International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 6, Nov-Dec-2020, ISSN (Online): 2395-566X

Online Toxic Speech: Automatic Detection Methods and Techniques

Asmi P, Sanaj M S
Dept. of Computer Science & Engineering
MGM College of Engineering and
Pharmaceutical Science, Kerala, India
azzmi287@gmail.com, mssanaj@gmail.com

***Abstract* – Nowadays, due to the increase of social media individuals are freely to communicate and also to express their thoughts of viewspublicly.It may be a text including blog post, updated status or comments posted on the social media. Some of them misuse the freedom of speech by harassingothers. With theexponential growth of online communication, the hate speech behavior also goes unchecked to alarming proportion. Speech may be normal, hate or offensive. Hate speech or toxic speech is an antisocial behavior. Thus, it is important to detect and remove the toxic or hate speech from the social media. Which is also a challenging problem. This paper givesa brief explanation about different toxic speech detection methods.**

***Keywords*– Natural Language Processing, Hate Speech, Deep Learning.**

I.INTRODUCTION

In social media, detecting toxic speech is a challenging problem. Definition of hate speech as language that's wont to expresses emotion towards a targeted cluster or it meant to be derogative, to bruise, or to insult the members of the cluster. Toxic speech may be hateor offensive. Toxic comments square measure quite common and have negatively impacts the dynamics of the web community still because the user experiences of the targeted people. The identification of hate speech is the sub filed of Natural Language Processing that studies the automated abstract thought of offensive language and hate speech from textual information.

Hate speech has specific target. Humor also sometimes be considered as hate speech. Some toxic can be identified by using the lexicons of the speech where as others can only identify by analysing the semantic of a speech,In many counties hate speech is punishable. Toxic speech may be targeted towards an individual or a company or to a specific community. The project is mainly concentrated on text classification models.Natural Language Processing, (NLP), which is a part of artificial intelligence that handle with the interaction between computers and user (human) using the natural language. A definitive target of NLP is to read interpret, understand, and make feeling of the human language in a way that is important. Text characterization is a significant issue in NLP. The proposed approach for automatic toxic detection will convert the entire text into numerical form after feature extraction and on them using some classification models. Text cleaning is a difficult step in text classification technique Which remove the implicit noise such as stop words, stemming etc and allowing a informative featurization. The simplest form of feature extraction is weighted words. Which consider the semantic meaning. Where as in word embedding feature extraction both semantic as well as syntax will considered. The paper is categorized as follows. Section II describes the literature survey of the previous methods which is explaineddifferent text classification methodsto detect various kind of speech detection technique. Section III explain the proposed method. Finally, the Section IV gives the conclusion.

II.LITERATURESURVEY

In recent, hate speech study is a point of view in computer science. This survey describes the current state of field of previous approaches, methods and main features used. There have been several studies about the different methods of text classification methods in different speeches like abusive, hate or offensive.

1.Abusive Language detection in Online user content
The proposed method [1], develop a state-of-the-art ethod for detecting abusive language in user comments, while also addressing the above deficiencies in the field. They develop a supervised classification methodology with NLP features to outperform a deep learning approach. They use and adapt several of the features used in prior art in an effort to see how they perform on the same data set. And also extend this feature set with features derived from distributional semantics techniques. They make public a new data set of several thousand user comments collected from different domains. Three judgments per

comment included in this set and for comments which are labeled as abusive, a more fine-grained classification on how each is abusive. Fixed and static data set was evaluated during work. Which provide curated public dataset and NLP features has been evaluated Different types of embedding as well as syntactic features are experimented. And then these features combine with powerful standard NLP features. The model is based on deep learning to avoid retrain embedding during every iteration. After that the model uses to perform an hate speech analysis over one year. Providing how much and the important data is necessary data for the task. The method focused on abuse words on English and show powerful n grams features were in English. Thus, require enough training data. Due to more than simple keyword spotting, Difficult to track all racial and minority insults detecting abusive language is more difficult. And also, Abusive language is more fluent and grammatically.

2. Inducing a lexicon of abusive words:Feature-based approach

In this paper [2], it addresses the task of detecting abusive words for example dumbass, bimbo, scum etc. The main assumption of in this paper is that abusive words form a subset of negative polar expression. And calibrate a supervised classifier to study various features. which is trained on that base lexicon which contains 551 abusive word then applied to a very large list of unlabeled negative polar expressions to extract an expanded lexicon of 2989 abusive words. New abusive words constantly enter natural language.
They present the first work that systematically describes the automatic construction of a lexicon of abusive words. Next examine novel features derived from various textual resources. And show that the knowledge learn can't be equally derived from a an outsized dataset with labeled microposts. The effectiveness of our expanded lexicon is demonstrated on cross-domain detection of abusive microposts. This is also the first work to address this task in general. Which is an binary classification problem. The given expression is to be classified as either abusive or not.We study this problem on English. Thus many of the features should also be applicable to other languages.

3. GloVe: Global Vectors for Word Representation
In this work, [3]they analyze the model properties necessary to produce linear directions of meaning and argue that global log-bilinear regression models are appropriate for doing so. In unsupervised methods for learning word representations the primary source of information is available from the statistics of word occurrences in a corpus. In this they merge the regularity vectors to make explicit the model properties. A new global log bilinear regression model was the result. For training on global word-word co-occurrence counts they proposed a specific weighted least squares model and thus make efficient use of statistics.The model creates a word vector space with important foundation, as proven by its best in class execution of 75% precision on the word similarity dataset. It also outperforms related models on similarity tasks and named entity recognition. Likewise,they show that their techniques beat other current techniques on a few word comparability errands, and furthermore on a typical named element acknowledgment (NER) benchmark.

3.Deep learning for hate Speech Detection in tweets
Detecting hate speech on twitter is difficult for applications. In this paper [4] authors defined the task as being able to classify a tweet as racist, sexist or neither. The task is very challenging due to the complexity of the natural language constructs. To manage this complexity, perform extensive experiments with multiple deep learning architectures which will learn semantic word embedding.Logistic Regression, Random Forest, SVMs, Gradient Boosted Decision Trees (GBDTs) and Deep Neural Networks (DNNs) are the multiple classifiers used in this experiment.

The paper donates three things. Firstly, for hate speech detection task they investigate the application of deep learning methods. Secondly, explore various tweet semantic embeddings forexample Global vectors, Bag of words,Term Frequency-Inverse Document Frequency, FastText, CNN and LSTM and finally the proposed methods beat state of the art methods by a large margin. The best accuracy values obtained from Embeddings learned from deep neural network models when combined with gradient boosted decision trees. In this method user network features are not included.

4. Multilingual Hate Speech Detection on Twitter with MUSE and ELMo Embeddings

The proposed method [5]assess the feasibility multilingual learning for the task of hate speech detection and also investigate with the adversarial learning as a means of creating a multilingual model. Word embeddings is the most popular input features in deep learning. MUSE and ELMo word embeddings are the two models used in the input layer.MUSE embeddings available in different languages and which is based on FastText. ELMo give a deep representation of words formed on output of a three-layer pre-trained neural network. The representation for a word is based on the context in which the word is used.

That the standard of word representations used incorporates a vital impact on the performance of a model. dynamic between MUSE and ELMo resulted in a very three - four-dimensional increase in accuracy even once MUSE based mostly models may gain advantage from multilingual coaching. The discourse nature of ELMo representation build them far more flexible and fewer domain unnatural than ancient word embeddings.

6.Supervised Classifiers to Identify Hate Speech on English and Spanish Tweets

In this paper [6] authors examine the performance of various supervised classifiers in the process of identifying hate speech on Twitter. The main objective of this paper is to look at the effectiveness and limitations of supervised classifiers to spot hate speech detection in twitter focused on 2 specific targets, girls and immigrants in 2 languages: English and Spanish.This paper contributes a report on a wide set of experiments aimed at calculating the effectiveness of the e most influence linguistic features in a supervised classification task. Concerning the correlation between machine learning techniques in this specific undertaking, Support Vector Machine, complement Naive Bayes, and Random Forest unmistakably beats all the rest classifiers and show stable execution with all features. One of the limitations of this paper is that comparison of other types of classifiers with additional compound linguistic features, by considering the new deep learning approaches based on neural networks.

7. Bi-directional LSTM for Hate Speech Detection

The proposed algorithm [7], outcome obtained by using a Logistic Regression (LR) and bi-directional long short-term memory (BiLSTM) with and without attention models. Attention and without attention mechanism is completely same except the global max pooling layer was replaced in attention mechanism. For the first test BiLSTM model without attention performed best where as in second test LR model with character n-grams performed best. This paper presents the outcomes got for Subtask A for English language. The result shows that the BiLSTM model obtained an F1 score on the test set and get an formal ranking of 8/71.

III. PROPOSED SYSTEM

In the literature survey we explained various existing methods of hate speech detection and also word representation technique. some of the papers can only detect abusive words and some other can find hate speech. Such that our proposed method will automatically detect whether a given online text is hate, offensive or neither. In this proposed approachesperform toxic speech classification and combining the features by using three methods namely Termfrequency and Inverse document frequency (TF-IDF), FastText embedding and BERT Embedding by using a deep neural network classifier such as CNN and BiLSTM. The proposed Embedding methods is very powerful classification technique. And also, we performed BERT pre trained fine-tuning model

IV. CONCLUSION

In this paper, we proposed automatic classification of toxic speech using powerful word Embedding representations and deep learning techniques. The classification performs binary as well as multi-class using twitter corpus or Facebook. Classification occur between toxic and non-toxic speech. In non-toxic it will be considered hate speech, offensive speech and neither.

REFERENCES

[1]. C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, “Abusive Language Detection in Online User Content”, in Proceedings of the 25th International Conference on World Wide Web, Republic and Canton of Geneva, Switzerland, 2016, pp. 145–153.

[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. 1046–1056.

[3]. 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.

[4]. 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

[5]. 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.

[6]. https://link.springer.com/chapter/10.1007/978-3-030-34058-2_3

[7]. 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.

[8]. https://www.aclweb.org/anthology/W17-1.pdf

[9].https://web.stanfrd.edu/class/archive/cs/cs2n/cs224n.1194/reports/custom/15744362.pd.

About Author

Asmi Preceived the B.Tech degree in Computer Science and Engineering from University of Calicut, Kerala, India, in 2014. Presently, she is doing her Post Graduation in the Department of Computer Science and Engineering, at MGM College of Engineering and Pharmaceutical Science, Valanchery, Kerala. Her latest research include Image processing, Cloud Computing,and Machine Learning

Sanaj M S Currently pursuimgPh.D in Computer Science major in Cloud Computing from Sathyabama Institute of Science & Technology, Chennai, Tamilnadu, India. Currently he is working as Assistant Professor in the department of Computer Science and Engineering,MGM College of Engineering and Pharmaceutical Science, Valanchery, Kerala. His Specialization include Cloud Computing, Ad-hoc network and Machine Learning.