HATE CLASSIFY: A SERVICE FRAME WORK FOR HATE SPEECH IDENTIFICATION ON SOCIAL MEDIA

PILLUTLA GAYATRI[1], M. SAI SARVANI[2], KROVI JASWITHA[3], MUTNURI SAI KEERTHANA[4]

ASSISTANT PROFESSOR[1], UG SCHOLAR[2, 3&4]

DEPARTMENT OF CSE, MALLA REDDY ENGINEERING COLLEGE FOR WOMEN,MAISAMMAGUDA, DHULAPALLY KOMPALLY, MEDCHAL RD, M, SECUNDERABAD, TELANGANA 500100

ABSTARCT: It is indeed a challenge for the existing machine learning approaches to segregate the hateful content from the one that is merely offensive. One prevalent reason for low accuracy of hate detection with the current methodologies is that these techniques treat hate classification as a multi-class problem. In this work, we present the hate identification on the social media as a multi-label problem. To this end, we propose a CNN-based service framework called "HateClassify" for labeling the social media contents as the hate speech, offensive, or non-offensive. Results demonstrate that the multi-class classification accuracy for the CNN based approaches particularly Sequential CNN (SCNN) is competitive and even higher than certain state-of-the-art classifiers. Moreover, in the multi-label classification problem, sufficiently high performance is exhibited by the SCNN among other CNN-based techniques. The results have shown that using multi-label classification instead of multi-class classification, hate speech detection is increased up to 20%

INTRODUCTION

SOCIAL MEDIA has emerged as a great platform to share feelings and emotions. However, the widespread acceptance of social media has also resulted in dissemination of hate content in the name of freedom of expression. The hate content on the social media has increased around 900% from year 2014 till year 20161 . According to a report, 73% of Internet users have seen online harassment and 40% personally experienced the online harassment2 . The term "hate speech" is defined by Council of Europe's Protocol to the Convention on Cybercrime as the speech to "spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin". However, under the free speech provisions of the First Amendment, hate speech is

protected in the United States. Online social media sites, such as Google, Facebook and Twitter have their own policies for deciding "what is hate speech?" in their online social media. There exists a disagreement among the social media sites about dealing with the hate and offensive speech. Among, Google, Facebook, and Twitter, Twitter is the only one that does not ban hate speech at all. Twitter differentiates between the hate speech and direct specific threats. The twitter only considers hateful behavior of accounts whose primary purpose is to target others and their reported behavior is "one-sided". Although, Twitter claims that nobody is above their rules, it still faces criticism due to the vague nature of the company rules. As of May 31, 2016, Facebook, Twitter, Google's YouTube, and Microsoft have agreed to voluntary code of conduct to remove hate speech as defined by European Union. Most recently, the issue of hate speech on social media gained significant attention when the Facebook CEO was questioned about the company's policy about the flagging and identifying the hate speech or hateful content. The remarks of the company's CEO depict that the current approach being used by the Facebook for flagging the hateful content is not effective to deeply identify the emotions at varying levels of intensities. The reason is that there is difference in defining the hate speech content by different individuals. Several previous works, for example [1] considered the offensive and hate speech as one problem. However, the authors in [2] differentiated hate speech from the offensive speech. The authors of the study argued that people often use highly offensive terms in their normal routines. Therefore, the problem of hate speech classification was presented as multi-class classification problem among the hate, offensive, and non-offensive speech. We agree with the categorization of speeches provided by [2]. However, we consider the hate speech problem as multi-label problem instead of multiclass problem. There is a very minute difference between offensive and hate speech and drawing a distinction between offensive and hate speech has confused human experts as well. Therefore, strictly labeling only one class can never resolve the conflicts between two arguing parties. Our results demonstrate that presenting the problem as multi-label problem increases the accuracy in detecting offensive and hate speech. The proposed service framework called HateClassify is a combination of a crowd-source and machine learning techniques to detect the offensive and hate speech in online social media platforms. The main contributions of the paper are as follows: • We present a framework for detection of hate and offensive speech as a service for social media companies • Contrary to the social media platforms where the policies regarding hate speech are regulated by the specific organizations,

the proposed framework employs a crowd-sourced approach for hate speech identification • The problem of hate speech detection is presented as multi-label classification problem and sufficiently high classification accuracy is achieved • The multi-label classification used in HateClassify framework yields 20% improvement in detection of hate speech on social media

RELATED WORK The work on the hate speech detection mostly revolves around finding the best features that can be used in text classification algorithms. The basic features that are used by most of the authors in their studies are n-grams and Bag-ofWords (BoW). Warner et.al. [3] argued that hatred against different groups can be categorized with the usage of small set of high frequency words. The authors in [4] used n-grams with syntactic rules, such as user's writing style. In Ref. [5], n-grams were used along with the number of comments for the images. Length of a tweet, geographical location, and gender information of the tweeting person were used along with the n-grams for hate speech detection in [6]. Finding the grammatical usage of hate content has also gained popularity among the researchers. The authors in [7] used the sentiment features along with the ngrams and the BoW for studying and detecting hate speech. In Ref. [8], the authors used n-grams with the Part-Of-Speech tagging (POS tagging) to study bullying traces on the social media. In [2], the authors used TF-IDF weighted unigram, bigrams, trigrams, sentiment score of the tweet, number of hashtags, retweets, URLs, characters, words, and syllables in each tweet as the feature set. To overcome the problem of sparsity due to short length of texts in tweets or online comments during hate detection, numerous researchers have utilized the concept of word generalization. In [3], the authors used Brown Clustering technique for word generalization. Unlike Brown Clustering that assigns word to exactly one cluster, Latent Dirichlet Allocation (LDA) predict the probabilities of word in different clusters. Ref. [9] used the LDA for word generalization. Recently, several distributed word representations, termed as the word embedding have been developed for word generalizations. The word embedding takes the large text as the input and develops a vector space of words. The word vectors are placed in such a manner that words with similar context are placed closer to each other. In [10], the authors used word2vec (a word embedding technique) along with the BoW and hate effectiveness score to detect the hate speech. Paragraph2vec another word embedding technique was studied for hate speech detection against the BoW approach in [11]. For classification, State Vector Machine (SVM) [12][3][4][5][7][8][9] and Logistic Regression (LR) [2][6][9] have outperformed the other techniques for the hate speech detection studies. In [13], the authors

preferred Vowpal Wabbit's regression model over other models. In [14], the authors have used Recurrent Neural Network (RNN) models for hate speech detection. In this paper, we proposed a crowd-sourced and neural network-based hate speech detection framework that can be adopted by the online social media websites. We have used word vectors embedding as input features and used the CNN models for classification in the proposed service framework. Moreover, the previous works has considered the hate speech problem as multi-class classification problem. We have identified and presented the problem as multi-label classification

EXISTING SYSTEM

The work on the hate speech detection mostly revolves around finding the best features that can be used in text classification algorithms. The basic features that are used by most of the authors in their studies are n-grams and Bag-of-Words (BoW). Warner et al.3 argued that hatred against different groups can be categorized with the usage of small set of high frequency words. Chen et al.4 used n-grams with syntactic rules, such as user's writing style. Hosseinmardi et al.5 used n-grams along with the number of comments for the images. Length of a tweet, geographical location, and gender information of the tweeting person were used along with the n-grams for hate speech detection by Waseem and Hovy.6 Finding the grammatical

usage of hate content has also gained popularity among the researchers. Van Hee et al.7 used the sentiment features along with the n-grams and the BoW for studying and detecting hate speech. Xu et al.8 used n-grams with the Part-Of-Speech tagging (POS tagging) to study bullying traces on the social media. Davidson et al.2 used TF-IDF weighted unigram, bigrams, trigrams, sentiment score of the tweet.

number of hashtags, retweets, URLs, characters, words, and syllables in each tweet as the feature set. To overcome the problem of sparsity due to short length of texts in tweets or online comments during hate detection, numerous researchers have utilized

the concept of word generalization. Warner and Hirschberg3 used Brown Clustering technique for word generalization. Unlike Brown Clustering that assigns word to exactly one cluster, latent Dirichlet allocation (LDA) predict the probabilities of word in different clusters.

Xiang et al.9 used the LDA for word generalization. Recently, several distributed word representations, termed as the word embedding have been developed for word generalizations.

The word embedding takes the large text as the input and develops a vector space of words. The word vectors are placed in such a manner that words with similar context are placed closer to each other. Zhong et al.10 used word2vec (a word embedding technique) along with the BoW and hate effectiveness score to detect the hate speech. Paragraph2vec another word embedding technique was studied for hate speech detection against the BoW approach by Djuric et al. For classification, state vector machine12,3-5,7-9 and logistic regression (LR)2;6;9 have outperformed the other techniques for the hate speech detection studies. Nobata et al.13 preferred Vowpal Wabbit's regression model over other models. Mehdad and Tetreault14 have used recurrent neural network (RNN) models for hate speech detection.

DISADVANTAGES

- In the existing work, the system did not implement Multilabel Classification Results.
- This system is less performance due to lack of CNN model which is for
- hate classification is sequential convolutional neural network model (SCNN).

PROPOSED SYSTEM

- The system presents a framework for detection of hate and offensive speech as a service for social media companies.
- Contrary to the social media platforms where the policies regarding hate speech are regulated by the specific organizations, the proposed framework employs a crowd-sourced approach for hate speech identification.
- The problem of hate speech detection is presented as multi label classification problem and sufficiently high classification accuracy is achieved.
- The multi label classification used in Hate Classify framework yields 20% improvement in detection of hate speech on social media.

ADVANTAGES

- an offline training module- The offline training is a periodic job that takes the tweets and labels the tweets tagged by different people.
- online hate and offensive speech detection module.

MODULES

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Browse Tweets Data Sets and Train & Test, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Of Tweet Classify Type, View Tweet Classify Type Ratio, Download Tweet Classify Predicted Data Sets, View Tweet Classify Type Ratio Results, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like register and login, predict tweet classify type, view your profile.

CONCLUSION

In this article, we presented a service framework called Hate Classify for hate speech detection on social media. The HateClassify framework employs a crowed-sourced approach that permits the social media users to vote about any textual speech or content that is deemed inappropriate. To evaluate the performance in terms of classification, the CNNs were employed and experimental results demonstrate that the classification accuracy achieved through the CNN models, particularly the SCNN is significantly competitive and even better than several state-of-the-art approaches. An important contribution of this article is that it presents the problem of hate speech classification as the multilabel classification problem. The experimental results attained by employing the CNN approaches both for the multiclass classification and multilabel classification are sufficiently encouraging and signify the feasibility of these approaches for hate speech classification on social media.

REFERNECES

1. F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," in Proc. 1st Italian Conf. Cybersecurity, 2017, pp. 86–95.

2. T. Davidson, D. Warmsley, M. Macy, and I.Weber,"Automated hate speech detection and the problem of offensive language," in Proc. 11th Int. AAAI Conf. Web Social Media, 2017, pp. 512–515.

3. W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proc. 2nd Workshop Lang. Social Media, 2012, pp. 19–26.

4. Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in Proc. IEEE Int. Conf. Privacy, Secur., Risk Trust Int. Conf. Soc. Comput., 2012, pp. 71–80.

5. H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing labeled cyberbullying incidents on the instagram social network," Social Inform., T. Y. Liu, C. N. Scollon, and W. Zhu, Eds., 2015, pp. 49–66.

6. Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter," in Proc. NAACL Student Res. Workshop, 2016, pp. 88–93.

7. C. Van Hee et al., "Detection and fine-grained classification of cyberbullying events," in Proc. Int. Conf. Recent Adv. Natural Lang. Process., 2015, pp. 672–680.

8. J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., 2012, pp. 656–666

