Automatic labeling of twitter data for developing COVID-19 sentiment dataset

K. M. Azharul Hasan
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna, Bangladesh
az@cse.kuet.ac.bd

Sajal Das Shovon
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna, Bangladesh
shovon030cse.kuet@gmail.com

Naimul Hoque Joy
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna, Bangladesh
joy15naimul@gmail.com

Md. Saiful Islam
Department of Computer Science
King Saud University
Riyadh-11451, Saudi Arabia
saislam@ksu.edu.sa

***Abstract*—The COVID-19 has started expanding through the world and has become a pandemic since january 2020 . With the rise of new cases daily along with mass death, nation and society are becoming fearful of it. People from all over the world are expressing their thoughts and views about this pandemic in many social media platforms. Nowadays social media is one of the most common ways to express idea or verdict on something. With the improvement of modern computing technology, machines are constantly being conducted to interpret what people express in social media like Twitter, Facebook, Instagram etc. These thoughts or views can be categorized and analysed based on sentiment. In this paper, we have analysed the sentiment of people what they express in social media by using tweets gathering from Twitter. We have categorized the sentiment of the tweets into five classes namely 'Strongly Negative', 'Negative', 'Neutral', 'Positive' and 'Strongly Positive'. Initially we use the Textblob of python for classification. This classification does not show good results and needs massive change as there lies many new terms related to COVID-19 which effects the sentiment of tweets. We have labeled automatically by creating Regular Expression rules with our new corrected word library which is created by analysing the tweets manually. We have trained a model with the updated labeled dataset with Long short-term memory(LSTM), Logistic Regression(LR), Multinomial Naive Bayes(MNB) and analysed the performance. We found that our data labelling shows better performance comparing to standard dataset.**

***Index Terms*—Sentiment Analysis, COVID-19, Twitter, Textblob, Automatic Labelling, TF-IDF Vectorizer, Logistic Regression, Multinomial Naive Bayes, LSTM.**

I. Introduction

Information technology provides profound opportunities to fight infectious disease outbreaks and have a remarkable role, especially in sentiment analysis for social media and this is important due to their tremendous role in analysing public sentiment. Research shows that many outbreaks and pandemics could have been promptly controlled if experts considered social media data[1]. The COVID-19 pandemic created a sensational loss of human life worldwide and presents the challenge to global health, food systems, and the work universe[2]. The social media like facebook, twitter, instagram etc provides the opportunity to share or exchange ideas and information of COVID-19. People express their feelings and share their opinions and their activities about the pandemic and social crisis all over the world [3] [4]. Facebook alone has over 1.2 billion monthly active users, Twitter has over 1 billion registered users and Instagram over 400 millions approximately. Therefore, social media is an important platform to share fun, ideas and information and plays a vital role in connecting people and developing relationships through the world. The social media creates opportunity for mass people to share their views and hence connect the people in a single platform. The rapid growth of COVID-19 creates the demand for understanding the sentiment [5] [6] of mass people about what they express in social media about this pandemic. While some initiatives like medical care, preventive care or vaccination are ongoing, there must be an emphasis on social media communications to predict the aspects of COVID-19's effect. People may often express negative or positive sentiment about such pandemic. To understand the overall sentiment, it's obvious to analyse the sentiment.There lies enormous data in social media about COVID-19 as it frightens people of all classes. Sentiment classification from tweets uses label dataset from different sources [7]. The accuracy of such result depends on the labeled accuracy of dataset. In this paper, we develop a sentiment dataset using twitter data to analyse the multi class sentiments. The sentiment data of COVID-19 has some uniqueness. For example the word "positve" carries positive sense in normal text but positve for COVID-19 carries negative sentiment because "tested COVID positive" shows negative sense . Initially we used Textblob of python for classification. But it does not work well. Finally we recalculated the sentiment label of the dataset automatically by creating Regular Expression rules with our word library which is created by analysing the tweets manually. Finally the sentiments were annotated and trained using the Long short-

978-1-6654-0906-3/21/$31.00 ©2021 IEEE

term memory(LSTM), Logistic Regression(LR), Multinomial Naive Bayes(MNV) and analysed the performance.

The rest of the paper is organized as follows; In section II, some related works of recent years have been explained. Section III describes the development of the dataset. Section IV illustrates the training and classification with the dataset. The comparison result is analyzed in section V and finally section VI outlines some conclusions.

II. RELATED WORKS

With the development and availability of social media trending topics can be often discussed as negative or positive ways. Sometimes it can mislead people with misinterpretation of the topic like COVID-19. But with the assist of machine learning algorithms we can define the impact of COVID-19 correctly. Hence lots of work is ongoing on sentiment analysis. [7] uses deep learning method to detect sentiment using labeled data. The accuracy of such result depends on the labeled accuracy of dataset. The textblob of python also used in [8] for sentiment analysis on twitter data that are collected by COVID-19 and using Twitter API through python library. They classified the sentiments in optimistic, neutral and negative sentiment using Naive Bayes method. They also categorized the emotion of tweets for an one week duration based on the polarity. Naseem et al. [9] constructed a methodology to classify the sentiment with a large scale Twitter data and find if there lies any misinformation about COVID-19 through the social media. Their study contains several ML based algorithm like SVM, Naive Bayes,BERT, and CNN. Their findings show that people had a positive sentiment in the lockdown of February 2020 and by mid-March sentiment is got shifted. Radaideh et al.[10] introduced a procedure to calculate the sentiment analysis using Naive Bayes and RNN method on a real time scenario. They have collected the data from December 2019 to July 2020 and classified them into positive, neutral and negative sentiment where most of the tweets came out positive. Kaur et al.[11] have analysed sentiment of tweets that are collected based on keyword like coronavirus, deaths, new case etc. They used R programming language for their research. By using RNN and SVM they classified the sentiment in positive, neutral and negative classes. In their research, SVM shows maximum number of tweets as neutral whether RNN describes maximum number of tweets as positive. Samuel et al. [12] have proposed a methodology to identify fear sentiment involvement. They have gathered almost 900,000 data using twitter API and rTweet package in R. They have demonstrated a comparison between fear sentiment and negative sentiment on geo-tagged data too. After researching they have noticed that Naive Bayes has more accuracy than Logistic Regression in terms of both shorter or longer tweets. Chakraborty et al. [13] presented the aspects and research of social media analysis in many ways. This paper also deals with social network involved items. This very article indicates many future development of research that are based on social media's data.

III. DEVELOPING COVID-19 SENTIMENT DATASET FROM TWITTER DATA

Huge attention is necessary to acquiring data from social networks because the ideology of the social media user behind a post is diverse. Moreover the collected data from social media has other technical issues. For example it may or may not be noisy, it may be homogeneous or heterogeneous, it may or may not have diverse range. So generating data from social media demands huge care. Popular social media sites like Facebook, LinkedIn and Twitter provide media for the community to post, share, like and comment along with other friends within the same network. On the other hand, sites like Youtube, Flickr and Digg also are gearing up in providing various facilities for enhancing the connection with social friends. Fig. 1 shows the proposed methodology.

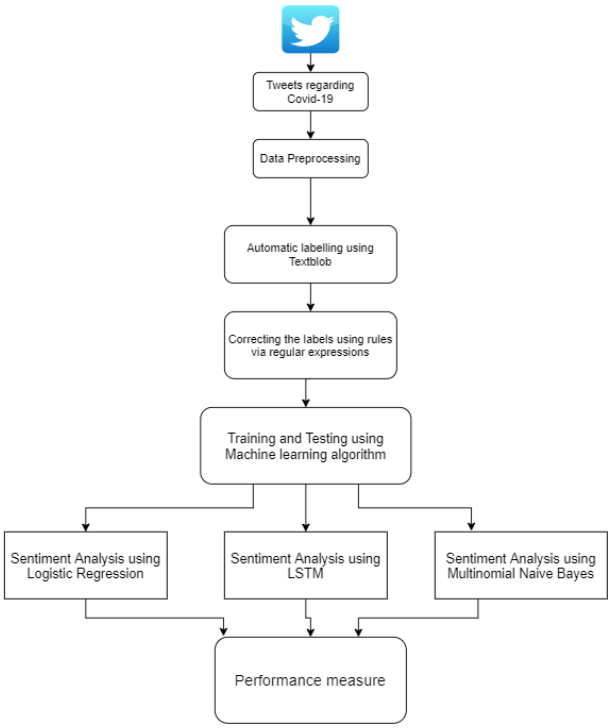

Fig. 1. Flow diagram of proposed methodology

A. Dataset Acquisition and Description

Without a clean and processed dataset research on machine learning is difficult. After having the clean dataset we have labeled our processed data using a python library called Textblob and classified. Our target is to generate a dataset containing COVID-19 tweets/sentences. Tweets contains a vast amount of texts that are relatable with COVID-19 based posts/ comments from CrisisNLP [14]. This resource consists of Twitter data collected during 19 natural and human-induced disasters. Each dataset contains tweet-ids and human-labeled tweets of the event. [15]. We used a data file consists of almost 1,10,000 tweets. Initially we have used python library twarc to hydrate tweets. Twarc is a toolkit for python to collect twitter

data by resembling twitter ID. Twarc provides several different methods for collecting Twitter data. We use twarc dehydrate method to fetch the tweets. It creates a list of tweet ID from a file of tweets. It has a collection of numerous tweets related to COVID-19. This tweets can be fetched even with geographic location. We gathered twitter ID from CrisisNLP[14] and using those IDs we hydrated tweet. The required dataset of tweets is stored in a CSV file.

For the best accuracy we have introduced automatic labelling and corrected the wrong labelling. Then we have used the labeled dataset to train in Ordinal Logistic Regression, Multinomial Naïve Bayes and LSTM. An analogy among these three methods and their effectiveness is discussed later to observe which model works better.

B. Data Pre-processing

For the sentiment data we need some related keywords to predict the sentiment. Therefore preprocessing is necessary. It will make the dataset clean to analyse the sentiment more precise way. A clean dataset is more likely to lead the relevant results. Preprocessing of dataset is a series of steps like removal of retweet, extra blank spaces, emojis,punctuation,unnecessary keywords, tokens etc. It consists of normalization of text data to improve text matching. The steps to preprocess the dataset are as follows.

- **Removal of retweet:** It means eliminating the retweet of the particular tweet. Retweets are unnecessary as they contain a little relevant sentiment decider word.
- **Removal of numbers and punctuation:** A tweet consists of lots of numbers and punctuation, we must remove them as they hold no important words which can be a part for analysing the sentiment.
- **Stop words removal:** In a sentence stopwords are very common and often have repetition. They don't carry an important meaning. Hence, we can remove stopwords to save computing time and efforts in processing large volumes of text. It is possible to remove stop words using Natural Language Toolkit(nltk).
- **Stemming:** Stemming is a technique that used to extract the base form of the words by removing affixes from them. There are various stemming algorithms in NLTK. We use the Snowball Stemmer to perform the stemming.
- **Noise removal:** Removing unnecessary keywords such as HTML tags, white spaces, emojis or any special character is noted as noise removal. These are not a part of sentiment analysis.
- **Letter casing:** Converting all letters to either small letter or capital letter. It is mainly done to avoid the duplication.
- **Tokenizing:** Turning the whole tweet into tokens is an important step in data preprocessing. Tokenization is the process of splitting the given text into smaller pieces called tokens. Words, numbers, punctuation marks, and others can be considered as tokens. We have used the Natural language toolkit (NLTK) library for tokenization. Tokens are mainly text words that are separated by spaces.

C. Data Labeling

1) Initial labelling by textblob: Sentiment analysis is the process of determining the attitude or emotion of a writer whether it's positive or negative or neutral. We have used Textblob to label our data. TextBlob is a python library and offers a simple API to access its methods and perform basic NLP tasks. TextBlob is used to initial data labelling. We use multi class classification (5 classes) namely 'Strongly Negative', 'Negative', 'Neutral', 'Positive', and 'Strongly Positive'. The sentiment function of Textblob returns two properties, polarity and subjectivity. Polarity is a float number ranging from [-1,1]. Table I shows the classification criteria according to the polarity.

TABLE I
TEXTBLOB SENTIMENT CLASSIFICATION CRITERIA

Polarity Value(P)	Sentiment
-1 $\leq$ P $\leq$ -0.5	Strongly Negative
-0.5 $\leq$ P $\leq$ 0	Negative
0	Neutral
0 $\leq$ P $\leq$ 0.5	Positive
0.5 $\leq P \leq$ 1	Strongly Positive

2) Label Correction: Textblob works as lexicon-based approach to predict the sentiment from a bag of words. This python library labels the data based on polarity and subjectivity. But in our real time data form twitter, Textblob misplaces some informal languages or word that are unknown to Textblob. This kind of tweets can't find the real sentiment through Textblob. Besides the new terms (new cases, PCR tests,corona positive/negative) that are related with COVID-19 are not familiar to Textblob. To make this labelling right, we introduce automatic labelling. It works based on regular expression that we create from dataset. We have made the word library and give them the right sentiment. For the time being, we ignore the "Neutral" sentiment from our dataset as it doesn't show any sentiment and causes trouble in correctly labelling the data. Therefore, our revised labelled dataset contains 4 classes instead of 5 namely as "Strongly Negative", "Negative", "Positive" and "Strongly Positive". Table II shows some example of labling data.

Regular expression library is used to match the sentiments related to covid-19 with the help of identifier words collected from our dataset. We use these identifiers(shown in Table II) to correctly label the public sentiments. We have used regular expression library (re) and used the findall function (re.findall) to search the specific group of identifiers to to categorize the sentiment correctly for Covid situation. For example, " My family is tested Covid positive ", is labelled " Negative" as it means bad news in Covid-19 situation.

Example 1: To examine a tweet:

- *Tweet: My cousin got Covid positive yesterday.*
- *TextBlob labeling: Positive.*

TABLE II
SENTIMENT CLASSIFICATION

Sentiment Type	Sentiment Identifiers
Strongly Negative	Fake news, new hotspot, Joe Biden syndrome, horrible situation, growing threats, threatened, hate this pandemic, ridicule, worst virus, Oxygen Crisis
Negative	Covid positive, Closing border, community spread, forced quarantine, ban dog meat, politicization of coronavirus can't afford, high risk countries, complete chaos, coronavirus infection, terrified, total coward, closes churches, new protests trump bashing, panic spreads, homelessness, bigger crisis, touching face, not using sanitizer, coronavirus spreading, shameful comments, sad place, coronavirus vs humanity, coronavirus mess, publishing false stories
Positive	Covid negative, washing hands, won't get coronavirus, donating fund, shop closed over corona concern, preparing hospitals, hygiene practice, prevent coronavirus, donating, closing borders, well being of people, quarantined, wear surgical mask, self isolation, avoid physical contact, avoid large crowds, Response to coronavirus, response to coronavirus, testing kits, productive meeting, protective equipment
Strongly positive	like a boss, recovered from COVID-19, stand up, love it, What a great idea, stay compassionate, stay safe, good friends, good news, vaccine tested, vaccinated stand up

- It's very clear that this tweet holds a positive sentiment for TextBlob because of the word 'positive' in the sentence. As it's a new term regarding to COVID-19 that are not known to Textblob, So we have corrected it as a negative sentiment in our word library and created the regular expression. From now every tweet that has the word 'positive' in it will return a negative sentiment. In the dataset, it is visible that Positive sentiment is greater than negative sentiment. But there is a substantial percentage of negative sentiment in public as COVID-19 has made people concerned, sad and worried.

D. Feature Extraction

As our machine can't deal with text data we need to transform the text into numerical data. For vectorization of text we have used TF-IDF vectorizer. The TF-IDF(Term Frequency-Inverse Document Frequency) implies the importance of a word in a dataset. It creates a TF-IDF matrix that represents the occurrence of a distinct word in our dataset. Bag of words method is used here for simpliied representation of words[16].

IV. TRAINING AND CLASSIFICATION

For training and classifying the dataset in python language, we have used Logistic Regression, Multinomial Naive Bayes and Long short-term memory (LSTM). These are some methods of supervised learning for sentiment analysis. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new input data. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen inputs. We compared with another COVID-19 [17] to measure our performance.

A. Ordinal Logistic Regression

Logistic Regression is a supervised discriminative model. The model builds a regression model to predict the probability. Logistic Regression converts output using the logistic sigmoid function to return a probability value. Ordinal Logistic regression uses a cost function namely sigmoid function. The sigmoid function is as follows:

$$\phi(t) = \frac{1}{1 + e^{-t}}$$

But for ordinal logistic regression the probability value looks like: $P(y = j/X_i)$ So the probability equation for Ordinal Logistic Regression will be like:

$$P(y \leq j/Xi) = \phi(\theta_j\text{-}w^{TX_i}) = \frac{1}{1 + \exp\left(w^{TX_i} - \theta_j\right)}$$

where θ *and* w *are two vectors to be calculated from dataset.* The w vector contains the model learned weights and t values are the feature values.

B. Multinomial Naive Bayes

Multinomial Naive Bayes is a learning algorithm frequently used in text classification problems. Naive Bayes is based on Bayes' theorem that calculates the probabilities of the classes for classifications.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

where A and B are two independent classes. As we need 5 classes for sentimental analysis of the covid-19 tweets, we use Multinomial Naive Bayes.

C. Long Short-Term Memory(LSTM)

LSTM is a deep learning architecture that works with sequential data to solve classification problem. A LSTM unit is comprised of LSTM cell, input gate, forget gate and output gate. LSTM can forget the previous values, add new values (new memory) to it's data. Fig. 2 shows the working procedure of LSTM for sentiment detection.

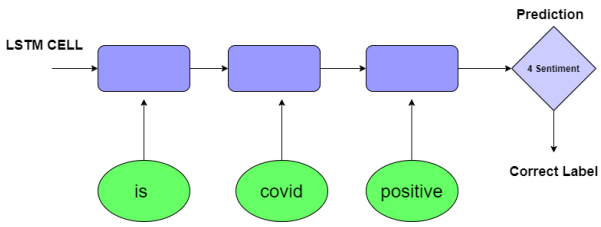

Fig. 2. Working function of LSTM

There are 1,10,000 tweets in our dataset where no tweets having more than 50 words. Therefore the maximum text

lenghth is limited to 50. We have used the LSTM with 128 hidden units with activation function 'softmax' to classify the 4 classes. The model is trained for 100 epochs with batch size 32. To prevent the overfitting, we have set the dropout rate to 0.3 with the learning rate of 0.05.

Example 2:

- *Raw tweet:* RT @ConradGoode: The missing six weeks: how Trump failed the biggest test of his life https://t.co/SAIHSF1fdF
- *Preprocessed data:* miss six week trump fail biggest test life
- Label(Manually) : Strongly Negative.

Example 3:

- *Raw tweet:* In an effort to slow the spread of the coronavirus, Governor Gavin New some announces all public schools in the state will remain closed for the rest of the year. I spoke to one high school senior who was really looking forward to a normal graduation. https://t.co/w2t91TUY1a
- *Preprocessed data:* effort slow spread coronavirus governor gavin newsom announc public school state remain close rest year spoke one high school senior realli look forward normal graduat
- Label : Positive

V. Performance Analysis

This section shows the performance comparison of of our data labelling. The dataset contains 1,10,000 tweets. The dataset is available at [18]. The sentiment distribution after our data labelling is shown in Fig. 3.

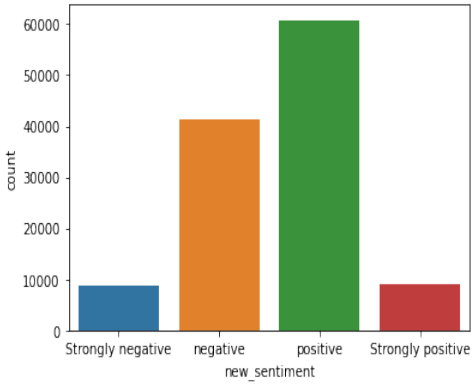

Fig. 3. Graph of distribution of sentiments of Covid-19 twitter data

To measure the correctness of the labelling we compare our work with the dataset [17]. It is a Twitter dataset of Indian users during the COVID-19 lockdown period in India. The dataset is used in sentiment detection research in [7]. There are 3090 cleaned tweets in the dataset on the topics coronavirus, lockdown, etc. Fig. 4 shows the sentiment distribution of the dataset. In the following we call our developed dataset [18] as dataset1 and dataset from [17] as dataset2.

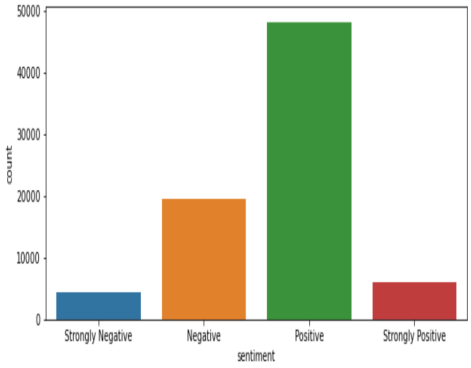

Fig. 4. Sentiment distribution of reference data

Table III shows the accuracy comparison of dataset1 and dataset2. The LSTM shows a better accuracy in comparison with Logistic Regression (LR) and Multinomial Naïve Bayes (MNB). LSTM works better on a large scale dataset. As our dataset is quite large, LSTM gets higher accuracy.

TABLE III
Accuracy Comparison

Method	Dataset1	Dataset2
MNB	79%	75%
LR	92%	90%
LSTM	96%	95%

Table IV and Table V shows the precision, recall and F1 score for comparison of dataset1 and dataset2. The LSTM shows a better performance than MNB and LR. LSTM works better on a large scale dataset. As our dataset is quite large, LSTM gets better performance. Adding to this, LR,MNB and LSTM are used as supervised machine learning methods needed.

In precision, recall and F1 score LSTM outscores both MNB and LR. As our dataset varies the data and not perfectly balanced, LSTM provides higher precision and recall. Even the F1 score varies for three models as our dataset is imbalanced. That means LSTM perfectly handled dataset1 and dataset2.

TABLE IV
Performance measures for Dataset1

Method	**dataset1**		
	Precision	**recall**	**F1 score**
MNB	0.87	0.62	0.68
LR	0.91	0.89	0.90
$LSTM$	0.96	0.96	0.96

VI. Conclusion

The twitter dataset was a mixture of different types of sentiment. We show that standard functions such as textblob of python has some miscalculations for labelling the sentiment. The correction against Textblob's labelling is done giving the

TABLE V
PERFORMANCE MEASURES FOR DATASET2

Method	dataset2		
	Precision	**recall**	**F1 score**
MNB	0.83	0.44	0.44
LR	0.88	0.83	0.85
$LSTM$	0.94	0.94	0.94

tweets correct sentimnet by creating specific regular expression regarding COVID-19. After correcting the label of dataset we have performed the sentiment analysis. We compaered with another standard dataset to measure the correctness of the data labelling. We found the LSTM shows better performance than Logistic Regression and Multinomial Naive Bayes. This work clarifies public opinion on COVID-19 pandemic and can guide authorities to overcome needless anxiety during COVID-19 pandemic.

REFERENCES

[1] R. Singh, R. Singh, and A. Bhatia, "Sentiment analysis using machine learning technique to predict outbreaks and epidemics," *Int. J. Adv. Sci. Res*, vol. 3, no. 2, pp. 19–24, 2018.

[2] S. Chawla, M. Mittal, M. Chawla, and L. Goyal, "Corona virus-sars-cov-2: an insight to another way of natural disaster," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 6, no. 22, p. e2, 2020.

[3] M. M. R. Hinnawi, "The role of social media in lifelong informal learning among members of society," in *2018 JCCO Joint International Conference on ICT in Education and Training, International Conference on Computing in Arabic, and International Conference on Geocomputing (JCCO: TICET-ICCA-GECO)*. IEEE, 2018, pp. 1–13.

[4] V. Kagan and V. S. Subrahmanian, "Understanding multi-stage, multi-modal, multimedia events in social media," in *2018 International Workshop on Social Sensing (SocialSens)*. IEEE, 2018, pp. 4–4.

[5] S. Vanaja and M. Belwal, "Aspect-level sentiment analysis on e-commerce data," in *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2018, pp. 1275–1279.

[6] A. J. Nair, G. Veena, and A. Vinayak, "Comparative study of twitter sentiment on covid-19 tweets," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2021, pp. 1773–1778.

[7] N. Chintalapudi, G. Battineni, and F. Amenta, "Sentimental analysis of covid-19 tweets using deep learning models," *Infectious Disease Reports*, vol. 13, no. 2, pp. 329–339, 2021.

[8] K. H. Manguri, R. N. Ramadhan, and P. R. M. Amin, "Twitter sentiment analysis on worldwide covid-19 outbreaks," *Kurdistan Journal of Applied Research*, pp. 54–65, 2020.

[9] 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*, 2021.

[10] A. Radaideh, F. Dweiri, and M. Obaidat, "A novel approach to predict the real time sentimental analysis by naive bayes & rnn algorithm during the covid pandemic in uae," in *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*. IEEE, 2020, pp. 1–5.

[11] H. Kaur, S. U. Ahsaan, B. Alankar, and V. Chang, "A proposed sentiment analysis deep learning algorithm for analyzing covid-19 tweets," *Information Systems Frontiers*, pp. 1–13, 2021.

[12] J. Samuel, G. Ali, M. Rahman, E. Esawi, Y. Samuel *et al.*, "Covid-19 public sentiment insights and machine learning for tweets classification," *Information*, vol. 11, no. 6, p. 314, 2020.

[13] K. Chakraborty, S. Bhattacharyya, and R. Bag, "A survey of sentiment analysis from social media data," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 450–464, 2020.

[14] Q. C. R. Institute. Crisisnlp. [Online]. Available: https://crisisnlp.qcri.org/

[15] M. Imran, P. Mitra, and C. Castillo, "Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA), may 2016.

[16] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, 2010.

[17] G. Preda. [Online]. Available: https://github.com/gabrielpreda/CoViD-19-tweets

[18] [Online]. Available: https://www.kaggle.com/naimuljoy/final-labeled-110k