sentiment of the tweets [2].

With the rise of coronavirus Pandemic twitter has influenced the powerful leaders of the world as well. They have also used their twitter platform to spread news related to coronavirus. A qualitative research is carried out with the tweets of 7 most followed world leaders. Their tweets were categorized using content analysis and the results showed that some of the tweets regarding Covid-19 were informative and the rest of them has other attributes [3-4]. As the use of twitter has risen with the evolution of coronavirus, the data generated from twitter has given an open platform to observe the activities on twitter all around the world in multiple disciplines. A dataset of 1.12 billion tweets were generated by the scientist from

January 2020 to June 2021 to give an opportunity to researchers from all fields to see the effect of tweets on people mental health and many others [5].

A comparison among Twitter and Facebook has started another discussion regarding low credibility sources that spread misinformation in both platforms and are somehow interlinked with each other [6]. Data about the quantity of posts, 'adherents' (number of users that are following the Twitter account), 'following' (number of different users following the Twitter account) and the social network between users could be utilized to decide the position of a particular user and hence its effect on the audience. [7– 9].

With the rise of Covid-19 Crisis in the world the social networking sites such as Facebook and Twitter is a source of information for the people sitting at homes. To control the spread of Infodemic on social networking websites is of key importance to overcome these crisis. Twitter is also one of the leading platform for misinformation regarding Covid-19.The author has presented the exploratory analysis to check the facts described in tweets from 92 different fact checked based organization in the time duration of two months (January to July 2020). The results shown 1274 false and 226 partially false tweets including the tweets from verified twitter accounts [10].

Another findings has used the NodeXL to gather the data from twitter with Covid-19 hashtag in Urdu language. They took two case studies with approximately 21k tweets regarding the Covid-19 pandemic and applied social framing theory, which proves that apparently twitter is not only spreading awareness to public about risks but also catering their personal as well as political gains [11].

Another paper has used fusion method and applied context based feature extraction techniques such as Bidirectional Encoder Representations from Transformers (BERT), Embeddings from Language Model (ELMo), and Generalized Auto-Regressive model for NLU (XLNet) for the detection of false news on social media websites such as twitter. Their proposed model has achieved the F1 score of 98% [12].

Nowadays the fight against Infodemic is of great concern for the world. The world health Organization (WHO) has spread awareness among people for Covid-19 vaccination on twitter along with mitigating the risk of spreading false information [13].

To curb the spread of false news three different projections are presented. First one is LDA to highlight the topics related to Covid -19 emerges as false and real news. Second step is to differentiate among real and fake news and the third step will identify the network based features [14]. To overcome the challenges faced on social media regarding misinformation another study focuses on the insights of the people on twitter regarding Covid-19. A wide range dataset is used consist of 90 thousand tweets labelled as COVID-SENTI in time duration of two months. The study gave the importance to the negative sentiment of the people due its adverse effects on people's mental health [15]. The risk of increase in Covid-19 cases the news regarding its surge circulate on social media platforms. A findings related to the impact of negative or positive news on social media is addressed with topic hunting strategy. An LSTM model is used to extract the public opinions in sentimental analysis. The model outperformed with the accuracy of 81% but it can be improved in future [16].

In a cafe survey, users share their experience around a few parts of the visit nourishment, mood, and administration, and so forth. As the quantity of tweets isn't restricted and every one notices an alternate opinion, getting a clue on the general feeling of these opinions from many surveys is bulky and tedious. Data mining, is a field for building a system to assemble and investigate emotions about the thing made in blog sections, comments, studies or tweets. Believe it or not, it has spread from programming building to the general sciences and humanistic systems. Different new organizations have created opinion mining systems which have found their applications in essentially every business and social life. [16– 18].

2.2. Covid-19 and Machine Learning

One of the significant issue that covid-19 has highlighted is the use of communicable and non-communicable devices such as Wi-Fi, and radar to track human motion and their health related issues such as chronic illness. Many Machine Learning (ML) based algorithms such as Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN) and multilayer perceptron are used to track and detect Covid-19 in humans [19]. A possible cure to curb another deadly variant of coronavirus (Omicron) is the antibody of monoclonal called sotrovimab. The AI models such as regressor learner have been utilized to fulfill the measures planned by the dynamical examination [20]. Another study regarding the novel virus has conducted in two highly infected countries India and Brazil. The findings used the Gaussian process regression (GPR) based model to accurately predict the redeemed and confirmed cases of Covid-19. The

results indicate a significant progress in using dynamic machine learning models as compared to others and gives mean percentage error of almost 0.1% and 95% predictions are within the confidence interval [21].

The covid-19 spread in countries like India, Brazil and America has always been on the peak. One of the studies conducted in India has used the ML based algorithm termed as Efficient Adaptive Migration Algorithm (EAMA) to efficiently predict the Covid-19 related parameters for long term and gives the predictions similar to the real time values. It is a hybrid approach which will focus on both present and past data obtained from Welfare India and Worldometer [22].

In predicting the Covid-19 epidemic in Jordan a method has already been used known as short term forecast (STF) as compared to the mathematical model labeled as long term forecast (LTF). A hybrid version (HF) of both models has used for the Covid-19 forecasting. The LTF was failed to accurately predict the Covid-19 related affairs due to the emergence of Omicron variant which spread faster than the previous variants. However by applying some changes to the LTF and HF model a better forecasting of Omicron variant has obtained [23-24].

2.3. NLP

Natural Language Processing (NLP) is a field of software engineering, intellectual reasoning, and computational semantics which is concerned about the collaborations between Computing systems and human (normal) dialects. All things considered; NLP is identified with the region of human computer interaction. Numerous difficulties in NLP include: common language understanding, empowering computing machines to get importance from human language information. In today's world NLP calculations depends on AI, particularly factual AI. The worldview of AI is not as same as that of earlier actions at language handling. Having the option to viably identify feelings in discussions prompts a wide scope of usage running from sentiment mining in online life stages.[25–28] Opinion mining of Twitter information is a field that has been given a lot of consideration in the course of most recent decade and includes dividing "tweets"(remarks) and the words of the sentences. Accordingly, this research investigates the different systematic analysis applied to twitter information and their results. [29–30].Twitter has been generally examined in settings of political, emergency, brand correspondence and client commitment around shared experiences, for example, TV screening and regular relational trades. [31–33].

3. Materials and Methods

In this work, the author consider a Twitter community as a set of Twitter users who share a common interest (e.g. news about Pandemic COVID - 19) the framework is made up of two principal components to deal with the User Generated Content (UGC) – in terms of topics, sentiment and emotions expressed - and the user's interaction behavior and posting patterns as shown in Fig. 2.

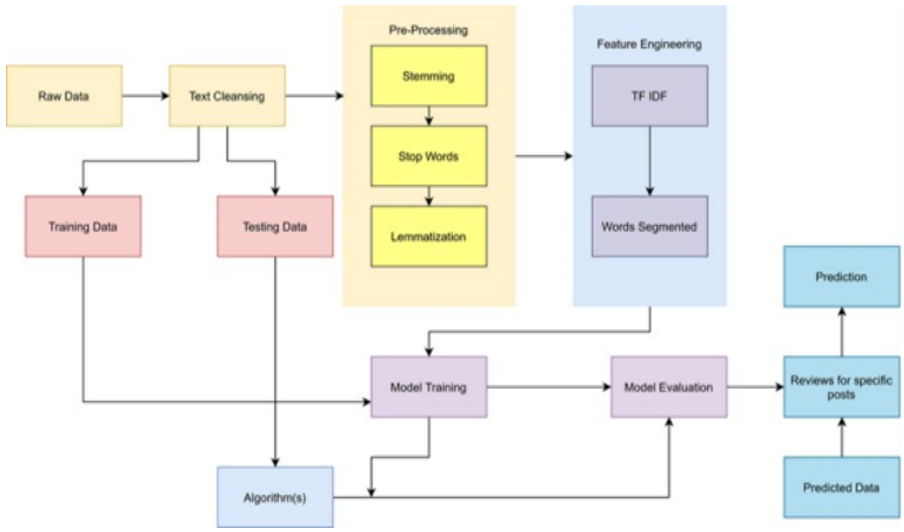

Figure 2. Methodology

3.1. Base Model for Data Mining

In Fig. 3 the model used collected the tweets which shows the hashtags of COVID-19. After that we have applied some data pre-processing for visualization, stemming and extracted feature from each dataset. We have Split the data into 20% testing and 80% training sets dataset. Training data will be tested and tested data will be predicted by the classifiers. Validation for the tested values compared by actual and predicted values in the graphs to show the efficiency of the classifiers. In the last the Identification and classification between fake news and correct news are predicted.

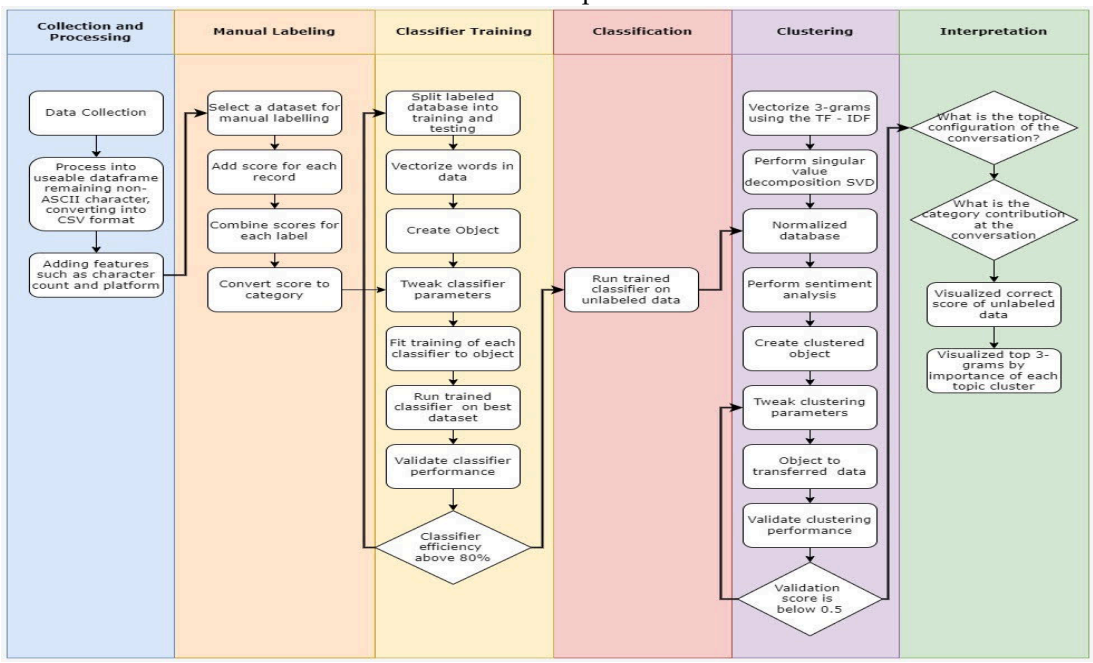

Figure 3. Base model interpretation

3.2. Data Collection and Processing

In Tab. 1 we have collected the data in the raw form, means its unlabeled data. We have converted our data into data frame for processing. We have added labels and features in order to process it. We have also added word count and weights, etc. These features are added into the csv data file.

Table 1. Sample tweets selected in dataset

Tweets	Labeled Sentiments
PM Imran Khan with Bill Gates today. Bill Gates highlighted Pakistan's great performance against COVID, under the leadership of PM IK.	Positive
CM Sindh Syed Murad Ali Shah presides over 31st meeting of Taskforce on Corona here at CM House	Neutral
Alarming : Actual coronavirus cases in Lahore 670,800 — According to this health ministry summary sent to the Punjab CM	Negative

3.3. Manual Labelling

We have selected dataset for manual labelling and for each data frame or feature we have added a score. Scores of each labels are combined. All combined scores are converted to categories. Further these categories are added to tweak classifier parameters.

Algorithm 1. Manual Labeling Steps

input: Unlabeled Tweet: U_A,
output: Labeled Tweet: L_B,

```
Positive: Lp← [];
Negative: Lneg ← [];
Neutral: Lneu ← [];
for each u ϵ U (u )do
    if (u is English): then Labeling: (SentiWordNet)
            Scorepos += positive score of u in the SentiWordNet
                Scoreneg += negative score of u in SentiWordNet
            If (Scorepos + Scoreneg > 0):
                Labeled as positive
              If (Scorepos + Scoreneg < 0):
                Labeled as negative
              Else:
                Labeled as Neutral
            endif
      Else:
      Delete u
    endif
endfor
return Labeled Tweets: LB = [Lp, Lneg, Lneu]
```

3.4. Classifier Training

In this step, we have split our dataset into training set and testing set. We have converted all Vectorized words into data. After this an object is created. All classifier parameters are placed in this object. Categories of all labelled data are also added into tweak classifier parameters. Each object is fitted on our model. After training it on training set, the model (classifiers) is mapped on best data set. We have validated the performance of our classifier on the basis of some parameters which are: If the efficiency of the classifier is more than 80%, then we will run our classifier on unlabeled data, Otherwise train it again on labelled data classification. As we have used supervised and unsupervised classifiers for our data validation, so in this step we have trained our unsupervised learning classifier with unlabeled data. And send the results to normalized database for clustering.

3.5. Clustering

In clustering the calculation of Term Frequency — Inverse Data Frequency (TF, IDF) of Vectorized data is performed. After this we have performed singular vector decomposition to find Eigen values. After performing semantic analysis, we will create an object for clustering. After creating an object the clustering algorithm is applied and fit that model to validate that object. If the validations score is below 0.5, it will be transferred for interpretation or visualization. The model is built for two classification tasks. A task is characterized into positive and negative classes and further that 3 possible ways of arranging the results are positive, negative and neutral classes.

3.6. Performance Parameters

Performance parameters such as accuracy, F score, precision and recall are calculated for authentication of the proposed technique. The visual and parametric outcomes of the proposed technique are compared with the existing literature. The performance parameters of the proposed method are computed as follows where TP represents true positive, TN represents True Negative, FP represents False positive and FN represents False Negative:

$$\textbf{Accuracy} = \frac{(\mathbf{TP}+\mathbf{TN})}{(\mathbf{TP}+\mathbf{TN}+\mathbf{FP}+\mathbf{FN})}$$

$$\textbf{Precision} = \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}} \quad (1)$$

$$\text{F score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

The accuracy is a measurement of a faithful representation of the truth, correctness. The accuracy can tell us immediately how the model is trained correctly. If the model is trained properly means that the system is accurate. In our research we are comparing the accuracy of these techniques to find out the efficiency of each algorithm.

4. Results and Discussions

This Models are arranged and cross validated by using the 2000 tweets as training data and 380 tweets for test data. Data is set up before applying machine learning algorithms to build up the models. Tokenization, is a process of converting a sentence into words, numbers, pictures, or other significant parts called tokens, is applied to the dataset. The series of tokens contributes for additional preprocessing. When everything is ready, the two fold cross validation is applied on every word (unigram) or per one of a kind pair of words (bigrams) in the corpus to such an extent that if that word or pair of words happens in a tweet, its relating property will get label of 1, else it will get a label of 0. Punctuation is treated as an element in deciding if the tweet is noised or not and yet not utilized as an element to decide sentiments. In Sentimental analysis for example, the question mark is a valuable marker for the sensitivity of a sentence. Moreover, the incorporation of a comma may stamp that a sentence is moderately more explained than one without a comma. For this situation, sentences in tweet information contain explanation, which are critical to distinguish in deciding if it is a fake sentence or an original sentence.

Our reference paper authors [16] developed a model to classify posts containing crosslink's from COVID-19 (i.e. links that point to another retweet) according to the sentiment contained in their texts. Indeed they expect that hidden negative words reveal a conflict intent in the source posts. In this work the text of comments are considered assuming that if a post receives attention and the response of other members agrees in the intent of the post we have further proof that a conflict may be generated by such activity.

The dataset used as a ground truth was built with the help of the Twitter website that manually annotated almost a thousand cross-links telling whether Source post sentiment toward target post was negative or positive / neutral (they got an inter-rater agreement of 0.95). They then used their model to expand the dataset.

Here, we suggest a better model created on which visibility of the text in terms of words without any use of additional extracted features such as the ones employed in the original paper. This is a trend in modern classification methods where the deep learning models substitute annotated feature engineering.

As we will see in the conclusions our model take the text of posts and comments achieves an improvement of 10 percent over the results of the research.

4.1. Infodemics Detection

In our case we have used four models for detection of Infodemics or rumors for COVID 19 in Twitter by using hashtags of profile having millions of followers and compare it with the lesser followers. In Fig. 4 and Fig.5 following are the number of followers of official profiles and having most COVID 19 tweets.

4.2. Data Cleansing

In data cleansing the dataset is cleansed to extract the desired features for our work. In this work, the following features are extracted which includes id (the reviewer who comment on a post), title (about the comment is posted), score (given by the user who comment), author (the id from the comment is posted), author_flair_text (comment), removed by(if the comment was removed), total_awards_received (collected awards by the reviewer), awarders, created_utc, full link, num_comments(total numbers of comments), over_18(how many comments having a length greater than 18), data is visualized as in Table 2.

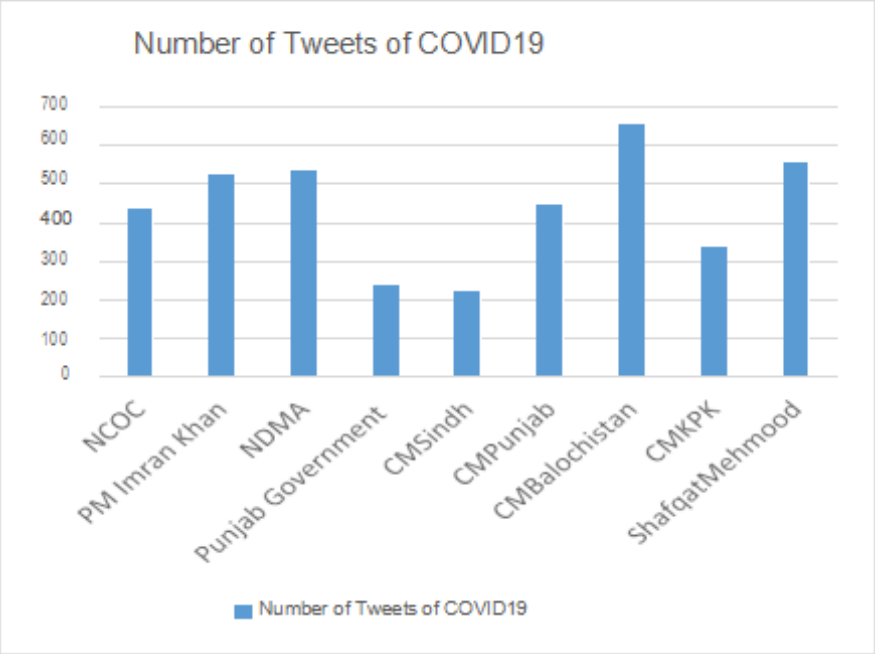

Figure 4. Number of tweet and COVID 19 of different user

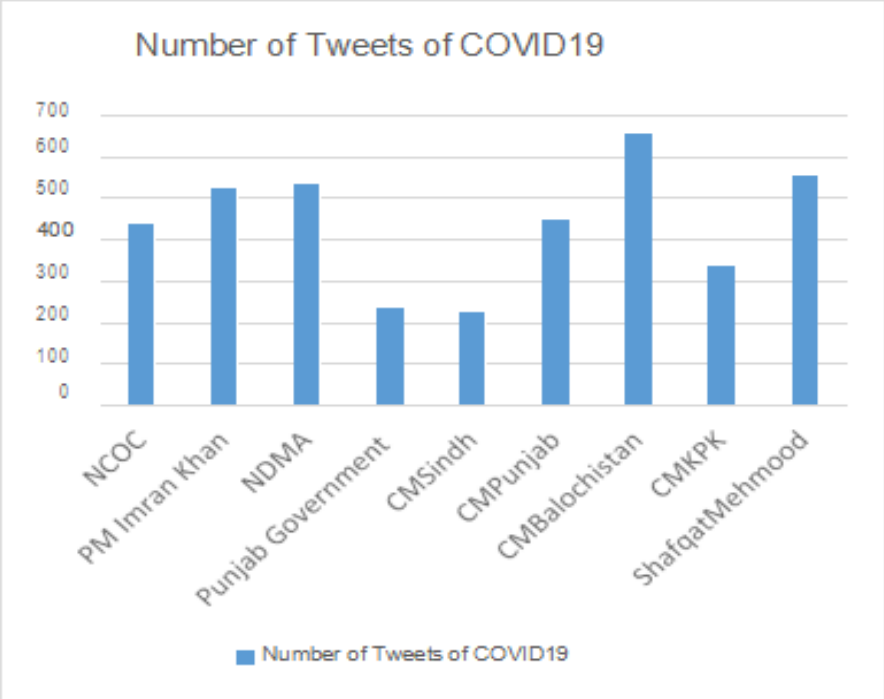

Figure 5. Number of followers of different user

Table 2. Features of data

Feature Name	Type	Description
"profile"	STR	Profile name
"Comments"	LIST	list of top comments(LIMIT 10)
"created_utc"	INT	timestamp of post
"link_flair_text"	STR	the flair of the post

"num_comments"	INT	number of comments on the post
"score"	INT	the score of the post (upvotes-downvotes)
"over_18"	BOOL	whether the post is sage-restricted or not
"selftext"	STR	description of the post
"title"	STR	title of the post
"url"	STR	url associated with the post

In Figure. 6 top 10 hashtags on twitter with Covid-19 label with given date is presented.

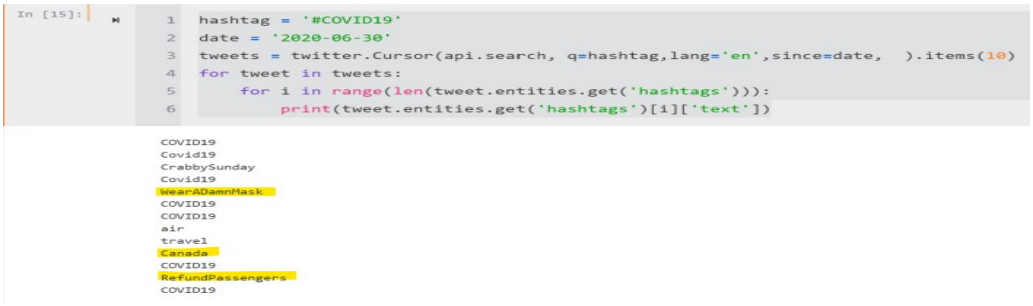

Figure 6. Extracted features

In Tab. 3 there are 183890 entries of each feature. In these features one is Boolean two are int64 and two are objects and these are using total memory of 7.2+ MB

Table 3. Calculated numeric columns

Feature Name	No _ of _entries	Integrity Constraints	Data Types
Title	183890	Non Null	Object
Score	183890	Non Null	Int64
Author	183890	Non Null	Object
Num_comments	183890	Non Null	Int64
Over_18	183890	Non Null	Bool

To check out numeric values we calculate count express a total number of comments.

Mean (m) and Standard Deviation:

$$\mathbf{m} = \frac{\text{sum of the terms}}{\text{numbers of terms}} \quad (5)$$

$$\boldsymbol{\sigma} = \frac{\sqrt{\sum(\mathbf{xi}-\boldsymbol{\mu})\mathbf{2}}}{\mathbf{N}} \quad \mathbf{(6)}$$

In Table 4 based on the scores and comment numbers and categorize them from minimum to the maximum level and visualize them in the form of a table:

Table 4. Visualization of numeric columns

Feature Name	Count	Mean	Std	Min	25%	50%	75%	Max
score	183890.0	186.653146	1969.840789	0.0	1.0	1.0	5.0	116226.0
Num_comments	183890.0	25.382386	195.623099	0.0	1.0	1.0	4.0	18801.0

4.3. Data Visualization

In Figure 7 and Figure 8 based on Table 4 we plotted the graph showing the number of digits and the number of comments.

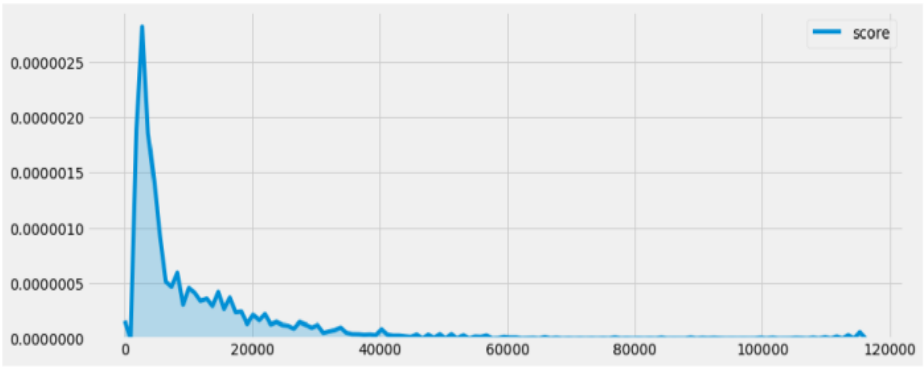

Figure 7. The graph for the score of comments

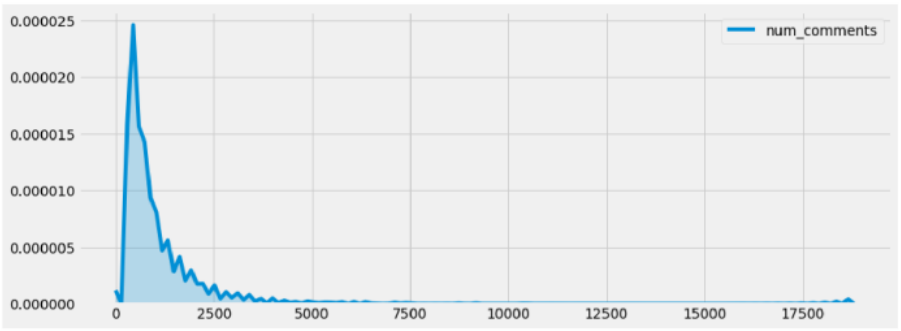

Figure 8. The graph for the number of comments

4.4. Word Cloud

Many times you may have seen a cloud with a collection of words in different sizes, expressing the frequency or significance of each word. This is called tag cloud or word cloud. For visualization purposes, we used the Matplotlib library which enables many other libraries to run and plot based on which include seaborne or words cloud.

4.5. Calculating TF-IDF

In a corpus, if we have to find frequency of the anticipated word in a document we use TF a.TF is calculated by formula having ratio the number of words appearing in a specific document to the total number of words in that document, which is given below. TF has direct relation with the number of word appearances within the associated document. So it is clear that TF depend upon document.

$$\text{TF}(t) = \frac{(\text{Number of times term t appears in a document})}{(\text{Total number of terms in the document})} \quad (3)$$

In the corpus, if we want to calculate weight of rare words in a document we use IDF. The term ID False indicates high score rare words in the corpus. The mathematical expression for IDF is given below.

$$\text{IDF}(t) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}\right) \quad (4)$$

Where t expresses the terms, d expresses each document and D expresses the collection of documents. Figure 9 shows the calculated frequency of each selected topic.

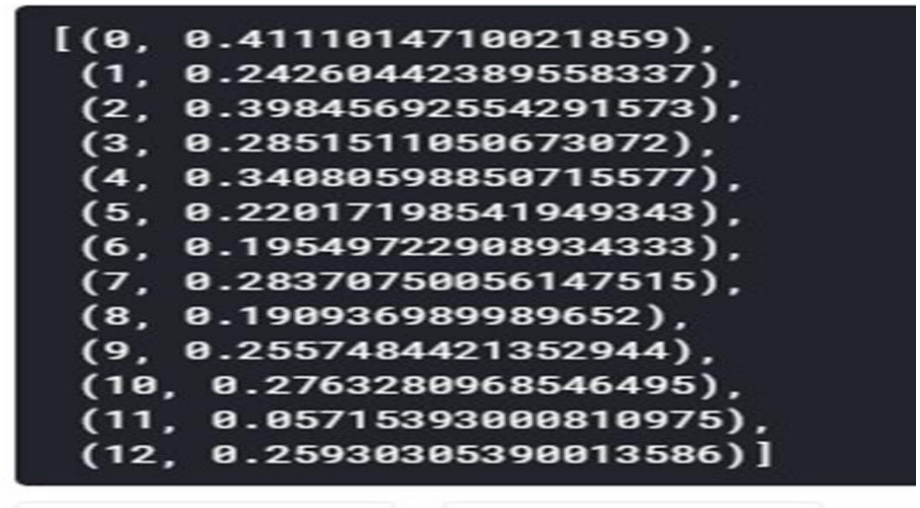

Figure 9. Calculated frequency of selected topics

4.6. Classification

In Figure 10 calculation of score of each selected topic is shown.

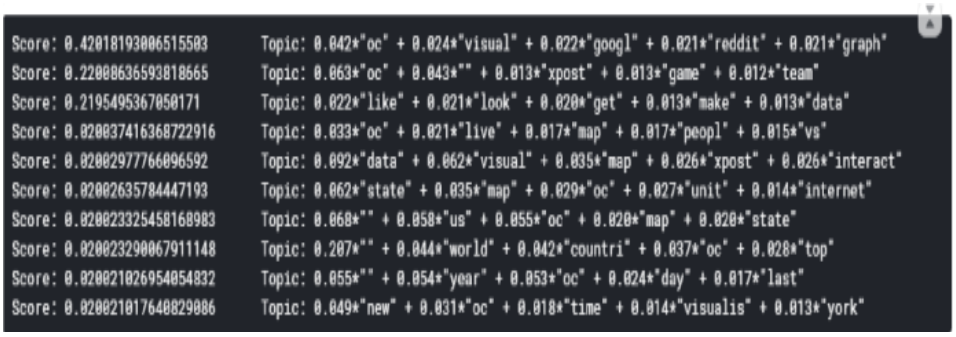

Figure 10. Score of each selected topic and profile

4.7. Naïve Bayes Classifier

It is a probability-based classifier. It does not work individually but with a whole family of algorithms.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad (5)$$

In the above, D is the available data and h is the hypothesis.

P (h): show the probability of hypothesis (without data). The term can also be called prior probability of h. P (D): show the probability of data (irrespective of hypothesis). The term can also be called prior probability. P (h|D): show the probability of hypothesis, by giving data to D. The term can also be called posterior probability.

P (D|h): show the probability of data d is the hypothesis h was true. The term can also be called progressive probability.

4.8. Linear Regression

Linear regression is a tool used to model the relationship between two variables using a linear equation. One variable is called the dependent while the other is called the independent variable. Here x is used as first and y is used as 2nd variable. In Fig. 12 the accuracy given by the linear regression model is given below:

In Figure 11 and Table 5 after running Naïve Bayes classifier the classification report for test data showing the precision, recall and F1 score values.

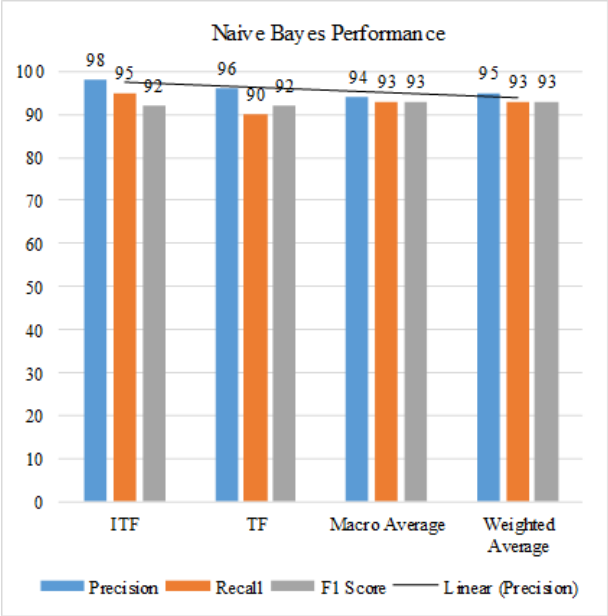

Figure 11. Naive Bayes performance

Table 5. Naive Bayes classifier performance on Covid KORONV dataset

Naïve Bayes Classifier				
	Performance Parameters	Precision	Recall	F1 score
Covid	TF	96%	90%	92%
KORONV	ITF	98%	95%	92%
Dataset	Micro Average Weighted	94%	93%	93%
	Weighted Average	95%	93%	93%

Figure 12. Performance of linear regression

Table 6. Linear regression classifier performance on Covid KORONV dataset

Linear Regression Classifier					
	Performance Parameters	Precision	Recall	Accuracy	F1 score
Covid					
KORONV	TF	96%	94%	97%	96%
Dataset	ITF	97%	95%	97%	97%
	Macro Average	94%	94%	97%	96%
	Weighted Average	97%	93%	97%	96%

In Table 6 results are shown with linear regression classifier on Covid KORONV dataset with TF-IDF.

4.9. KNN Classifier

KNN algorithm is categorized from one type of supervised ML algorithm. By using this we can find the solution of problems like classifications and regression. Though, the classifications problem has major applications in industry.

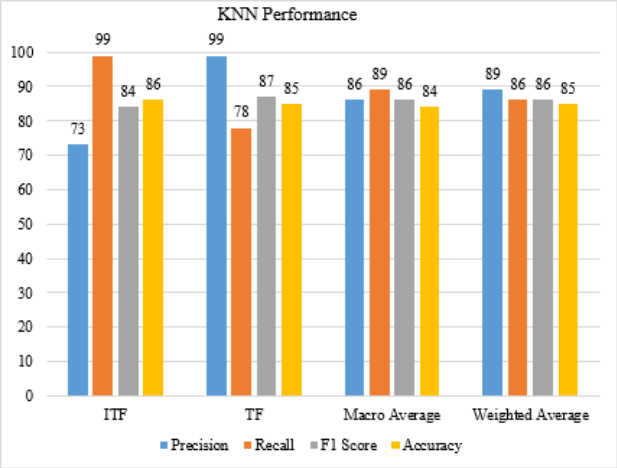

Figure 13. KNN performance

Table 7. KNN classifier performance on COVID KORONV dataset

	KNN Classifier				
	Performance Parameters	Precision	Recall	Accuracy	F1 score
Covid	TF	99%	78%	85%	87%
KORONV	ITF	73%	99%	86%	84%
Dataset	Macro Average	86%	89%	84%	86%
	Weighted Average	89%	86%	85%	86%

In Table 7 results with KNN classifier are shown on COVID KORONV dataset with TF-ITF as feature extraction technique.

4.10. SVM Classifier

In Figure 14 the support vector machine algorithm is categorized from one type of supervised ML algorithm. By using this we can find the solution of problems like classifications and regression. Though, the classifications problem has major applications in industry.

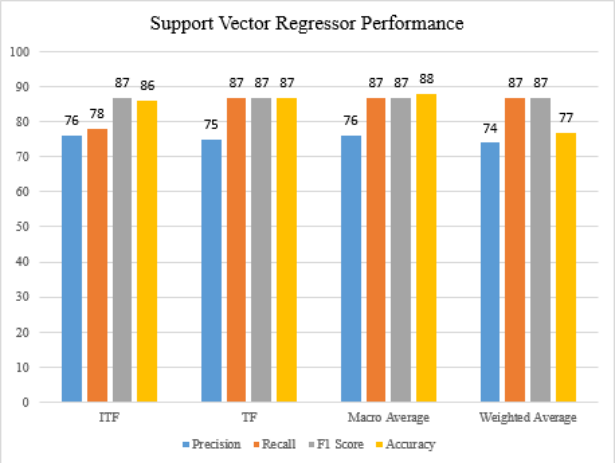

Figure 14. Support vector machine performance

Table 8. SVM classifier performance on Covid KORONV dataset

SVM Classifier					
	Performance Parameters	Precision	Recall	Accuracy	F1 score
Covid KORONV Dataset	TF	75%	87%	87%	87%
	ITF	76%	78%	86%	87%
	Macro Average	76%	87%	88%	87%
	Weighted Average	74%	87%	77%	87%

In Tab. 8 results are displayed with SVM classifier on Covid KORONV Dataset with TF-ITF as feature extraction technique.

4.11. Long-Short Term Memory

This model is essentially based on a combination of a convolutional neural network followed by an LSTM recurrent network. At the basis of the model, we use Glove word embedding's that allow us to represent each word in the text in a fixed-size, compact, and dense vector of 300 floats also taking into account similarity between words. The advantage of Glove vectors over the simple one-hot representation of words is that these vectors were trained using a neural network so that Words that have the same reference are closer in vector space. The purpose of this layer is to capture more general patterns in the data helping the network to better generalize to new examples without excessively specializing in the text of the training data. The result is given in input to the bidirectional LSTM layer which is responsible for learning the patterns. Something that recurrent networks were designed to do well. For each class we have calculated precision, recall and F1 score obtaining the following results.

Table 9. Bidirectional LSTM layer at threshold 18

LSTM Classifier				
Covid KORONV dataset	Precision	Recall	F1	Support
Balanced	0.87	0.93	0.97	769
Unbalanced	0.93	0.87	0.97	807

As mentioned in Table 9 above one of the main challenges was to solve the initial unbalancing of the dataset. We did so using an under-sampling technique that could have been random but we chose a different solution. The performance overall is pretty good especially concerning the results of the original paper (0.80) considering that we didn't use much data (about 8000 examples) and that these data were mostly automatically generated. As shown in Fig.15 often in deep learning having more data always improves performance especially if we want to add complexity to the network.

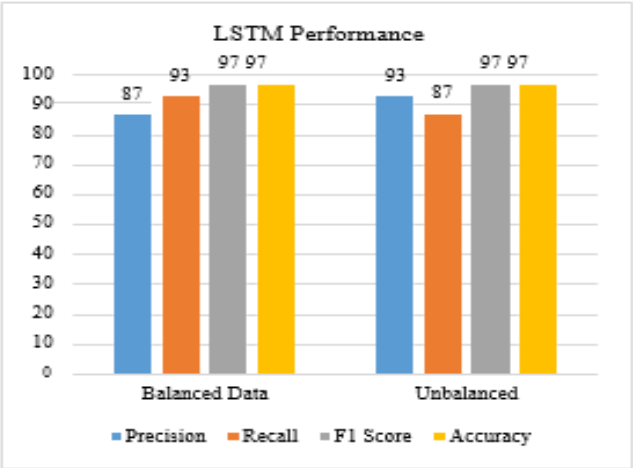

Figure 15. LSTM performance at test size 0.2 at threshold 18

In Figure 16 and Table 10 we are comparing the previous techniques with our proposed LSTM, as of literature as shown in previous paper [16]. The accuracy changes according to the ratio of data cleansing and other important preprocessing techniques, many authors have used KNN, SVM, and NB for discourse analysis. LSTM shows the best accuracy in terms of, Recall, AUC, F1 Score and Precision

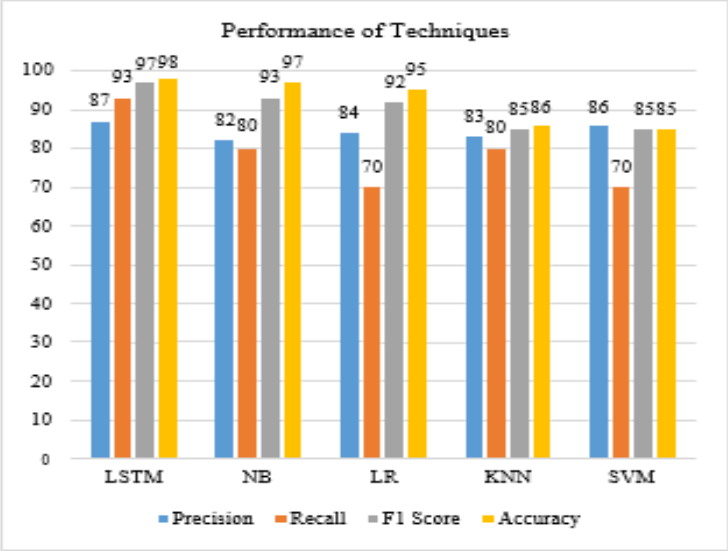

Figure 16. Performance of techniques

Table 10. Classifiers performance on Covid KORONV dataset

Performance of Techniques					
	Classifiers	Precision	Recall	Accuracy	F1 score
	LSTM	87%	93%	98%	97%
Covid	NB	82%	80%	97%	93%
KORONV	LR	84%	70%	95%	92%
Dataset	KNN	83%	80%	86%	85%
	SVM	86%	70%	85%	85%

5. Discussion

This Models are arranged and cross validated by using the 2000 tweets as training data and 380 tweets for test data. Data is set up before applying machine learning algorithms to build up the models. Tokenization, is a process of converting a sentence into words, numbers, pictures, or other significant parts called tokens, is applied to the dataset. The series of tokens contributes for additional preprocessing. When everything is ready, the two fold cross validation is applied on every word (unigram) or per one of a kind pair of words (bigrams) in the corpus to such an extent that if that word or pair of words happens in a tweet, its relating property will get label of 1, else it will get a label of 0. Punctuation is treated as an element in deciding if the tweet is noised or not and yet not utilized as an element to decide sentiments. In Sentimental analysis for example, the question mark is a valuable marker for the sensitivity of a sentence. Moreover, the incorporation of a comma may stamp that a sentence is moderately more explained than one without a comma. For this situation, sentences in tweet information contain explanation, which are critical to distinguish in deciding if it is a fake sentence or an original sentence.

Our reference paper authors [16] developed a model to classify posts containing crosslink's from COVID-19 (i.e. links that point to another retweet) according to the sentiment contained in their texts. Indeed they expect that hidden negative words reveal a conflict intent in the source posts. In this work the text of comments are considered assuming that if a post receives attention and the response of other members agrees in the intent of the post we have further proof that a conflict may be generated by such activity.

The dataset used as a ground truth was built with the help of the Twitter website that manually annotated almost a thousand cross-links telling whether Source post sentiment toward target post was negative or positive / neutral (they got an inter-rater agreement of 0.95). They then used their model to expand the dataset.

Here, we suggest a better model created on which visibility of the text in terms of words without any use of additional extracted features such as the ones employed in the original paper. This is a trend in modern classification methods where the deep learning models substitute annotated feature engineering.

As we will see in the conclusions our model take the text of posts and comments achieves an improvement of 10 percent over the results of the research.

6. Conclusions

Due to the nature of the data and the general interference, Twitter is becoming very useful for understanding and modelling various topics. KORONV aims to use the tweet as a sample theme for information shared by people during COVID-19 epidemics, to visualize what is being studied and to model human emotions. Its main purpose is to examine the psychology and behavior of large-scale communities to deal with the economic and social crises of the current major epidemic, and to learn what we call rumors called Infodemics. The corona virus infection (COVID-19) has forced people to stay at home to reduce the spread of the virus while maintaining social differences. However, social media encourages people both domestically and globally. People share useful information (personal opinions, some facts, news, status, etc.) on social media sites to understand the infection that occurs in various public behaviors such as emotions, feelings and dynamics. In this work, we are developing a live application for viewing tweets in COVID-19. Although explicitly prohibited by Twitter, cliques may have formed within a retweet to tweet each other's posts or down people they don't like. Another issue is that users who are generally disliked by the community (due to trolling behavior, coherent off-topic posts or other factors) can quickly see their posts, regardless of the topics covered in the post. These issues may be a part of reducing the accuracy of classification models. In proposed model we have compared different supervised models such as Naïve Bayes (NB), LR, KNN and LSTM. The results clearly depicts that LSTM has outperforms all other mentioned methods with precision 87%, Recall 93%, F1 score 97% and Accuracy 98%.

References

1. N. Khan and M. N. A. Khan, "A review of opinion mining in twitter streams," International Journal of Next-Generation Computing, vol. 9, no. 1, 2018.
2. J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng et.al. "Twitter discussions and emotions about the COVID-19 pandemic: machine learning approach", Journal of Medical Internet Research, vol. 22, no.11, 2020.
3. S. R. Rufai and C. Bunce, "World leaders' usage of twitter in response to the COVID-19 pandemic: a content analysis", Journal of Public Health, vol. 42, no. 3, pp. 510-516, 2020.
4. M. Haman, "The use of Twitter by state leaders and its impact on the public during the COVID-19 pandemic", Jounral of Heliyon, vol. 6, no. 11, 2020.
5. J. M. Banda, R. Tekumalla, G. Wang, J.Yu, T. Liu et.al., "A large-scale COVID-19 twitter chatter dataset for open scientific research—an international collaboration", Epidemiologia, vol.2, no. 3, pp. 315-324, 2021.
6. K.C. Yang, F. Pierri, P.M. Hui, D. xelrod and C. Torres-Lugo, "The covid-19 infodemic: twitter versus facebook", Big Data & Society, vol. 8, no.1, 2021.
7. H. K. Sul, A. R. Dennis and L. I. Yuan, "Trading on Twitter: using social media sentiment to predict stock returns," Decision Sciences, vol. 48, no. 3, pp. 454–488, 2017.
8. A. Shelar and C. Y. Huang, "Sentiment analysis of twitter data," in Proc. - 2018 Int. Conf. Comput. Sci. Comput. Intell. CSCI, pp. 1301– 1302, 2018.
9. G. K. Shahi, A. Dirkson and T. A. Majchrzak, "An exploratory study of covid-19 misinformation on twitter" Online Social Networks and Media, vol. 22, 2021.
10. S.H. Batool, W. Ahmed and K. Mahmood, "Social network analysis of twitter data from Pakistan during COVID-19", Information Discovery and Delivery, 2021.
11. S. Biradar, S. Saumya and A. Chauhan. "Combating the infodemic: COVID-19 induced fake news recognition in social media networks", Complex & Intelligent Systems, pp.1-13, 2022.
12. D. Muñoz-Sastre, L. Rodrigo-Martín and I. Rodrigo-Martín, "The Role of twitter in the WHO's fight against the Infodemic." International Journal of Environmental Research and Public Health, 2021.
13. A. Gupta, H. Li, A. Farnoush and W. Jiang, "Understanding patterns of COVID infodemic: A systematic and pragmatic approach to curb fake news." Journal of Business Research, vol. 140, pp. 670-683, 2022.
14. U. Naseem, I. Razzak, M. Khushi, P. W. Eklund and J. Kim, "COVIDSenti: A large-scale benchmark twitter data set for COVID-19 sentiment analysis." IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 1003-1015, 2021.
15. H. Jelodar, Y. Wang, R. Orji and S. Huang, "Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach." IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 10, pp. 2733-2742, 2020.
16. M. Anjaria and R. M. R. Guddeti, "Influence factor based opinion mining of Twitter data using supervised learning", in 6th Int. Conf. Commun. Syst. Networks, COMSNETS, 2014.
17. D. D. Gaikar, B. Marakarkandy and C. Dasgupta, "Using twitter data to predict the performance of bollywood movies," Industrial Management & Data Systems, vol. 115, no. 9, pp. 1604–1621, 2015.
18. U. Saeed, S. Y. Shah, J. Ahmad, M. A. Imran, Q. H. Abbasi et.al.,"Machine learning empowered COVID-19 patient monitoring using non-contact sensing: An extensive review", Journal of Pharmaceutical Analysis,vol. 12, no. 2, pp. 193-204, 2022.
19. A. Nutini, J. Zhang, A. Sohail, R Arif and T. A. Nofal, "Forecasting of the efficiency of monoclonal therapy in the treatment of CoViD-19 induced by the Omicron variant of SARS-CoV2", Results in Physics, vol. 35, 2022.
20. Y. Alali, F. Harrou, and Y. Sun, "A proficient approach to forecast COVID-19 spread via optimized dynamic machine learning models", Scientific Reports, vol. 12, pp. 1-20, 2022.
21. S. Mohan, A. Abugabah, S. Kumar Singh, A. Kashif Bashir and L. Sanzogni, "An approach to forecast impact of Covid-19 using supervised machine learning model", Software: Practice and Experience, vol. 52, no. 4, pp. 824-840, 2022.

22. T. Hussein, M.H. Hammad, O. Surakhi, M. AlKhanafseh, P.L. Fung et.al. "Short-term and long-term COVID-19 pandemic forecasting revisited with the emergence of OMICRON variant in Jordan", Jounal of Vaccines, vol. 10, no. 4, 2022.
23. T. Hussein, M.H. Hammad, P.L. Fung, M. Al-Kloub, I. Odeh et.al. "COVID-19 pandemic development in Jordan—short-term and long-term forecasting", Journal of Vaccines, vol. 9, no. 7, 2021.
24. H.M. Arifin, "The influence of competence, motivation, and organisational culture to high school teacher job satisfaction and performance", International Education Studies, no. 1, pp. 105–112, 2015
25. A. Antelmi, J. Griffith and K. Young, "Towards a more systematic analysis of twitter data: A framework for the analysis of twitter communities," CEUR Workshop Proc., vol. 2259, pp. 303–314, 2018.
26. Y. Garg and N. Chatterjee, "Sentiment analysis of twitter feeds", in Int. Conf on Big Data Analytics, pp. 33–52, 2014.
27. R. Attu and M. Terras, "What people study when they study Tumblr," Journal of Documentation, vol.73, no. 3, pp. 528–554, 2017.
28. S. A. El Rahman, F. A. Alotaibi and W. A. Alshehri, "Sentiment analysis of twitter data", in Int. Conf. Comput. Inf. Sci. ICCIS, 2019.
29. E. M. Cody, A. J. Reagan, P. S. Dodds and C. M. Danforth, "Public opinion polling with twitter,"arXiv preprint, pp. 1– 15, 2016.
30. R. Soni and K. J. Mathai, "Improved twitter sentiment prediction through cluster-then-predict Model," arXiv preprint, vol. 4, no. 4, pp. 559–563, 2015.
31. K. Z. Bertrand, M. Bialik, K. Virdee, A. Gros and Y. Bar-Yam, "Sentiment in New York City: A high resolution spatial and temporal view," arXiv preprint, pp. 1–12, 2013.
32. T. Rao and S. Srivastava, "Twitter sentiment analysis: how to hedge your bets in the stock markets," arXiv preprint, pp. 227–247, 2014.
33. B. Metin, M. Atasoyu, E. Arslan, N. Herencsar and O. Cicekoglu, "A tunable immitance simulator with a voltage differential current conveyor", in Midwest Symp. Circuits Syst., (MWSCAS), pp. 739–742, 2017.
34. F. Galip, M. H. Sharif, M. Caputcu and S. Uyaver, "Recognition of objects from laser scanned data points using SVM," in Proc. Int. Conf. Multimed. Image Process. ICMIP, pp. 28–35, 2016.