CoCoNet
Third International Conference on Computing and Network Communications
December 18-21, 2019, Trivandrum, Kerala, India

Detection of Hate Speech Text in Hindi-English Code-mixed Data

Sreelakshmi k, Premjith B, Soman K.P
Center for Computational Engineering & Networking (CEN)
Amrita School of Engineering, Coimbatore,
Amrita Vishwa Vidyapeetham, India

AMRITA
VISHWA VIDYAPEETHAM

Outline

- Introduction
- Motivation
- Related works
- Proposed Method
- Experiments and Results
- Conclusion
- Future Work
- References

Introduction

- Hate speech is a form of verbal or non-verbal communication expressing prejudice and aggression. It is defined as an act of belittling a person or community based on their gender, age, sexual orientation, race, religion, nationality, ethnicity etc.

- Code-mixing is the usage of certain words, phrases or morphemes of one language in other language.

Eg: "***Look ye politicians suvar jaise baithe rahte hain sirf money ke liye kaam karte hain. They don't care about public"***

Motivation

- Need for use of hate speech identification in Chatbots "Microsoft's launched A.I.-powered bot called Tay, which was responding to tweets and chats, has already been shut down due to concerns with its inability to recognize when it was making offensive or racist statements."
- Investigating cyber bullying.
- Examining socio-political controversies.

Related works

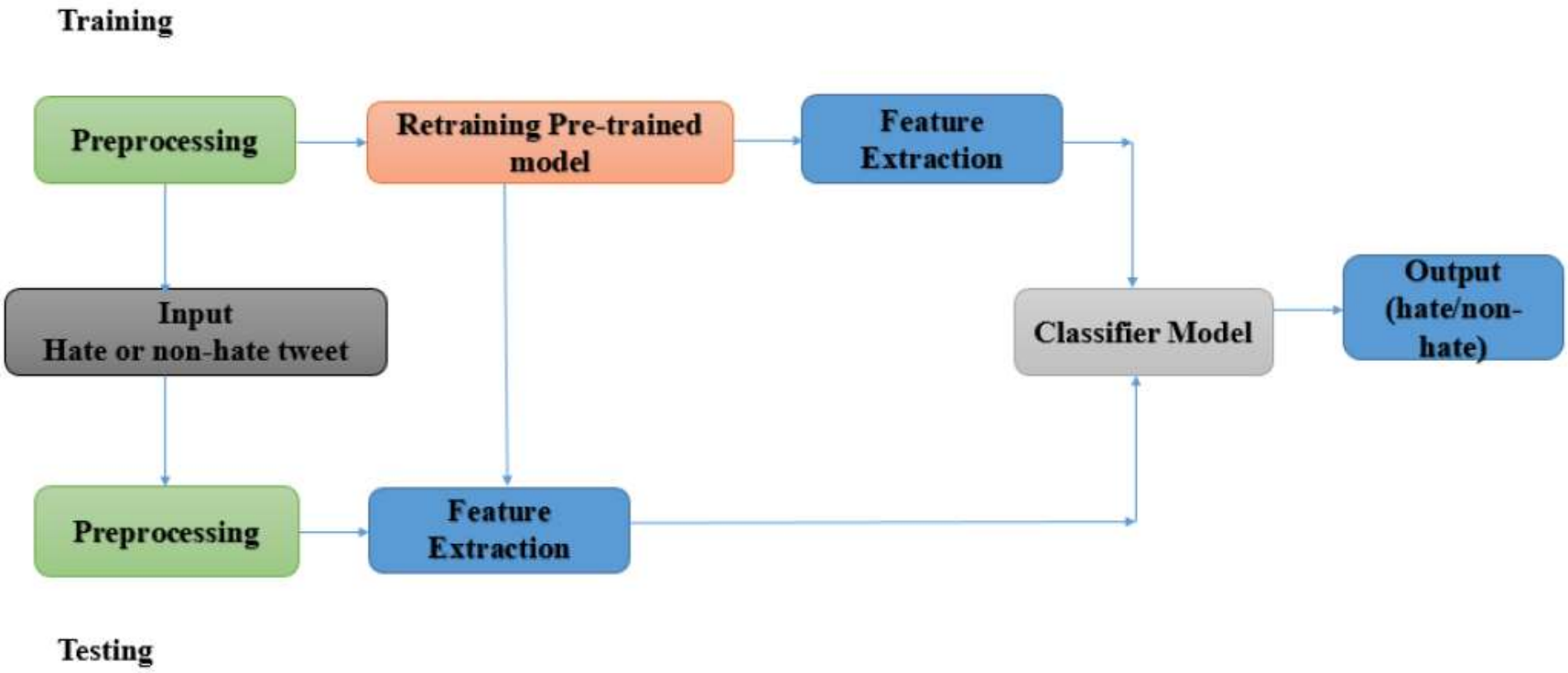
- Aditya et.al conducted the first experiment on Hindi-English Code-Mixed Social Media Text for Hate Speech Detection. They created a dataset of size 4578 and annotated it. Classifiers such as SVM and Random Forest were used [1].
- Santosh et.al developed a deep learning approach for hate and non hate classification of texts. Deep leaning algorithms such as sub-word level LSTM and LSTM with attention mechanism were applied [2].
- Puneet et.al developed a deep learning approach for the classification of Offensive Tweets in Hinglish Language. Transfer learning approach was used [3].
- Satyajit et al published a paper on hate speech detecting using deeplearning. They used a word2vec model trained on a huge corpus as feature and the classification was done my CNN [4].

Dataset Description

I used a dataset of size 10000 which was collected from 3 different sources.

Classes	Number of texts in each class
Hate	5000
Non-hate	5000

Proposed Methodology



- **Pre-processing:** As our dataset consists of tweets, there will be lot of emoticons, punctuation, special characters, URLs, hashtags and usernames. So the first step is pre-processing which involves removing them.

- **Retraining the model:** The cleaned data is then used for retraining the pre-trained model. In our work, we made use of two pre-trained model namely fastText and domain specific word embedding. In the case of fastText, each sentence along with the label is passed to the pre-trained model. Sentences are modified to ‘ label <L> <Text>’ format where L is the class label and Text is the tweet to be classified.

Eg: label <Hate> <Look ye politicians suvar jaise baithe rahte hain sirf money ke liye kaam kartehain. They don’t care about public>

For retraining the domain specific word embedding, each sentence is tokenised and passed to Gensim’s word2vec command which has an option to train.

- **Feature Extraction:** The most complex part of a NLP task is grabbing the right feature. Any deep learning or machine learning algorithm can only take number as input, so it is very important to convert the textual data to vectors. For this, the vector representation of each word is obtained from the retrained model in the case of fastText and domain specific word embedding. Where as for doc2vec the vector representation for a whole sentence is obtained. This forms the feature vector for the task.

- **Classification:** The obtained feature vectors were passed to the classifier models which predict whether a sentence is hate inducing or not. For our experiments we considered the classifiers such as linear SVM, SVM-RBF and Random Forest. Finally the performance of the classifier was evaluated by finding the parameters such as accuracy, precision, recall, F1-Score.

Experiments and Results

Performance matrix for Doc2vec as feature

Machine learning algorithms	Accuracy (%)	Precision	Recall	F1-Score
SVM -linear	0.6375	0.6375	0.6375	0.6374
SVM-rbf	0.613	0.6150	0.613	0.6112
Random Forest	**0.6415**	**0.6415**	**0.6415**	**0.6414**

Performance matrix for Word2vec as feature

Machine learning algorithms	Accuracy (%)	Precision	Recall	F1-Score
SVM -linear	0.7281	0.7288	0.7281	0.7278
SVM-rbf	**0.7511**	**0.7517**	**0.7511**	**0.7509**
Random Forest	0.7267	0.7267	0.7267	0.7266

Performance matrix for Fasttext as feature

Machine learning algorithms	Accuracy (%)	Precision	Recall	F1-Score
SVM-linear	0.8144	0.8147	0.8144	0.8143
SVM-rbf	**0.8581**	**0.8586**	**0.8581**	**0.8580**
Random Forest	0.7834	0.7848	0.7834	0.7831

Conclusion

- In this work we collected hate and non-hate annotated data from different sources and built a Fasttext model with different machine learning classifiers.
- The performance of the proposed method was compared with word level and sentence level features.

Future Work

The proposed method will be extended to the finer classification of hate tweets to offensive, abusive and profane.

References

[1] Bohra, A., Vijay, D., Singh, V., Akhtar, S.S. and Shrivastava, M.,2018, June. 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)

[2] Santosh, T.Y.S.S. and Aravind, K.V.S., 2019, January. Hate Speech Detection in Hindi-English Code-Mixed Social Media Text. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data (pp. 310-313). ACM.

[3] Mathur, P., Shah, R., Sawhney, R. and Mahata, D., 2018, July. Detecting offensive tweets in hindi-english code-switched language. In Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media (pp. 18-26)

[4] Kamble, S. and Joshi, A., 2018. Hate Speech Detection from Code-mixed Hindi-English Tweets Using Deep Learning Models. arXiv preprint arXiv:1811.05145