JOURNAL OF AERONAUTICAL MATERIALS
ISSN: 1005-5053
Vol. 43, Issue-01, 2023
pp. 906-912

COVID-19 DATA TWITTER SENTIMENT ANALYSIS USING SUPERVISED DEEP LEARNING TECHNIQUES

Mekala Susmitha
Research Scholar, Department of Computer Science, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India
183030081@kluniversity.in

Dr Shaik Razia
Associate Professor, Department of Computer Science, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India

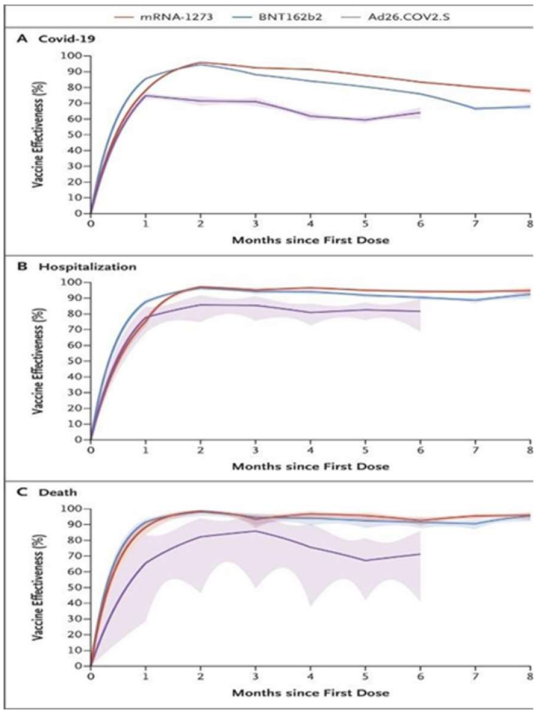
ABSTRACT
Nowadays, social media is a great source and a very common platform for each and every one to interact and massive amount of data is collected. Different opinions and experiences are being shared on various platforms. With, an unexpected outbreak has been occurred which affected people worldwide. This paper mainly focuses on analyzing sentiments (sentiment analysis) of COVID-19 using twitter data. This article provides how different people react and express their various opinions regarding on covid like how it affected them and various effects of pandemic. Word embedding is applied, and comparison of accuracy will be done between applied algorithms. They give out particular and effective information in form of tweets. Understanding all these public thoughts and their tweets posted helps the various health agencies and volunteers.
Keywords: COVID19, sentiment analysis, tweets

1. INTRODUCTION
COVID-19, a proceeding disease which emerged in China. It is believed that it is created, and the strain is first identified in bats which was scientifically proven. Some cases are reported from seafood and animal market. This transmission from animal to human lead to a cause of disease. This disease is more epidemic when person manifestation is at summit. According to one survey, roughly 16 percent of 1,099 patients in China with verified symptoms turned alarming [1].

SARS-CoV-2 is a virus that has caused 453 million illnesses and 6 million deaths worldwide. Alpha,Beta, Gamma, Delta, and Omicron are the five varieties that the World Health Organization (WHO) is currently tracking. According to new research, the first substantial proof of omicron and delta recombinant virus has been discovered, and the World Health Organization has stated that it would be monitoring and addressing the situation. The backbone of the Delta variation is combined with the spike protein of the Omicron variant in the hybrid coronavirus, according to the researchers. The process of genetic recombination occurs when two variations contaminate the same host cell. The World Health Organization (WHO) recently said that it "highly advocates broad access to existing COVID-19 vaccines for booster doses," citing that vaccines and boosters guard against dreadful disease in the face of

Omicron [2]. All countries are adopting various measures to combat Covid-19, which has taken on many forms and is wreaking havoc on the situation. To avoid a pandemic, governments began to impose bans on a variety of societal restrictions.

The main aim was to analyze how psychologically covid is affecting people for government. There is even a large impact of vaccine when vaccine for covid-19 is introduced, and its effectiveness increased to a larger curve.



This graph describes the first dose (the effectiveness) within few months. Specifically with twitter, it is the most powerful and popular platform which includes various opinions and decisions of different individuals.

So, analyzing this continued generated data is informative and helps various health organizations. As Twitter is a mini blogging social media platform which contains 217 million monetizable daily active users. The age group mostly between 30 to 60 of users who post tweets with short messages or quotes. Our research paper main aim is sentiment analysis which well-studied and applied on Twitter data of Covid19 tweets and dividing those tweets into positive, negative, or neutral by applying various techniques [4].

Sentiment Analysis helps to understand the emotions of different people which they are expressed. Hence

tweets are all collected from twitter data and sentiment analysis is applied using machine and deep learning algorithms. Different algorithms are applied like Logistic Regression, Naïve Bayes, CNN and RNN which are machine learning and deep learning algorithms, and comparison of accuracy will be verified.

CNN- A Convolutional Neural Organize (CNN) may be a Profound Learning calculation that can take in an input picture and is generally utilized in picture examination errands such as picture distinguishing proof, protest discovery and division.

RNN - Recurrent neural networks (RNN) are the foremost progressed calculation for sequential information and the primary to hold its input much obliged to an inside memory, making them perfect for machine learning challenges including successive information.

2. RELATED WORK

Four different algorithms are used for classification on covid1.csv dataset which includes opinions on
covid tweets. Categorization of people's viewpoints into positive, negative and neutral was made on 1296 data points.
An amalgam approaches is developed for sentimental analysis that makes use of machine learning algorithms like Naïve Bayes and Logistic Regression and Deep Learning Algorithms like CNN and RNN. This paper reveals several opinions applied to Twitter data and its results. We described various methods of sentiment analysis on Twitter, such as machine learning, ensemble approach, and lexicon (dictionary) based approach. Twitter sentiment analysis and hybrid sentiment analysis methods based on the ensemble method were observed. Compare emotions to see which algorithm works much better.

A stochastic machine learning method is naive Bayes. This approach of text classification is straightforward but extremely effective. Tweets are recorded for this survey's test data. The categorisation of documents happens in two steps. The initial stage is to train the category-specific data. On the other hand, the process of classifying data into unknown categories is the second stage. The general formula of the Bayes' theorem is as follows:

$$P(H \mid X) = P(X \mid H)\, XP(H)\, P(X)$$ Where:

1. P (H | X)-The final probability that the hypothesis H will come true is when the evidence E is presented.
2. P (X | H)-The likelihood that evidence E will materialise has an impact on hypothesis H.
3. P (H)—The probability hypothesis H from before occurs without regard to the data.
4.P (X)-A previous probability proof of E, independent of a hypothesis or other proof. The two variables used to use Bayesian theory are the hypothetical aspect / characteristic (H) and the evidence emotion (E) [5].

Assumption Examination Twitter Information utilizing Calculated Relapse could be a web-based application which takes tweets as an input and gives opinion esteem as a yield. It is viable on text information and the basic calculation is additionally simple to get it.

Murthy et al. [6] first preprocess the tweets. Then They take away phrases that aren't found in dictionaries, which include slang terms, however they keep directly to factors precise to Twitter, like emoticons, hashtags, and usernames, additionally accurate misspelled phrases. For classifying the sentiment, they use a listing of emoticons which have been manually elucidate for fine or terrible sentiment, and that they look for those emoticons withinside the enter tweet.

Word embedding is used random initialization and supervised getting to know, RNN for the long-time period dependencies, CNN with numerous filters in place[8]. Apart from acquiring a version with better accuracy in comparison to different base models, the authors have been additionally worried approximately maintaining the computational value in affordable limits, the cause for combining CNNs with RNNs.

This allows to achieve excessive accuracy among the models. We diagnosed exclusive processes on studying sentiment primarily based totally on twitter data. To boom the performance and carry out higher on ML and Deep getting to know methods.

3. METHOLOGY

This research paper explains about different tweets tweeted in twitter platform about covid19 experiences on which sentimental analysis are applied and dataset is taken from Kaggle which consists of item id, sentiments, and text. Firstly, we found different emotions used in dataset then stop words are removed and we preprocess the tweets and, we get the tweets in form of positive, negative, or neutral.

From the representation of tweets, we extract useful features. Using Logistic Regression and Naïve bayes, we applied semantic analysis and for deep learning methods we apply word embedding. As there is a disadvantage of sematic analysis, as it has an inability to capture multiple meanings of the word, word embedding is used in progress along with techniques CNN and RNN.

Logistic Regression is a statistical technique for predicting the outcome of a Logistic regression is a good approach for classification problems. The logistic function, also known as the sigmoid function, was designed by statisticians to represent population increase, with the population rapidly increasing until it meets the environment's carrying capacity.

Multinomial Bayes, Naive A classifier may classify distinct things using specific properties of the object using a machine learning technique. In natural language processing (NLP), the nave Bayes algorithm is commonly used. It is based on the probability idea. The Bayes theorem is the basic foundation of the naive Bayes classifier. It is predicated on the idea that words are conditionally independent.

Collecting data

Before and after the presidential election, data was collected across two periods of time. We crawled the tweet five times in the months leading up to the presidential election, in January, February, March, and April. The data was crawled the day following the different experiences of covid of different people.

Data Labeling

Manual data labelling is expensive and time-consuming. To speed up the labelling process, we can use pseudo-labelling to label our data. A semi-supervised labelling technique called pseudo-labeling still needs labelled data. The development of models for pseudo-labelling employs machine learning techniques.

Pre-processing

On our dataset, we used certain pre-processing techniques. We removed ASCII characters, usernames, hashtags, URLs, and retweets from the Twitter data, as well as punctuations and duplicates.

Tweet duplication, excessive space, stop words, and characters in a word In addition, we put on a show. case folding, word standardization, and stemming are examples of other pre-processing activities.

Variations in Sentiment Analysis Models

We developed our sentiment analysis model in this study by running multiple experiments. variants. We used three typical machine learning techniques and deep learning techniques to create the first model variations, such as Logistic Regression, and Naive Bayes are some examples. We also compared the results of the accuracy test by not using TFIDF on those three algorithms and instead using Term Frequency- Inversed Document Frequency (TF-IDF) as a feature.

Convolutional Neural Network (CNN) and RNN were two deep neural network techniques used in the construction of the second iterations of the sentiment analysis models.

4. RESULT

We collected tweets on different experiences of people and implemented machine learning techniques which are logistic regression and Naïve Bayes and also implemented deep learning techniques which are CNN and RNN which includes glove and keras embeddings and we obtained accuracies of individual algorithms and classified tweets into positive, negative and neutral.

LOGISTIC REGRESSION AND NAÏVE BAYES

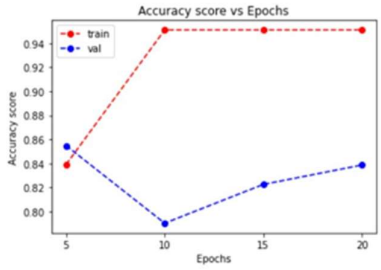
```
=== logistic regression ===
scores =  [0.90789474 0.93877551 0.95172414 0.93333333 0.94520548 0.90322581
 0.92105263 0.93150685 0.90789474 0.90666667]
mean =  0.9247279888617012
variance =  0.0002834237649933908
score on the learning data (accuracy) =  0.9007717750826902

=== bernoulliNB ===
scores =  [0.91729323 0.87692308 0.96350365 0.9352518  0.93233083 0.95035461
 0.90909091 0.95652174 0.91780822 0.90909091]
mean =  0.9268168971689054
variance =  0.0006173541411051059
score on the learning data (accuracy) =  0.9206174200661521

=== multinomialNB ===
scores =  [0.92086331 0.91970803 0.95714286 0.96402878 0.94366197 0.94520548
 0.91275168 0.95035461 0.92517007 0.91666667]
mean =  0.9355553446429218
variance =  0.0003106254457095086
score on the learning data (accuracy) =  0.9239250275633958
```

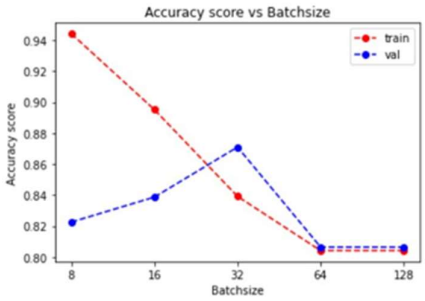
CNN

```
2/2 [==============================] - 0s 5ms/step - loss: 0.4206 - accurac
2/2 [==============================] - 0s 10ms/step - loss: 0.4726 - accura
2/2 [==============================] - 0s 11ms/step - loss: 0.5937 - accura
2/2 [==============================] - 0s 5ms/step - loss: 0.4298 - accurac
best activation function is  selu
2/2 [==============================] - 0s 7ms/step - loss: 0.4373 - accurac
2/2 [==============================] - 0s 6ms/step - loss: 0.5482 - accurac
2/2 [==============================] - 0s 6ms/step - loss: 0.4813 - accurac
2/2 [==============================] - 0s 5ms/step - loss: 1.3109 - accurac
best optimizer is  adam
5/5 [==============================] - 0s 5ms/step - loss: 0.2872 - accurac
2/2 [==============================] - 0s 5ms/step - loss: 0.4311 - accurac
5/5 [==============================] - 0s 4ms/step - loss: 0.1276 - accurac
2/2 [==============================] - 0s 5ms/step - loss: 0.6446 - accurac
5/5 [==============================] - 0s 4ms/step - loss: 0.1152 - accurac
2/2 [==============================] - 0s 7ms/step - loss: 0.6378 - accurac
5/5 [==============================] - 0s 5ms/step - loss: 0.1154 - accurac
2/2 [==============================] - 0s 5ms/step - loss: 0.7073 - accurac
best epoch is  5
5/5 [==============================] - 0s 4ms/step - loss: 0.1512 - accurac
2/2 [==============================] - 0s 6ms/step - loss: 0.4993 - accurac
5/5 [==============================] - 0s 5ms/step - loss: 0.3126 - accurac
2/2 [==============================] - 0s 6ms/step - loss: 0.4597 - accurac
5/5 [==============================] - 0s 5ms/step - loss: 0.3704 - accurac
2/2 [==============================] - 0s 6ms/step - loss: 0.4435 - accurac
5/5 [==============================] - 0s 4ms/step - loss: 0.4601 - accurac
2/2 [==============================] - 0s 7ms/step - loss: 0.4592 - accurac
5/5 [==============================] - 0s 6ms/step - loss: 0.4702 - accurac
2/2 [==============================] - 0s 6ms/step - loss: 0.5395 - accurac
best batchsize is  32
```

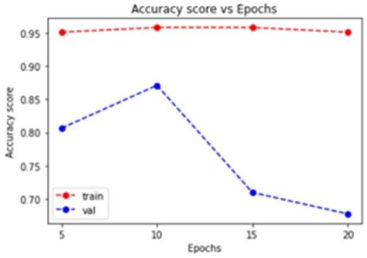
RNN

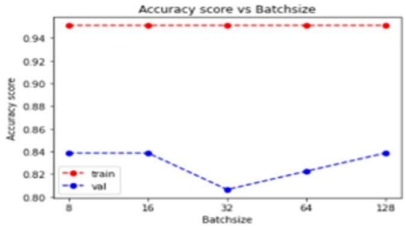
```
test accuracy score =  0.8709677419354839
test f1 score =  0.8413381123058543
time taken is  1.8855109214782715
```






```
best batchsize is  8
test accuracy score =  0.7419354838709677
time taken is  4.360903263092041
```





5. CONCLUSION

This work has been accomplished on primarily based totally on Coronavirus Outbreak using twitter data where the virus proliferated around the whole world. Different sentiments are collected and done research on people opinion and implemented machine learning and deep learning techniques and obtained an accuracy of 91% on average on machine learning techniques and 80% on average of deep learning techniques. This study helps the various government officials to gain more knowledge on people experiences and covid can be reduced by mainly following social distancing which can be a most preferred solution.

6. FUTURE SCOPE

The basic comparisons are done using deep learning techniques, we can use BERT, LSTM etc. to increase accuracy and more fine-grained analysis of opinions can be considered for better semantic analysis of opinions. Many opinions are more specific to positive reviews, so more analysis has to be done in negative reviews

7. REFERENCES

[1] L. Kong et al., "Leveraging multiple features for document sentiment classification," Information Sciences, vol. 518, pp. 39–55, May 2020, doi: 10.1016/j.ins.2020.01.012.

[2] H. Lin et al., "Detecting stress based on social interactions in social networks," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 9, pp. 1820–1833, Sep. 2017, doi: 10.1109/TKDE.2017.2686382.

[3] R. Alfrjani, T. Osman, and G. Cosma, "A hybrid semantic knowledgebase-machine learning approach for opinion mining," Data and Knowledge Engineering, vol. 121, pp. 88–108, May 2019, doi: 10.1016/j.datak.2019.05.002.

[4] A. Abdi, S. M. Shamsuddin, S. Hasan, and J. Piran, "Deep learning-based sentiment classification of evaluative text based on Multifeature fusion," Information Processing and Management, vol. 56, no. 4, pp. 1245–1259, Jul. 2019, doi: 10.1016/j.ipm.2019.02.018.

[5] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," Expert Systems with Applications, vol. 72, pp. 221–230, Apr. 2017, doi: 10.1016/j.eswa.2016.10.065.

[6] A. Da'U and N. Salim, "Sentiment-aware deep recommender system with neural attention networks," IEEE Access, vol. 7, pp. 45472–45484, 2019, doi: 10.1109/ACCESS.2019.2907729.

[7] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding emotions in text using deep learning and big data," Computers in Human Behavior, vol. 93, pp. 309–317, Apr. 2019, doi: 10.1016/j.chb.2018.12.029.

[8] N. Tsapatsoulis and C. Djouvas, "Opinion mining from social media short texts: does collective intelligence beat deep learning?," Frontiers in Robotics and AI, vol. 5, no. JAN, Jan. 2019, doi: 10.3389/frobt.2018.00138.

[9] V. Karthik, D. Nair, and J. Anuradha, "Opinion mining on emojis using deep learning techniques." Procedia Computer Science, vol. 132, pp. 167–73, 2018. doi: 10.1016/j.procs.2018.05.200.

[10] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1–2, pp. 1–135, 2008.

[11] M. Susmitha and R. L. Pranitha, "Performance assessment using supervised machine learning algorithms of opinion mining on social media dataset," in Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems ICACECS, 2022, pp. 419–427, doi: 10.1007/978-981-16-7389-4_41

[12]. Utilization of deep learning and semantic analysis for opinion mining in information extraction: a review in Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), p-ISSN: 2502-4752, e-ISSN: 2502-4760 , April 2023 Vol 30, No 1
Mekala Susmitha, Shaik Razia