Toxic Speech Classification via Deep Learning using Combined Features from BERT & FastText Embedding

Asmi P
Computer science and Engineering
M G M College of Engineering and Pharmaceutical Science
Valanchery, Kerala, India

Sanaj M S
Computer science and Engineering
M G M College of Engineering and Pharmaceutical Science
Valanchery, Kerala, India

Abstract— **With the growing internet usage rate, people are more likely to express their opinion or ideas openly on social media. A lot of discussion platforms are available nowadays. But some are misused the freedom of speech by spreading online toxic speech. The hate speech that is intended not just to insult or mock, but to harass and cause lasting pain by attacking something uniquely dear to the target. Thus, the necessary of automatically detecting and removing toxic speech in social media is very important. We proposed a feature-based method that combining the features of TF-IDF, FastText Embedding and BERT Embedding and by using a DNN classifier. We compare the individual features of these three methods with the combined features as a performance analysis.**

Keywords:- Social media, hate speech, word embedding, DNN classifier.

I. INTRODUCTION

The use of social media is increased nowadays. Social Media is an open platform for communication between the people. People communicate online to attach to family friends, collaborate with co-workers, acquire news, and participate in their communities. It's a useful thing in a situation where people cannot see each other. But some of them misuse the online platform. When a person decides to interact in an online discussion on social media, they are exposing themselves to the risk of being harassed by trolls and offensive comments. Thus, spreading verbal violence. In the online community offensive comments quite make a negative impact and it may target to an individual, company, religion etc.

Online toxic speech is an antisocial behaviour which is punishable by the law in many countries. In this paper we considered *offensive, abusive,* and *hate* speech as toxic speech. Toxic speech can be expressed either by Explicitly or Implicitly. Explicit toxic speech can identify by using lexicons whereas implicit toxic speech [1] identified by semantic analysis of the sentence.

Regulators have a tough time regulating social media because it is a private space for public expression. Online toxic speech is the expressions of tensions between various groups within and across communities. Toxic speech on the internet is a type of speech that is directed at a person or a group of people based on their race, ethnic origin sexual orientation or gender [3]. In case of an organization may lose the reputation of its product.

Detecting online hate speech is a challenging task. Before spreading the toxic content, we want to remove the same. It is difficult to find the given comment is hate or not manually because on the internet large amount of data are circulating. Thus, an effective method of automatic detection is necessary. In *Natural Language Processing* (NLP), automatic detection of toxic speech is a challenging problem. Toxic speech can be classified by deep learning which is a very powerful technique.

In this paper, we proposed a new technique to automatically detect toxic speech. Classification of toxic speech can be performed by using two powerful word representation namely fastText and BERT embedding and also, we use *Term Frequency Inverse Document Frequency* (TF-IDF). These words representation is used to *Deep Neural Network* (DNN) classifiers inputs. For fastText embedding we use *Convolutional Neural Network* (CNN) and for BERT embedding we use *Bidirectional Long Short-Term Memory* (BiLSTM) classifiers. Further we compare the individual features of these methods with the combined features for performance analysis.

FastText and BERT embedding is a feature-based approach. In feature-based performance two step. Firstly, every comment is represented as a sequence of words or word piece. This embedding sequence will be the DNN classifier input. CNN and BiLSTM are used as the DNN classifiers.

The paper is organized as follows. Section II explains the methods of different types of toxic speech detections. Section III illustrates the proposed system classification of toxic speech by combined features of BERT, FastText Embedding and TF-IDF using deep learning. Section IV discusses performance evaluation. Finally, section V concludes the paper.

II. RELATED WORKS

Due to the increase in the use of online platform, spread of online toxic speech also increased immensely. Thus, it is important to automatically detect and remove the toxic speech in online platform. There have been several studies for the same. In this section we discuss about the previous methods used for detection of online toxic speech.

A state-of-the-art method for detecting abusive language in user comments by using a supervised classification methodology with NLP features to outperform a deep learning approach [2]. *Global Vectors for Word Representation* (GloVe) [4] used for analyse the model properties to necessary to

produce linear directions of meaning and argue the global log-bilinear regression models are appropriate for doing so.

Another powerful technique for classifying toxic speech is deep learning [5]. Random embedding as input to DNN classifiers has been compared. Local patterns in text are captured by CNN [8] and long-range dependencies are captured by *Long Short Term Memory* (LSTM) [9] model. Sentence embedding [7] used as input to classifiers for toxic comment classification. *Embeddings from Language Models* (ELMo) [6] give a deep representation of words formed on output of a three-layer pre-trained neural network.

In this paper, we study two case: (a) binary classification: considered two classes toxic versus non-toxic speech; (b) multi-class classification: considered three classes hate speech, offensive speech and
neither. We proposed an effective methodology to automatically detect online toxic speech by using TF-IDF and two effective word representations namely BERT and fastText embedding.

III. PROPOSED ARCHITECTURE AND IMPLEMENTATION SETUP

A. Data set

For toxic classification, we use Kaggle twitter corpus. The data set contain 25296 tweets comments. At least three people comment on each tweet. There are three classes namely *hate speech*, *offensive language* and *neither*.

During our experiments, we grouped hate-speech and offensive speech into a single class during **binary classification** (toxic speech). As a result, we have a toxic speech class. as well as a non-toxic speech class. We use a **multi-class classification** scheme. Make use of the three classes and labels included in the data set: There are three types of speech: hate speech, offensive speech, and neither.

B. Text preprocessing

Text preprocessing contain many steps depend on the task and the given text for example segmentation, cleaning, normalization, analysis etc. We want to remove all the special characters.

We perform the preprocesing steps. Before removing the special character, we want to tokenize the document. Next, we want to remove the special characters like '@', '!' etc. Then we want to remove the stop words like 'a'. 'the', 'is' etc. Finally convert into lower case.

C. TF-IDF

Tf-idf weighting scheme mainly used for scoring and ranking of term relevance in a given document. This weight is a statistical measure used to find how important is to a document in a corpus. The weight is based on its *term frequency* (tf) and *inverse document frequency* (idf). The highest weight score of word considered to be more important.

The tf-idf value of a particular term *t* in a given document *d* across all documents in the corpus *D* will be the product of tf and idf. Thus, the term frequency can be calculated as the frequency of the term *t* is present in a given document *d*. Then tf(t,d) can be calculated as:

$$tf(t,d) = \frac{frequency\ of\ term\ t\ in\ d}{total\ no\ of\ terms\ in\ d} \quad (1)$$

Inverse document frequency can be calculated as the inverse fraction of the number of documents containing the term *t*. Then idf(t,D) can be defined as:

$$idf(t,D) = \frac{|D|}{|d \in D, t \in d|} \quad (2)$$

Thus, the tf-idf value of a term *t* in a document *d* in the given corpus *D* becomes:

$$tf - idf(t,d,D) = tf(t,d).idf(t,D) \quad (3)$$

D. FastText Embedding

FastText embedding is a feature-based model [12]. FastText embedding created by Facebook team for efficient learning of sentence classification and word representation. FastText is a extension to Word2Vec which will breaks words into several n-grams. Where each word contains bag of words or characters [10]. In fastText, it helpful to find the vector representation of rare words which may not be present in the dictionary. For a given comment we generate one embedding for each word by using a pre-trained fastText embedding model. Figure 1 shows the simple fastText model.

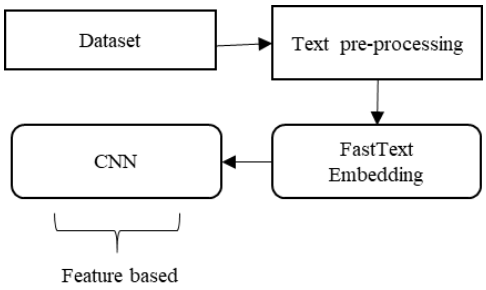

Fig. 1. Simple fastText embedding model.

E. BERT

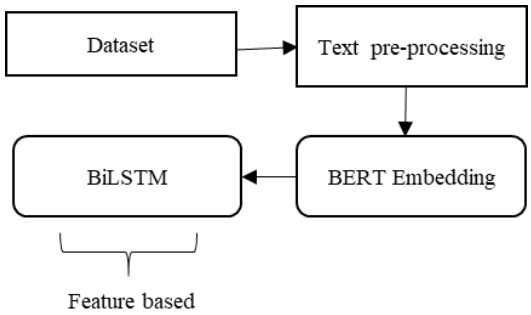

Fig. 2. Simple BERT embedding model.

Bidirectional Encoder Representations from Transformers (BERT) is one of the most powerful word representation model [11][12]. Which is based on the technique of uses attention mechanism and transformers. BERT is designed to condition both left and right background in all layers to pretrain deep bidirectional representations from unlabelled text. BERT is both conceptually and empirically powerful. On eleven natural

Identify applicable funding agency here. If none, delete this text box.

language processing tasks, it obtains new state-of-the-art results. Figure 2 shows a simple BERT model. BERT uses masked language models to allow deep bidirectional representations that have been pre-trained.

It's important to remember that depending on the meaning, the same word may have different embeddings Bert can overcome this. Like fastText, Bert can embedding rare words.

F. Architecture

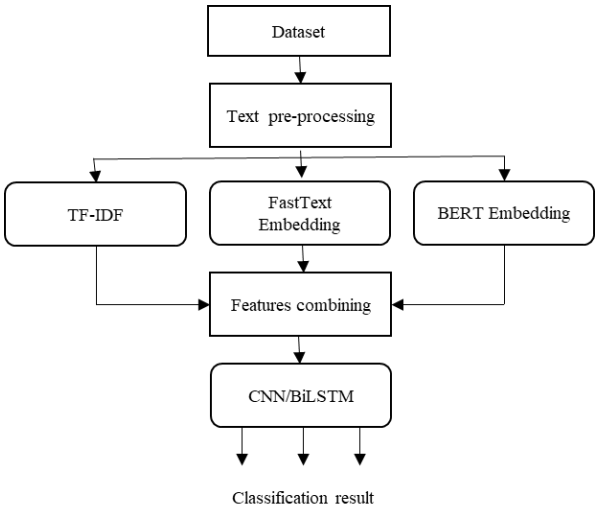

Fig. 3. Proposed methodology

The overall architecture is shown in figure 3. We proposed a methodology by combining the three technique features together that is TF-IDF, fastText and BERT embedding. Thus, can improve the efficiency of detection. In the present system there is a confusion between hate speech and offensive speech. This problem can be overcome by the proposed system. The classification results consist whether the given data is a offensive speech, hate speech or neither.

IV. PERFORMANCE EVALUATION

The macro-average F1-measure is used to determine the efficiency of our approaches. The F1-measure is a statistical metric for determining classification accuracy. This number ranges from 0 to 1, with 1 indicating the best results. The following formula is used to determine the F1-measure:

$$f1 = \frac{2*(precission*recall)}{(precision+recall)} \quad (4)$$

where precision is the ratio of correctly predicted class A samples to total number of samples predicted as class A by the classifier, and recall is the ratio of correctly predicted class A samples to total number of samples that should be predicted as class A by the classifier.

$$macro\ f1 = \frac{1}{c}\sum_{i=1}^{c} f\ 1_i \quad (5)$$

Where c is the average no of classes.

Table I gives the macro average F1 result for binary classification task using CNN and BiLSTM classifiers. We give the input features as fastText and BERT embedding and also the combined features of TF-IDF, BERT and fastText. Table II shows the macro average F1 result of binary classification.

It is clear from the table I and II that fastText and BERT embedding have almost same result. But when we combine the features of TF-IDF, BERT and fastText embedding have better performance. And also, when we compare table I (binary classification) and table II (multi-classification), binary classification has higher performance. Bi-LSTM performs marginally better than CNN among the classifiers in both cases.

TABLE I. MACRO-AVERAGE F1-MEASURE FOR DIFFERENT CLASSIFIERS AND DIFFERENT EMBEDDINGS. BINARY CLASSIFICATION

	CNN	**Bi-LSTM**
fastText Embedding	91.25	92.51
Bert Embedding	92.59	93.04
TF-IDF + fastText embedding + BERT embedding	95.42	96.02

TABLE II. MACRO AVERAGE F1-MEASURES FOR DIFFERENT CLASSIFIERS AND DIFFERENT EMBEDDINGS. MULTI-CLASS CLASSIFICATION

	CNN	**Bi-LSTM**
fastText Embedding	71.25	72.51
Bert Embedding	72.59	73.04
TF-IDF + fastText embedding + BERT embedding	78.34	79.69

TABLE III. CONFUSION MATRIX FOR FEATURE-BASED BI-LSTM WITH BERT EMBEDDINGS . MULTI-CLASS CLASSIFICATION

True Label	**Hate**	**Offensive**	**Neither**
Hate	29	29	4
Offensive	98	1838	96
Neither	16	52	316

TABLE IV. CONFUSION MATRIX FOR FEATURE-BASED CNN-FASTTEXT WITH BERT EMBEDDINGS. MULTI-CLASS CLASSIFICATION

True Label	**Hate**	**Offensive**	**Neither**
Hate	1042	51	2
Offensive	321	18912	82
Neither	67	227	4079

A confusion matrix is a table that shows how well a classification model performs on a collection of test data for which the true values are known. Table III and table VI shows the confusion matrix based on CNN and BiLSTM.

CONCLUSION

In this paper, we introduce a new methodology for online toxic speech detection by using combined features from TF-IDF, fastText embedding and BERT embedding. The

classification is based on binary classification and multi-class classification. Binary classification consists toxic and non-toxic speech, and multi-class classification consist offensive speech, hate speech and neither. BERT and fastText embedding is a feature-based method. The embedding features are input to the classifiers namely CNN and BiLSTM.

There was confusion between hate speech and offensive speech in the current method, that problem can be overcome by using the proposed methodology. In future work, we can combine the proposed system with student sentimental analysis method to get a new single model.

REFERENCES

[1] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding Abuse: A Typology of Abusive Language Detection Subtasks," in Proceedings of the First Workshop on Abusive Language Online, 2017, pp. 78–84.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] M. Wiegand, J. Ruppenhofer, A. Schmidt, and C. Greenberg, "Inducing a Lexicon of Abusive Words–a Feature-Based Approach," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 2018, pp. 1046–1056K. Elissa, "Title of paper if known," unpublished.

[3] R. Delgado and J. Stefancic, "Hate Speech in Cyberspace", Social Science Research Network, Rochester, NY, 2014..

[4] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation", in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[5] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets", in Proc. 26th Int. Conf. World Wide Web Companion - WWW 17 Companion, pp. 759–760, 2017.

[6] M. Bojkovský and M. Pikuliak, "STUFIIT at SemEval-2019 Task 5: Multilingual Hate Speech Detection on Twitter with MUSE and ELMo Embeddings", in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 464–468.

[7] V. Indurthi, B. Syed, M. Shrivastava, N. Chakravartula, M. Gupta, and V. Varma, "FERMI at SemEval-2019 Task 5: Using Sentence embeddings to Identify Hate Speech Against Immigrants and Women in Twitter", in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 70–74.

[8] Y. Kim, "Convolutional Neural Networks for Sentence Classification", in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746–1751.

[9] A. Baruah, F. Barbhuiya, and K. Dey, "ABARUAH at SemEval-2019 Task 5: Bi-directional LSTM for Hate Speech Detection", in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 371–376

[10] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "FastText.zip: Compressing Text Classification Models", ArXiv Prepr. ArXiv161203651, 2016

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186

[12] Ashwin Geet D'Sa, Irina Illina, Dominique Fohr," BERT and fastText Embeddings for Automatic Detection of Toxic Speech " in theUniversité de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France.