Twitter based Data Analysis in Natural Language Processing using a Novel Catboost Recurrent Neural Framework

V. Laxmi Narasamma[1], Dr. M. Sreedevi[2]
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,Vaddeswaram, Guntur, Andhra Pradesh, India-522502

***Abstract*—In recent years, the sentiment analysis using Twitter data is the most prevalent theme in Natural Language Processing (NLP). However, the existing sentiment analysis approaches are having lower performance and accuracy for classification due to the inadequate labeled data and failure to analyze the complex sentences. So, this research develops the novel hybrid machine learning model as Catboost Recurrent Neural Framework (CRNF) with an error pruning mechanism to analyze the Twitter data based on user opinion. Initially, the twitter-based dataset is collected that tweets based on the coronavirus COVID-19 vaccine, which are pre-processed and trained to the system. Furthermore, the proposed CRNF model classifies the sentiments as positive, negative, or neutral. Moreover, the process of sentiment analysis is done through Python and the parameters are calculated. Finally, the attained results in the performance parameters like precision, recall, accuracy and error rate are validated with existing methods.**

***Keywords*—Natural language processing; sentiment analysis; twitter data; Catboost; recurrent neural network**

I. Introduction

In recent, several Artificial Intelligence (AI) [1] techniques are worn in NLP for many purposes like, sentiment analysis, question and answering system and so on [2]. The major reason of using NLP in big data is to reduce the time complexity. Moreover, the big data is applicable in all online application [3], so to handle the big data is the great deal. In this, the sentiment analysis is a key topic to evaluate sentiment values in customer suggestion in online application [4].Thus the sentiment analysis are processed in three levels that are document, sentence and feature level. Here the advanced level of sentiment analysis is feature level [5], which is proved in all research implementation, because it achieved high accuracy than document as well as sentence level [6]. In that, one of the broad social networking sites is twitter, a person uses the twitter for short message communication called tweets [7]. Twitter is defined as the online platform where publics can develop the messages, post, read, and update the text, which is called tweets. Moreover, the sentiment investigation based on the twitter statistics is mentioned as the scientific study of the tweets semantic parts. Subsequently, the sentiment analysis is the method of attaining data from numerous sources that are classified based on the sentiments. Generally, the tweets are reflecting the opinion from public based on the particular data about product, or any topic.

The public opinion is normally categorized into positive, negative and neutral tweets. However, the categorizations of tweets are very difficult for large quantity of data. In addition, if any one continuously following your tweets then your message is liked or attracted by the particular person [8], the twitter analysis is shown in Fig. 1. Even it has lot of facilities, the analyses of data in twitter is challenging task because of large volume of data [9]. This reason turned the interest of researchers towards this area [10]. Thus several researchers found much solution but it is not applicable for long time due to data complexity [11]. Also, in NLP text summarization is one of the schemes to identify the uniqueness of each document [12]. So, in NLP text summarization frame work is elaborated in better way: several machine learning techniques and vector based word embedding models were studied for the better classification [13, 30]. But for the complicated data these approaches are misbehave because the error removing model is not available in all machine learning model [14].

So, it failed to prune the error, this cause the difficulties to specify the sentiment value of data. The computer does not know the people language [15], so to make the human machine interaction machine learning is the advanced topic. In addition, the data can train the system in the form of 0 or 1 [16]. Because the machine only knows the binary value 0's and 1's, so the classification of sentiment value is in the form of decision making [17]. The sentiment analysis using the large quantity of data is done through the machine learning approaches [18]. Several machine learning approaches are found but still the issues are not end [19]. Thus, the present research work aimed to develop an efficient machine learning model to classify tweets data based on their sentiment values.

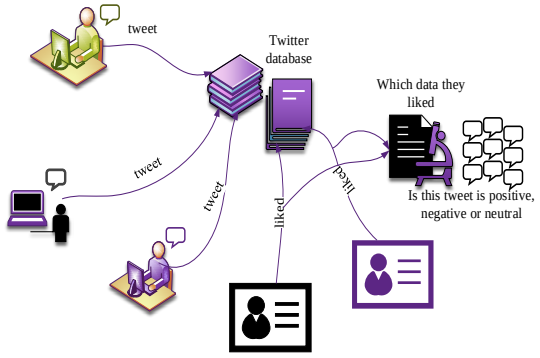

Fig. 1. Twitter Data Analysis.

This research work is organised as follows. The recent literature works based on sentiment analysis using twitter data is detailed in section 2. Also, the system model and problem statement is mentioned in section 3. Moreover, the developed methodology is elaborated in section 4 and the attained outcome of the proposed work is declared in section 5. Thus, section 6 detailed the end of the entire research.

II. Related Works

Several literature works correlated to the twitter data analysis is summarized below and detailed in Table I.

Ruz et al [21] introduced the bayes aspect manner to produce high real network. When comparing to the random forests and vector support in machine it gives competitive sentiment prediction result. However, approach cannot able to differentiate it behave in Spanish or English in RF and SVM. Finally conclude that, this network also allow to determine relation in the words, historically it gives interested quality data and catch socially the main headline of the dynamic act, accuracy in result and also reduce the exposure of misinformation.

Prediction of visiting next location using machine leaning by utilizing twitter information developed ensemble classified approach (ESA) has proposed by Kumar et al [22]. Moreover, this proposed work is utilized the twitter data for predicting the next visiting location of the user. Also, the developed ESA model attained the outcome of the prediction based on various classification models. For this prediction model, the voting technique is adopted to enhance the accurate sentiment calculation. This approach predicts accurate result with high desirable but it lack in security.

In recent, to assign a text for an emotion in classes automatically based on soft classified approach Hasan et al [23] developed a learning framework. That includes two tasks i.e., online and offline task. The result shows that the 90% of correct emotion of text can be created for real time. Finally, it gives best performance comparing to other approaches and also it doesn't depend on other system. However, it attained high error rate.

TABLE I. Summary of Literature Survey

Author and year	Technique	Merits	Demerits
Ruz et al [21], 2020	Bayes aspect	Better information recall	It takes more time
Kumar et al [22], 2019	Ensemble classification approach	Accurate prediction	It attained Less privacy rate.
Hasan et al [23], 2019	Soft classification approach	Maximum Probability	High error rate
Barker, J. L. P., and Christopher JA Macleod [24], 2019	Prototype social geodata	Awareness during flood	Some time it lack in signal to predict the rain fall rate
Li al [25], 2019	patent analysis and twitter data mining	It required less time to process the mechanism	More complex

Barker, J. L. P., and Christopher JA Macleod [24] created prototype based social geo data from twitter to make the people aware from flood or huge disaster. Also, the decision mechanism is used to specify the rain fall and flood based data. Also, this model establishes the sentiment analysis using the twitter data based on pipeline extract tweets that involves 420000 tweets. Moreover, this supports a people lot to get aware about flooding.

Analysing the data is important in big data, Li et al [25] proposed patent analysis to determine the trends change in perovskite solar tech, and to identify response, expectation and sense is monitor using information from twitter mining. Finally, the comparison is made to identify the development of trends how the twitter users interested, this offer better in understanding and also it helps to find the development of trends in future but it may weak in capturing signals.

The key metrics of the proposed model is mentioned as follows,

- Initially, the twitter data based on the user opinion about COVID-19 is collected and trained to the system.
- Moreover, the novel CatBoost Recurrent Neural Framework (CRNF) is developed for analysing the sentiment value of twitter data.
- Subsequently, the developed CRNF model is utilized to remove the error while removing the leaf node layer that has increased the classification rate.
- Thus, the proposed approach effectively classifies the sentiments as positive, negative, or neutral.
- Additionally, the performance metrics such as recall, F-measure, accuracy, precision, and error rate are calculated and validated using existing approaches.

III. System Model and Problem Defination

Normally, the sentiment analysis or data analysis in natural language processing is done over big data dataset such as Facebook, twitter, etc. Moreover, sentiment analysis for large volume of data is some more difficult as because of its complexity and part of speech classification [20]. In addition, the sentence which contains positive words may also end with negative sentence. Thus the opinion or sentiment classification is one of the important tasks in NLP, which is mostly helpful for online service because the success of online business is based up on the customer review. Moreover, the process of sentiment analysis using twitter data is explained in Fig. 2.

Also to predict the uniqueness of each sentence, thus the classification of sentiment measure is more important. This motivate this research to find the scientific solution to enhance twitter data analytics using sentiment analysis in Natural Language Processing to reduce all kinds of issues

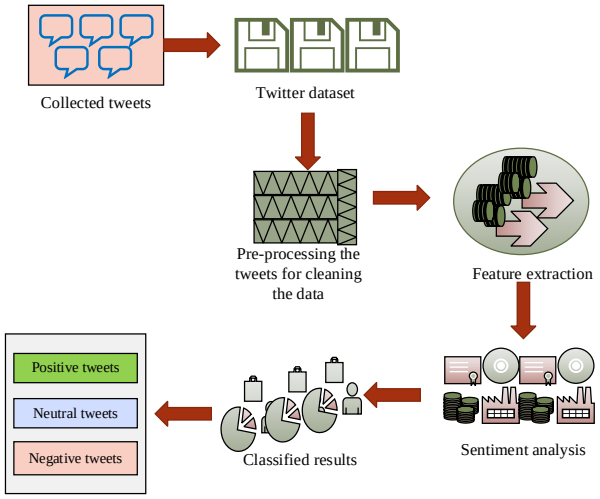

Fig. 2. System Model for Sentiment Analysis using Twitter Data.

IV. Proposed CRNF Methodology

In general, the sentiment analysis is one of the predictive modeling tasks that are trained with sentiments or textual data. However, the sentiment analysis using large data is the difficult task that provided lower efficiency for classifying the sentiments. In this research, the novel CatBoost Recurrent Neural Framework (CRNF) is developed for analysing the sentiments using twitter data. Here, the tweets are collected based on the COVID-19 vaccine from twitter that is utilized for training.

Moreover, the procedure of the developed CRNF approach is detailed in Fig. 3. Also, the developed CRNF approach is pre-processed the trained dataset and finally classifies the sentiment while removing the error. Thus, the proposed CRNF model is to remove the leaf node layer using pre-processing function and to enhance sentiment classification rate.

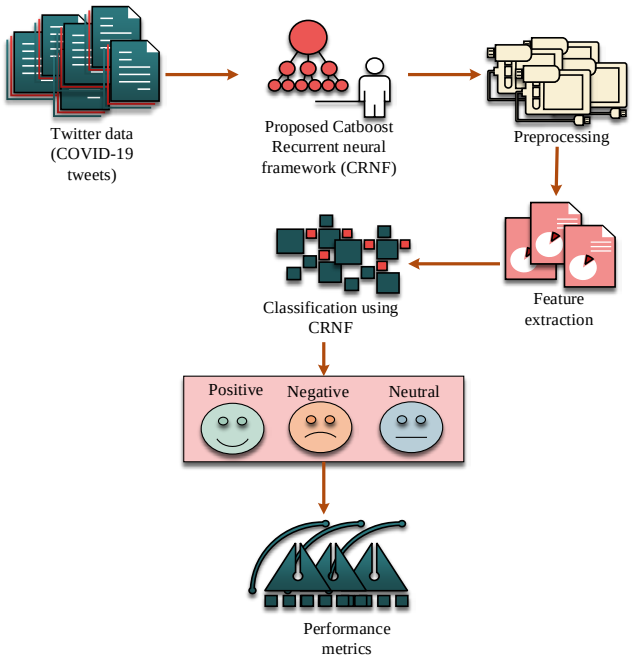

Fig. 3. Proposed CRNF Methodology.

A. Dataset Description

The proposed approach utilized the Twitter data for analysing the sentiments. Here, the twitter data based on COVID-19 vaccine tweets are collected from the kaggle.com that is processed in this research. The utilized dataset details about the 38460 numbers of tweets that involves the user name, location, description, friends, followers, favourites, text, tweet date, hashtags, and so on. In this, the collected tweets are based on the category of positive reviews, negative reviews and neutral reviews. Moreover, the collected dataset is given to the developed CRNF model for further processing.

B. CRNF Process for Sentiment Analysis

The proposed CRNF approach is processed on the twitter dataset for analyzing sentiments. Here, the developed catboost recurrent neural framework can reduce the error in the dataset, which is utilized for enhancing the classification accuracy [26, 27]. Here, the developed model performs the functions are pre-processing, feature extraction, and classification. The proposed CRNF model is the neural architecture that can reduce the training error, which is utilized to classify the sentiments in an effective manner. Furthermore the occurrence of CatBoost in the recurrent neural model can attain the enhanced classification accuracy as well as precision rate. Primarily, the dataset is initiated in the input layer of the network that is mentioned in eqn.(1),

$$d_T = \{(P_k)\}, k = 1,2,...,N \quad (1)$$

Where, P_k denotes the N^{th} quantity of tweets in the dataset d_T that involves positive, negative and neutral tweets. In this work, the dataset training process is done in the input layer. Here, P_k is the input and the output of the input layer is h^k that is given to the next layer. Moreover, the attained dataset having several errors or noise that should remove for attaining better results. So, the proposed model performing the pre-processing function.

- Pre-processing

This process is carried on the next layer of the network that is necessary for the dataset to remove the unnecessary data by cleaning the tweets, which involves the functions such as normalization, stop words removal and tokenization. Here, normalization process is utilized to remove the special characters, URLs, and emojis from the dataset. Also, the stop word deletion is processed to split the tweets that are compared using the stop words library, which involves the words not affects the original meaning the utilized sentences. Moreover, the raw is fragmented into the sentences or words with the use of tokenization process, which is employed to understanding the original meaning of text. Thus, the training errors in the tweets $(k = 1,2,3,....,N)$ are mentioned as H^{k-1} using additive manner that is represented by $H^k = H^{k-1} + \alpha h^k$ through the step size α and h^k utility removed the mistakes that is mentioned in eqn.(2),

$$h^k = d_T\left[\arg\min_{h\in H}\left(R^{k-1} + r_k\right)\right] \tag{2}$$

Where, the errors are mentioned as $\left(R^{k-1} + r_k\right)$ the repeated words r_k and the value of h^k is removed the errors and repeated words in a sentence, which are done in the hidden layer.

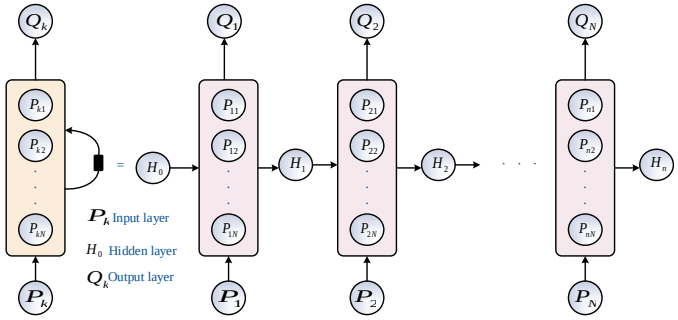

Fig. 4. Process of CRNF Network Model.

The dataset is given to the input layer of the proposed CRNF network that can be processed through the system. Moreover, the pre-processing and feature extraction function are done in the hidden layer of the network. Thus, the errors in the sentences are removed in the pre-processing process that is removed the leaf node layer. Additionally, the classification layer is utilized for classify the sentiments and finally, the output is obtained from the output layer of the network, which are represented in Fig. 4.

- Feature extraction

The pre-processed dataset is processed the feature extraction method that is done using the layer of the CRNF model. Here, the feature extraction is utilized for extracting the features from the dataset. Moreover, these features are utilized to identify the polarity of the sentences. In this approach, the feature extraction process is done using the factor φ_k and the activation function $\tanh$ that calculation is mentioned in eqn.(3),

$$f_k = \tanh\min\sum_{k=1}^{N}\left\{\left(P_k\varphi_k\right)\right\} \tag{3}$$

Thus, the leaf node layer is removed while completing the pre-processing and feature extraction process. Moreover, the attained output is given to the classification layer that is performed the sentiment analysis process.

- Classification

The proposed method classifies the tweets like positive tweets P_t, negative tweets N_t, and neutral tweets N_l. The proposed model CRNF classifies the sentiments using the aspect terms in the sentences. Finally, the classification of sentiments is done using eqn.(4),

$$Q_k = d_T \rightarrow \sum_{k=1}^{N} P_k\left(A_W\left(P_t, N_t\right)\right) \tag{4}$$

Algorithm: CRNF for sentiment analysis

Input: COVID-19 Tweets from Twitter data

Output: classified output $\left(Pt, Nt, Nl\right)$

//where, Pt -positive, Nt -negative, and Nl -neutral

Start

{

Initialization ()

Initialize the dataset d_T // COVID-19 Vaccine Tweets

Import the dataset

Preprocessing()

For all d_T do

Remove error, repeated words, noise, urls, numbers, special characters, stop words

If misspelled (word) then

Replace the word by correct word

End if

End for

Feature extraction()

For all d_T do

Extract the features of the words

End for

Classification()

Mentioned the aspect terms $A_W\left(P_t, N_t\right)$

If $A_W\left(P_t\right) > A_W\left(N_t\right)$ then

$Sentiment \leftarrow P_t$ //(+1) positive tweet

Else if $A_W\left(N_t\right) > A_W\left(P_t\right)$ then

$Sentiment \leftarrow N_t$ //(-1) negative tweet

Otherwise

$Sentiment \leftarrow N_l$ // (0) neutral tweet

End if

Classified sentiments

}

Stop

In this, the positive and negative aspect terms are saved in the layer that words are utilized to analyze the sentences. If the sentence having positive aspects then it is the positive tweet and if the sentence has negative aspects then it is negative tweets otherwise that sentence is considered neutral tweets. Finally, the results are attained from the output layer of the proposed network. Moreover, the positive tweets are represented as (+1), negative tweets are represented as (-1), and neutral tweets are represented as (0).

Thus, the sentiments have classified by the proposed catboost recurrent neural framework. The complete procedure of the proposed CRNF technique is detailed in the algorithm 1 and the flow chart is represented in Fig. 5.

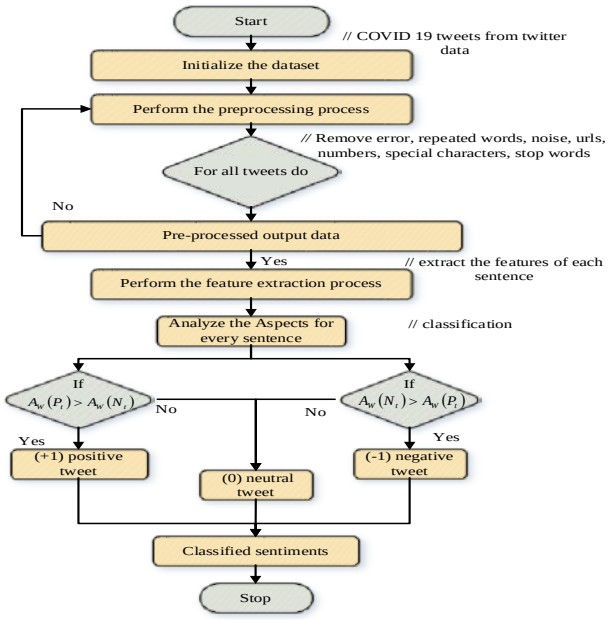

Fig. 5. Flow Chart for CRNF Model.

V. Results and Discussion

In this work, the developed CRNF approach is simulated with the use of Python; moreover, the efficiency of the proposed strategy is evaluated with prevailing manners. Here, the comparison is carried out in the performance metrics like accuracy, recall, F-measure, precision, and error rate. In this, the developed is effectively classifies the sentiments using the proposed CRNF approach.

A. Case Study

In this paper, the sentiment analysis is done using the twitter data. Here, the tweets for COVID-19 are collected and processed using the proposed CRNF model. Several samples are for COVID-19 tweet and the classified results are mentioned in Table II. Here, the utilized dataset is initially pre-processed and feature extracted in the layers of the proposed CRNF network.

Subsequently, the positive and negative aspects are mentioned for classifying the sentiments. For example, vaccine, immunity, protect, etc., are considered as the positive aspects and sick, side effects, death, spread, etc., are considered as the negative aspects. Based on the considered aspects, each sentences are classified as positive tweet (+1), negative tweet (-1), and neutral tweet (0).

B. Performance Matrics

This research work performed the sentiment analysis using the developed CRNF approach, which is implemented using Python. Moreover, the performance metrics are calculated that are compared using existing methods for identifying the efficiency of the developed approach. Thus, the parameters like accuracy, precision, and error rate of the proposed model is validated with prevailing methods like Tree Augmented Naive Bayes (TAN) [21], Bag of words using machine learning (BOW-ML) [28], and Attention using Bidirectional CNN-RNN Deep Method (ABCDM) [29].

TABLE II. Classified Results for Twitter Dataset about COVID-19

S.No	Text about COVID-19	Positive	Negative	Neutral
1	There are presently more than fifty COVID-19 vaccine contenders in the trials.	+1	-	-
2	The developed vaccine may cause the various side effects, which is related to the symptoms and signs of COVID-19.	-	-1	-
3	The Vaccine about COVID-19 is manufactured in Australia, which is supplied to the citizens at no cost. AFP quotes Prime Minister	+1	-	-
4	Got my CovidVaccine today. Ready to end this pandemic Protect your families.	+1	-	-
5	Got my covid vaccine! Tired, mild headache - work those antibodies, immune system	-	-1	-
6	Presently more than 50 numbers of COVID-19 vaccine candidates in trials.	+1	-	-
7	COVID-19 affected people develop mild to moderate disorder and recover without hospitalization.	-	-	0
8	Third stage of Russia's Covid-19 vaccine may initiate in seven to ten days	+1	-	-
9	COVID-19 is easily spread from one person to another like friends, family, and surrounding peoples.	-	-1	-
10	Masks are used to protect the people from COVID-19.	-	-	0

1) Accuracy: validation is utilized for determining the efficiency of the proposed framework. Also, it is identified the effectiveness of the developed model for classifying the sentiments, which is computed using eqn.(5),

$$Acc = \left(\frac{T_P + T_N}{T_P + F_P + F_N + T_N} \right) \quad (5)$$

Where, T_P represents the true positive that is the calculation for the total quantity of properly classified positive tweets, T_N is the true negative that represents the total quantity of properly classified negative tweets, F_P is the false positive that

symbolizes the total quantity of improperly classified positive tweets, and F_N is the false negative that represents the total quantity of improperly classified negative tweets.

The accuracy calculation of the proposed CRNF model is compared with existing methods like TAN, BOW-ML, and ABCDM that are mentioned in Table III. The existing approaches TAN and BOW-ML approaches are attained lower accuracy as 80.8% and 85%.

Also, the ABCDM approach achieved nearly 93% accuracy. Thus, the proposed CRNF approach has attained high accuracy as 99.34% than other models while considering tweets data from twitter, which is represented in Fig. 6.

2) Precision: The calculation of precision is utilized for identifying the effectiveness of the proposed classifier. Here, the lower precision value denotes the high false positives and high precision rate denotes the less number of false positives. Moreover, the precision value of the proposed model is calculated using eqn. (6),

$$P = \left(\frac{T_P}{T_P + F_P} \right) \tag{6}$$

The precision value of the proposed CRNF model is validated with existing methods and the values based on the quantity of tweets are mentioned in Table IV.

TABLE III. COMPARISION OF ACCURACY

No. of tweets taken	Accuracy (%)			
	TAN	BOW-ML	ABCDM	CRNF [proposed]
100	80.8	85.1	93.40	99.34
200	79.96	84.87	92.68	98.65
300	76.48	84.05	91.97	97.89
400	75	83.79	89.63	95.76
500	73.27	83.27	87.67	93.56

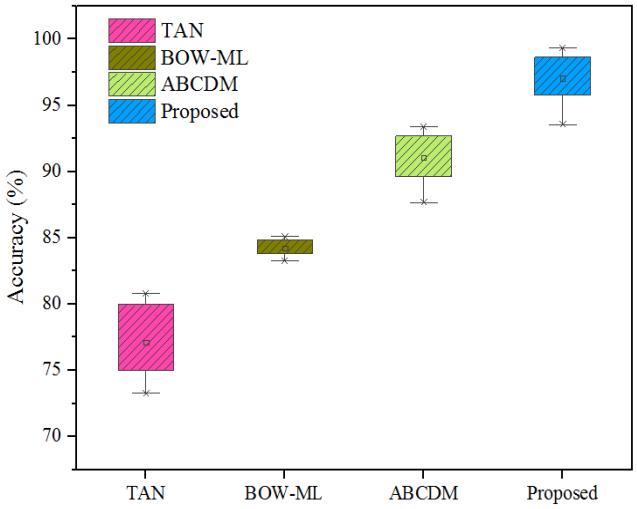

Fig. 6. Comparison of Accuracy.

TABLE IV. COMPARISON OF PRECISION

No. of tweets taken	Precision (%)			
	TAN	BOW-ML	ABCDM	CRNF [proposed]
100	90.6	83.6	95.70	98.38
200	88.04	82.87	93.78	97
300	85.58	80.96	93.27	96.78
400	83.47	79.67	92.09	96.07
500	82.97	77.86	90.83	95.36

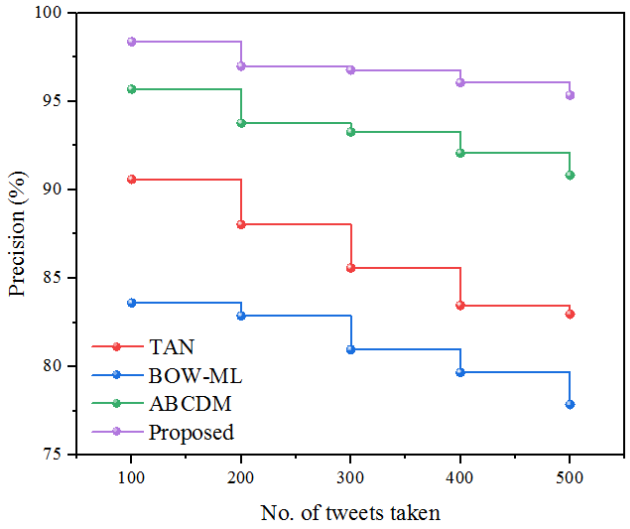

Fig. 7. Comparison of Precision.

Also, the existing BOW-ML model attained lower precision as 83.6%, TAN approach achieved 90.6% precision, and ABCDM model attained 95.7% precision value. Hence, the proposed CRNF model has achieved high precision rate as 98.38% than other methods that is represented in Fig. 7.

3) Recall: The calculation of recall is utilized for identifying the sensitivity or the completeness of the proposed classifier. In this, the lower recall value denotes the high false negatives and high recall rate denotes the less number of false negatives. Moreover, the recall value of the proposed model is calculated using eqn.(8),

$$R = \left(\frac{T_P}{T_P + F_N} \right) \tag{7}$$

The recall value of the proposed CRNF model is validated with existing methods and the values based on the quantity of tweets are mentioned in Table V.

Also, the existing BOW-ML model attained lower recall as 88%, TAN approach achieved 85.4% recall, and ABCDM model attained 90.88% recall value. Therefore, the proposed CRNF model has achieved high recall rate as 97.45% than other methods that is represented in Fig. 8.

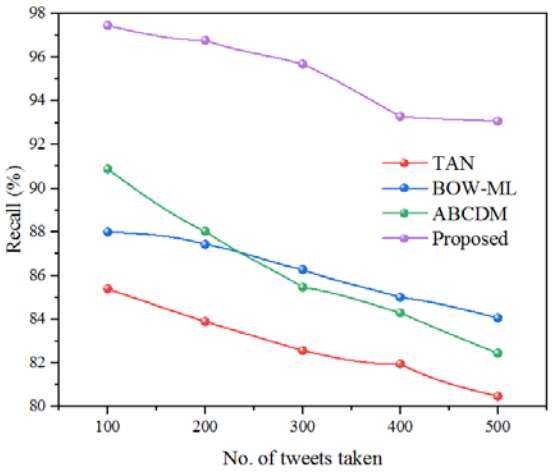

Fig. 8. Comparison of Recall.

TABLE V. COMPARISON OF RECALL

No. of tweets taken	Recall (%)			
	TAN	BOW-ML	ABCDM	CRNF [proposed]
100	85.4	88	90.88	97.45
200	83.9	87.43	88.04	96.78
300	82.57	86.28	85.47	95.69
400	81.96	85	84.30	93.29
500	80.48	84.07	82.45	93.08

4) F1-measure: The calculation of F1-score is defined the combination of the calculated precision and recall values, which is computed using eqn.(8),

$$F1-score=\left(2\frac{P*R}{P+R}\right) \tag{8}$$

The F1-measure value of the proposed CRNF model is validated with existing methods and the values based on the quantity of tweets are mentioned in Table VI.

Also, the existing TAN approach achieved 87.9% F1-measure value, BOW-ML model attained lower F1-measure value as 85.8%, and ABCDM model attained 92.22% lower F1-measure value. Moreover, the proposed CRNF model has achieved high F1-measure value as 97.91% than other prevailing approaches that are characterized in Fig. 9.

5) Error Rate: This calculation is utilized to identify the classification error of the proposed model, which is computed using eqn.(9).

$$Error\_rate=\left(\frac{F_P+F_N}{T_P+T_N+F_P+F_N}\right) \tag{9}$$

The error rate value of the proposed CRNF model is validated with existing methods and the values based on the quantity of tweets are mentioned in Table VII. These prevailing methods are attained higher error rate for classifying sentiments using twitter data.

TABLE VI. COMPARISON OF F1-MEASURE

No. of tweets taken	F1-measure (%)			
	TAN	BOW-ML	ABCDM	CRNF [proposed]
100	87.9	85.8	93.22	97.91
200	86.35	84.08	92.76	96.88
300	87.39	82.78	92.94	96.23
400	85.28	83.49	91.67	94.65
500	84.37	82.08	91.06	94.20

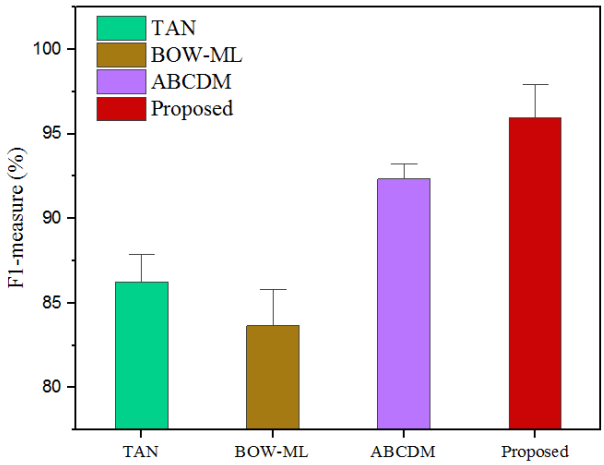

Fig. 9. Comparison of F1-Measure.

TABLE VII. COMPARISON OF ERROR RATE

No. of tweets taken	Error rate (%)			
	TAN	BOW-ML	ABCDM	CRNF [proposed]
100	19.2	14.9	6.6	0.66
200	20.04	15.13	7.32	1.35
300	23.52	15.95	8.03	2.11
400	25	16.21	10.37	4.24
500	26.73	16.73	12.33	6.44

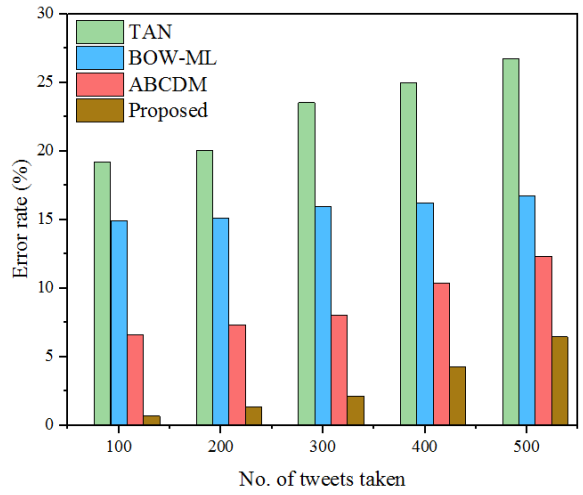

Fig. 10. Comparison of Error Rate.

Also, the existing TAN approach achieved 19.2% high error rate value, BOW-ML model attained 14.9% error rate value, and ABCDM model attained 6.6% error rate value. The comparison of the error rate value is characterized in Fig. 10. Moreover, the proposed CRNF model has achieved lower error rate value as 0.66% than other existing methods for classifying the sentiments using twitter data for COVID 19.

VI. Conclusion

In this research, the novel Catboost Recurrent Neural Framework (CRNF) is developed for performing sentiment analysis in the twitter dataset. Here, the tweets about COVID-19 are considered as the dataset that is utilized for classifying the sentiments as positive tweets, negative tweets, and neutral tweets. The noise, error, url, repeated words, stop words, numbers, special characters are removed by the pre-processing process. Also, the feature extraction method is utilized to extract the characteristics of each sentence. Subsequently, the classification of sentiments is done in the layer of the proposed CRNF model using the aspects words. Hence, the proposed model has achieved high accuracy as 99.34% with lower error rate as 0.66% than other existing approaches.

References

[1] Bigsby, K. G., Ohlmann, J. W., & Zhao, K. (2019). The turf is always greener: Predicting decommitments in college football recruiting using Twitter data. Decision Support Systems, 116, 1-12.

[2] D. Kandé, F. Camara, S. Ndiaye. "FWLSA-score: French and Wolof Lexicon-based for Sentiment Analysis", In 2019 5th International Conference on Information Management (ICIM), IEEE, 2019.

[3] Kursuncu, Ugur, et al. "Predictive analysis on Twitter: Techniques and applications." Emerging research challenges and opportunities in computational social network analysis and mining. Springer, Cham, 2019. 67-104.

[4] Liu, Xia. "Analyzing the impact of user-generated content on B2B Firms' stock performance: Big data analysis with machine learning methods." Industrial marketing management 86 (2020): 30-39.

[5] Kumar, Sachin, and Mikhail Zymbler. "A machine learning approach to analyze customer satisfaction from airline tweets." Journal of Big Data 6.1 (2019): 62.

[6] Mandloi, Lokesh, and Ruchi Patel. "Twitter Sentiments Analysis Using Machine Learninig Methods." 2020 International Conference for Emerging Technology (INCET). IEEE, 2020.

[7] Ahmed, Hager, et al. "Heart disease identification from patients' social posts, machine learning solution on Spark." Future Generation Computer Systems 111 (2020): 714-722.

[8] Tahmasebi, Hossein, Reza Ravanmehr, and Rezvan Mohamadrezaei. "Social movie recommender system based on deep autoencoder network using Twitter data." Neural Computing and Applications (2020): 1-17.

[9] Hasan, Mahmud, Mehmet A. Orgun, and Rolf Schwitter. "Real-time event detection from the Twitter data stream using the TwitterNews+ Framework." Information Processing & Management 56.3 (2019): 1146-1165.

[10] Gupta, Aakansha, and Rahul Katarya. "Social Media based Surveillance Systems for Healthcare using Machine Learning: A Systematic Review." Journal of Biomedical Informatics (2020): 103500.

[11] K. Sivakumar, N.S. Nithya, O. Revathy. "Phenotype Algorithm based Big Data Analytics for Cancer Diagnose", Journal of medical systems, 43(8), pp. 264, 2019.

[12] K. Weiying, D.N. Pham, Y. Eftekharypour. "Benchmarking NLP Toolkits for Enterprise Application", Pacific Rim International Conference on Artificial Intelligence, Springer, Cham, 2019.

[13] D. Benarji Tharini, V.V. Bulusu. "Development of a Micro Telugu Opinion WordNet and Aligning with TELOWN Ontology for Automatic Recognition of Opinion Words from Telugu Documents".

[14] M. Trupthi, S. Pabboju, N. Gugulotu. "Deep Sentiments Extraction for Consumer Products Using NLP-Based Technique", Soft Computing and Signal Processing, Springer, Singapore, pp. 191-201, 2019.

[15] B.A. Hammou, A.A. Lahcen, S. Mouline. "A Distributed Ensemble of Deep Convolutional Neural Networks with Random Forest for Big Data Sentiment Analysis", International Conference on Mobile, Secure, and Programmable Networking, Springer, Cham, 2019.

[16] K. Negi, A. Pavuri, L. Patel, C. Jain. "A novel method for drug-adverse event extraction using machine learning", Informatics in Medicine Unlocked, pp. 100190, 2019.

[17] H. Yang, L. Luo, L.P. Chueng, D. Ling, F. Chin. "Deep Learning and Its Applications to Natural Language Processing", Deep Learning: Fundamentals, Theory and Applications, Springer, Cham, pp. 89-109, 2019.

[18] J. von Bloh, T. Broekel, B. Özgun, R. Sternberg. "New (s) data for entrepreneurship research? An innovative approach to use big data on media coverage", Small Business Economics, pp. 1-22, 2019.

[19] Yang, Chao, et al. "Aspect-based sentiment analysis with alternating coattention networks." Information Processing & Management 56.3 (2019): 463-478.

[20] Chandra, Nidhi, Sunil Kumar Khatri, and Subhranil Som. "Natural Language Processing Approach to Identify Analogous Data in Offline Data Repository." System Performance and Management Analytics. Springer, Singapore, 2019. 65-76.

[21] Ruz, Gonzalo A., Pablo A. Henríquez, and Aldo Mascareño. "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers." Future Generation Computer Systems 106 (2020): 92-104.

[22] Kumar, Sachin, and Marina I. Nezhurina. "An ensemble classification approach for prediction of user's next location based on Twitter data." Journal of Ambient Intelligence and Humanized Computing 10.11 (2019): 4503-4513.

[23] Hasan, Maryam, Elke Rundensteiner, and Emmanuel Agu. "Automatic emotion detection in text streams by analyzing twitter data." International Journal of Data Science and Analytics 7.1 (2019): 35-51.

[24] Barker, J. L. P., and Christopher JA Macleod. "Development of a national-scale real-time Twitter data mining pipeline for social geodata on the potential impacts of flooding on communities." Environmental modelling & software 115 (2019): 213-227.

[25] Li, Xin, et al. "Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology." Technological Forecasting and Social Change 146 (2019): 687-705.

[26] Huang, Guomin, et al. "Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions." Journal of Hydrology 574 (2019): 1029-1041.

[27] Liu, Fagui, et al. "Combining attention-based bidirectional gated recurrent neural network and two-dimensional convolutional neural network for document-level sentiment classification."Neurocomputing 371 (2020): 39-50.

[28] Soumya, S., and K. V. Pramod. "Sentiment analysis of malayalam tweets using machine learning techniques." ICT Express (2020).

[29] Basiri, Mohammad Ehsan, et al. "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis." Future Generation Computer Systems 115 (2020): 279-294.

[30] V.Laxmi Narasamma, and M. Sreedevi. "Modeling of Tweet Summarization Systems using Data Mining Techniques: A Review Report." Indian Journal of Science and Technology 9 (2016): 44.