OffTamil@DravideanLangTech-EACL2021: Offensive Language Identification in Tamil Text

Disne Sivalingam
Eastern University, Sri Lanka
disnesiva777@gmail.com

Sajeetha Thavareesan
Eastern University, Sri Lanka
sajeethas@esn.ac.lk

Abstract

In the last few decades, Code-Mixed Offensive texts are used penetratingly in social media posts. Social media platforms and online communities showed much interest on offensive text identification in recent years. Consequently, research community is also interested in identifying such content and also contributed to the development of corpora. Many publicly available corpora are there for research on identifying offensive text written in English language but rare for low resourced languages like Tamil. The first code-mixed offensive text for Dravidian languages are developed by shared task organizers which is used for this study. This study focused on offensive language identification on code-mixed low-resourced Dravidian language Tamil using four classifiers (Support Vector Machine, random forest, k- Nearest Neighbour and Naive Bayes) using χ^2 feature selection technique along with BoW and TF-IDF feature representation techniques using different combinations of n-grams. This proposed model achieved an accuracy of 76.96% while using linear SVM with TF-IDF feature representation technique.

1 Introduction

Offensive language is the key concern of technical companies nowadays due to exponential growth in number of internet users around the world and since these people are from different culture, race, religion, origin, gender and nationality (Chakravarthi et al., 2021). Internet gives more freedom to people to express their opinions freely in different forms such as blogs, forums and social media platforms (e.g., Facebook, Twitter, and YouTube) (Suryawanshi and Chakravarthi, 2021). It is noted that the usage of social media among the people has increased rapidly since last decade (Hande et al., 2020). People can express their opinion in a positive way as well as negative way (Chakravarthi and Muralidaran, 2021). So the offensive comments are now avoidable in those platforms, the problem has to be solved. It has a negative impact on society and individuals (Puranik et al., 2021; Hegde et al., 2021; Yasaswini et al., 2021; Ghanghor et al., 2021b,a). There are huge amount of researches are found in identifying the offensive words in English language. There are so many publicly available corpora in English language.

People from multilingual society will add comments and reviews with the mixing of vocabulary and syntax of multiple languages in the same sentence (Priyadharshini et al., 2020; Jose et al., 2020). So it is a big challenge to identify the offensive words from the Dravidian languages (Mandl et al., 2020; Chakravarthi et al., 2020c). With a history stretching back to 600 BCE, the Tamil language is one of the world's longest-surviving classical languages. Poetry, especially Sangam literature, which is made up of poems written between 600 BCE and 300 CE, dominates Tamil literature. All the Dravidian languages evolved from Tamil language. The first attempt to create this offensive language has done by Chakravarthi et al. (2020b). We used code-mixed texts from the YouTube reviews as a corpus. In this paper, we proposed a method using linear SVM with chi^2 feature selection based approach to find offensive language in Tamil language (Thavareesan and Mahesan, 2019, 2020a,b).

The rest of the paper is organised as follows. Section 2 describes related works. Section 3 presents proposed method. Section 4 presents experimental setup and the results. Finally, discussion and conclusion are in Section 5.

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 346–351

2 Related work

Detecting offensive language is not an easy task. Many researchers has proposed many methods and algorithms to detect offensive language content on the web.

Alakrot et al. (2018) trained a Support Vector Machine (SVM) classifier using world-level features (with and without preprocessing), SVM (with and without the normalisation), n-gram level features. They used Arabic language corpus which was pre-processed with tokenization, filtering and normalisation. Their classifier achieved an accuracy of 90.05 upon using 10-fold cross-validation. It has been observed that the n-gram features improves the classifier's performance. On the contrary, the combination of stemming and n-gram features harms precision.

Ibrohim and Budi (2018) used machine learning approach with simple word n-gram and character n-gram features and trained Naïve Bayes, Support vector machine, and Random Forest Decision Tree Classifiers. Discussed abusive language detection in the Indonesian language corpus. 10-fold cross validation technique is used to evaluate the classification result. If the corpus labeled with three classes(non-abusive language, abusive language but not offensive and offensive language) Naïve bayes classifier with the combination of word unigram and bi-grams features gives the best result 70.06% of F1-score. If the corpus labeled with two classes(abusive language or non-abusive language) then all the classifiers gives higher results. The have Concluded that the classifying into three classes is more difficult than just classifying abusive language or non-abusive language.

Waseem and Hovy (2016) analyzed the impact of various extra-linguistic features in conjunction with character n-grams for hate speech detection. Corpus is normalized for pre-processing. Uni grams, bi grams, tri grams and four grams were Collected for each tweets. Logistic regression classifier and 10-fold cross validation were used to test the influence of various features on prediction. Found that character n-grams of length up to 4 along with gender as an additional feature provides the best results.

Nayel and Shashirekha (2019) In this research they have used corpus in three languages (English, Germany and Hindi). TF-IDF vectors has been computed for all the posts in the training set. For pre-processing all un-informative tokens such as urls, digits and special characters have been removed from all the posts. Trained using linear classifier, SVM and MLP classifier, tested with 5-fold cross-validation approach. Classified as three tasks as A, B and C respectively. Unlabeled instance into one of the two predefined categories classified for task A, unlabeled instance into one of the three predefined categories classified for task B and same instance into one of the two predefined categories classified for task C. Concluded that for English language SVM outperforms for all three tasks. For German language MLP out performs for task A and B. For Hindi language MLP outperforms for task A and SVM gives better results for task B.

Chakravarthi et al. (2020b) created a corpus containing 15744 YouTube comments and posts as a code-mixed dataset. Following classifiers are used to classify the corpus Support Vector Machine(SVM), Logistic Regression (LR), k-Nearest Neighbour (k-NN), Decision tree, Random Forest, Multinomial Naive Bayes,BERT Multilingual, 1DConv-LSTM, DME, CDME. They have concluded that Random Forest classifier, Logistic regression and decision tree gives the best results.

Chakravarthi et al. (2020a)analysed in code-mixed Dravidian text from social media that aims at classifying YouTube comments. The hundred and nineteen teams partici-pated in the task, and a total of 32 teams for Tamil and 28 teams Malayalam submitted the results.They trained on the unbalanced dataset. The methods proposed by participants ranged from traditional machine learning models with features based approaches to using state-of-the-art embedding methods in deep learning models.The best performing run achieved weighted F1-score of 0.65 and 0.74 for Tamil and Malayalam respectively.

3 Methodology

Offensive language identification aims to identify the offensive text written in Tamil language. In this paper we experimented a method using four classifiers and χ^2 feature selection technique to identify the offensive text written in Tamil documents. The overall framework of Offensive language identification is shown in Figure 1.

The methodology applied in this research is divided into four parts. Subsection 3.1 describes the corpus used, subsection 3.2 describes the preprocessing applied, subsection 3.3 describes the feature selection and representation techniques and

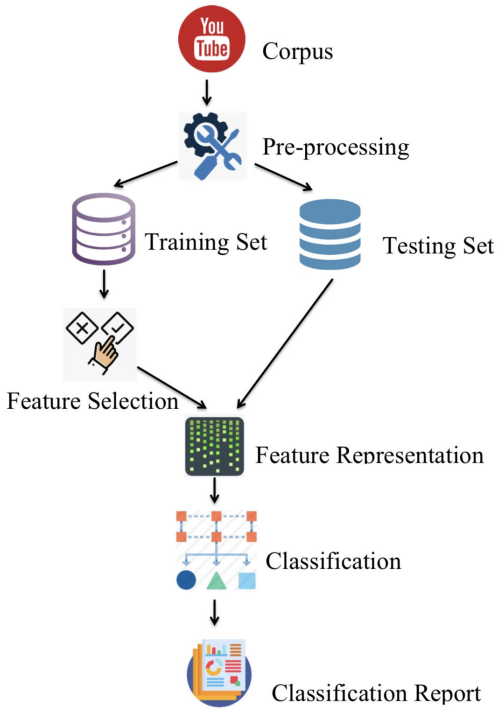

Figure 1: Overall framework of Offensive Language Identification

the subsection 3.4 describes classifiers used.

3.1 Corpus

Offensive language identification is a natural language processing (NLP) task which aims to moderate and minimise offensive content in social media. It is an active research area in both fields: academic and industry. There is an increasing demand for offensive language identification on social media texts written in code-mixed. Code-mixing is the text written in two or more languages or language varieties in speech. The shared task Organized by *dravidianlangtech* presents a new gold standard corpus for offensive language identification of code-mixed text in Dravidian languages such as Tamil-English (Chakravarthi et al., 2020b), Malayalam-English (Chakravarthi et al., 2020a) and Kannada-English (Hande et al., 2020). As far as we know, this is the first shared task on offensive language identification in Dravidian languages. The goal of this task is to identify offensive language content of the code-mixed corpus of comments in Dravidian Languages collected from social media. This task aims to classify the given comment into Not-offensive, offensive-untargeted, offensive-targeted-individual, offensive-targeted-group, offensive-targeted-other, or Not-in-indented-language. Description of corpus used for this research is shown in Table 1 and Table 2.

Label	Tamil
Not offensive	25,425
Not in indented language	1,454
Offensive-Targeted-Insult-Individual	2,343
Offensive Targeted Insult Group	2,557
Offensive Targeted Insult Other	454
Offensive Untargeted	2,906

Table 1: Class distribution of Tamil corpus.

Language	Train	Dev	Test
Tamil	35139	4388	4392

Table 2: Corpus statistic of Tamil.

3.2 Pre-processing

In the pre-processing phase all un-informative tokens such as symbols, numbers, URLs and non-Tamil words in other language fonts are removed.

3.3 χ^2 Feature Selection and Representation

A χ^2 test is used in statistics to test the independence of two events which is used as the feature selection method in this proposed method. Given the data of two variables, we can get observed count O and expected count E. χ^2 measures how expected count E and observed count O deviates each other.

BoW and TF-IDF are used as the feature representation techniques in this proposed method. BoW is used to represent the number of times a word appears in a comment. Equation of BoW is shown in Equation 1.

$$\texttt{BoW} = \texttt{No. of times word w occurred} \quad (1)$$

TF-IDF is the multiplication of Term Frequency (TF) and Inverse Document Frequency (IDF) scores whereas TF algorithm is the ratio of number of times the word appeared in a comment compared to the total number of words in that comment and IDF is a scoring of how rare the word is across comments.

$$\texttt{TF} = \frac{\texttt{No. of times w appeared in a document}}{\texttt{Total no. of words in that comment}} \quad (2)$$

Classifier	BoW	TF-IDF
k-Nearest Neighbour	75.34	74.73
Linear SVM	75.57	75.62
Logistic Regression	31.22	41.82
Random Forest	75.34	73.66

Table 3: Results of feature-set-1 with BoW and TF-IDF

Feature set	50	100	500	1000	1500	2000	2500	3000	3500
Feature-Set-1	74.77	75.39	75.57	75.93	76.17	76.16	76.14	**76.48**	75.98
Feature-Set-2	74.77	75.07	75.59	75.59	75.57	75.93	76.19	**76.25**	75.18
Feature-Set-3	74.77	74.93	75.58	75.24	75.34	75.73	76.16	**76.18**	74.78

Table 4: Results of linear SVM with BoW

Feature set	50	100	500	1000	1500	2000	2500	3000	3500	4000
Feature-Set-1	74.77	74.68	75.61	76.2	76.12	76.25	76.55	76.19	**76.96**	76.34
Feature-Set-2	74.64	74.59	74.77	75.09	75.84	76.09	76.23	76.28	**76.78**	76.12
Feature-Set-3	74.51	74.48	74.65	74.92	75.56	75.68	76.01	76.25	**76.74**	75.88

Table 5: Results of linear SVM with TF-IDF

$$\texttt{IDF(w)} = \frac{\texttt{No. of comments}}{\texttt{No. of comments containing word w}} \quad (3)$$

3.4 Training the Classifier

In this training phase the classifiers such as linear Support Vector Machine(SVM), random forest, k-Nearest Neighbour (k-NN) and Naive Bayes are used along with BoW and TF-IDF feature representation techniques using different combination of n-grams.

Firstly, we selected the most relevant features using χ^2 feature selection technique and represented them using BoW and TF-IDF feature representation techniques. We used different combinations of uni gram, bi gram and tri grams of words in training corpus to create the vocabulary of the proposed method. We used three feature sets to experiment this proposed method. They are listed below:

- Feature-Set-1: Word Uni gram.
- Feature-Set-2: Word Uni gram and bi gram.
- Feature-Set-3: Word Uni gram, bi gram and tri gram.

These three feature sets are represented using BoW and TF-IDF feature representation techniques and trained using four classifiers mentioned above. We performed six experiments per each classifier. Moreover, we have repeated these six experiments for linear SVM by selecting varying number of features using χ^2 feature selection technique.

3.5 Evaluation

Evaluation of these experiments are performed by calculating accuracy as in equation 4.

$$\texttt{Acc} = \frac{\texttt{No. of correctly classified comments}}{\texttt{Total no. of comments in the Corpus}} \times 100 \quad (4)$$

4 Experimental Setup and Results

Test results of Feature-Set-1 with BoW and TF-IDF feature representation techniques of four classifiers are listed in Table 3.

It is observed that from the table that linear SVM performs better than other three classifiers for this corpus while using Feature-Set-1. Therefore we continued our experiments using linear SVM.

Results of BoW feature representation technique with varying values of features for all three feature sets are shown in Table 4.

Test results of linear SVM with TF-IDF feature representation technique are shown in Table 5

5 Discussion and Conclusion

In this paper we proposed χ^2 feature selection technique based Offensive language identification

method. We compared results of four classifiers and observed that linear SVM outperformed all other classifiers tested here. Moreover we have checked the influence of different feature sets and found that Feature-Set-1 performs better than other two feature sets for both feature representation techniques. Another finding of this research is that we can be able to get better results with least number of features while using χ^2 feature selection technique. The highest accuracy of 76.96% is obtained while using TF-IDF feature representation with linear SVM. More over we obtained highest accuracy while using 3500 features as vocabulary for both BoW and TF-IDF feature representation techniques.

References

Azalden Alakrot, Liam Murray, and Nikola S Nikolov. 2018. Towards accurate detection of offensive language in online communication in arabic. *Procedia computer science*, 142:315–320.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan V, Elizabeth Sherly, and John Philip McCrae. 2021. Findings of the shared task on Offensive Language Identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. 2020c. Overview of the Track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 21–24, New York, NY, USA. Association for Computing Machinery.

Nikhil Kumar Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021a. IIITK@DravidianLangTech-EACL2021: Offensive Language Identification and Meme Classification in Tamil, Malayalam and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, Online. Association for Computational Linguistics.

Nikhil Kumar Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021b. IIITK@LT-EDI-EACL2021: Hope Speech Detection for Equality, Diversity, and Inclusion in Tamil, Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, Online.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63, Barcelona, Spain (Online). Association for Computational Linguistics.

Siddhanth U Hegde, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. UVCE-IIITT@DravidianLangTech-EACL2021: Tamil Troll Meme Classification: You need to Pay more Attention. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Muhammad Okky Ibrohim and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222–229.

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A Survey of Current Datasets for Code-Switching Research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Forum for*

Information Retrieval Evaluation, FIRE 2020, page 29–32, New York, NY, USA. Association for Computing Machinery.

Hamada A Nayel and HL Shashirekha. 2019. Deep at hasoc2019: A machine learning framework for hate speech and offensive language detection. In *FIRE (Working Notes)*, pages 336–343.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Mani Vegupatti, and John P. McCrae. 2020. Named Entity Recognition for Code-Mixed Indian Corpus using Meta Embedding. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 68–72.

Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@LT-EDI-EACL2021-Hope Speech Detection: There is always hope in Transformers. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on Troll Meme Classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based Part of Speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@DravidianLangTech-EACL2021: Transfer Learning for Offensive Language Detection in Dravidian Languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.