Corpus and Baseline System for Hate Speech Detection in Telugu-English Code-Mixed Tweets

Koushik Reddy Sane[1]*, Sairam Kolla[2]*, Sushmitha Reddy Sane[1], Vamshi Krishna Srirangam[1], and Radhika Mamidi[1]

[1] International Institute of Information Technology, Hyderabad, India
[2] Microsoft, Hyderabad, India

Abstract. The rapid increase in the use of social media by people led to an increase of hate speech on platforms like Twitter, Facebook and other blogging sites as it enabled flexibility in communication by being anonymous sometimes. This propagation has drawn significant importance to detect Hate Speech on social media texts to take effective measures. Automatic hate speech detection has applications in natural language processing like sentiment analysis and text recommendation for detecting cyber bullying. In this paper, we present the first Telugu-English code-mixed corpus consisting of tweets posted on Twitter for the presence of hate speech. The tweets are annotated with the language at word level and the class they belong to at tweet level (Hate Speech or Normal Speech). We also proposed a supervised classification system for detecting hate speech in the text using the same corpus, which achieved an average F-score of 0.81 using Support Vector Machine with Radial Basis Function kernel classifier and validated with 10-fold cross-validation.

Keywords Language detection, Linguistics, Code-mixing, RBF Kernel SVM, Random Forest, Hate-Speech.

1 Introduction

There is a surge in the amount of data on social media and online platforms due to the rapid increase in data being posted on these platforms. Many individuals are taking advantage of these platforms to put hateful content on other individuals and groups. This led to a massive increase in the amount of hate speech being posted online on Twitter, Facebook and other sites. Hate speech is generally used by users to post offensive and abusive content on individuals or groups. Hateful content may be harmful to people exposed to it in many ways and can also be used to mislead them, influence and attract them towards wrongful activities. The task of hate speech detection is thus important for authorities and social media platforms to discourage any wrongful activities such as cyber bullying.
Telugu is a Dravidian language native to India. There are about 81 million native Telugu speakers. Telugu ranks fifteenth in the Ethnologue

* These authors contributed equally to this work.

list of most-spoken languages world-wide[3]. Most of these people on social media often use more than one language to express themselves. So, they generally use code-mixed Telugu-English text to express their opinions online. Code-mixing refers to the mixing of two or more languages or language varieties in speech. This phenomenon is extended to writing as well. It is the embedding of various linguistic units such as axes, words, phrases and clauses of one language in the other [1–5].

There are no resources currently available for the detection of hate speech in Telugu-English code-mixed text. So, in this paper we present the first Telugu-English corpus for Hate Speech detection and also propose a supervised classification system for detecting hate speech in the text using the same corpus and using machine learning techniques like RBF Kernel, SVM classifier and Random Forest classifier. We did not make any distinction between borrowing and mixing in this paper. This corpus can be used by researchers working in this domain to further improve the system.

The structure of the paper is as follows. In Section 2, we review related research in the area of code mixing and hate speech detection. In Section 3, we describe the corpus creation and annotation scheme. In Section 4, we present our system architecture which includes the pre-processing steps and classification features. In Section 5, we present the results of experiments. In the Section 6, we conclude our paper.In Section 7 , we write about future work followed by references.

2 Related Work

Even though the research in code-mixed texts started decades ago, there is still a lot of scope in it [6]. POS annotated corpus [7] of code-mixed Telugu-English data was proposed and various statistical methods for automatic POS tagging of code-mixed social media data were attempted. [7] also proposed methods for language identification by creating a standard dataset for building supervised models of shallow parsing on code-mixed Telugu-English text. [8] addressed the problem of language identification in code-mixed facebook comments using character embeddings. [9] presented an end-end web-based factoid question-answering system for code-mixed Telugu-English text among others. [10] presented an evaluation dataset consisting of more than crowd-sourced questions along with their answers in Telugu-English code-mixed text among others. [11] took the first step towards understanding code-mixing in dialog processing, by recognizing intention of the code-mixed utterance in Telugu-English code-mixed corpus. [12] used code-mixed features to enhance the sentiment prediction in code-mixed Telugu-English songs data.

The previous research work on hate speech detection has mainly been focused on monolingual texts [13–15]. [13] examined methods to detect hate speech in social media. Only limited work in detection of hate speech [16, 17] in Hindi-English code-mixed texts has been done.

[3] https://www.ethnologue.com/statistics/size

3 Dataset

3.1 Data Collection

The dataset is created by extracting the tweets posted on twitter using the Twitter Scraper API[4] and manually selecting the Telugu-English code-mixed tweets from them. We used certain hashtags and keywords to get tweets from different domains using the Twitter Scraper API. We made sure that there are both ’normal speech’ and ’hate speech’ tweets from all the domains so that the classification is unbiased and that it does not lead to a biased classification system. The twitter scraper API collects each tweet in json format after which we extract the tweet content and tweet id from it.

3.2 Data Processing and Annotation

We retrieved 2,05,118 tweets from Twitter in json format which consists of information such as timestamp, URL, text, user, re-tweets, replies, full name, id and likes. The noisy tweets were initially removed from the data. The Noisy tweets consist of all the tweets which comprise only of Hashtags, mentions and URLs. The tweets which have words belonging to languages other than English and Telugu are also removed. Furthermore, the tweets which have words belonging to only Telugu or English are also removed. As a result of manual filtering, a corpus containing only code-mixed Telugu-English data is created. The code-mixed Telugu-English corpus consists of 3677 Tweets. Newly created corpus and code is available online at Github.[5]

Hate Speech Annotation Each of the tweets is manually annotated with one of the following tags: YES and NO. ’YES’ tag is given to those tweets which have hate speech in them and ’NO’ tag is given to those tweets which do not contain any hate speech in them. Following are some instances of Telugu-English code-mixed texts along with their translation in english, annotated for the presence of hate speech:

T1: *”ee movie lo war scenes chala baagunayi.... @ssrajamouli did a great job!! #bahubali1”*
Translation : ”The war scenes in the movie are very good.... @ssrajamouli did a great job!! #bahubali1”
Tag: NO

T2 :*”petrol inka diesel rates dhaarunanga perigayi , @narendramodi i hate this bloody government #gobackmodi”*
Translation : ”petrol and diesel rates have increased rapidly , @narendramodi i hate this bloody government #gobackmodi”
Tag : YES

[4] https://pypi.python.org/pypi/twitterscraper/0.2.7
[5] https://github.com/koushiksane/HateSpeech-Telugu-English

T3: *"Prati government adikarini kalchi champali.. Andaru corrupted ee!! #Corruption"*
Translation : "Every government officer has to be shot to death.. All are corrupted!! #Corruption"
Tag : YES

Here T1 is normal speech whereas T2 and T3 contain hate speech.

Language Annotation Each tweet is tokenized using white spaces as delimiters and taking into account the data set trends such as using multiple consecutive punctuation marks, mentions, etc. Each token is annotated with a language tag. One of the following tags is assigned for language: eng, tel and other, where eng stands for English, tel for Telugu and other for punctuation marks, emoticons, named entities, urls, hashtags, etc. eng is assigned to English words such as cold, good, etc. and tel is assigned to Telugu words transliterated in English such as nenu, enduku,etc. At first each token is assigned a language tag using online dictionaries and hashtags, URLs and mentions are assigned 'other' tag. Each language tag and token is checked manually to correct any language tag errors and tokenization errors. Table 1 shows an example of a tweet with language tokens:

Table 1. A tweet with token level language annotation

Token	**Language**
Prati	tel
government	eng
adikarini	tel
kalchi	tel
champali	tel
Andaru	tel
corrupted	eng
and	eng
waste	eng
fellows	eng
ee	tel
#Corruption	other

Inter Annotator agreement Two human annotators who are fluent in both English and Telugu carried out the annotation of the tweets in the dataset for the presence of hate speech. Both annotators were provided with a sample annotation set consisting of 40 tweets (20 Hate Speech and 20 Normal Speech) randomly selected from the corpus in order to have a baseline reference to distinguish between Hate Speech and Normal Speech. To validate the annotation quality, we used Cohens Kappa as a measure of inter-annotator agreement and it was calculated to be 0.82.

Dataset Analysis and Structure The dataset consists of 3677 Telugu-English code-mixed tweets out of which 1445 tweets are marked as Hate Speech and the remaining tweets are marked as Normal Speech. The corpus is structured into three files. The first file contains tweet ids followed by the corresponding tweet text. The second file consists of tweet ids followed by language annotated tweets. The third file has the annotation for presence of Hate Speech for each tweet where each tweet id is followed by its corresponding class label.

4 System Architecture

We present a baseline classification system for hate speech detection in Telugu-English code-mixed tweets using various features. We compare different machine learning models which use these features to detect hate speech.

4.1 Pre-processing

Pre-processing of the code mixed tweets is carried out as follows: All the links and urls are removed from the tweets. It is observed that most of the tweets often contain mentions which are directed towards certain users. We make a count of all those mentions before removing them since we use this count as a feature in our model. We also make a count of all the hashtags in the tweet before removing them from the tweet as we use this count as a feature in our model. All the emoticons used in the tweets are also removed. All the punctuation marks in a tweet are also removed.

4.2 Classification features

Character N-Grams Character n-gram refers to the presence or absence of contiguous sequence of n characters in a tweet. These are useful in cases when when the text is not normalised and suffers from spelling errors [18–20]. Character n-Grams are also language-independent and proved to be very effective for text classification [21]. It can be observed from previous experiments [13, 22] that character n-grams play an important role in hate speech detection. Character n-grams also help in capturing semantic meaning code-mixed language. We consider all Character n-grams for values of n ranging from 1 to 3.

Word N-Grams Word n-gram refers to presence or absence of contiguous sequence of n words or tokens in a tweet. Word n-grams have proven to be useful features for emotion detection [23] and also for hate speech detection in previous experiments [24]. We consider all n-grams for values of n ranging from 1 to 3.

Lexicon When both the training and testing dataset are drawn from the same corpus, previous research on topics like emotion detection and hate speech detection have shown that using lexicon features [25] increase the accuracy of the classification model, thus making the model better. We identified from the dataset 151 Telugu and English hate words and used them as a classification feature.

Tweet Structure The tweets which contain hate speech in our dataset are often longer than the ones containing normal speech. To capture this structure we use a group of features like the number of characters present in the tweet, the number of words in the tweet and the average word length in the tweet.

Hashtags and Mentions The tweets which contain hate speech in our dataset is observed to contain more hashtags and mentions than the ones in normal speech. To capture this we use the count of hashtags and mentions as a feature.

Table 2. F1 Score for each feature using RBF Kernel SVM classifier.

Features	**F1 Score**
All Features	**0.82**
Char N-Grams	0.81
Word N-Grams	0.62
Lexicon	0.73
Structural features	0.61
Hashtags and Mentions	0.61

5 Experiments and Results

In various experiments, feature selection algorithms have been observed to significantly improve the performance of machine learning models. We use chi square feature selection algorithm which uses chi-squared statistic to evaluate individual feature with respect to each class. To extract the best features and reduce the feature vector size to 1000[6], this algorithm was used. We use two classification techniques: Support vector machine with Radial basis function kernel and Random forest classifier.

For the detection of hate speech we use scikit-learn implementation of these methods [26]. On the corpus created to develop the system, we also perform 10-fold cross validation. For each of the individual features, 10-fold cross validation is carried out separately to observe the effect of

[6] The size of feature vector was decided after empirical fine tuning

Table 3. F1 Score for each feature using Random Forest classifier.

Features	F1 Score
All Features	0.68
Char N-Grams	0.66
Word N-Grams	0.64
Lexicon	0.67
Structural features	0.62
Hashtags and Mentions	0.61

each feature on classification. Our system achieves a best average F-score of 0.81 after running 10-fold cross validation using the RBF Kernel SVM classifier on the dataset. Table 2 and Table 3 show the F-scores achieved by each of the systems for each feature separately as well as with all the features combined. It can be observed that each feature affects each technique differently.

6 Conclusion

In this paper, we present the first Telugu-English code-mixed dataset for Hate Speech detection collected from twitter consisting of tweet ids and the corresponding annotations. The data is annotated at token level for language and at tweet level for the presence of hate speech. The corpus consists of 3677 code-mixed tweets which are annotated with hate speech or normal speech. The same dataset is used to present a baseline supervised classification developed which uses two different machine learning techniques and 10-fold cross validation. RBF Kernel SVM performed better than random forest for the classification task in our experiments. The features used in our classification system are character n-grams, word n-grams, lexicon, tweet Structure, hastags and mentions . Best F1 score of 0.81 is achieved when all the features are incorporated in the feature vector using RBF Kernel SVM as the classification system.

7 Future Work

As part of future improvements to this work, the corpus can be normalized at token level which will improve the performance of the classification system. Morphological analysis can also be done on the corpus at token level for the tokens annotated with Telugu language which is expected to improve the performance of the classification system. For better results, each word in the texts can be tagged with part-of-speech tags. Similar corpus can be created for different language pairs with the presence of other emotions such as irony, sarcasm, etc. Moreover, the annotations and experiments described in this paper can also be carried out for code-mixed texts containing more than two languages from multilingual societies, in future. The system can also be improved by exploring language based features, word embeddings and also by using other machine learning and deep learning techniques.

References

1. Ayeomoni, M.O.: Code-switching and code-mixing: Style of language use in childhood in yoruba speech community. Nordic journal of African studies **15** (2006)
2. Myers-Scotton, C.: Common and uncommon ground: Social and structural factors in codeswitching. Language in society **22** (1993) 475–503
3. Gysels, M.: French in urban lubumbashi swahili: Codeswitching, borrowing, or both? Journal of Multilingual & Multicultural Development **13** (1992) 41–55
4. Duran, L.: Toward a better understanding of code switching and interlanguage in bilinguality: Implications for bilingual instruction. The Journal of Educational Issues of Language Minority Students **14** (1994) 69–88
5. Muysken, P., Díaz, C.P., Muysken, P.C., et al.: Bilingual speech: A typology of code-mixing. Volume 11. Cambridge University Press (2000)
6. Sridhar, S.N., Sridhar, K.K.: The syntax and psycholinguistics of bilingual code mixing. Canadian Journal of Psychology/Revue canadienne de psychologie **34** (1980) 407
7. Nelakuditi, K., Jitta, D.S., Mamidi, R.: Part-of-speech tagging for code mixed english-telugu social media data. In: International Conference on Intelligent Text Processing and Computational Linguistics, Springer (2016) 332–342
8. Veena, P., Kumar, M.A., Soman, K.: An effective way of word-level language identification for code-mixed facebook comments using word-embedding via character-embedding. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE (2017) 1552–1556
9. Chandu, K.R., Chinnakotla, M., Black, A.W., Shrivastava, M.: Webshodh: A code mixed factoid question answering system for web. In: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer (2017) 104–111
10. Chandu, K., Loginova, E., Gupta, V., van Genabith, J., Neuman, G., Chinnakotla, M., Nyberg, E., Black, A.W.: Code-mixed question answering challenge: Crowd-sourcing data and techniques. In: Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching. (2018) 29–38
11. Jitta, D.S., Chandu, K.R., Pamidipalli, H., Mamidi, R.: nee intention enti? towards dialog act recognition in code-mixed conversations. In: 2017 International Conference on Asian Language Processing (IALP), IEEE (2017) 243–246
12. Reddy, G.R.R., Mamidi, R.: Addition of code mixed features to enhance the sentiment prediction of song lyrics. arXiv preprint arXiv:1806.03821 (2018)

13. Malmasi, S., Zampieri, M.: Detecting hate speech in social media. arXiv preprint arXiv:1712.06427 (2017)
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. (2017) 1–10
15. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Eleventh International AAAI Conference on Web and Social Media. (2017)
16. 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 Peoples Opinions, Personality, and Emotions in Social Media. (2018) 36–41
17. Kamble, S., Joshi, A.: Hate speech detection from code-mixed hindi-english tweets using deep learning models. arXiv preprint arXiv:1811.05145 (2018)
18. Cavnar, W.B., Trenkle, J.M., et al.: N-gram-based text categorization. In: Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval. Volume 161175., Citeseer (1994)
19. Huffman, S.: Acquaintance: Language-independent document categorization by n-grams. Technical report, DEPARTMENT OF DEFENSE FORT GEORGE G MEADE MD (1995)
20. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C.: Text classification using string kernels. Journal of Machine Learning Research **2** (2002) 419–444
21. Bermingham, A., Smeaton, A.F.: Classifying sentiment in microblogs: is brevity an advantage? In: Proceedings of the 19th ACM international conference on Information and knowledge management, ACM (2010) 1833–1836
22. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: 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) 71–80
23. Purver, M., Battersby, S.: Experimenting with distant supervision for emotion classification. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics (2012) 482–491
24. Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: Proceedings of the Second Workshop on Language in Social Media, Association for Computational Linguistics (2012) 19–26
25. Mohammad, S.M.: # emotional tweets. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, Association for Computational Linguistics (2012) 246–255
26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.,

et al.: Scikit-learn: Machine learning in python. Journal of machine learning research **12** (2011) 2825–2830