Amrita_CEN at HASOC 2019: Hate Speech Detection in Roman and Devanagiri Scripted Text

Sreelakshmi.K, Premjith.B, and Soman K.P

Center for Computational Engineering & Networking (CEN)
Amrita School of Engineering, Coimbatore,
Amrita Vishwa Vidyapeetham, India
ammaslakshmy@gmail.com

Abstract. Nowadays the usage of social media sites like Facebook and Twitter has increased rapidly which has lead to huge flooding of data in the social media sites. Though these social media sites give free opportunities to people to express and share their thoughts they also end up in spread of huge amount of hate content. In this paper we present a domain specific word embedding model for classification of English tweets to Non Hate-Offensive and Hate-Offensive and a fastText model for Hindi text classification. The classification is done using the dataset got from HASOC 2019 shared task. Deep learning algorithm is used as the classifier.

Keywords: FastText · Convolutionl Neural Network · Long short term memory, · Hate speech.

1 Introduction

Hate speech is a form of expressing aggression, profanity in verbal or non-verbal way. It can be like discriminating or using filthy language against a person or group just on grounds of their age, gender, sex, caste, economical status etc. this can even lead to huge violence or conflict between individuals or communities. So it is very important to detect them before it reaches a huge mass [1], [2], [3], [4].

For a country like India people tend to use regional language for texting or tweeting. Around half of the population speak Hindi. So the need to find hate speech in Hindi is very high. Not only human it can even corrupt chatbots. Since chatbots learn from conversation with human if it is not able to differentiate hate and non-hate content then it also starts to use it. So it is has become a huge

responsibility for the government as well as Twitter and Facebook to detect this hate speech content.

So for this,in the paper we developed two separate models to classify tweets in Hindi and English as hate or not. The English data is in roman script and the Hindi data in Devanagari script. The dataset is from HASOC 2019 shared task [5] Two samples of English data is given below **HATE**
"I love this bill, I think they should start printing them FuckTrump https://t.co/NY9CuyivGl"
Non-HATE
"All Indian spectators shd hv BalidanBadge in ground, DhoniKeepsTheGlove DhoniKeepBalidaanBadgeGlove DhoniKeepsTheGlove DhoniKeSathDesh"

2 Related Work

There are lot of works done in the area of hate speech detection, few of them are given below

Shervin et.al [6] in his paper developed a model using character n-grams, word n-grams and word skip-grams for the classification of English tweets to hate speech (HATE), offensive and no offensive content. The system used SVM as the classifier with an accuracy of 78%.

Georgious et.al [7] in his paper presents a model to detect hateful content in so-cial media. They made use of of Recurrent NeuralNetwork (RNN)classifiers and fed various features associated with user-related information, such as the users' tendency towards racism or sexism. They made use of a publicly available corpus with 16000 tweets.

Satyajith et.al [8] collected around 250000 tweets using Twitter API and trained a word2vec model and obtained the domain specific word embedding. Using these embeddings they extracted the features for 4500 Hindi-English code-mixed data and classified it as hate and non-hate. They used CNN, LSTM and BiLSTM as classifiers.

3 Proposed methodology

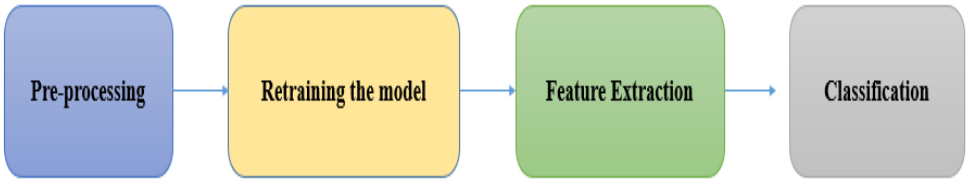

Fig. 1. Proposed Methodology

The steps used for our proposed methodology is as follows:

- **Pre-processing:** The data consists of usernames, hashtags, urls and unwanted characters. The first step was to remove these usernames, hashtags, urls , unwanted characters ,punctuations. Then the whole text was converted to lower case.
- **Retraining the model:**Once the text data is cleaned we tokenized the data and segemented it to the level of words. Each tokeninzed sentence is given to a bilingual model which is already trained on 250K code-mixed sentences. We retrained that model using gensim's word2vec with our data and generated word embedding as feature vectors from the retrained model.
- **Feature Extraction:** For the Hindi corpus fastext features were extracted. FastText consists of pre-trained model for hindi. Each sentence was tokenised and the wordvector of each word was taken from fastText model and the average of each words of a sentence was taken. The vector size for fastText was specified as 300. For english data teh vector representation for each data was taken using bilingual word embedding and the average of each words of a sentence was taken. For this word2vec was used and the vector isze was specified to be 300.
- **Classification:** For deep learning model which consists of CNN, LSTM layers were used for classification. The feature extracted matrix was fed to a embedding layer then to CNN and then LSTM. The flow diagram is given in Fig. 2

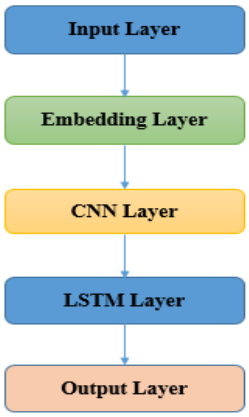

Fig. 2. Proposed Methodology

4 Conclusion

In many applications like chatbot building, content recommendation and sentiment analysis the need for hate speech detection is high. Especially for a country like India with diverse culture and language the usage of Hindi in Twitter is also high. So this paper presents a deep learning model which makes use of two different features to classify tweets in English and Hindi to hate and non-hate.

References

1. M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 75–86.
2. M. Wiegand, M. Siegel, and J. Ruppenhofer, "Overview of the germeval 2018 shared task on the identification of offensive language," 2018.
3. R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, "Benchmarking aggression identification in social media," in *Proceedings of TRAC*, 2018.
4. M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the Type and Target of Offensive Posts in Social Media," in *Proceedings of NAACL*, 2019.
5. S. Modha, T. Mandl, P. Majumder, and D. Patel, "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.
6. G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Detecting offensive language in tweets using deep learning," *arXiv preprint arXiv:1801.04433*, 2018.
7. S. Malmasi and M. Zampieri, "Detecting hate speech in social media," *arXiv preprint arXiv:1712.06427*, 2017.
8. S. Kamble and A. Joshi, "Hate speech detection from code-mixed hindi-english tweets using deep learning models," *arXiv preprint arXiv:1811.05145*, 2018.