2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)
Pune, India. Oct 29-30, 2021

2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) | 978-1-6654-2503-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SMARTGENCON51891.2021.9645877

A Context Aware Embedding for the Detection of Hate Speech in Social Media Networks

Sanjana Kavatagi
Assistant Professor
AITM, Belagavi
kawatagi.sanjana@gmail.com

Dr. Rashmi Rachh
Associate Professor
VTU, Belagavi

***Abstract*- Proliferation of social media platforms in recent past has resulted into upsurge in the number of users. Advent of these sites have paved way for the users to easily express share and communicate. In such a scenario, it is imperative to analyze the content and identify nasty content so as to avoid unpleasant situations. Machine learning techniques are extensively used for this purpose. In this paper, we propose a language model for the identification of hate speech in twitter data. Distil-BERT, a context aware embedding model along with Support Vector Machine (SVM) for the classification of hate speech has been used. SVM with a 10-fold cross validation and linear kernel has been found to provide better accuracy as compared to existing models. Results show that accuracy is improved with the use of context aware embedding model.**

Key words: hate speech, Distil-BERT, Support vector machine, Cross validation

I. INTRODUCTION

The extensive use of social media by the people is becoming a large platform to share their opinion on the internet. Social media connects millions of people with different background, religion and culture. People tend to discuss everything that is happening within the entire world whether it may be related to some religion, racism, politics, government or even about any small posts made by the common man. In discussion of such topics on social media like Twitter or Facebook which a large network of people, knowingly or unknowingly they make use of hate speech.

Hate speech pertains to a statement in which the individual or several people who belong to the group are denigrated, which is typically represented by social class, ethnic background, gender preference, gender expression, impairment, religious belief, political persuasion or opinions[1]. This hate speech can be a threat to an individual or group, it can also be abusing somebody or something online and cyber bullying. It can be in the form of words, images or videos that call for or glorify violence against a group. And these may encourage others to commit hate crimes.

Detection of such hate speech is an interesting topic of research. Several researchers have carried out extensive study on identification of these with different techniques. It was clear that manual detection and labeling these types of texts is difficult as it requires man power also the cost of time. It is also known that these types of texts contain noise and ambiguity to label manually. To avoid these issues researchers have formulated automated method for detection of hate speech. They include mainly two categories: first by making use of propagation networks. Second with content within these texts. Propagation networks have hierarchical structures one at the macro level where these propagation network illustrates the path of specifying hate speech from the person writing it on the twitter to other tweets that repost these tweets. On the other hand, micro level propagation network demonstrates the conversation of users on the tweets in the form of replies or comments [2-3]. When content of the tweets is considered for the detection of hate speech, several neural networks and deep neural networks approaches have been proposed as a way to learn automatically and fuse different multimodal features [4]. This text data can be modelled as a time series, later with a convolutional neural network or recurrent neural network the latent textual representation is learned and the accuracy when compared to traditional classifier-based methods shows a greater improvement.

In section II we have described literature survey, section III contains methodology, results are discussed in section IV. In the last section, we describe the conclusion and future scope.

II. LITERATURE SURVEY

In preventing text posting or blacklist users, simple wordbased approaches do not only fail to identify delicate offensive content, it also affects the freedom of speech and expression. The high false positivity rate is mainly due to the problem of the word ambiguity – a word can have different meanings in different contexts. In spite of their popularity[5], regular NLP approaches are cumbersome to detect extraordinary orthography in the comments generated by users. It's just a problem of orthographic variation. This is due to the deliberate or careful substitution of individual characters in a token to distort the detectors. The complexity of natural language structures generally makes the task difficult.

978-1-6654-2503-2/21/$31.00 ©2021 IEEE

Regardless of the application of NLP approaches, in the existing hate speech problem, two main categories exist: unsupervised learning and supervised learning.

In the beginning, n-gram with a supervised classification, regular expression, and contextual characteristics were used in abusive language identification [6]. n-grams with other character functionalities to detect various forms of foul language and intimidation was used [7]. Used twitter user profiles to distinguish militant communities through an analysis of links among members of militant communities based on cross-country interactions [8]. For racist identification, Kwok et al. used BOW and Naive Bayes [9]. Wulczyn et al. used CrowdFlower on Wikipedia text to obtain human annotations (individual attacks and bullying) [10].

Classification of sentiment based on aspects is used to classify opinions as positive or negative in significant ways. In the aspect-based classification of sentiment, the primary task is to determine the opinion words and the polarity of those words towards their context. Extraction of adjectives, verbs and adverbs from the tagged phrases is used as a method by the researchers [11-14]. The language rules in this method [15] were used to determine the word opinion guidelines. In this process, a group of adjectives are also retrieved near every aspect rather than using the semantic rule. The techniques of word embedding overlap the lexical boundary, because word representation can also capture semantical and syntactic aspects. Two recognized and accepted methodologies used as a real value vector in the encryption of words are Word2Vec[16] and GloVe[17]. Word2Vec is a model that comprises the continuous bag of word (cbow) and skip-gram architectures of the "artificial neural network predictive" model. GloVe is a "count-based" model, on the other hand. The current word is taken from the context words of the CBOW model. Instead, the model Skip-gram uses the current word to estimate its context words. Otherwise, the GloVe model uses a combined global word-word corpus co-occurrence statistics. Thus, Word2Vec and GloVe, are the two approaches that are capable of encrypting various language aspects. Moreover, if different corpora are used, then different word vectors will appear to be created, in which each vector consists of highly complementary information regarding others.

The Deep Neural Networks (DNNs) have been widely explored to handle NLP tasks, because they can identify data representations useful for classification. The two major DNN architectures that NLP has taken advantage of are Convolution Neural Nets (CNNs)[18] and Recurrent Networks (RNNs)[19]. CNNs are suitable for periodically sampled multi-dimensional input data that incorporate several of the nearby inputs into one of the network's next layer. In the course of a training process, the RNN can be considered to add loops to the architecture to update the network weights in each stratum. Long Short-Term Memory Networks (LSTMs) are special RNNs that arbitrarily spread signals over the network and are therefore susceptible to the range of values [20].

As discussed in above section, various models have been used for identification of hate speech.

In this paper, we have explored context aware embedding techniques for the tweets using the content-based approach.

III. METHODOLOGY

The proposed methodology is described as follows:

A. Data set:

In this paper we have taken the dataset of twitter which is consisting of sentence level annotation of textual hate speech. Also, annotation at the sentence level work with the hate speech contained in the minimum unit and it has the capacity of reducing the noise introduced by other clean sentences [21].

A dataset consisting of 4000 tweets was considered from Stromfront. The tweets are classified as ‘hate’ or ‘non-hate’ which can be identified by the labels assigned to them. Label 1 is given for hate speech and label 0 is given for non-hate speech. The dataset is consisting a total number of 1430 tweets of hate type and 2570 tweets of non-hate type. Hate speech tends to occur at a very lower rate than non-hate speech on a social media platform like Twitter. Even though the data sets signify this imbalance to some aspect, because of training requirements they do not map the actual figure [22].

B. Preprocessing:

The next step is to pre-process the data after data collection. It must be trained to use the data in any model. Preprocessing is a part of training the data. We should provide clean data for our model as an input. We imply by the word clean that there must be no redundant information, i.e. all duplicate data should be removed. And if it has null data to be deleted because we are not processing null data. Data cleaning steps involves :- lower casing: all words are converted to lower case to avoid creation of two different vectors for the same word, deletion of HTML characters: the information that is generally obtained contains many HTML characters which is combined with the original data, data decoding: the complex symbols employed in the data must be decoded to make things simple, and the information must be decoded in standard form to achieve better analytical results, punctuation marks to be deleted: any punctuation marks and apostrophes that can result in word ambiguity may be deleted from the data, remove stop words: since the data analysis must take place at the word level, in this step it is essential to delete the most frequently encountered words that are stop words.

C. Padding and Masking:

Sentences of fixed length is given as input to the DistilBERT model. The sentence of maximum length is identified in the dataset and the maximum length of sentence differs with different dataset. Then we have added paddings of value 1 in sentences that are shorter than that maximum length to make up the length.

After padding, all the tweets are of same length which can be passed as input to the model. By making use of masking layer the padded data is made as value 0 ignored by the DistilBERT embedding model.

D. Embedding:

For embedding, we use the DistilBERT model in this paper. DistilBERT is a lighter version of the general-purpose BERT that has 97 percent language understanding capabilities, 60 percent faster speed, and 40 percent smaller size than the original BERT model. This general-purpose language model is trained successfully with distillation.

The architecture of our model is shown in the Fig1

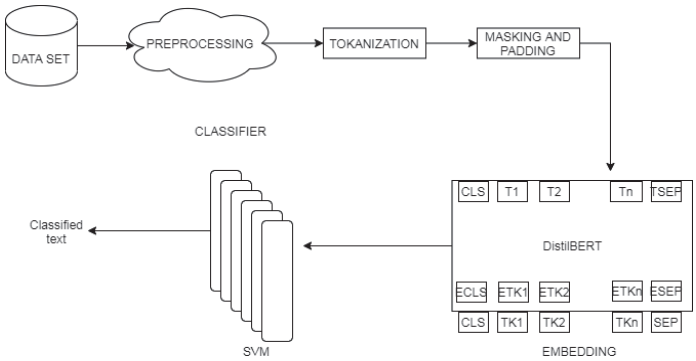

Fig. 1. Architecture of the model

We have implemented BERT model for the detection of hate speech. This bidirectional encoder model is pre trained by the twitter dataset. We have considered max length of tweets as 60 characters for training the model. All other sentences with shorter length than this are padded. We have used BERT tokenizer in the implementation of the model. This model is consisting of 12 layers of transformer block. Initially each token was assigned by a token embedding, BERT consists of [CLS] token which is added at the beginning of every statement in the input and it contains the special type of embedding. The last hidden layer of BERT corresponding to this token is used to represent the sequence aggregated for classification purposes. To separate two sentences a [SEP] token is used. BERT model is making use of 748 dimensions for the embedding. Then every token is supplemented by an embedding segment to determine which statement it belongs to. Atlast, a position with learned embedding is added to make the model aware of information.

E. Classification:

After the embedding process, this data is passed to the classifier. We are making use of Support Vector Machine(SVM) with linear kernel which makes the leaning process faster for the classification of hate speech and non-hate speech. Cross validation of SVM is done with 10-fold. The training data in this model is split into 10 smaller sets and each data point is assigned to one of the smaller sets which are almost of same size. For each of these smaller sets an individual model is built by applying the method to the training data set. Then the average of the outcomes of all these model evaluation gives the cross-validation value. The resulted model is used as the test set for the computation of performance measures like accuracy. In k-nearest neighbor, with Support vector machine we have tested 24 different values for n-neighbors and we have developed the grid matrix according to it.

IV. Results and Discussion

In this section we first describe the evaluation parameters and results obtained from our context aware Distil-BERT embedding model and then we have shown the comparison of results with other traditional models used for hate speech detection.

Hate speech detection is a binary classification problem. We have used two parameters: Accuracy and F1 score for the testing of our model. We describe the evaluation parameters as follows: 'S' indicates the phrase 'hate', and if the phrase is only 'hate', it is counted as true positive (TP) value. 'T' is used to indicate when the phrase is 'non-hate', but if the phrase is 'hate' in real, it is considered as false positive value (FP). 'U' is used to describe 'hate' phrase, if its 'non-hate' in real it is accounted as false negative (FN) value. 'V' shall indicate the phase 'non-hate' even when the real phrase is 'non-hate', it is considered as true negative (TN) value.

Then we have used the standard performance metrics-precision (P), recall (R) and accuracy (A) as follows:

$$\text{Precision (P)} = S/S+T \quad (1)$$

$$\text{Recall (R)} = S/S+U \quad (2)$$

$$\text{Accuracy} = S+V/N, \quad (3)$$

where $N=S+T+U+V > 0$

$$\text{F-1 Score} = 2*P*R/P+R \quad (4)$$

In the implementation of our model, the classifier we have used is support vector machine with cross validation to classify the tweets containing hate speech and those that contains non hate speech. We have split up the dataset into training dataset and testing dataset. 70% of dataset is considered as training dataset and 30% of dataset is considered as testing dataset.

With this for the dataset we have considered with Distil-BERT embedding and Support vector machine with 10-fold cross validation model we have obtained F-1 score for both hate speech as well as non-hate speech. F-1 score for hate speech is 76% and the same for non-hate speech is 89%. We have achieved an accuracy of 85% and this result is compared with the accuracy achieved by other traditional embedding methods as shown in the table below.

TABLE I. COMPARISON OF PERFORMANCES OF VARIOUS MODELS

Citations	**Method**	**F-1 score in %**	**Accuracy in %**
Yu Quing Lim, et al[23]	IF-TDF	71	71
	Word2Vec	76	77
Sean MacAvaney, et al[22]	ELMo	73	73
Nur Indah Pratiwi, et al[24]	Fasttext	48	52
Akanksha Bisht, et al[25]	GloVe	69	69
Our model	**Distil-BERT**	**84**	**85**

Thus, by comparing the results of our approach with other state-of-the-art embedding models which do not make use of

the context for embedding, we say that our context aware embedding model has achieved improved accuracy.

V. Conclusion and Future Scope

Detection of hatefulness within the texts in an efficient manner can help the social media platforms to prevent these types of chaos arising.

In the current work, we have used a language model with a context aware embedding method called Distil-BERT for the detection of hate speech in twitter data. This embedding technique using DistilBERT model has been able to provide better results compared to the previous techniques like GloVe and Word2Vec embedding methods used by other researchers. This approach has shown improved accuracy when compared to other state-of-the-art models which made use of context unaware embedding techniques. Further, there is huge scope for development of multi-lingual language models for the detection of hate speech in social media.

References

[1] W. Warner and J. Hirschberg, 'Detecting hate speech on the world wide web," *Proceeding LSM '12 Proc. Second Work. Lang. Soc. Media,* no. Lsm, pp. 19-26,2012.

[2] Shao, C., Ciampaglia, G.L., Varol, O. et al. The spread of low-credibility content by social bots. Nat Commun 9, article no. 4787 (2018).

[3] Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science* 359(6380):1146–1151.

[4] Gilbert, C. H. E. Vader: A parsimonious rule-basedmodel for sentiment analysis of social media text. Vol. 8 No.1(2014): Eighth International AAAI Conference on Web blogs and Social Media, pp 659-688.

[5] Schmidt A, Wiegand M A survey on hate speech detection using natural language processing. In Proceedings of the 5thinternational workshop on natural language processing for socialmedia. Association for Computational Linguistics, pp 1–10, 2017.

[6] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. Detectionof harassment on web 2.0. Proceedings of the Content Analysis in the WEB, 2:1-7, 2009.

[7] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting o_ensive language in social media to protectadolescent online safety. In Proceedings of the 2012 ASE/IEEE International Conference on Social Computingand 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, SOCIALCOM-PASSAT'12, pp 71-80, Washington, DC, USA, 2012.

[8] Derek O'Callaghan, Derek Greene, Maura Conway, Joe Carthy, and P_adraig Cunningham. An analysis ofinteractions within and between extreme right communities in social media. In Ubiquitous social media analysis,pp 88-107. Springer, 2012.

[9] Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In Twenty-seventh AAAIconference on artificial intelligence, 2013.

[10] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In Proceedingsof the 26th International Conference on World Wide Web, pp 1391-1399. International World Wide WebConferences Steering Committee, 2017.

[11] W. Zhang, H. Xu, W. Wan, Weakness finder: Find product weakness from chinese reviews by using aspects-based sentiment analysis, Expert Systems with Applications (2012) pp 10283 – 10291.

[12] H. H. Lek, D. Poo, Aspect-based twitter sentiment classification, in: Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on, 2013, pp. 366–373.

[13] J. Jmal, R. Faiz, Customer review summarization approach using twitter and sentiwordnet, in: Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, WIMS '13, ACM, New York, NY, USA, 2013, pp. 33:1–33:8.

[14] H. Kansal, D. Toshniwal, Aspect based summarization of context dependent opinion words, Procedia Computer Science 35 (0) (2014) 166 – 175, knowledge-Based and Intelligent Information & Engineering Systems 18th Annual Conference, KES-2014 Gdynia, Poland, September 2014 Proceedings.

[15] E. Marrese-Taylor, J. D. Vel squez, F. Bravo-Marquez, A novel deterministic approach for aspect-based opinion mining in tourism products reviews, Expert Systems with Applications 41 (17) (2014) 7764– 7775.

[16] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, CoRR, 2013.

[17] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014,Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1532–1543.

[18] Yann LeCun, Leon Bottou, Yoshua Bengio, Patrick Haffner, Gradientbasedlearning applied to document recognition, Proceedings of theIEEE, 2018, pp 2278–2324.

[19] Marzeih Mozafali, RezaFarah Bhaksh and Noel Crespi,"A BERT-Based Transfer learning Approach for Hate Speech Detection in Social Media", in proceedings of the International conference on Complex networks and their Applications, pp 928-940, 2019.

[20] Georgios K. Pitsilis, Heri Ramampiaro, Helge Langseth, Effective hatespeechdetection in Twitter data using recurrent neural networks,Department of Computer Science, Norwegian University of Science andTechnology (NTNU), NO-7491 Trondheim, Norway, 03 Jul 2018.

[21] Ona de Gilbert, Naiara Perez,Aitor Garcia-Pablos, Monte Cuads, " Hate Speech Dataset from a White Supremacy Forum, 2018,arXiv:1808.0444v1.

[22] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian,Ophir Frieder, "Hate speech detection: Challenges andSolutions", Plos one,2019.

[23] Yu Quing Lim, Chun Ming Lim, Keng Hoon Gan, Nurhana Samsudin, "Text sentiment Analysis on Twitter to Identify Positive or Negative Context in Addressing Inept regulations on Social Media", in IEEE 10th Symposium on Computer Applications and Industrial Electronics (ISCAIE), 2020.

[24] Nur Indah Pratiwi, Indira Budi, Ika Alfina,"Hate Speech Detection on Instagram Comments using FastText Approach", in International Conference on Advanced Computer Science and Information Systems(ICACSIS), 2018.

[25] Akanksha Bisht, Annapurna Singh, H.S. Bhaduria, Jitendra Virmani and Kriti,"Detection of Hate Speech and Offensive Language in Twitter Data using LSTM model", in Recent Trends in Image and Signal processing in Computer Vision, pp 243-264, 2020.