SENTIMENT ANALYSIS OF REVIEW TWEETS USING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

[1]A VISHNUVARDHAN REDDY, [2]G VENKATA RAMANA REDDY, [3]M.MADHUSUDHAN REDDY, [4]S.SRIHARI REDDY

[1,2]Assistant Professor, Department of Computer Science and Engineering, G Pulla Reddy Engineering College (Autonomous): Kurnool, Andhra Pradesh, India

[3]Assistant Professor, Department of Electronics and Communication Engineering, G Pulla Reddy Engineering College (Autonomous): Kurnool, Andhra Pradesh, India

[4]Assistant professor, Department of Computer Science and Engineering, Brindavan Institute of Technology & Science, Pedda Tekur, Andhra Pradesh, India

ABSTRACT: The new era is trending with micro blogging websites such as Twitter and Facebook, these social networking sites are used to share the opinions with the internet citizens and to store the data.Twitter is one of the most popularly known blogging site to the society. It is the social networking platform where people share their opinions in the form of tweets. Hence, it is a best sentiment analysis platform. Generally, the opinion of users are categorized into three types, they are good which indicates positive, bad indicates negative, and the process of analyzing differences and variations in the mind of users and grouping this data collected is known as sentiment analysis. This study represents a comparative analysis on various Deep Learning techniques that are helpful in sentiment analysis on twitter data. In this research Deep Learning methods are known for the efficiency in solving various issues. The main aim of this research is to emphasis the data collected from twitter. In current study machine learning based deep learning techniques are utilized such as Convolutional Neural Networks (CNN) which is performed on image processing and Recurrent Neural Networks (RNN) as they are efficiently applied on Natural Language Processing (NLP) tasks. The combination of CNN and a part of RNN technique called Long Short Term Memory (LSTM) networks are used in our research. Large number of testshave been conducted on the data collected from twitter platform. This method obtains best score and high efficiency in terms of performance metrics. From the experiment our proposed CNN-LSTM gives best results and proved with high standard performance in classification of twitter data for Sentiment analysis which gives accuracy in f-measure in contrast to the various contrast techniques.

KEYWORDS: Sentimental analysis, LSTM, deep learning, convolutional neural networks, Twitter data.

I.INTRODUCTION

In this current generation social networking sites have high-time and billions of users being active on number of social networking sites such as twitter and facebook where huge amounts of data is generated regularly. The users share their opinions and feelings on these sites which are designed by the IOT system. Large amounts of data are reserved on Twitter.Current issues in the society are discussed on the networking sites. The political activities and initiatives taken by the government are discussed in the twitter. Each user shares their views regarding the current issues happening in that particular nation. Here, the accounts of famous personalities are followed by the users of twitter and their opinions are shared with twitter followers. Raw data extracted from tweets is the input to our model. This data is processed and categorized into positive and negative tweets. Opinion shared on the twitter may be in relation with various kinds of products, events, people or current political affairs in the nation. Following are the reasons expressed to be best for performance of sentiment analysis on twitter data.

1. Brief expressions are given in Tweets which has fixed length of 280 characters.

2. Sentiment analysis is performed in real-time.

3. Vast data of tweets is available for performing the analysis on raw data.

There is growing demand for social media platforms which gave access for sharing the personal opinions of users on various aspects and to discuss on different subject matters. Therefore, sentiment analysis on twitter data and opinion mining are the aspects that attracted number of researchers all over the globe. In olden days various works are publishes by authors who worked on the various techniques used in the previous days [1]. From the recent study Deep Neural Networks are proven to be best for performing sentiment analysis. Out of these techniques Convolutional Neural Networks (CNN) [2][3] and Recurrent Neural Networks [4] are utilized at large implementations. Dimensionality reduction problems can be solved effectively by using CNN technique and a variety of RNN methods i.e.,LSTM networks [5] can handle sequential data.

The architecture designed by the CNN can be used effectively in case of sentence classification as presented by the authors in previous research works [6][7]. Moreover, these CNN techniques are proved to be effective when compared with the conventional practices [8]. In [9] the efficiency of RNN is clearly presented as it exhibits better results and LSTM networks with combination of CNN networks was presented [10], using these two kinds of neural networks obtains significant result and precise accuracy showing the significant advantages.

The existing techniques are having variant terms of network configuration, tuning, classifications etc., yet as they are not so effective. Hence, there is a need and great demand for implementation of new techniques for evaluation of the research for understanding the challenges and limitations on sentiment analysis.Hence, the twitter data is evaluated under single network framework in the research by using deep learning networks.

II. LITERATURE SURVEY

From the past research studies the Gated Recurrent Unit (GRU)networks were introduced in 2014 [11] is also an effective method which is more similar with that of LSTM networks [12]. A survey is conducted on deep learning methods which deals with sentiment analysis [13] observed to have a word embedding technique. It is done by using two methodssuch as Word2Vec [14] and another one isGloVe [15] as these techniques make the analysis basing on the work count of raw data of tweets.

All over the world [16] in present era twitter is a trending social media platform which shares the information dealing with current issues in the nation. Therefore, public opinions are extracted from the tweets on various subject matters which give analysis of different subject matters and classification of opinions is known as sentiment analysis that became a subject matter having great interest in the account holders of twitter. The early works such as bi-grams, unigrams, POSspecific polarity featuresare used for sentiment analysis. The machine learning classifiers like the Bayesian networks or support vector machines [17][18] were also used for sentiment analysis which are not proven to be so effective. Today, high scores are gained by the deep learning methods as they are having high demand in present day technology. In the current study various neural networks combinations are used to obtain expected accuracy.

This study propose smachine learning based deep learning techniques for sentiment analysis and it is the combination of CNN (Convolutional Neural Network) and LSTM (Long Short Term Memory) networks. Raw data is taken as test case and experiments are conducted to gain high accuracy and precise analysis.Finally, a variation of CNN'sthe RCNN'swere successfully applied. Despite of all these features and the performance evaluation became difficult for comparison between various datasets, network configurations, tuning and specific setup.The main intension behind development of this model is to design a single framework for comparison of these methods and estimation of advantages and drawbacks of each configuration within the datasets.

III. SENTIMENT ANALYSIS APPROACH ON TWEETS

This paper clearly presents the configurations of raw data on twitter by using advanced CNN-LSTMbased deep neural networks application. As we already know some traditional practices such as GRU network and RNN (Recurrent Neural Network) methods, they generate almost same kind outcome, those methods are not used in our research work instead of this advanced LSTM technique is used.

Data Collection: By setting batch streaming configuration apache flume gathers data on streaming tweets, this collected data is sent to Hadoop Distributed File System (HDFS) formation gathered is filtered by using keywords.

Data Preprocessing: The nested JSON format is used to represent the data with the help of flume. But this format consists of duplicate data, in such case that data has to be cleaned and preprocessed for analysis of data effectively.

Class Labeling:Manual labeling may lead to some errors. So that, Senti Strength lexicon based approach is utilized for annotating the trained data among the classifiers.

3.1 $Tf - idf$vectorization:At the time of retrieving information, $Tf - idf$ or TFIDF (Term Frequency– Inverse Document Frequency) process is used. This method analyses the data for each and every word of dataset and understands the importance of words used in that context. Text mining and information retrieval is done by using weights in $Tf - idf$vectorization method. It mainly considers the number of words present within the given dataset with more focus on number of times the particular word is used within the dataset for enhancing the information retrieval system as it regulates the frequently appearing words.When a user gives query, these queries are searched in the documents by using $Tf - idf$ weight. This is used to search queries in the relevant documents when the user rise query which is done by the search engine for ranking the documents.Term Frequency (TF) represents the frequency of the word used in that particular document whereas Inverse document frequency (idf) represents the importance of that word in the document. This can be explained with the help of an equation containing a dataset D, w denotes word and $d \in D$denotes document records, This can be computed with the help of equation below:

$$w_d = f(w, d) * \log(|D|/f(w, D)$$

Where D is the collection of records, f (w, d) denotes the number of times that a word is used in the document. $w_d$is used to represent the importance of the word used. The$Tf - idf$ orwordsearch query.

The following flow chart represents the workflow of the proposed system for obtaining the sentiment analysis by using deep learning technique involving single and multiple CNN, LSTMnetworks with the help of $Tf - idf$ vectorization method as shown in the following fig.1as follows. The tweets are processed by obtaining the data sets of tweets a social networking site as a platform for generation of sentiment analysis on twitter reviews.
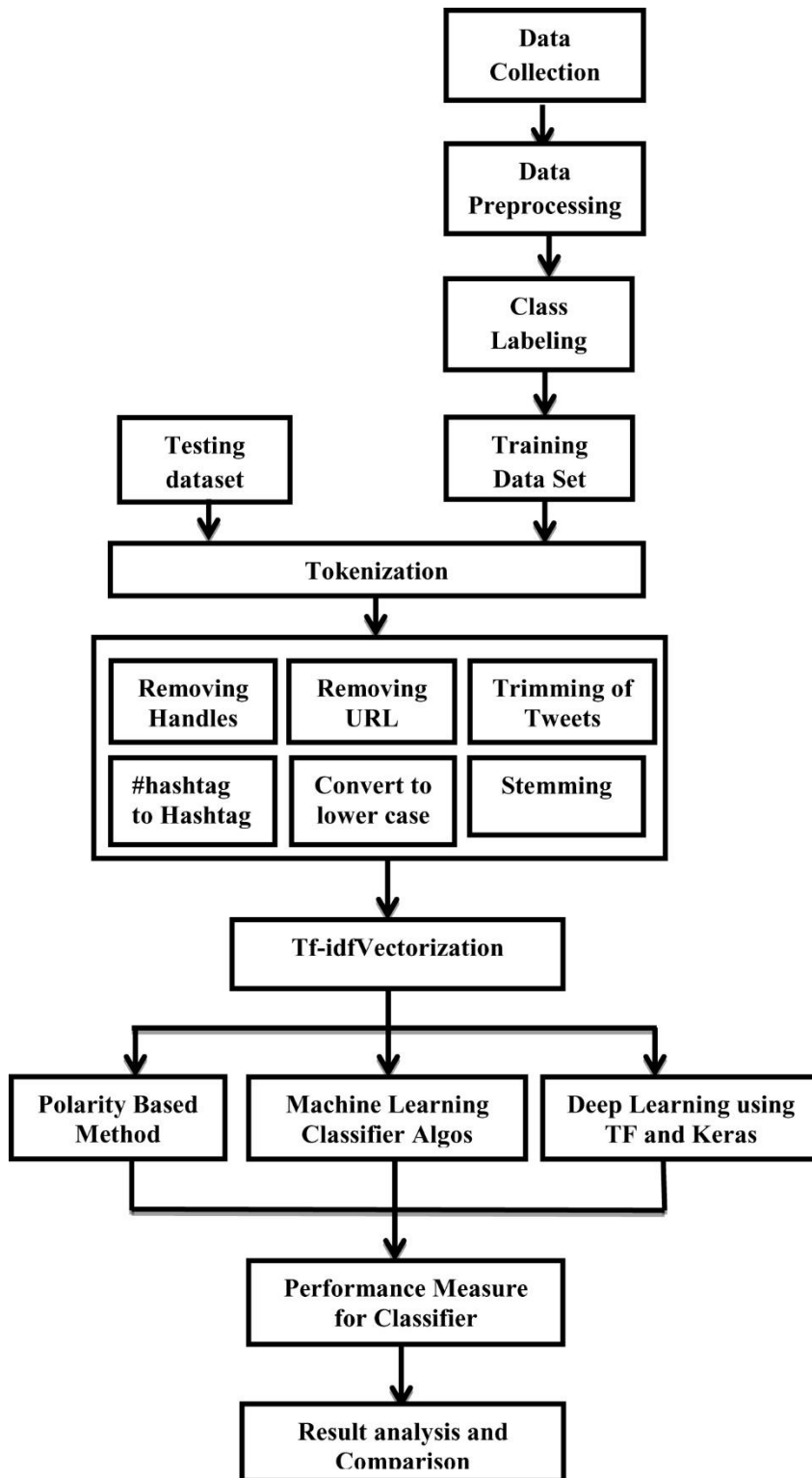

Fig.1: WORKFLOW DIAGRAM FOR SENTIMENT ANALYSIS ON TWITTER DATASET REVIEWS.

3.2 Neural Networks:

In the proposed system usage of the neural network configurations is highly recommended for evaluation of twitter data basing on CNN and LSTMnetworks. Regional and Non- regional datasets are used for testing the proposed system. RCNN and GRUmethods are not applied here as they generate similar result as that of CNN and LSTMnetworks correspondingly.

3.2.1 Single CNN network: Single dimensional CNN layer is used in this network. Fig.1 represents the flowchart for functioning of proposed sentiment analysis on twitter data. The method also use Tf-idf tokenization method on the twitter data collected from the network. It consists of 12 kernels having 1x3 size and the max pooling size is also 1x3. The same parameters are same for CNN configuration. Finally the output is predicted in terms of positive, negative or neutral.

3.2.2 Single LSTM network:A single LSTMlayer is used in the configuration which is having a dropout of 20%. In order to predict the polarity as positive, neutral or negative an output of 1x3 frame is used.

3.2.3 Individual CNN and LSTM networks: The main aim of this configuration is accomplished by considering the individual outputs of CNN and LSTMnetworksand finally the results are evaluated together. The outcome of network can be predicted based on soft voting. Fig.1 represents the structure of this configuration whereas the CNN and the LSTMnetworks have same configurations as that of previously used methods such as having CNN12 kernels with size 1×3 and a max pooling layer with a size of 1×3.

3.2.4 Multiple Networksbasing on CNN's and LSTM networks: The input is divided into basic elements in current setup. The words are used for non-regional inputs and regions are used for regional inputs. These terms can be taken as input to individual CNN's.The output of every CNN node is considered as an input to a single LSTM network.

IV. RESULT

In this section it is explained about the performance standards of various networks in terms of Accuracy, Precision Recall, and F-measure (F1). All these metrics are explained with the help of following equations:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}(2)$$

$$Precision = \frac{TP}{TP+FP} (3)$$

$$Recall = \frac{TP}{TP+FN}(4)$$

$$F_{Measure} = \frac{2*Recall*Precision}{Recall+Precision}(5)$$

The performance of the system can be improved by using multiple CNN with LSTM networks instead of using single and plain configurations. As we tested the data sets by using various kinds of methods on sentiment analysis it is proved that deep neural network system which is machine learning and deep learning based techniques gives out best results when compared with other networks.

Word embedding system is also used for counting the number of words used in the document by using vectorization method. The main aim of this machine learning based sentiment analysis is to obtain the single framework for extracting the twitter data. The extracted raw data is filtered based on the flow chart mentioned in this research work which finally gives out the sentiment analysis either as positive, negative or to be neutral opinion expressed by the twitter users in the nation on the current issues and the welfare scheme being implemented in the country.

In previous study number of researchers has been conducting research on the sentiment analysis of the twitter reviews basing on the word count. Basing on the network system used the word count is considered where labeled tweets are considered as datasets and the each network system has their respective accuracy standards which are shown in the below table.1 as follows.

Table.1 ACCURACY STANDARDS OF VARIOUS NETWORK SYSTEMS ON SENTIMENT ANALYSIS OF TWITTER REVIEWS

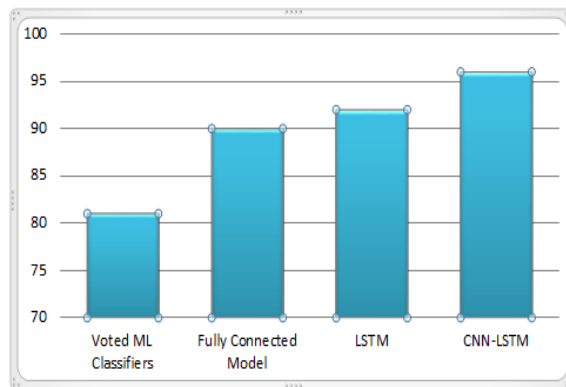| Study	Network System	Dataset (labeled Tweets)	Accuracy
Baziotis	bi-LSTM	~50.000	0.62
Cliche	CNN+LSTM	~50.000	0.65
Deriu	CNN	~300.000	0.65
Rouvier and Favre	CNN	~20.000	0.60
Wange	CNN+LSTM	~8.500	1.340^{∂}
Current Study	CNN+LSTM	~31.000	0.89

Fig.2 COMPARATIVE ANALYSIS OF VARIOUS CLASSIFIERS ON TWITTER REVIEWS.

V. CONCLUSION

This paper presents the various combinations of deep learning based CNN and LSTM techniques are utilized for sentiment analysis on twitter data which gave out an accurate result to the optimum level of 97%. This CNN-LSTM based evaluation is slightly different but gives similar results as that of state of the art methods. Hence, it gives access to extract credible information from various setups. Basing on all these factors from this research work it can be observed that the combination of CNN and LSTM networks performs better when compared with single technique application. The CNN has effective dimensionality reduction process along with the usage of LSTM networks that can preserve the word dependency. Moreover, usage of CNN and LSTM networks in combination increase the performance standards of the system.

When compared with conventional practices these enhanced features and functionalities of combined neural networks designed a hybrid model improved the overall classification accuracy and f-measure of sentiment prediction. Consequently, this proposed method is an effective method when compared with the f-measure scores of other methods. The summary of this paper is to evaluate various machine learning based deep learning, neural network configurations which experimented with single dataset for development of ideal framework for sentiment analysis. This evaluation framework shed more illusion on the limitations and challenges related with sentiment analysis.

VI.REFERENCES

[1] L. L. Bo Pang, "Opinion Mining and Sentiment Analysis Bo", Found. Trends® Inf. Retr., vol. 1, no. 2, pp. 91–231, 2008.

[2] K. Fukushima, "Neocognition: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position", Biol. Cybern., vol. 202, no. 36, pp. 193–202, 1980.

[3] Y. Lecun, P. Ha, L. Bottou, Y. Bengio, S. Drive, eta R. B. Nj, "Object Recognition with Gradient-Based Learning", in Shape, contour and grouping in computer vision, Heidelberg: Springer, 1999, pp. 319–345. [4] D. E. Ruineihart, G. E. Hinton, eta R. J. Williams, "Learning internal representations by error propagation", 1985.

[5] S. Hochreiter eta J. UrgenSchmidhuber, "Ltsm", Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[6] Y. Kim, "Convolutional Neural Networks for Sentence Classification", arXivPrepr. arXiv1408.5882., 2014.

[7] C. N. dos Santos eta M. Gatti, "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts", in Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 69–78.

[8] P. Ray eta A. Chakrabarti, "A Mixed approach of Deep Learning method and Rule-Based method to improve Aspect Level Sentiment Analysis", Appl. Comput. Informatics, 2019.

[9] S. Lai, L. Xu, K. Liu, eta J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification", Twenty-ninth AAAI Conf. Artif. Intell., pp. 2267–2273, 2015.

[10]D. Tang, B. Qin, eta T. Liu, "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification", in Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 1422–1432.

[11]K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, eta Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation", arXivPrepr. arXiv1406.1078, 2014.

[12]J. Chung, C. Gulcehre, K. Cho, eta Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling", arXivPrepr. arXiv1412.3555, 2014.

[13]L. Zhang, S. Wang, eta B. Liu, "Deep Learning for Sentiment Analysis: A Survey", Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 8, no. 4, pp. e1253, 2018.

[14] T. Mikolov, K. Chen, G. Corrado, eta J. Dean, "Efficient Estimation of Word Representations in Vector Space", arXivPrepr. arXiv1301.3781, 2013.

[15] J. Pennington, R. Socher, eta C. Manning, "Glove: Global Vectors for Word Representation", Proc. 2014 Conf. Empir. Methods Nat. Lang. Process., pp. 1532–1543, 2014.

[16] H. Kwak, C. Lee, H. Park, eta S. Moon, "What is Twitter, a social network or a news media", in Proceedings of the 19th international conference on World wide web, 2010, pp. 591–600.

[17] A. Go, R. Bhayani, eta L. Huang, "Proceedings - Twitter Sentiment Classification using Distant Supervision (2009).pdf", CS224N Proj. Report, Stanford, vol. 1, no. 12, 2009.

[18] A. Srivastava, V. Singh, eta G. S. Drall, "Sentiment Analysis of Twitter Data", in Proceedings of the Workshop on Language in Social Media, 2011, pp. 30–38.