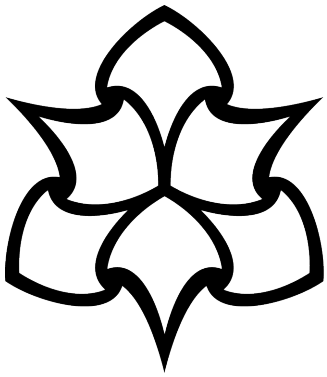

HateClassify: A Service Framework for Hate Speech Identification on Social Media

Muhammad U. S. Khan
COMSATS University Islamabad, Abbottabad Campus, Pakistan

Assad Abbas
COMSATS University Islamabad, Islamabad Campus, Pakistan

Attiqa Rehman
COMSATS University Islamabad, Abbottabad Campus, Pakistan

Raheel Nawaz
Manchester Metropolitan University, UK

***Abstract*—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 2016[1]. According to a report, 73% of Internet users have seen online harassment and 40% personally experienced the online harassment[2]. 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

[1]https://www.usatoday.com/story/news/2017/02/23/hate-groups-explode-social-media/98284662/

[2]https://www.pewresearch.org/internet/2014/10/22/online-harassment/

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 multi-class 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.

The rest of the paper is organized as follows. Section II discusses the related work. The service framework is presented in Section III. The results of multi-class and multi-label classification and comparisons with state-of-the-art techniques are presented in Section IV whereas Section V finally concludes the paper.

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-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. 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 n-grams 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 problem.

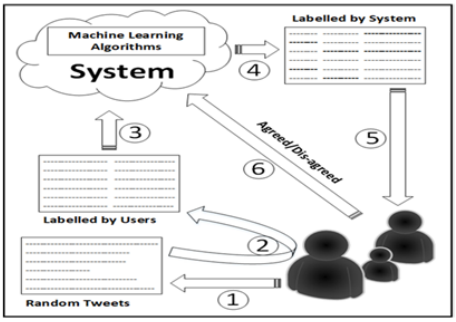

Figure 1. The proposed hate speech detection framework for social media

FRAMEWORK FOR HATE SPEECH DETECTION

In terms of functionalities, the framework has two components: (i) an offline training module and (ii) online hate and offensive speech detection module. The offline training is a periodic job that takes the tweets and labels the tweets tagged by different people, shown as Step 1 and Step 2 in Fig. 1. The offline training procedure trains the deep neural network to learn the features in the tweets. The online procedure is responsible for using the model trained in the offline procedure and predicts the labels for the new tweets. The social media users are allowed to agree or disagree with the automatic labels. The online procedure is shown through Step 4, Step 5, and Step 6 in the Fig. 1. The tweets labeled in online procedure and new tweets labeled by Twitter users are again fed to the offline procedure to re-train the algorithm for optimization of the automatic labeling task.

CROWD-SOURCED POLICY

Contrary to the social media platforms where the policies of social media organizations about hate speech are regulated by the specific organizations, the proposed framework involves the social media users in deciding about the nature (hate or otherwise) of the tweets. The people are encouraged to participate to vote and train the machine to decide about hate speech. Moreover, depending upon the judgments regarding the hate speech using the majority votes, the tweets can remain visible or hidden in a certain geographical region. Modifying the social media sites with this methodology will not enforce the bias of a certain

Table 1. HYPERPARAMETER VALUES

	Parameter	Value
Embedding	Embedding dimension	256
ConvID	Filters	512
(Ist layer)	Kernel Size	3
(2nd Layer)	Kernel Size	4
(2nd Layer)	Kernel Size	5
	Activation Function	Relu
MaxPooling ID All three layers	Pool Size	2
Dropout	Rate	0.2
Dense	Activation Function	Sigmoid
Model.compile	Optimizer	adam
Model.fit	Batch Size	30

group on others. People of different geographical regions will be able to train the machines for their regions according to their likeness and laws in a democratic manner. We implemented only one iteration of this proposed methodology. We did not retrain the models again with new votes or labels. Instead of enforcing the models to get trained again according to our own bias that could have been added with our votes, we tried to find the most stable model that can consistently perform better with the bias within the different dataset. The model that can re-learn and can consistently perform good on different dataset will be able to adjust themselves easily for the changing bias in the votes of different geographical regions.

THE CNN MODEL

The stable CNN model we have proposed for hate classification is Sequential Convolutional neural network model (SCNN). A SCNN is sequential model having embedding, three convolutional 1D layers with three maxpooling 1D layers, Dropout, Flatten, and dense layers. Table 1 presents the values of the parameters that we obtain after hyperparameter tuning our SCNN model.

We used the technique of splitting the dataset into three portions, training, development, and test datasets. We have used 60% dataset as training data, 20% as validation data, and 20% as test data. The validation set is used in the hyperparameters tuning and test set is used in the model testing and comparison with other models.

TWEETS CLASSIFICATION FOR HATE SPEECH DETECTION

We studied several machine learning models and compared them with the SCNN approach. The following techniques were compared: (i) n-grams with SVM [12], (ii) Logistic Regression with multiple features list [2], (iii) Long short-term Memory (LSTM), (iv) CNNLSTM, (v) CNN-non-static, (vi) CNN2D, (vii) ATTCNN, and (viii) ATTCNN with max. We used n-grams up to n=4 in the first two schemes and uses the filter sizes ranges from 2 to 4 in neural networks, for fair comparison. We consider n-grams with SVM and logistic regression with multiple features list as our baseline models for comparison. Brief narratives of the aforesaid models employed for comparison are given below:

n-grams with SVM: The technique is presented in [12][15]. The unigrams, bigrams, and trigrams are taken as features. The features are provided to the SVM classifier. We used this simple model as our baseline model for performance comparison of the other models in multi-class classification.

Logistic Regression with multiple features list: The technique is presented in [2]. This technique uses TF-IDF weighted unigram, bigrams, and trigrams, sentiment score of the tweet, number of hashtags, retweets, URLs, characters, words, and syllables in each tweet as the feature set.

Long short-term Memory (LSTM): The LSTM is Recurrent Neural Network (RNN) architecture that is being used for text classification. For comparison, we created model of a single layer LSTM after embedding layer with one dense layer.

CNNLSTM: The abovementioned model was modified to use LSTM layer before the dense layer. CNN-non-static: The model was presented in [16]. The original technique was used to find sentiment analysis in the text. The technique uses an embedding layer and three layers of convolutional 1D maxpooling ID and Flatten that are concatenated before the output dense layers. The technique was modified to use three classes instead of two classes. The weights of vectors in the embedding layers are fine-tuned in each task to obtain a better classification result from the model that do not use fine-tuning of the weights.

CNN2D: The model in [16] was modified to use convolutional 2D neural layers instead of 1D layers and consequently the modified model is CNN2D[3]. The output from the embedding layer is reshaped so that output of embedding layers can be used in the convolutional 2D layers.

ATTCNN: The model uses the attention mechanism in convolutional layer as described in [17]. The model has attentive convolutional layer, flatten layer, and dense layer. ATTCNN with max: The ATTCNN model is modified by adding an additional maxpooling layer after the attention convolutional layer. The model has attentive convolutional layer, maxpooling layer, flatten layer, and dense layer.

EXPERIMENTAL RESULTS

To determine the efficacy of the CNN-based approach, comparisons with the existing techniques described in Section III were made on a set of tweets: (i) Dataset 1 is a CrowdFlower[4] hate vs. offensive dataset [12], (ii) Dataset 2 is previously used in [2], and (iii) Dataset 3 is Sexism vs. Racism dataset used in [6]. The CrowdFlower dataset comprises of a total 14,509 tweets. The Dataset 2 [2] contains a total of 24,783 tweets. The third dataset, Dataset 3 [6] contains total of 6,492 tweets. The Dataset 3 is the most unbalanced dataset among the three and the Dataset 1 is the least unbalanced. In Dataset 3, around 86% of the tweets belong to one class and rest are divided into three classes. In Dataset 2, offensive tweets are comprised of 77% alone. However, in Dataset 1, the share of offensive tweets is 50%, neither 33%, and about 16% are the tweets labelled as hate speech. Experiments are conducted on Amazon EC2 cloud using Keras, Tensorflow, and Sklearn libraries of Python. The classification accuracy for both the multi-class classification and multi-label classification was determined. Precision, recall, and F-measure are used as the evaluation metrics to determine the accuracy.

[3]https://github.com/bhaveshoswal/CNN-text-classification-keras

[4]https://data.world/crowdflower/hate-speech-identification/workspace/file?filename=twitter-hate-speech-classifier-DFE-a845520.csv

MULTI-CLASS CLASSIFICATION RESULTS

The multi-class classification results are given in Table 2, Table 3, and Table 4. One important observation, we made during the analysis of the results is that the neural network-based models except RNN performed better in precision scores in identifying individual classes, especially the hate class in Dataset 1 and Dataset 2, and Sexism class in Dataset 3 as compared to the baseline model n-grams with the SVM and the LR with multiple features. However, the recall results of neural network-based models appear lower than the baseline, especially in Dataset 2, and Dataset 3. Therefore, the neural network-based models except RNN produced slightly better results than the baseline beyond expectations. Another important observation, we have made that the more unbalanced the dataset is, the more F-measure scores of neural network-based models are affected in identifying the minority class. For example, for hate speech detection in Dataset 2 neural network-based models performs slightly less than the baseline model and LR with multiple features in F-measure scores performs better in Dataset 1. However, they remain higher in precision scores in all the datasets. Recent, researches have shown that attention mechanism in convolutional layers performs better than the model without attention mechanism in text classification [17]. However, in our experiments, we observed that using the attention mechanism in convolution layer, especially with maxpooling, increases the precision score but decreases the recall score. Therefore, overall F-score of attention convolutional models remains low.

The high precision and low recall results signify that the neural network-based models are comparably stringent in classification as compared to baseline model. To understand the reasons for low recall results, we compare the percentage of similar words among the different classes in the datasets. In Dataset 1, 45.7% of the words in class labelled as hate speech are also present in offensive tweets class. Similarly, in Dataset 2, 65.2% of words in tweets labelled as hate speech are also present in offensive tweets class. The same scenario is also present in the Dataset 3. In Dataset 3, 43.6% of words in racism tweets are also present in tweets labelled as sexism. The high percentage of similar words af-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MIC.2020.3037034, IEEE Internet Computing

Table 2. MULTI-CLASS CLASSIFICATION RESULTS ON DATASET 1

Dataset 1									
	Precision			Recall			F-measure		
Classification Technique	Hate	Offensive	Neither	Hate	Offensive	Neither	Hate	Offensive	Neither
Multi-Features LR	0.39	0.95	0.7	0.53	0.92	0.83	0.449348	0.934759	0.759477
n-gram SVM	0.39	0.94	0.7	0.48	0.93	0.82	0.430345	0.934973	0.755263
RNN	0	0	0.51	0	0	1	0	0	0.675497
CNNLSTM	0.48	0.58	0.87	0.46	0.66	0.8	0.469787	0.617419	0.833533
SCNN	0.47	0.65	0.88	0.56	0.58	0.89	0.511068	0.613008	0.884972
CNN-non-static	0.43	0.63	0.94	0.72	0.78	0.7	0.538435	0.697021	0.802439
CNN2D	0.61	0.68	0.86	0.32	0.73	0.95	0.419785	0.704113	0.902762
ATTCNN	0.53	0.65	0.85	0.34	0.66	0.93	0.42	0.65	0.89
ATTCNN-with max	0.65	0.63	0.81	0.08	0.72	0.96	0.14	0.67	0.88

Table 3. MULTI-CLASS CLASSIFICATION RESULTS ON DATASET 2

Dataset 2									
	Precision			Recall			F-measure		
Classification Technique	Hate	Offensive	Neither	Hate	Offensive	Neither	Hate	Offensive	Neither
Multi-Features LR	0.21	0.95	0.87	0.53	0.92	0.83	0.300811	0.934759	0.849529
n-gram SVM	0.3	0.94	0.88	0.48	0.93	0.82	0.369231	0.934973	0.848941
RNN	0	0.78	0	0	1	0	0	0.876404	0
CNNLSTM	0.37	0.91	0.74	0.28	0.92	0.75	0.318769	0.914973	0.744966
SCNN	0.2	0.9	0.71	0.45	0.86	0.67	0.276923	0.879545	0.68942
CNN-non-static	0.52	0.94	0.92	0.09	0.89	0.15	0.153443	0.914317	0.257944
CNN2D	0.58	0.91	0.8	0.16	0.96	0.8	0.250811	0.934332	0.8
ATTCNN	0.47	0.9	0.78	0.14	0.95	0.76	0.22	0.93	0.77
ATTCNN-with max	0.58	0.89	0.81	0.06	0.97	0.7	0.11	0.83	0.75

Table 4. MULTI-CLASS CLASSIFICATION RESULTS ON DATASET 3

Dataset 3									
	Precision			Recall			F-measure		
Classification Technique	Sexism	Racism	Neither	Sexism	Racism	Neither	Sexism	Racism	Neither
Multi-Features LR	0.76	0.25	0.93	0.59	0.6	0.98	0.67	0.1	0.95
n-gram SVM	0.75	0.33	0.94	0.64	0.12	0.97	0.69	0.17	0.96
RNN	0.3	0	0.86	0.07	0	0.98	0.12	0	0.91
CNNLSTM	0	0	0.81	0	0	1	0	0	0.92
SCNN	0.67	0.33	0.87	0.17	0.15	0.98	0.28	0.21	0.92
CNN-non-static	0.81	0	0.94	0.44	0	0.91	0.57	0	0.92
CNN2D	0.78	0.33	0.91	0.48	0.15	0.98	0.6	0.21	0.95
ATTCNN	0.8	0	0.89	0.28	0	0.99	0.41	0	0.94
ATTCNN-with max	1	0	0.92	0.02	0	1	0.04	0	0.92

fected the performance of strict classifiers (neural network-based models) during the recall scores calculations. We also visualized the datasets using the scattertext. Fig. 2, F.g. 3, and Fig. 4 show the scattertext of all the three datasets. In Fig. 2 and Fig. 3, the right top corner is more cluttered with words as compared other parts of the graphs. The phenomena represent that there are more words that are in high frequency both in hate and offensive labelled data. However, in Fig. 4, the number of words are greater in the center of the graph. The phenomena describe that there are more words that are medium frequent both in the sexism and racism categories. Therefore, it is clear that it is hard to separate both the hate and offensive data.

MULTI-LABEL CLASSIFICATION RESULTS

Distinguishing between hate speech and offensive speech often becomes difficult for humans as well due to the same usage of the words and very slight distinction between the semantics. The similar problem occurs in machine learning as well. Due to the overlapping nature of vocabulary used in all the three classes, we re-evaluated the hate speech detection problem as the multi-label problem. The multi-label classification is evaluated using the α -Evaluation metric [18]. The α -Emulation performance metric evaluates each prediction using the following equation:

$$Score(P_x) = (1 - \frac{|\beta M_x + \gamma F_x|}{|Y_x \cup P_x|})^{\alpha}, \qquad (1)$$

Figure 2. ScatterText plot of Dataset 1

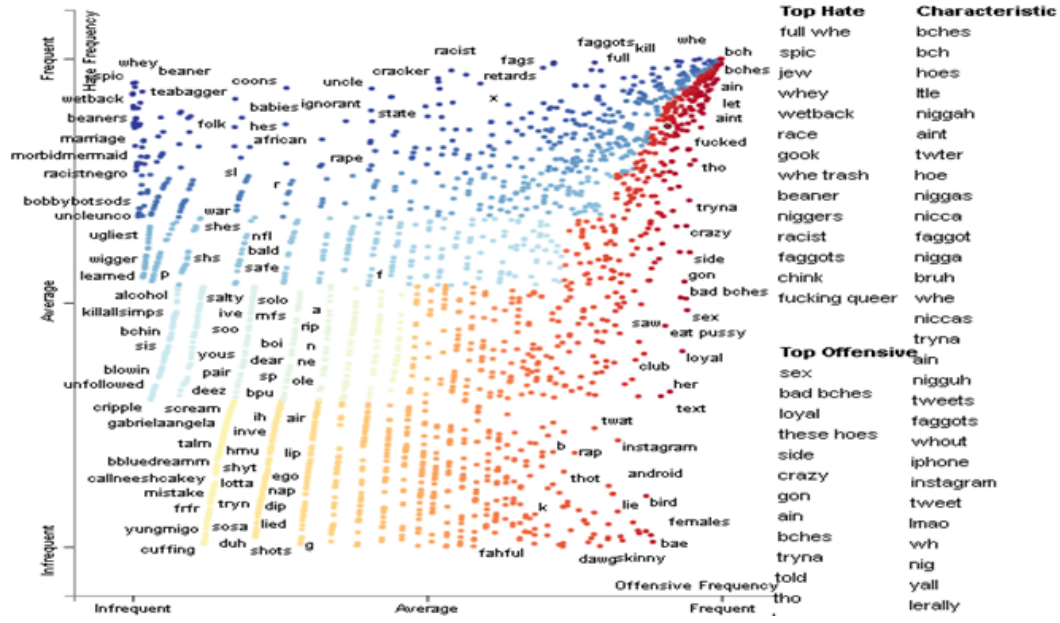

Figure 3. ScatterText plot of Dataset 2

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MIC.2020.3037034, IEEE Internet Computing

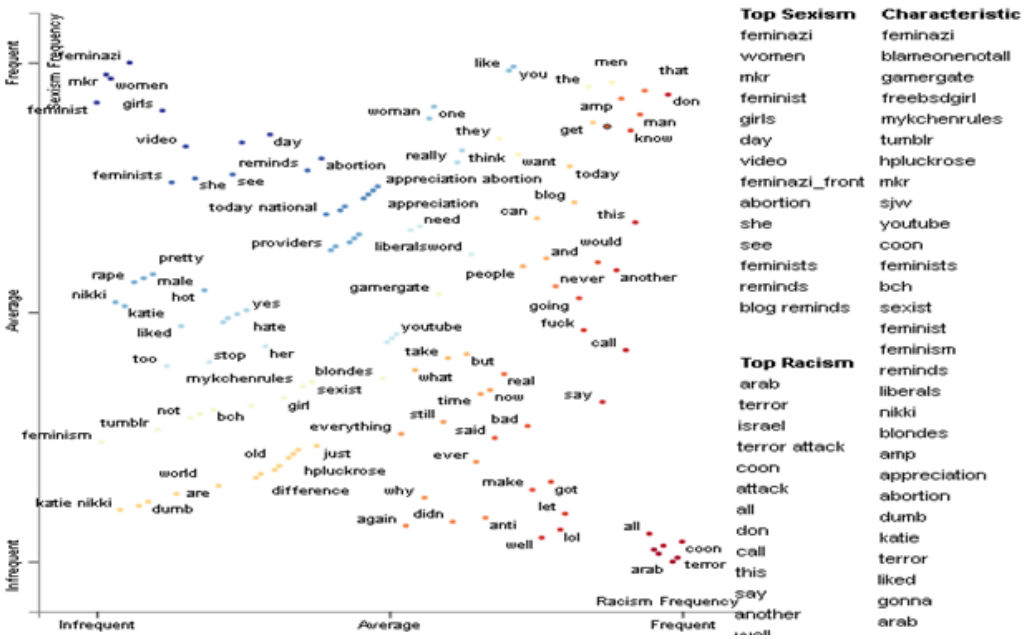

Figure 4. ScatterText plot of Dataset 3

where Y_x represents a set of actual label(s) while P_x represents the predicted label(s) against the test case x. Moreover, $M_x = Y_x - P_x$ represents the missed labels that model fails to predict and $Y_x = P_x - Y_x$ is a set of false positive labels in the above equation. The parameters β and γ are to penalize the missed labels and false positives in the multi-label classification. The parameter α is the forgiveness rate parameter. The three parameters in the equations α, β, and γ are restricted to keep the score of prediction as non-negative. The restrictions are:

$$\alpha \geq 0, \beta \geq 0, \gamma \leq 1, \beta = 1 | \gamma = 1, \qquad (2)$$

The neural network models predict the probabilities of all the classes and the class with the highest probability is considered as the predicted class. For multi-label scenario, we considered all the classes as labels. However, to restrict ourselves from getting all the labels in each scenario, we used the threshold of 0.5. The classes with the prediction score of above the threshold were assigned to the text as labels. The precision and recalls scores are evaluated using the following formulas for the multi-label classification,

$$precision_c = \frac{1}{|D_c|} \sum_{x \in D_c} score(P_x), \qquad (3)$$

where $D_c = \{x | C = P_x\}$.

Similarly, the recall score is calculated using equation given below:

$$recall_c = \frac{1}{|D_c|} \sum_{x \in D_c} score(P_x), \qquad (4)$$

where $D_c = \{x | C = Y_x\}$.

Table 5 presents the results of SCNN model with multi-label classification by varying different parameters in Equation 1 and Table 6 presents comparison of results of different classifiers. The results demonstrate that the F-measure is highly affected by being tolerant to the false positives — decreasing the value of γ — and by getting strict on the missing the labels —increasing the value of the β. We have found $\beta = 1, \gamma = 0$, and $\alpha = 1$ as the best case in the results. The case is extreme relaxed on false positives and extreme strict on missing the labels. We obtained the precision score of 1 under all the models except RNN and CNNLSTM. This is due to the reason that we are too relaxed in false positive. Moreover, the recall scores of the convolutional neural network-based models have shown significant improvement that has affected the F-measure scores as well. The results show an average increase of 0.095 in F-measure score for the convolutional neural network-based models

against all the classes. However, the hate class in Dataset 1 has shown the maximum increase of 0.2 in F-measure score as compared to the results in multi-class classification. It is clear from the results that the high number of similar words in the different classes and strict nature of convolutional neural network-based models resulted in low recall in multi-class classification but still they predict the correct classes with probabilities higher than the 0.5. Overall, the SCNN has performed consistently well than the other models in the three datasets.

CONCLUSIONS

In this paper, we presented a service framework called HateClassify 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 the paper is that it presents the problem of hate speech classification as the multi-label classification problem. The experimental results attained by employing the CNN approaches both for the multi-class classification and multi-label classification are sufficiently encouraging and signify the feasibility of these approaches for hate speech classification on social media.

REFERENCES

1. F. Del Vigna12, A. Cimino23, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook," in *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, 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 *Eleventh international aaai conference on web and social media*, 2017.
3. 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.
4. 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.
5. H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of cyberbullying incidents on the instagram social network," *arXiv preprint arXiv:1503.03909*, 2015.
6. 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.
7. C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste, "Detection and fine-grained classification of cyberbullying events," in *International Conference Recent Advances in Natural Language Processing (RANLP)*, 2015, pp. 672–680.
8. J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics, 2012, pp. 656–666.
9. G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012, pp. 1980–1984.
10. H. Zhong, H. Li, A. C. Squicciarini, S. M. Rajtmajer, C. Griffin, D. J. Miller, and C. Caragea, "Content-driven detection of cyberbullying on the instagram social network." in *IJCAI*, 2016, pp. 3952–3958.
11. 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*, 2015, pp. 29–30.
12. P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy & Internet*, vol. 7, no. 2, pp. 223–242, 2015.
13. 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*, 2016, pp. 145–153.
14. Y. Mehdad and J. Tetreault, "Do characters abuse more than words?" in *Proceedings of the 17th Annual Meeting*

Table 5. MULTI-LABEL CLASSIFICATION RESULTS FOR DIFFERENT PARAMETERS

Parameters Settings

Dataset 1

	Precision			Recall			F-measure		
	Hate	Offensive	Neither	Hate	Offensive	Neither	Hate	Offensive	Neither
$\alpha = 1, \beta = 1/4, \gamma = 1$	0.64	0.74	0.91	0.70	0.75	0.87	0.67	0.74	0.89
$\alpha = 1, \beta = 1/8, \gamma = 1$	0.68	0.76	0.91	0.73	0.78	0.89	0.70	0.77	0.90
$\alpha = 1, \beta = 1, \gamma = 1/4$	0.66	0.76	0.92	0.68	0.74	0.87	0.67	0.75	0.89
$\alpha = 1, \beta = 1, \gamma = 1/8$	0.69	0.79	0.93	0.71	0.76	0.88	0.70	0.77	0.90
$\alpha = 1, \beta = 1, \gamma = 0$	1	1	1	0.7	0.87	0.92	0.83	0.93	0.96

Dataset 2

	Precision			Recall			F-measure		
	Hate	Offensive	Neither	Hate	Offensive	Neither	Hate	Offensive	Neither
$\alpha = 1, \beta = 1/4, \gamma = 1$	0.55	0.94	0.83	0.55	0.94	0.83	0.55	0.94	0.83
$\alpha = 1, \beta = 1/8, \gamma = 1$	0.59	0.94	0.84	0.60	0.95	0.85	0.60	0.94	0.85
$\alpha = 1, \beta = 1, \gamma = 1/4$	0.57	0.94	0.84	0.53	0.93	0.82	0.55	0.94	0.83
$\alpha = 1, \beta = 1, \gamma = 1/8$	0.61	0.95	0.86	0.57	0.94	0.84	0.59	0.94	0.85
$\alpha = 1, \beta = 1, \gamma = 0$	1	1	1	0.59	0.94	0.73	0.74	0.97	0.85

Dataset 3

	Precision			Recall			F-measure		
	Sexism	Racism	Neither	Sexism	Racism	Neither	Sexism	Racism	Neither
$\alpha = 1, \beta = 1/4, \gamma = 1$	0.64	0.42	0.94	0.74	0.57	0.91	0.69	0.49	0.93
$\alpha = 1, \beta = 1/8, \gamma = 1$	0.68	0.47	0.95	0.76	0.63	0.92	0.72	0.54	0.93
$\alpha = 1, \beta = 1, \gamma = 1/4$	0.66	0.53	0.95	0.72	0.47	0.91	0.69	0.50	0.93
$\alpha = 1, \beta = 1, \gamma = 1/8$	0.69	0.59	0.95	0.74	0.51	0.92	0.72	0.55	0.94
$\alpha = 1, \beta = 1, \gamma = 0$	1	1	1	0.44	0.09	0.98	0.61	0.17	0.99

Table 6. MULTI-LABEL CLASSIFICATION RESULTS COMPARISON

Parameters Settings

Dataset 1

	Precision			Recall			F-measure		
	Hate	Offensive	Neither	Hate	Offensive	Neither	Hate	Offensive	Neither
RNN	0	1	1	0	1	1	0	1	1
CNNLSTM	1	1	1	0.45	0.65	0.79	0.62	0.79	0.88
SCNN	1	1	1	0.7	0.87	0.92	0.83	0.93	0.96
CNN-non-static	1	1	1	0.72	0.78	0.74	0.84	0.87	0.85
CNN2D	1	1	1	0.32	0.71	0.94	0.48	0.87	0.85

Dataset 2

	Precision			Recall			F-measure		
	Hate	Offensive	Neither	Hate	Offensive	Neither	Hate	Offensive	Neither
RNN	0	1	0	0	1	0	0	1	0
CNNLSTM	1	1	1	0.22	0.88	0.37	0.33	0.93	0.55
SCNN	1	1	1	0.59	0.94	0.73	0.74	0.97	0.85
CNN-non-static	1	1	1	0.09	0.89	0.45	0.17	0.94	0.62
CNN2D	1	1	1	0.32	0.71	0.94	0.25	0.98	0.87

Dataset 3

	Precision			Recall			F-measure		
	Sexism	Racism	Neither	Sexism	Racism	Neither	Sexism	Racism	Neither
RNN	1	1	1	0.34	0.14	0.97	0.51	0.24	0.98
CNNLSTM	0	0	1	0	0	1	0	0	1
SCNN	1	1	1	0.44	0.09	0.98	0.61	0.17	0.99
CNN-non-static	1	0	1	0.45	0	0.88	0.62	0	0.94
CNN2D	1	1	1	0.5	0.09	0.97	0.67	0.17	0.99

of the Special Interest Group on Discourse and Dialogue, 2016, pp. 299–303.

15. 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.
16. 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.
17. W. Yin and H. Schütze, "Attentive convolution: Equipping cnns with rnn-style attention mechanisms," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 687–702, 2018.
18. M. S. Sorower, "A literature survey on algorithms for multi-label learning," *Oregon State University, Corvallis*, vol. 18, pp. 1–25, 2010.