AI_ML_NIT Patna at HASOC 2019: Deep Learning Approach for Identification of Abusive Content *

Kirti Kumari[1] and Jyoti Prakash Singh[1]

National Institute of Technology Patna, Patna, India
{kirti.cse15, jps}@nitp.ac.in

Abstract. Social media is a globally open place for online users to express their thoughts and opinions. There are numerous advantages of social media but some severe challenges are also associated with it. Anti-social and abusive conduct has become more common due to the emergence of social media. Identification of Hate Speech, Cyber-aggression, and Offensive language is a very challenging task. The nature of structures of the natural language makes this task even more tedious. Being a challenging task, we are fascinated to propose a deep learning system based on Convolutional Neural Networks to identify Hate Speech, Offensive language, and Profanity. We have done experiments with three different embeddings. These experiments have been associated with comments of code-mixed Hindi-English and multi-domain social media text. We have found that One-hot embedding performed better than pre-trained fastText embedding for the code-mixed Hindi dataset.

Keywords: Hate Speech · Offensive Language · Convolutional Neural Network · GloVe · fastText

1 Introduction

In social media, anyone is free to post their ideas and views without declaring his/her identity. Detection of Cyber-aggression [9], Hate Speech [14], Offensive language and Profanity used by social media users have become one of the major challenges of the current scenario. Social media users are being targeted by Hate Speech and Offensive language such as abusive, hurtful, derogatory or unlawful user-generated content by some mischievous users. These online platforms provide an open place to discuss and comment on different matters but abusive comments and online violence on individuals have turned this into a very important social issue. As a result of the misuse of online interactions, a large number of people have fallen into depression, anxiety, and other mental health problems. A survey undertaken by Feminism in India has noted that online abuse has been

faced by more than 50% of females in major cities of India[1]. During the study[2] (July 2017), 66% of online abused people reported that they feel powerlessness in their capacity to react to Internet violence or harassment. These statistics emphasize the necessity of an automated system for the detection of abusive comments as well as the moderation of the system. As a result, several research efforts across the world have emerged over the past few years to identify abusive content [1, 2, 4, 12, 15] using machine learning and natural language processing.

Hate Speech detection becomes a challenging task because it can not be addressed simply by filtering words. In addition to the meaning of words, a lot of other factors such as context information, characteristics of the user, the gender of individual people have to be considered for the detection of Hate Speech. Abuse is a term that includes many varying forms of fine-grained adverse expressions in the framework of natural language. For example, Nobata et al. [12] concentrated on Hate Speech, Derogatory language, and Profanity while Wassem and Hovy [15] focused on racism and sexism types of abuse. Definitions tend to be overlapping and ambiguous for different types of abuse. The Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) organizer team defines Hate Speech as describing negative attributes or deficiencies toward groups because of race, political opinion, sexual orientation, gender, social status, health condition or similar [10]. A large number of works [1, 5, 6, 15] are reported by the researchers on Hate Speech detection for English language only and very few works [2, 8] have been reported for mixed languages such as English-Hindi, English-Bengali, and other languages.

In this paper, we have used the multi-lingual HASOC Corpus [10] and proposed a deep learning model based on Convolutional Neural Networks (CNN) to identify Hate Speech and Offensive content on multi-domain social media platforms collected from Facebook and Twitter. We have used three types of embeddings of text and transliteration tools to normalize Devanagari to Roman script for code-mixed Hindi corpus.

The rest of the paper is structured as follows. Section 2 presents associated works for the detection of Hate Speech and Offensive Language while Section 3 presents our suggested framework for identification of Hate Speech, Offensive language, and Profanity. Section 4 presents the finding of the suggested scheme. Finally, in Section 5, we have concluded the paper and have discussed the future directions for these tasks.

2 Related Works

As social media and online platforms have grown in terms of impact and acceptance of users, various problems such as Hate Speech, Profanity and Offensive language on these platforms have increased drastically. Several systems have been proposed by researchers for automatic detection and classification of these problems.

[1] https://blog.ipleaders.in/cyber-stalking
[2] https://www.statista.com/statistics/784838/online-harassment-impact-on-women/

Burnap and Williams [3] identified Hate Speech on the Twitter network focusing mainly on racism. Nobata et al. [12] used character n-gram features and reported that character n-gram features are the most predictive features for the detection of Hate Speech. Wassem and Hovy [15] have focused on Hate tweet detection related to racism and sexism. They have used Logistic Regression as classifier and character n-gram features to classify the tweets. Davidson et al. [5] have found that racist and homophobic tweets are generally Hate Speech and sexist tweets are in general offensive. They have used Logistic Regression, Naive Bayes, Decision Trees, Random Forests and Support Vector Machines (SVM) to classify the tweets. Mehdat and Tetreault [11] also found that character n-gram features are more predictive than token n-gram features for Hate Speech detection. Badjatiya et al. [1] detected Hate Speech using deep learning models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM). They have experimented with several embeddings named Random, GloVe and fastText embeddings and have found that combination of LSTM, Random embedding and Gradient Boosted Decision Trees (GBDTs) had performed the best for classifying the Hate tweets. Del Vigna et al. [6] have classified the Hate Speech of Facebook comments into fine-grained classes. They have used two different approaches with SVM and LSTM to identify the Hate comments. Bohra et al. [2] have identified code-mixed Hate tweets, especially for Hindi and English language on Twitter. Kamble and Joshi [8] have also focused on Hindi and English code-mixed tweets and have detected Hate Speech using various deep learning models such as CNN, LSTM, and Bi-directional LSTM. A lot of research works have been done for the English language but very few works have been done for the other languages and code-mixed languages. In this paper, we have focused on multi-lingual text, especially for Hindi and English languages and used a deep neural network model to detect Hate Speech, Offensive language, and Profanity.

3 Methodology

This section describes the details of datasets and proposed approaches. The description of the datasets used in the experiments has been given in sub-section 3.1 and the details of the proposed approaches to identify Hate Speech, Offensive language and Profanity are presented in sub-section 3.2.

3.1 Description of Datasets

In this paper, the multilingual datasets [10] provided by Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC)[3] have been used. The shared tasks of HASOC have been provided for three languages (English, code-mixed Hindi, and German) and each language, there are three sub-tasks (Sub-task1, Sub-task2, and Sub-task3). The provided comments have been collected from Twitter and Facebook. The details of the sub-tasks are: Sub-task1

[3] https://hasoc2019.github.io

is a coarse-grained binary classification that needed respondents to classify tweets into two groups: Hate and Offensive (HOF) and Not Hate-Offensive (NOT). (i) HOF: This post includes hateful, offensive or profane contents and (ii) NOT: This post contains neither Hate Speech nor offensive content. Sub-task2 is a fine-grained classification. Hate Speech and Offensive posts from the Sub-task1 are further classified into three categories: (i) Hate Speech (HATE): Posts under this class contain Hate Speech contents. (ii) Offensive (OFFN): Posts under this class contain offensive contents. (iii) Profane (PRFN): These posts contain profane words. In Sub-task3, the category of abuse is checked and includes only the posts marked as HOF in Sub-task1. Sub-task3 is further grouped into two classes: (i) Targeted Insult (TIN): Such posts contain humiliating/insulting or threatening content. (ii) Untargeted Insult (UNT): Posts that contain untargeted swearing and profanity, those posts of particular profanity which are not targeted at anybody but contain language that is not acceptable. The sizes of the training datasets for English and Hindi corpus are 5852 and 4665 posts, respectively. Test data for English and Hindi corpus are 1153 and 1318 posts, respectively. The detailed description of the datasets used in this work is given in Table 1.

Table 1. Description of datasets

Corpus	**Sub-tasks**	**Class**	**#training_samples**	**#testing_samples**
English	Sub-task1	HOF	2261	288
		NOT	3591	865
	Sub-task2	HATE	1143	124
		OFFN	451	71
		PRFN	667	93
	Sub-task3	TIN	2041	245
		UNT	220	43
Hindi	Sub-task1	HOF	2469	605
		NOT	2196	713
	Sub-task2	HATE	556	190
		OFFN	676	197
		PRFN	1237	218
	Sub-task3	TIN	1545	542
		UNT	924	63

3.2 Proposed Approach

The proposed methodology is based on the Convolutional Neural Networks (CNN) model, a block diagram of which is shown in Figure 1. At first, we have removed the stopwords from the comments by using Natural Language Toolkit[4]. The embedding layer is the representation of inputs in the deep neural

[4] www.nltk.org

network models. The embedding layer encodes the word used in the comments. We have done experiments with three different embeddings including One-hot embedding, GloVe embedding [13] and fastText embedding [7]. In the case of the Hindi dataset, we have used only One-hot and fastText embeddings. For One-hot embedding, we have transliterated Devanagari to Roman script by using transliteration tools[5]. These tools identify the Unicode patterns and transliterate the Devanagari script to Roman script. The dimensions of a word embeddings are kept 300 for pre-trained (GloVe and fastText) and 100 for One-hot embeddings. This embedded comment is fed to the CNN layer. In our case, we have used four layers of convolution and one layer of the max-pooling in between the 3rd and 4th convolution layers. At last, we have used the flatten layer followed by a dense layer. Within the layer, we have used sigmoid and softmax activation function at the dense layer for binary class and multi-class problems, respectively. In every hidden layer, we have used the Rectified Linear Unit (ReLU) activation function. Number of filters used in 1st, 2nd, 3rd and 4th convolutional layers are 8, 16, 64 and 64, respectively. The filter size and max-pooling size in both cases are used as 4. We have used 80% of the samples for training and the remaining are used for validation. The details of the hyper-parameters of our experiments are shown in Table 2. In all the experiments we have used Keras library[6].

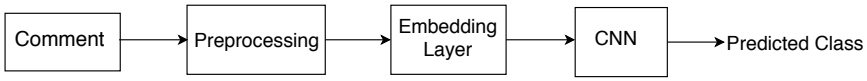

Fig. 1. Overview of the proposed model

Table 2. Hyper-parameters used in the proposed approach

Parameter description	**Values**
Maximum length_of_comment	100
Size of filters	4
Number of filters	8, 16, 64, 64
Pooling size	4
Activation function	ReLU, Sigmoid, Softmax
Number of Convolutional layers	4
Learning rate	0.001
Batch size	32
Loss function	Binary cross-entropy
Optimizer	Rms-prop
Epoch	100

[5] https://pandey.github.io/posts/transliterate-devanagari-to-latin.html
[6] http://keras.io

4 Results and Discussions

This section describes the results obtained on the English and code-mixed Hindi languages for all three Sub-tasks. For both datasets, we have used several embeddings followed by a Convolutional Neural Networks (CNN) model to classify the comments into their output classes. Using a macro-averaged F1-score and weighted F1-score, classification models have been evaluated for all the tasks. Table 3 shows the results obtained by proposed approaches to test samples with different combinations of embeddings, which have been used for training and testing. Our system achieved an approximate weighted F1-score of 69%, 55% and 60% for English Sub-task1, Sub-task2, and Sub-task3, respectively with fastText embedding. For the Hindi Sub-task1, Sub-task2 and Sub-task3, our system achieved an approximate weighted F1-score of 78%, 52%, and 66%, respectively with One-hot embedding. It is clear from the Table 3 that fastText embedding is performing better than GloVe and One-hot embeddings in case of the English dataset and One-hot embedding is performing better than fastText embedding for the Hindi dataset. Our results are ranked 17^{th}, 20^{th}, 12^{th} among participants of shared tasks for Hindi Sub-task1, Sub-task2 and Sub-task3, respectively and the results on English dataset are positioned 56^{th}, 35^{th}, 30^{th} among participants of shared tasks for Sub-task1, Sub-task2 and Sub-task3, respectively.

The proposed model has performed better for Sub-task1 and Sub-task3 which can be seen in Table 3. Table 3 also shows that misclassified instances are more for Sub-task2 in both datasets. The main reason for the misclassification of Sub-task2 is that it is a fine-grained classification task. Even for the human being, it is very difficult to differentiate among the Hate Speech, Offensive language, and Profanity; not only due to very fine but also very fade differences among these classes. Just filtering the keywords will generally result in many false-positive cases because context plays a major role in the detection of the Hate Speech, Offensive language, and Profanity. Another important reason for the misclassification of classifiers is that the datasets are very unbalanced which can be seen in Table 1.

5 Conclusion and Future Work

Hate Speech and Offensive language identification is a challenging task. Many numbers of research have been carried out in the domain of Hate Speech detection for the English language but very few researches are reported for the other languages and multi-lingual text. This research work has been focused on multi-lingual text classification, especially for Hindi and English code-mixed text. In this paper, a deep learning model for the identification of Hate Speech, Offensive contents, and Profanity on multi-domain platforms have been proposed. Three types of embeddings: One-hot, pre-trained GloVe and fastText embeddings have been used in the experiments. It has been found that fastText embedding has performed better than the other two embeddings for the English dataset and One-hot has performed better for the Hindi dataset.

Table 3. Results of classification with different Embeddings

Test Dataset	Sub-task	Embedding	Macro-F1	Weighted-F1
English	Sub-task1	One-hot	0.4803	0.6477
		GloVe	0.5308	0.6485
		fastText	0.5921	**0.6854**
	Sub-task2	One-hot	0.2425	0.5475
		GloVe	0.2049	0.4636
		fastText	0.3405	**0.5548**
	Sub-task3	One-hot	0.3335	0.5983
		GloVe	0.3585	0.5917
		fastText	0.3607	**0.5979**
Hindi	Sub-task1	One-hot	0.7827	**0.7834**
		fastText	0.5907	0.5898
	Sub-task2	One-hot	0.3486	**0.5208**
		fastText	0.1103	0.0656
	Sub-task3	One-hot	0.4878	**0.6588**
		fastText	0.4464	0.6444

Hate Speech detection is an open challenge for the research community. The social media post contains not only text but also image followed by text and even in the case of text, code-mixed languages are used. Therefore, future works on Hate Speech detection might address multi-lingual cases of several languages and consideration of multi-modal forms of social media posts to make the system more robust.

Acknowledgements

The first author would want to acknowledge the Ministry of Electronics and Information Technology (MeitY), Government of India for the financial support during the research work through the Visvesvaraya Ph.D Scheme for Electronics and IT.

References

1. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for Hate Speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 759–760. International World Wide Web Conferences Steering Committee (2017)
2. Bohra, A., Vijay, D., Singh, V., Akhtar, S.S., Shrivastava, M.: A dataset of Hindi-English code-mixed social media text for Hate Speech detection. In: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. pp. 36–41 (2018)
3. Burnap, P., Williams, M.L.: Cyber Hate Speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & Internet **7**(2), 223–242 (2015)

4. Chen, J., Yan, S., Wong, K.C.: Verbal aggression detection on Twitter comments: convolutional neural network for short-text sentiment analysis. Neural Computing and Applications pp. 1–10
5. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated Hate Speech detection and the problem of Offensive language. arXiv preprint arXiv:1703.04009 (2017)
6. Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M., Tesconi, M.: Hate me, hate me not: Hate Speech detection on facebook. In: Proceedings of the First Italian Conference on Cybersecurity (ITASEC17). pp. 86–95 (2017)
7. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 (2016)
8. Kamble, S., Joshi, A.: Hate Speech detection from code-mixed Hindi-English tweets using deep learning models. arXiv preprint arXiv:1811.05145 (2018)
9. Kumari, K., Singh, J.P., Dwivedi, Y.K., Rana, N.P.: Aggressive social media post detection system containing symbolic images. In: Conference on e-Business, e-Services and e-Society. pp. 415–424. Springer (2019)
10. Modha, S., Mandl, T., Majumder, P., Patel, D.: Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (2019)
11. Mehdad, Y., Tetreault, J.: Do characters abuse more than words? In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 299–303 (2016)
12. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th international conference on world wide web. pp. 145–153. International World Wide Web Conferences Steering Committee (2016)
13. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
14. Schmidt, A., Wiegand, M.: A survey on Hate Speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. pp. 1–10 (2017)
15. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for Hate Speech detection on Twitter. In: Proceedings of the NAACL student research workshop. pp. 88–93 (2016)