Hate Speech Detection using Fusion Approach

1st Muhammad Sajjad
Al-Khawarizmi Institute of Computer Science (KICS)
Pakistan
muhammad.sajjad@kics.edu.pk

2nd Fatima Zulifqar
Al-Khawarizmi Institute of Computer Science (KICS)
Pakistan
fatima.zulifqar@kics.edu.pk

3rd Muhammad Usman Ghani Khan
Al-Khawarizmi Institute of Computer Science (KICS)
Pakistan
usman.ghani@kics.edu.pk

4th Muhammad Azeem
Al-Khawarizmi Institute of Computer Science (KICS)
Pakistan
muhammad.azeem@kics.edu.pk

***Abstract*—Detection of hate speech in user-generated online content has become an issue of increasing importance in recent years and is discerning for applications such as disputed event identification and sentiment analysis. Text classification for online content is a bit challenging task due to the natural language complexity and hastily generated online user microblogs including a plethora of informality and mistakes. This work introduces a system to classify tweets in three categories (i.e., racism, sexism and none). In our classification strategy, we integrate deep features extracted from Convolutional Neural Network(CNN) trained on semantic word embedding with state-of-the-art syntactic and word n-gram features. We perform comprehensive experiments on a standard dataset containing 16k manually annotated tweets. Our proposed approach outperform all other state-of-the-art approaches with a significant increase in accuracy.**

***Index Terms*—hate speech detection, deep learning, fusion approach, convolutional neural network, logistic regression**

I. Introduction

Research on well being and security in online social life has developed generously in the most recent decade. A especially applicable part of this work is distinguishing and avoiding the utilization of different types of harsh language in online journals, micro-blogs, and social media networks.

Now a days, social media is becoming more popular not only for social interaction and communication but also for recreation and information searching. Internet users surf their majority of the time on social media. Users of social media across the world, are increasing rapidly and exponentially. Social media is being widely used as a communication system and way of interaction between people across the world. On the other hand, negative usage of social media is on its peak. Hate speech on the basis of race, sex, caste, religion, color etc., has increased on all these platforms which exploits such infrastructures. Not only hate speech, but abusive language and offensive speech is also getting increased day by day. This is a serious issue which needs to be tackled. Some well known social media sites have taken steps towards blocking such contents but they failed to do so primarily because of lack of general definition of hate speech. Hate contents are something which varies with change in demographics. We need an enough intelligent system which is capable of detecting hate speech on social media.

Manually filtering out of hateful content is not scalable because a huge no. of messages are daily sent and received on these social media sites e.g. WhatsApp handles 4 billion messages per day. This need of automated system motivates the researcher to think towards designing an artificial intelligence based system.

Hate speech may be one of aforementioned types but in this research we are focused on filtering out tweets which contains hateful contents based on sexism, racism from neutral tweets. This task is quite challenging because of the ambiguity of the natural language in which a single word can have multiple semantics in various contexts. For example, a word 'Fan' sometimes referred as an electronic appliance but sometimes, this word is referred as a follower of a celebrity. Similarly, slang some slang words have the same case. Spelling variation is another issue, where a user intentionally or unintentionally changes a character or more than one character of words which is not detected by the system as hate speech if, simple word bases approach is used.

As far as earlier work is considered, most of it revolves around manual feature extraction or use representation learning which is followed by any linear classifier. However, deep learning models have shown great improvement in accuracy of task relevant to text. So, we are aimed to use deep learning models for classification of tweets combined with simple machine learning models. Ensemble learning is very popular now a days which exhibits good results then earlier methods.

In our proposed strategy, we used neural network solution to overcome the natural language complexity by using a fusion approach for classification of tweets (racism, sexism and none). Our main contribution in this paper is 1) by combining deep neural network features with the other syntactic (POS-Tagging and Dependency features) and word n-gram (word bi-gram, word tri-gram and TFIDF) features. 2) feed the fused features to the traditional machine learning. 3) experimental results of the machine learning model outperforms on the Twitter data-set. Our strategy gains a total of 5% accuracy compared with the state-of-the-arts methods.

978-1-7281-2353-0/19/$31.00 ©2019 IEEE

The paper is organized as follows. Section 2 is describe literature review. In Section 3, we tell about the feature fusion in more detail. Section 4, describe the classifier we used in our work while section 5, we refer the experimental setup and dataset used. Results are described in section 6. Finally, we summarize the proposed work in conclusion section.

II. Literature Review

Word based approach is the most simple approach to detect offensive/abusive text on social media which is not enough efficient to detect inappropriate/offensive speech by a user. Therefore, this simple word based approach cannot be used for real time detection and blocking or blacklisting the users posting such comments/posts on social media. Furthermore, freedom of speech and expression is affected by this approach[1]. Natural languages are ambiguous as a single word may have different meaning in different places in a single text. This property of natural language, called word ambiguity problem, is responsible for bad performance of word based approach as it increases the false positive rate in such problems e.g. hate speech detection problem. On the other hand, Natural Language Processing(NLP) approaches also failed to detect unusual spellings used by a user while posting anything on social media, which consequently unable to detect hate/abusive speech. Using this type of unusual spelling is referred as 'spelling variation' problem. Sometimes, spelling variation is unintentional but many of the times it is intentional to replace a single or more characters of an abusive/hate word, so that a user would not be detected/blocked by the system. Hence, the ambiguity/complexity of natural language words makes this hate speech detection problem quite challenging and difficult.

Using supervised learning, a machine learning technique, for classification of hate speech is not new and has been used since long. Vigna et al.[2] (2017) worked on hate speech detection using supervised learning methods to classify text into one of the two category(Binary classification problem) using simple LSTM(Long short term memory) classifier and the performance was not better than an ordinary SVM(Support vector machine). They evaluated the model on a small Facebook data-set to classify each post/comment into 'Hate' or 'No-Hate'.

Davidson et al.[3] (2017) also used some supervised learning classification model and described another way to detect hate/offensive speech on social media. They used twitter data-set for this purpose. They used Naive-Bays(NB), Decision Tree(DT), and SVM classification models to make differentiation between offensive language and hate speech. Nobata et al.[4] (2016) detect abusive contents using supervised classification model, by merging various syntactic and linguistic features in the text. He attempted using character uni-gram, bi-gram level, and used the data-set of Amazon. By looking at aforementioned work by authors, we can conclude to point out the main weaknesses of NLP models because of their non-language agnostic nature which consequences in low score in process of detection of such contents.

A lot of work has been done for detection of hate speech using unsupervised learning. NLP concepts are applied in unsupervised learning approach to detects offensive messages/posts in text. NLP uses the lexical syntactic features of text as reported by Chen et al.[5] (2012). NLP may also use artificial intelligence and Bag-of-words(BOW) based representation of text(Warner and Hirschberg, 2012)[6]. These two approaches are not very efficient because of 'Spelling variation' problem, unintentionally or intentionally by users not to get detected by system.(Djuric et al., 2015)[7] applied paragraph2vec on amazon data-set but, it only well performed for binary classification.

Waseem and Hovy in 2016 [8] also worked on hate speech detection using unsupervised learning. In their work, they have purposed a criteria that a tweet should fulfill in order to be declared as hate/offensive tweet. In their work, they showed that difference in geographical location of users have marginal effect on performance of detection. Although hate speech is growing exponentially, but its growth is not uniform in all demographics. Despite of all above mentioned limitations and observations, we tried to explore some other features which possibly can improve the performance(Accuracy, Precision, F1-score) of the system.

Waseem[9] worked in 2016 to extend the existing corpus by creating an annotated data-set for hate speech detection by crowd sourcing. He, then investigated the experience of annotators in task of classification. Jha and Mamidi(2017)[10] worked as well to provide a solution for classification of tweets but their focus was on sexism. They distinguished the 'sexism' into 'Hostile', 'Benevolent', or 'Other'. They also used data-set of waseem and Hovy[8] which they created in 2016. They changed the title of 'sexism' tweets to 'Hostile' but they have collected their own tweets for the class 'Benevolent and finally they applied FastText by Joulin et al. And SVM classification.

Pinkish tiya[11] used supervised learning approach and provided solution using Neural network(NN). His model achieved higher performance(Accuracy) on the same data-set than any others unsupervised classification model so far. In his solution, he used LSTM model, extracted features by character N-gram and assisted by GDBT(Gradient Boost) decision trees. A potential solution is also provided by CNN(convolution Neural Network) Model in hate speech detection in tweets using character N-gram and word2vector pre-trained vectors. In 2017, Park and Fung[12] converted this problem into 2-step classification problem. First, he differentiated the abusive speech from Non-abusive speech and then sub-divide the abusive tweets into Sexism or Racism. Pre-trained CNN vectors were used by Sikdar and Gamback[13] in their work to predict four classes and achieved a little bit better F1-score then character N-gram.

We know that NLP approaches are highly popular for hate speech detection, but we believe, there is a high potential still to use deep learning models as they can further contribute to this issue. A limited work on hate speech detection can be found using NLP primarily because of no standard definition of hate speech. Moreover, hate speech is something which

depends on demographics. At this point, it is also important to note the challenging nature of the task. No model is able to get F1-score greater than 93.

III. Methodology

The performance, in a complex classification task, can be altered by intense amount of data, feature space having larger dimensions, count of class, and mutually exclusive classes. For a single classifier, it is worst to rely on a single feature, and handle of variety of data and inconsistencies. The most latest techniques used for classification employ a combination of appropriate features and decision is fused. It is very important to select useful set of features very carefully. In feature fusion strategy following three different techniques are widely used:

1) **Early Fusion:**
   In this technique of fusion, features are included/added at the input of the network and then sent for task of classification. Early fusion is exceedingly expressive as it incorporates complex feature space, where include with various semantic dimensions are connected at a similar entry level.
2) **Late Fusion:** In late fusion, ultimate choice of each kind of model is summed by casting a vote. one the other hand, late combination does not count potential connection between features from various ideal models.
3) **Mid-Level Fusion:** This sort of fusion is great exchange off among expressiveness and simplicity. Final classifier is trained on each group of aforementioned features. In this mid level fusion, we combine the intermediate outcome of the several classifiers/estimators. Then the network is given the resultant vector for the purpose of classification.

IV. Classifiers

Following machine learning models have been used for evaluation of the performance:

1) **Logistic Regression (LR):** This model has a linear decision boundary and it is a Descriminative model. To optimize its probabilistic results, logistic loss is used for its training also known as cross entropy. The final layer of neural network with binary result can be considered as logistic regression which operates on the latent space induced by the top section of the network. The probabilistic labels of the Logistic Regression (LR) is computed by training and fitting of model by this equation:

$$\log(\frac{P(A=1|X)}{1-P(A=1|X)}) = B0 + B1X1 + ... + BNXN \quad (1)$$

   where A is a binary variable and its value is 1 only if it is greater than the reference, otherwise its value remains equal to 0(zero). Whereas, X is a descriptive variable and B is coefficient of regression.
2) **Random Forest (RF):** They make multiple random decision trees which are trained using bagging. Each of the tree is made up on a randomly sampled subset of training set and at each decision level, model considers a random subset of features. We may comprehend this collection of models where the training process disregards subsets of features (activation's) as a type of dropout, a strategy that has been broadly utilized in the deep learning network these days.
3) **Support Vector Machine (SVM):** support vector machines (SVMs) is a kernal based statistical technique which is genuinely non-parametric and supervised. Input is defined in portions which is mapped to the high-dimensional feature space. These models attempt to maximize the difference between the decision boundary and the nearest vectors from each class being a family of Descriminative non-probabilistic models. SVM is defined as the linear model by its base formulation, and most ofenly they are merged with kernels that induce much more complex decision boundaries. SVM tries to compute the partial differentiation by using language multiplier, with respect to separate feature for getting best/optimal results. For instance, having N instances of data input points, whereas, u and v are real number constants, x and y are input points and ψ is used as hyper parameter, then SVM used following equation to classifies the data:

$$y(x) = sin[\sum_{K=1}^{N} \mu_K y_K \Psi(x) x_K + \upsilon] \quad (2)$$

V. Proposed Approach

In our proposed approach we make use of the feature fusion technique specifically late fusion to alter the performance of complex classification task for hate speech detection. For this: 1) We first initialized glove embedding and embedding with random weights; 2) trained them on CNN and LSTM inspired by the work of Badjatiya et al. 2017 [11]; 3) saved the feature map of training instances; 4) extract other baseline features including POS tag features and dependency features as syntactic features, n-grams including word bi-grams and tri-grams and TF-IDF features; 5) generate a feature space having larger dimension using automatic and manual features of training data; and, 6) finally train and test a few classifiers using these features with a focus on Logistic Regression, Random Forest and Support Vector Machine described in the previous section.

VI. Dataset and Experimental Settings

For experimentation, we have used data-set which is created by 'Wassem and Hovy' [8]. This data-set contains 16k tweets in total. Authors did not provided the tweets instead, they have provided the IDs of the tweets used in their work. We have used Twint (a python library for scrapping tweets) to get tweets text against these tweets IDs. In this data-set of 16k tweets, no. of tweets labeled as as sexist are 3383, Racist tweets are 1972, and remaining are labeled as Neither. Unfortunately, we are unable to scrap all these tweets because of many accounts

has been blocked by twitter. We could only scrap 7704 tweets for Neither label, 1302 for Racism, and 2901 for sexism I.e total of almost 12k tweets. We have used Glove(Global Vector) word embeddings which is pre-trained for the embedding based method. This word embedding model is trained on 2 billion(having 27 billion tokens and 1.2 million distinct words) tweets. We tested the performance of the system by taking different embedding dimension sizes. Results were almost similar with all dimension sizes. Our reported results are taken using embedding dimension 25 because of lack of space. We have used 10 epochs and K-fold cross validation with value of k=10. Precision, Recall, F1-score have been used as evaluation measure of the system. 'Adam' optimizer is used for CNN and LSTM. A batch size of 128 was set for both CNN and LSTM.

VII. Evaluation and Results

The results of the experiments we performed using different methods are shown in Table I. We performed our system evaluation using macro average of precision, recall and f1-score and compared it with the results of the system proposed by Badjatiya e. al. [11] in 2017. We performed classification experiments with multiple classifiers and mark out the three top classifiers that show better results with almost all settings. These classifiers include Logistic Regression, Random Forest and Support Vector Machine.

Our proposed methods outperform all other methods used by the authors of [11] as shown in the table. Although our all proposed methods yield good results the best of all method is "CNN + Glove Embedding + Baseline Features+ Logistic Regression". CNN learned glove embedding and then used to train Logistic Regression as classifier model with other baseline features including POS tag features, dependency features, TF-IDF and n-grams including word bi-grams and tri-grams.

As the Table I shows, our aforementioned best method outperforms the previous work on hate speech detection by authors of [11] with a significant deviation of 0.056 in precision, 0.035 in recall and 0.046 in f1-score. However, the performance of the other methods varies from 0.002 to 0.0031 from the best method in f1-score.

VIII. Conclusion

Identification of hate/abusive speech has been under focused for researcher since a long ago. In this paper, we have applied the deep neural network architecture with base line classifiers for the detection of hate speech on social media platform. We found that our methodology shows extra-ordinary results compared to earlier models. We employed deep learning for feature extraction, fused with other syntactic and n-gram feature and then used simple baseline classifier (SVM, LR, RF) for training and prediction purpose. Using of deep learning architecture as feature extraction and combining with baseline methods, resulted in best accuracy values.

	Method	Precision	Recall	F1
Badjatiya et al. [11]	CNN + Random Embedding + GBDT	0.813	0.816	0.814
	CNN + Glove + GBDT	0.864	0.864	0.864
	LSTM + Random Embedding + GBDT	0.930	0.930	0.930
	LSTM + Glove + GBDT	0.849	0.848	0.848
Our Proposed Models	CNN + Random Embedding + Baseline Features + LR	0.962	0.955	0.959
	CNN + Glove + Baseline Features + LR	0.984	0.965	0.974
	CNN + Random Embedding + Baseline Features + RF	0.971	0.932	0.950
	CNN + Glove + Baseline Features + RF	0.980	0.962	0.967
	CNN + Random Embedding + Baseline Features + SVM	0.950	0.937	0.943
	CNN + Glove + Baseline Features + SVM	0.974	0.960	0.967
	LSTM + Random Embedding + Baseline Features + LR	0.964	0.956	0.960
	LSTM + Glove + Baseline Features + LR	0.983	0.962	0.972
	LSTM + Random Embedding + Baseline Features + RF	0.960	0.933	0.946
	LSTM + Glove + Baseline Features + RF	0.975	0.960	0.967
	LSTM + Random Embedding + Baseline Features + SVM	0.959	0.958	0.958
	LSTM + Glove + Baseline Features + SVM	0.979	0.960	0.970

TABLE I: "Comparision of our proposed methods"

Acknowledgment

This work is funded by HEC, TDF grant with "Semantic Engine for Text Management using Ontology and Knowledge Engineering (SETMOKE)" having project code "TDF-077".

References

[1] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Detecting offensive language in tweets using deep learning," *arXiv preprint arXiv:1801.04433*, 2018.

[2] F. Del Vigna12, A. Cimino23, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook," 2017.

[3] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[4] 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*. International World Wide Web Conferences Steering Committee, 2016, pp. 145–153.

[5] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 2012, pp. 71–80.

[6] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Proceedings of the second workshop on language in social media*. Association for Computational Linguistics, 2012, pp. 19–26.

[7] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *Proceedings of the 24th international conference on world wide web*. ACM, 2015, pp. 29–30.

[8] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL student research workshop*, 2016, pp. 88–93.

[9] Z. Waseem, "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter," in *Proceedings of the first workshop on NLP and computational social science*, 2016, pp. 138–142.

[10] A. Jha and R. Mamidi, "When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data," in *Proceedings of the second workshop on NLP and computational social science*, 2017, pp. 7–16.
[11] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2017, pp. 759–760.
[12] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on twitter," *arXiv preprint arXiv:1706.01206*, 2017.
[13] B. Gambäck and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," in *Proceedings of the first workshop on abusive language online*, 2017, pp. 85–90.