2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) | 978-1-7281-8337-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICACCCN51052.2020.9362853

# Sentiment Analysis on Twitter Data Using Deep Learning approach

**Vishu Tyagi**
Department of CSE
Graphic Era Deemed to be University
Dehradun, India
tyagi.vishi@gmail.com

**Ashwini Kumar**
Department of CSE
Graphic Era Deemed to be University
Dehradun, India
ashwinipaul@gmail.com
ORCID: 0000-0001-9908-7430

**Sanjoy Das**
Dept. of Computer Science
Indira Gandhi National Tribal
University-RCM, Imphal, India
sdas.jnu@gmail.com
ORCID: 0000-0001-8018-0870

*Abstract*— **The recent developments of many social networking websites have created large collections of product reviews and polarity of opinions etc, data for customer around the world. Data collection from these social media can be utilized to solve objectives such as market prediction, product recommendation, and reviewer's sentiment. It is very difficult task to manage of unstructured data available on social media. To handle such type of data, Deep learning algorithms is an appropriate solution for analysis these challenges. In this paper, we propose a CNN-LSTM based deep learning method with pre trained embedding approach learns to extract feature automatically for analysing sentiments and classification of reviews or opinions labelled into two polarity as positive or negative. Our proposed model to implement the result gives better performance on benchmark dataset. The performance of CNN-LSTM based deep learning method has compared with baseline machine learning methods.**

***Keywords— Deep learning, Sentiment analysis, CNN, LSTM.***

## I. INTRODUCTION

In the last few years, Social media become the most common platform for sharing data and information across the Universe. Social media plays a vital role for the person who wants to put their opinions about any current or real actions. People can share information through platforms like Facebook, Twitter and other social media application commonly used to share information along with their sentiment by using some tweets or text emotions [17, 18]. The tweets may belong to any services, factors or any products which reflect their views or emotions, on a product or service which may be positive or negative. The customer views can be identified using sentiment analysis which permits businesses to recognize customer emotion toward goods with their features in online reviews and feed backs [2, 3]. The main motive this type of analysis is differentiating text sentiment division, as a classification problem. The main source of sentiment analysis comes from social media sites [19, 22]. Continuously many social networking websites generate complex information related to sentiments. Many peoples share opinions in social media, tweets, and micro blogs to explore reviews on online products [4]. These data needs to be analysed to detect the polarity on movie reviews, products, country political issues, etc. Recently, many researchers have been proposed deep learning methods to solving different tasks on NLP. In Deep Learning Neural networks [16], we can easily apply our automatic features on supervised learning and no need to require any manual features [5, 6].

In this paper, we have proposed a CNN-BiLSTM method for the sentiment analysis of Twitter data having 1.6 million annotated tweets. The accuracy is 81.02 for classifying tweets as Negative or Positive on benchmark dataset. Our novel approach uses LSTM based neural network with few of parameters attains modest results and outperformed against statistical machine learning methods.

The remainder of this paper is organized as follows. We discussed related work in Section 2. Description of our proposed CNN-BiLSTM architecture and Sentiment140 dataset in Section 3. Section 4 discusses our experiments and results. In, Section 5 paper is concluded.

## II. RELATED WORK

In last few years, most of the research works have been discussed on sentiment analysis on social media's data. By Deep learning neural network, the author of [8] had improve the accuracy of sentiment analysis prediction and using text mining to handling the tweets and huge data . They used deep learning neural network with multiple hidden layers and analysed the different tweets or comments. Subsequently, author [9] described some of the different approaches used in sentiment analysis research. They discussed various techniques of machine learning such as traditional models, deep learning models etc. In recent years, deep learning models appeared as a dominant tool for natural language processing and traditional models mostly suitable for sentiment analysis. The performance of the deep learning models, mainly CNN and LSTM is used on various datasets.

In author [10], Chinese text is considered for sentiment analysis which could be thought as two-classification problems to find the polarity of text, with positive and negative emotional tendencies. They performed data cleaning, word segmentation, removing stop words, feature selection and classification. The weights of features were calculated by the TF-IDF(Term Frequency–Inverse Document Frequency). In their work, SVM model is used for text representation and used SVM and ELM to analysing text emotions.

ISBN: 978-1-7281-8337-4/20/$31.00 ©2020 IEEE

In Author [20], proposed a neural network based model, that mainly focused on the association between the behavioural data of the account user in a given tweets, which can be responsible to upgrading its performance and accuracy by using some indicators. Based on existing research, mostly have used classic baseline machine learning methods such as Logistic regression, Na¨ıve Bayes, Decision Trees, Support Vector Machine (SVM) and Neural Networks with small amount of data [12, 13, 14, 15, 21]. In current scenario, using deep learning neural networks for handing huge data and improving the accuracy.

## III. METHODOLOGY

We applied a LSTM deep neural architecture [1] and used automatic feature extraction for this experiment. The Proposed deep learning approach is described in details in Section 3.1. We use the annotated Sentiment140 Twitter dataset from Stanford University [11]. It was collected using Twitter API and contains 1.6 million annotated English tweets and each tweet is categorize them into Negative or Positive class.

The dataset is pre-process first using the following text normalization process. Removing unnecessary attributes such as Twitter users name, URLs and re-tweets. We also apply Stemming (same linguistic root) and remove stop words, numbers, extra white spaces, punctuation and then convert all the texts to lowercase.

### *A. The CNN-LSTM Architecture*

The Proposed CNN-LSTM neural network architecture is shown in Fig.. 1. We used Glove 6B 300D as word embedding model in first layer, in which our embedding learns all words from Sentiment140 training datasets. The size of the vocabulary is 15k and the output dimension of embedding layers is 30 x 300 matrix, in which the maximum length of each batch is 30. After embedding layer passes features into Spatial drop out layer with the rate to 0.2 to avoid over fitting.

The output feds into first 1-Dimensional CNN layer and the size of each filter is 5 x 5. Further, we will define 64 filters. This allows us to train 64 different features on the second layer of the network. Thus, the output of the second neural network layer is a 26 × 64 matrix, and the result will be fed into the Bi-LSTM layer of size 64 to capture long range dependencies to extract feature and then feed into the Fully Connected Dense layer has a size of 128 × 512 neurons .Next, the output of Dense Layer passes to drop out with a rate of 0.5, the purpose of which to drop some randomly weights of the matrix. Finally, fully connected dense layer with sigmoid function that producing the vector as input to predict with 2 units (positive, negative). The output of vector has a size of 512 × 1 neurons.

We used binary cross entropy loss function and Adam optimizer with learning rate of 0.001 to train the model using 10-epochs on a batch size of 1024. In Fig. 2, we are shows the each Layer details of CNN-LSTM model.

## IV. EXPERIMENT AND RESULTS

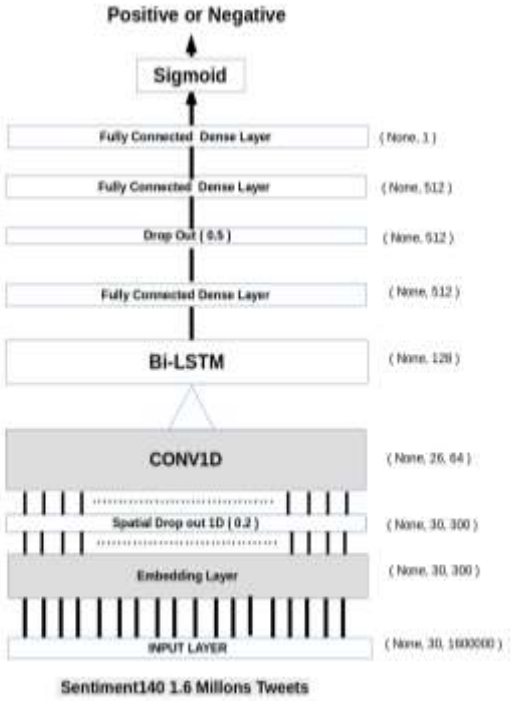



Fig. 1. Proposed Model for Sentiment analysis (CNN-LSTM)

| Layer (type) | Output Shape | Param # |
|---|---|---|
| embedding_1 (Embedding) | (None, 30, 300) | 175969800 |
| spatial_dropout1d_1 (Spatial | (None, 30, 300) | 0 |
| conv1d_1 (Conv1D) | (None, 26, 64) | 96064 |
| bidirectional_1 (Bidirection | (None, 128) | 66048 |
| dense_3 (Dense) | (None, 512) | 66048 |
| dropout_1 (Dropout) | (None, 512) | 0 |
| dense_4 (Dense) | (None, 512) | 262656 |
| dense_5 (Dense) | (None, 1) | 513 |

Total params: 176,461,129
Trainable params: 491,329
Non-trainable params: 175,969,800

Fig. 2. Details of CNN-LSTM model.

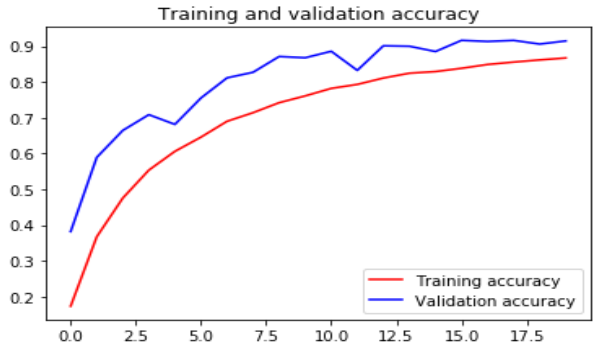
The sentiment140 twitter dataset of 1.6 million annotated tweets provided by Stanford University [11] is used in our experiment. To annotated tweets, equally 0.8 million tweets are labelled as positive and negative. We used python Keras with TensorFlow as a backend and using Sckit-learn toolkit CNN-LSTM method is implemented. The dataset is, split into 75% for training using 10-epochs and test the optimized model on 25% for measure performance after tuning.

We report our performance on the 25% validation data for benchmark dataset using proposed model against the baseline methods as shown in Table.I.

TABLE I. COMPARISON OF OUR PROPOSED MODEL WITH MACHINE LEARNING METHODS

| Models | Accuracy(%) |
|---|---|
| SGDClassifier | 65.87 |
| LogisticRegression | 68.37 |
| LogisticRegressionCV | 68.37 |
| LinearSVC( SVM) | 65.80 |
| RandomForestClassifier | 64.62 |
| CNN | 78.81 |
| **Proposed CNN-LSTM** | **81.20** |

An observation that is clearly seen in Table I is our proposed CNN-LSTM model achieved highest performance against several baseline methods. The overall accuracy of our model is 81.20% on the task of sentiment analysis of large benchmark dataset. In CNN-LSTM model, we have used Bi-Directional LSTM stacked layers of 128 along with dense layer to extract features from twitter dataset to enhance the performance in the term of accuracy values. As part of the analysis, we show training and validation behaviour of our model performance and loss for the 10 epochs in Fig. 3 and Fig.4.




Training and Validation performance

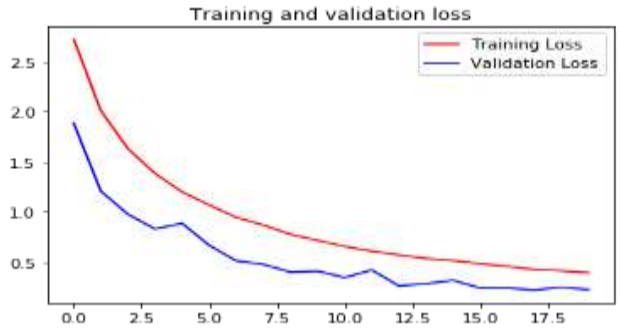



Fig. 3. Training and Validation loss

## V. CONCLUSION

In this paper, we provide an overview of sentiment analysis for Sentiment140 Twitter dataset. We proposed a novel approach for analyzing tweets as a Negative or Positive class from annotated twitter dataset using deep learning neural networks combined with CNN-LSTM methods. In this model, we have used efficient deep learning architecture with tuned hyper parameters on CNN layers followed by Bidirectional LSTM neural networks having long range dependencies. Our model performed better than all baseline methods on our benchmark dataset. The accuracy of the proposed model is 81.20% In future work; we plan to explore other sentiment twitter dataset and hybrid deep learning models for sentiment analysis.

## REFERENCES

[1] Sepp Hochreiter and J¨urgen Schmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735–1780

[2] Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews. In Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 42–49.

[3] Tai, K.S., Socher, R. and Manning, C.D., 2015. Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075.

[4] Zhai, S. and Zhang, Z.M., 2016, February. Semisupervised autoencoder for sentiment analysis. In Thirtieth AAAI Conference on Artificial Intelligence.

[5] Dou, Z.Y., 2017, September. Capturing user and product information for document level sentiment analysis with deep memory network. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 521-526).

[6] Babaie, M., Shiri, M.E. and Bahaghighat, M., 2018, April. A new descriptor for UAV images mapping by applying discrete local radon. In 2018 8th Conference of AI & Robotics and 10th RoboCup Iranopen International Symposium (IRANOPEN) (pp. 52-56). IEEE.

[7] Dragoni, M. and Petrucci, G., 2018. A fuzzy-based strategy for multidomain sentiment analysis. International Journal of Approximate Reasoning, 93, pp.59-73.

[8] Ramadhani, A.M. and Goo, H.S., 2017, August. Twitter sentiment analysis using deep learning methods. In 2017 7th International annual engineering seminar (InAES) (pp. 1-4). IEEE.

[9] Singhal, P. and Bhattacharyya, P., 2016. Sentiment analysis and deep learning: a survey. Center for Indian Language Technology, Indian Institute of Technology, Bombay.

[10] Zhang, X. and Zheng, X., 2016, July. Comparison of text sentiment analysis based on machine learning. In 2016 15th International Symposium on Parallel and Distributed Computing (ISPDC) (pp. 230 233). IEEE.

[11] Available online:http://help.sentiment140.com/ site-functionality (accessed on 12 june 2020).

[12] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[13] Johnson, R. and Zhang, T., 2015. Effective use of word order for text categorization with convolutional neural networks. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2015 (p. 103).

[14] Johnson, R. and Zhang, T., 2015. Semi-supervised convolutional neural networks for text categorization via region embedding. In Advances in neural information processing systems (pp. 919-927).

[15] Dos Santos, C. and Gatti, M., 2014, August. Deep convolutional neural networks for sentiment analysis of short texts. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (pp. 69-78).

[16] Nio, L. and Murakami, K., 2018, March. Japanese sentiment classification using bidirectional long short-term memory recurrent neural network. In Proceedings of the 24th Annual Meeting Association for Natural Language Processing (pp. 1119-1122).

[17] Zhang, L., Wang, S. and Liu, B., 2018. Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), p.e1253. Natural Language Processing (pp. 1119-1122).

[18] Socher, R., Lin, C.C., Manning, C. and Ng, A.Y., 2011. Parsing natural scenes and natural language with recursive neural networks. In Proceedings of the 28th international conference on machine learning (ICML-11) (pp. 129-136).

[19] Long, H., Liao, B., Xu, X. and Yang, J., 2018. A hybrid deep learning model for predicting protein hydroxylation sites. International Journal of Molecular Sciences, 19(9), p.2817.

[20] B. K. Bhavitha, A. P. Rodrigues and N. N. Chiplunkar, "Comparative study of machine learning techniques in sentimental analysis," *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore, 2017, pp. 216-221, doi: 10.1109/ICICCT.2017.7975191.

[21] A. Kumar, S. Das and V. Tyagi, "Anti Money Laundering detection using Naïve Bayes Classifier," 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, 2020, pp. 568-572, doi: 10.1109/GUCON48875.2020.9231226.

[22] Saroj Kushwaha and Sanjoy Das, “Sentiment Analysis of Big-Data in Healthcare: Issue and Challenges”, 2020 5th IEEE International Conference on Computing, Communication and Automation is being jointly organized by Aurel Vlaicu University of Arad, Romania & Galgotias University, India on October 30-31, 2020 at Galgotias University Greater Noida (NCR New Delhi) India.

.