Detection and Classification of Toxic Comments by Using LSTM and Bi-LSTM Approach

Akash Gupta[1](✉), Anand Nayyar[2], Simrann Arora[1,3], and Rachna Jain[1]

[1] Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, India
rachna.jain@bharatividyapeeth.edu
[2] Graduate School, Duy Tan University, Da Nang 550000, Viet Nam
anandnayyar@duytan.edu.vn
[3] Faculty of Information Technology, Duy Tan University, Da Nang 550000, Viet Nam

Abstract. With the advancement in the technology, a lot of comments has been produced on a regular basis through the various online communication platforms like Wikipedia, twitter, Glassdoor etc. Although, many of these comments really benefit the people, but the various high toxic comments are also responsible for the increasing online harassment, mental depression and even personal attacks. Toxic Comment Classification is one of the active research topics at present. In the following study, a multi-label classification model is presented to classify the various toxic comments into six classes namely toxic, severe toxic, obscene, threat, insult and identity hate. The proposed classification model has been built using deep learning algorithms explicitly Long Short-Term Memory (LSTM) and Bi-Directional Long Short-Term Memory (Bi-LSTM) along with the word embeddings by adapting insights from previous proposed works. The dataset for this research is obtained from the Kaggle and is provided by the Conversation AI team (a research ingenuity co-founded by Google as well as Jigsaw). The accuracy score of both the proposed techniques is evaluated and compared. Finally, the empirical results show that Bi-LSTM algorithm achieved better in comparison to LSTM with an increased accuracy of 98.07%.

Keywords: Toxic comments classification · LSTM · Bi-LSTM · Word embeddings · Multi-label classification · Online harassment · Personal attack

1 Introduction

As the world is progressing with an ever-increasing rate, the surge in technological advancements is also at an all-time high [1]. This has caused more and more people around the globe to have access to several platforms on the internet and express their views and opinions on almost every other thing [2]. The social media sites are becoming easily accessible each passing day, thereby increasing the number of users and intensifying their vulnerability. On one side, this has helped many people to interact with each other, discuss over various eclectic issues around the globe while sitting at the comfort of their homes and provide information on certain topics in the form of comments, but

A. K. Luhach et al. (Eds.): ICAICR 2020, CCIS 1393, pp. 100–112, 2021.
https://doi.org/10.1007/978-981-16-3660-8_10

on the other hand it has increased the cases of online harassment and misconduct among people [3]. The social media sites especially, do have photographs of people and may provide a glimpse of their personal life as well. Inappropriate comments on such sensitive content can also harm mental or physical well-being of people and force them to take some inappropriate actions. Such comments are identified as being toxic in nature and they contain abusive words, foul language, aggression, hate, insulting remarks and threats of various types [4]. There is a need to help identifying these comments and stop them from causing any harm or loss of life further. Thus, this topic becomes an extensive and challenging area of research and might help in earlier and faster detection of typical comments in future.

Some of the key objectives of this research are:

- Detection and classification of toxic comments to prevent online harassment and misconduct to a large extent
- Development of a multi-label classification model using deep learning models, namely, LSTM and Bi-LSTM LSTM along with the word embeddings by adapting insights from previous proposed works into 6 different categories of abusive words, foul language, aggression, hate, insulting remarks and threats of various types.
- Achieved a high accuracy of 98.07% by using the Bi-LSTM which performed better than LSTM approach and facilitate research in this field.

The rest of the paper is organized in the following manner. Section 2 provides a detailed review of the various researches takes place in the world for the classification and detection of toxic comments. Section 3 discusses the algorithms and techniques used in this research work. Section 4 deliberates the proposed methodology steps along with the proposed model of the entire research work in detail. Later, Sect. 5 discusses about the experimental results and simulations along with the various evaluation plots used in this research. Finally, Sect. 6 concludes the paper with future scope.

2 Literature Review

The advent in technology has brought people closer by interacting through comments on various platforms. These comments mostly are neutral, but some comments include hate, aggression, abusive words which can seriously cause harm to the other person. Thus, toxic comment classification has been a major concern these days to prevent people from online harassment and mental breakdown. Many types of researches are being done on this issue. Here, are some of the researches listed below from all over the world.

van Aken et al. [7] worked on the comparison of various deep learning models along with the shallow approaches on a novel, huge dataset of comments and proposed an ensemble method that outshined all the other individual models. Subsequently, the findings were validated on another dataset. The results obtained by the ensemble method facilitated the authors for performing an all-embracing error scrutiny, which consequently revealed the encounters for the advanced approaches along with guidelines for the scope of future research. The challenges contained the inconsistency in dataset labels along with the missing paradigmatic context.

Srivastava et al. [8] proposed a solitary model capsule network which had a focal forfeiture to accomplish the chore of identifying the aggression as well as toxic comments. This approach is well suited for the production environment. The proposed model achieved an outstanding result as compared to other baselines models, showing its efficacy and depicting that the focal loss displays crucial improvement in the cases where the imbalance of classes is a major concern. Along with this the concerns regarding extensive data augmentation and processing are dealt with the proposed network. The model also tackles the transliteration problem in an effective manner, which had comments in both English and Hindi languages.

Saeed et al. [9] worked on several Deep Neural Network (DNN) techniques for classification of the overlapping sentiments with a high accuracy. Furthermore, the proposed framework used for classification did not necessitate a large volume of text pre-processing and is able to handle this concern implicitly. The pragmatic validation performed on a practical dataset supported the authors' claim by giving superior results.

Vaidya et al. [10] evaluated various advanced models with particular focus on the reduction of model prejudice towards the most vulnerable and attacked identity groups. The authors proposed a multi-task erudition framework with a consideration layer that with a joint focus, predicts the identities in the comment as well as its toxicity to reduce the bias. Then they compared the model to an arrangement of deep learning and shallow learning models by leveraging the metrics that have been devised for testing the bias within these groups of identity.

Deshmukh et al. [11] presented a novel approach which used RNNs as well as Capsule networks as the backbone and captured the information (contextual) to a greater extent while learning the representations of word in the text. Experiments were steered on Wikipedia's talk leaf controls. The results showed that the proposed model outperformed the conventional advanced models and displayed the efficacy of capsule networks. After discussing the various researches, this study is focused on the recognition and cataloguing of toxic comments by means of the Bi-LSTM and LSTM approaches.

3 Algorithms and Techniques Used

In this section, the algorithms and techniques used in this research are discussed in detail.

3.1 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) network is a specialized sort of Recurrent Neural Network (RNN) which is able to learn the dependencies that are long-term [11]. They perform exceptionally well on the sequence modelling problems and are devised to circumvent the problem of long-term dependency [12]. Their behavior is to retain the information for long periods of time. Figure 1 below displays the LSTM architecture.

With slight linear interactions across this path, cell state C allows information to be passed unchanged across the complete LSTM which allows LSTM to recognize multiple times steps, the context in the past [14]. There are many inputs and outputs throughout this line which enable us to add or remove the cell state information. Gates control the insertion or deletion of the information [15].

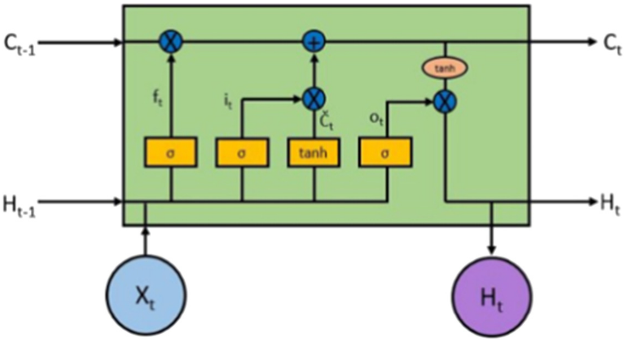

Fig. 1. LSTM architecture [13]

The Sigmoid layer outputs zero-to-one numbers, specifying how much of every part should be allowed through. A zero value suggests letting nothing at all in, whereas a one means getting everything into it [16]. The architecture of LSTM includes 3 gates, namely, the forget gate, the input gate as well as the output gate.

Some applications of LSTMs are in:

- Generation of Handwriting
- Image Captioning
- Language Modelling
- Chatbots for Question/Answering

3.2 Bi-directional Long Short-Term Memory (Bi-LSTM)

Bi-directional RNN basically means a combination of two individual RNNs [17]. This structure enables the network to contain forward and backward, both information regarding the sequence at each and every time step [18]. Utilizing the bidirectional approach, the input can be run in 2 behaviors that are from past to future as well as future to past [19]. It is different from the unidirectional LSTM in a way that backward run in this saves the future information and utilizing the latent states collated together, the information can be saved from both the future as well as the past [20]. Figure 2 below shows its architecture.

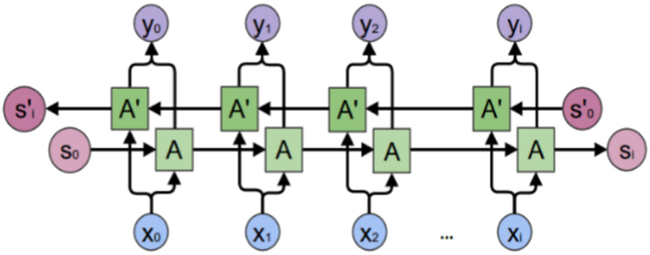

Fig. 2. Bi-LSTM architecture [21]

In this, the activation values are also used, not only the candidate values. Along with this, two outputs are obtained from the cell, a novel activation as well as a novel candidate value [22]. Its architecture also has 3 gates, that are, update gate, forget gate and the output gate. Bi-LSTMs have a major application in text related problems where the previous and future response generation comes into picture [23].

4 Proposed Methodology

In the following section, the research methodology is discussed in depth. The proposed prototype of this research is represented in Fig. 3. Initially, the toxic comments dataset obtained from the Kaggle is provided as the input to the model. After the analysis of the data, the dataset is pre-processed for further analysis. The pre-processing steps includes removal of punctuations, stopwords and null values followed by the stemming techniques. After the pre-processing, the dataset is converted to a suitable input matrix by employing tokenization, padding and the word embedding techniques. After this, the dataset is fragmented into the training and the validation set. 75% of the data is utilized for the training resolution while 25% of the data is used for the validation resolution. The proposed model is then trained by using LSTM and Bi-LSTM algorithms. Finally, the model is validated upon validation set and the presentation of both the algorithms is gauged and compared with each other.

4.1 Dataset Description

The dataset for the research is attained from Kaggle and is made available by the Conversation AI team, which is an examination initiative co-founded by Jigsaw and Google [5]. The dataset consists of a huge number of toxic Wikipedia remarks which had been categorised into the six classes explicitly toxic, sever toxic, obscene, threat, insult and identity hate. These are categorised by the professional ratters. The dataset consists of around 1,60,000 comments taken from the Wikipedia talk pages. Since, it is multi-label classification problem, hence, the comments can belong to more than one classes i.e., a particular remark can be toxic, threat or an insult at the same time. The dataset consists of Comment ID, Comment Text, and the Boolean entries against the corresponding toxic comment category.

4.2 Data Analysis

There are around 1,60,000 comments are present in the dataset. Figure 4 shows the distribution of these comments according to its length. After the analysis of Fig. 4, we conclude that, most of the comments (around 1,20,000) are generally short and having words in the range of 0–100. Also, the average length of the comments is calculated to be around 80 words.

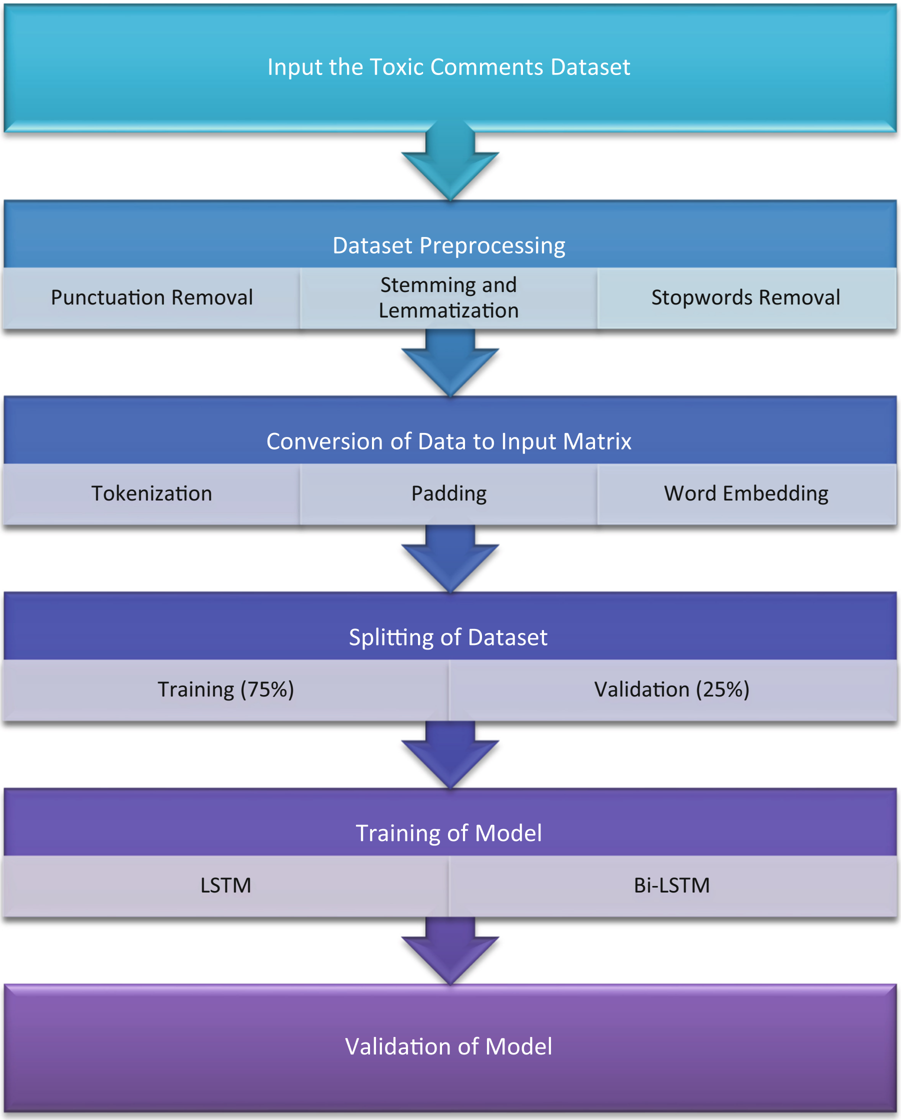

Fig. 3. Flow of proposed methodology

Figure 5 shows the further distribution of these comments into six labels according to its length. After the analysis of Fig. 5, we can conclude that large number of remarks fits to the toxic, obscene in addition to insult classes.

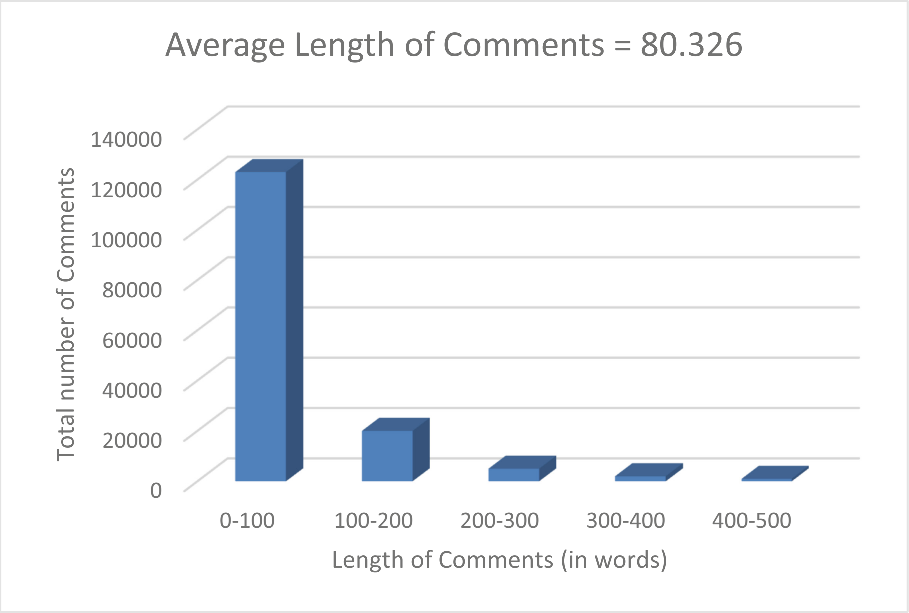

Fig. 4. Distribution of comments according to length

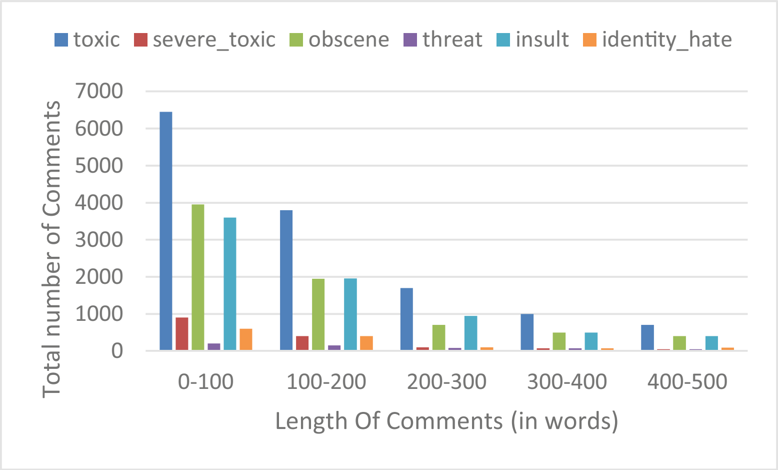

Fig. 5. Distribution of the comments into six labels according to length

4.3 Data Pre-processing

The following steps are employed in this research for the pre-processing of the dataset:

1. Initially, all the null values are removed from the dataset.
2. Then, all the punctuation signs and the numerical digits are removed from the dataset.
3. After that, all the Stopwords like for, this, in, the etc. are removed from the dataset.
4. Finally, stemming and lemmatization is performed which converts the various forms of verbs present in the comments to its base word.

4.4 Data Conversion to Input Matrix

The following steps are employed in this research for the conversion of the data to suitable input matrix.

1. **Tokenization:** It is employed to convert the comments into a series of tokens.
2. **Padding:** Since the average length of the comments is determined as 80, hence, the standard length is taken to be of 80 words.
3. **Word Embedding:** It is performed to get insights from the previous research works. In this research, Glove.6B.300D is used which contain 6 billion tokens and each token is represented by 300D vector representation. This glove dataset is obtained from the web [6].

4.5 Build LSTM and Bi-LSTM Model

After conversion of tokens data into a suitable input matrix, the data is fragmented into the training along with the validation set. 75% of the data is used for the training resolution, while the 25% of the data is utilized for the validation resolution. After splitting, the model is then trained by using Long Short-Term Memory (LSTM) and Bi-Directional Long Short-Term Memory (Bi-LSTM) algorithms.

5 Experimental Results and Analysis

In this section, experimental results are discussed and analysed in detail. The elementary approach of both the algorithms is same and is discussed here.

1. The LSTM or Bi-LSTM network is initialised with 100 neurons.
2. Four dense layers are used in these models in which, three of them having ReLu as its activation function with 100, 70 and 30 neurons respectively. The last layer has sigmoid as its activation function with 6 neurons, as comments are belonging to the 6 classes in total.
3. The dual cross-entropy is chosen as the loss function here.
4. The model is further optimized using Adam optimization technique with learning rate equals to 0.01.
5. Number of epochs is selected to be 2. Because more than 2 epochs overfits the model as shown in Fig. 6 and 7.

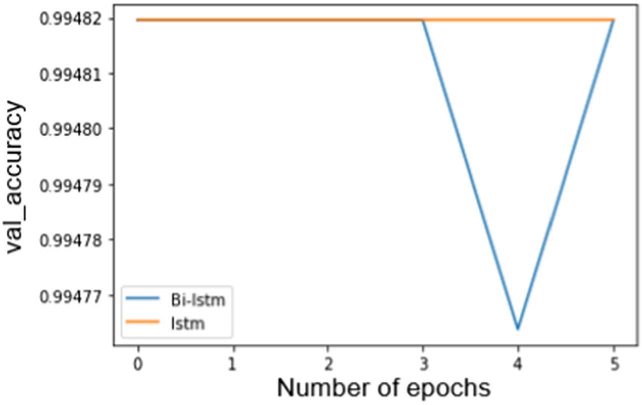

Fig. 6. Validation accuracy with increase in figure of epochs

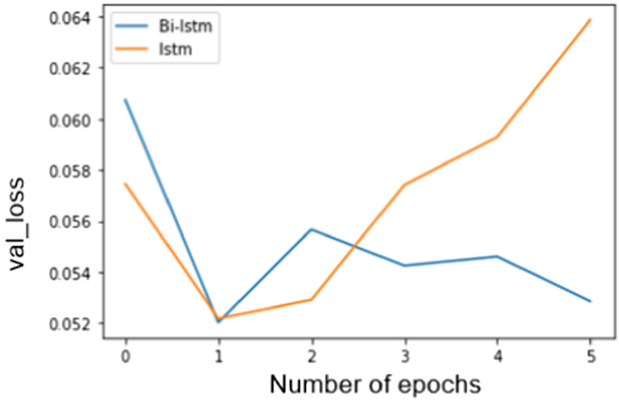

Fig. 7. Validation loss with surge in figure of epochs

Then, after selecting the epoch size as 2 and batch size as 128, the training and validation accuracy and the loss curves for the LSTM and the Bi-LSTM models are plotted then compared. Figure 8 and 9 represents the training accuracy and the training loss for the LSTM and Bi-LSTM neural networks.

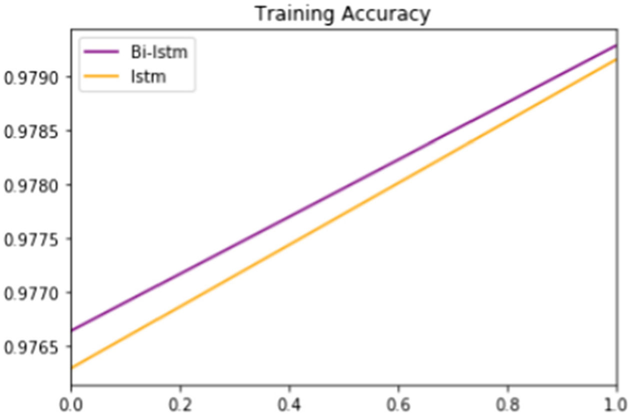

Fig. 8. Training accuracy plot for LSTM and Bi-LSTM networks

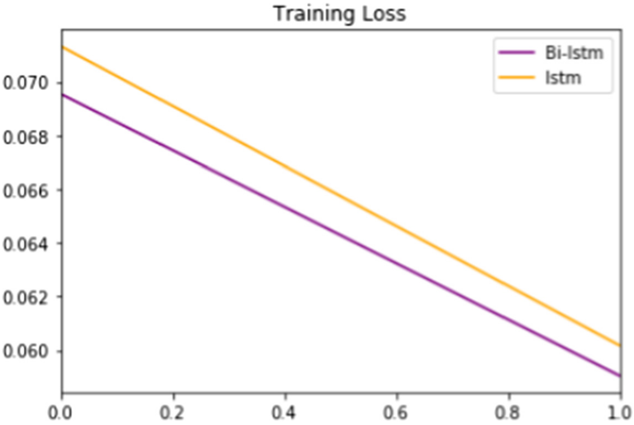

Fig. 9. Training loss plot for LSTM and Bi-LSTM networks

Figure 10 and 11 represents the validation accuracy and validation loss for the LSTM and Bi-LSTM neural networks.

After analysing the training and validation accuracy and loss curves, it is concluded that Bi-LSTM neural network shows improved recital as compared to LSTM network for both the training and the validation set. The validation accuracy score of Bi-LSTM is nearly equals to 0.9807 which is considerably higher than the LSTM network.

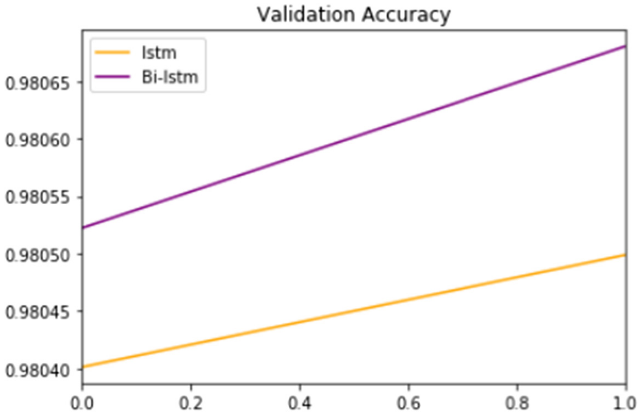

Fig. 10. Validation accuracy plot for LSTM and Bi-LSTM networks

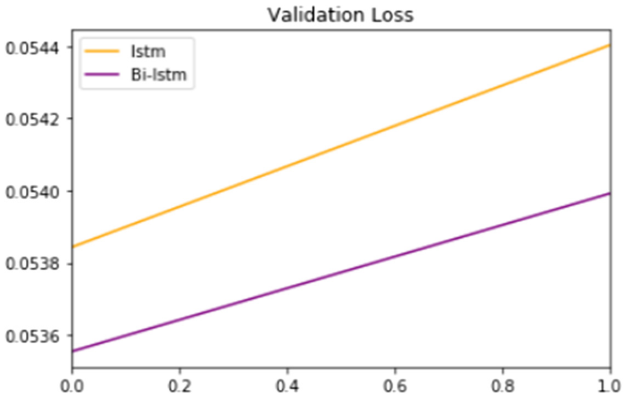

Fig. 11. Validation loss plot for LSTM and Bi-LSTM networks

6 Conclusion and Future Scope

This research work is carried out to offer a multi-label classification model for the classification of various toxic comments collected from the Wikipedia talk pages dataset. The dataset for the following research is obtained from Kaggle. Here, the comments are categorised into six classes namely toxic, severe toxic, threat, insult, obscene, and identity hate. The proposed classification model is build using deep learning algorithms namely Long Short-Term Memory (LSTM) and Bi-Directional Long Short-Term Memory (Bi-LSTM) along with the word embeddings by adapting insights from previous proposed works. In this research, both the proposed models are gaged and equated using the accuracy and the loss curves for the training and the validation datasets. Finally, the pragmatic results display that the Bi-LSTM network shows improved performance with an increased accuracy of 98.07%. The model can be enhanced in future either by developing a denser neural network by increasing the number of dense layers or by employing other RNN techniques.

References

1. Georgakopoulos, S.V., Tasoulis, S.K., Vrahatis, A.G., Plagianakos, V.P.: Convolutional neural networks for toxic comment classification. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence, pp. 1–6 (2018)
2. Mohammad, F.: Is preprocessing of text really worth your time for toxic comment classification?. In: Proceedings on the International Conference on Artificial Intelligence (ICAI), pp. 447–453. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) (2018)
3. Carta, S., Corriga, A., Mulas, R., Recupero, D.R., Saia, R.: A supervised multi-class multi-label word embeddings approach for toxic comment classification. In: KDIR, pp. 105–112 (2019)
4. D'sa, A.G., Illina, I., Fohr, D.: Towards non-toxic landscapes: automatic toxic comment detection using DNN. arXiv:1911.08395 (2019)
5. Toxic Comment Classification Challenge: Identify and classify toxic online comments (2018). Accessed from https://www.kaggle.com/datamunge/sign-language-mnist
6. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for Word Representation (2014). Accessed from https://nlp.stanford.edu/projects/glove/
7. van Aken, B., Risch, J., Krestel, R., Löser, A.: Challenges for toxic comment classification: an in-depth error analysis. arXiv:1809.07572 (2018)
8. Srivastava, S., Khurana, P., Tewari, V.: Identifying aggression and toxicity in comments using capsule network. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 98–105 (2018)
9. Saeed, H.H., Shahzad, K., Kamiran, F.: Overlapping toxic sentiment classification using deep neural architectures. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 1361–1366. IEEE (2018)
10. Vaidya, A., Mai, F., Ning, Y.: Empirical analysis of multi-task learning for reducing model bias in toxic comment detection. arXiv:1909.09758 (2019)
11. Deshmukh, S., Rade, R.: Tackling toxic online communication with recurrent capsule networks. In: 2018 Conference on Information and Communication Technology (CICT), pp. 1–7. IEEE (2018)
12. Sherstinsky, A.: Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D **404**, (2020)
13. Liu, J., Wang, G., Hu, P., Duan, L.Y., Kot, A.C.: Global context-aware attention LSTM networks for 3d action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1647–1656 (2017)
14. Kim, H.Y., Won, C.H.: Forecasting the volatility of stock price index: a hybrid model integrating LSTM with multiple GARCH-type models. Expert Syst. Appl. **103**, 25–37 (2018)
15. Ma, Y., Peng, H., Cambria, E.: Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
16. Zhao, Z., Chen, W., Wu, X., Chen, P.C., Liu, J.: LSTM network: a deep learning approach for short-term traffic forecast. IET Intell. Transp. Syst. **11**(2), 68–75 (2017)
17. Alzaidy, R., Caragea, C., Giles, C.L.: Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents. In: The World Wide Web Conference, pp. 2551–2557 (2019)
18. Tourille, J., Ferret, O., Neveol, A., Tannier, X.: Neural architecture for temporal relation extraction: A Bi-LSTM approach for detecting narrative containers. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers, pp. 224–230 (2017)

19. Minaee, S., Azimi, E., Abdolrashidi, A.: Deep-sentiment: Sentiment analysis using ensemble of cnn and Bi-LSTM models. arXiv:1904.04206 (2019)
20. Lin, J.C.W., Shao, Y., Zhou, Y., Pirouz, M., Chen, H.C.: A Bi-LSTM mention hypergraph model with encoding schema for mention extraction. Eng. Appl. Artif. Intell. **85**, 175–181 (2019)
21. Li, C., Zhan, G., Li, Z.: News text classification based on improved Bi-LSTM-CNN. In: 2018 9th International Conference on Information Technology in Medicine and Education (ITME), pp. 890–893. IEEE (2018)
22. Hua, Q., Qundong, S., Dingchao, J., Lei, G., Yanpeng, Z., Pengkang, L.: A character-level method for text classification. In: 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), pp. 402–406. IEEE (2018)
23. Zhang, Y., Liu, Q., Song, L.: Sentence-state lstm for text representation. arXiv:1805.02474 (2018)