CEN-Tamil@DravidianLangTech-ACL2022: Abusive Comment detection in Tamil using TF-IDF and Random Kitchen Sink Algorithm

Prasanth S N, R Aswin Raj, Adhithan P, Premjith B, Soman K P
Centre for Computation Engineering and Networking (CEN)
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
b_premjith@cb.amrita.edu

Abstract

This paper describes the approach of team CENTamil used for abusive comment detection in Tamil. This task aims to identify whether a given comment contains abusive comments. We used TF-IDF with char-wb analyzers with Random Kitchen Sink (RKS) algorithm to create feature vectors and the Support Vector Machine (SVM) classifier with polynomial kernel for classification. We used this method for both Tamil and Tamil-English datasets and secured first place with an f1-score of 0.32 and seventh place with an f1-score of 0.25, respectively. The code for our approach is shared in the GitHub repository.[1]

1 Introduction

Abusive speech refers to any form of communication done with the intention to humiliate, or spread hatred against a vulnerable individual or a vulnerable group on the basis of gender, race, religion, ethnicity, skin color or disability using abusive or vulgar words. It causes psychological effects on the targeted individual and leading them towards unrightful act.

In recent years, there has been significant growth in the volume of digital content exchanged by people through social media. Online social networks have grown in importance, becoming a source for acquiring news, information, and entertainment. Despite the apparent advantages of using online social networks, there is an ever-increasing number of malevolent actors who use social media to harm others.

The goal of the shared task is to identify abusive comments in Tamil and code-mixed Tamil-English data developed by collecting YouTube comments. The code-mix Tamil-English dataset consists of eight different classes namely, 'Counter-speech', 'Homophobia', 'Hope-Speech', 'Misandry','Misogyny', 'None-of-the-above', 'Transphobic', 'Xenophobia'. In addition to the aforementioned eight classes, the Tamil dataset consists of one more class, 'Not-Tamil'.

We used Random Kitchen Sink (RKS) (Sathyan et al., 2018) algorithms with character word-bound based Term Frequency-Inverse Document Frequency (TF-IDF) (Barathi Ganesh et al., 2016) for text representation and classification was performed using Support Vector Machines (SVM) classifier (Soman et al., 2009), (Premjith et al., 2019). The rest of the paper is organised as follows: Section 2 describes about the related works, Section 3 describes about the Datasets, Section 4 describes about the preprocessing and different methods used, Section 5 describes about the result and analysis and Section 6 concludes the paper.

2 Related Works

Analysis of Online Social Networks' content is an active research area with tasks like Offensive Language Identification and Hope Speech Detection. Recent work in Hope Speech Detection in Dravidan languages includes the shared task on hope speech detection in LT-EDI in EACL (Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021). Abusive language detection for other languages has been done in literature (Jahan et al.; Akhter et al.; Sundar et al.) but as far as we know, this is the first shared task on abusive detection in Tamil at this fine-grained level.

We used TF-IDF because it helps in understanding the importance of a word in the corpus(Sammut and Webb, 2010) and we used Random Kitchen Sink (RKS) on top of it because RKS helps in mapping the data from the feature space to a higher dimensional space(S et al.). We used SVM because of its ability to perform well in the higher dimensional data(Cortes and Vapnik, 1995).

[1] https://github.com/Prasanth-s-n/CEN-Tamil_Abusive_Comment_Detection

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 70 - 74

3 Dataset

The organisers of the Abusive Comment Detection shared task provided two datasets, where one contains Tamil comments and the another one contains code-mixed Tamil-English comments(Chakravarthi, 2020).

Table 1 shows the classwise distribution of data for Tamil dataset and Table 2 shows the classwise distribution of data for Tamil-English dataset. Table 3 shows the statistics of the datasets given for this task.

4 Methods

We started with preprocessing the YouTube comments in the datasets, and the preprocessed texts were converted into vectors. The classification of the YouTube comments was carried out by supplying the text vectors to a classifier, SVM. Figure 1 shows the pipeline of the methodology we followed for this task.

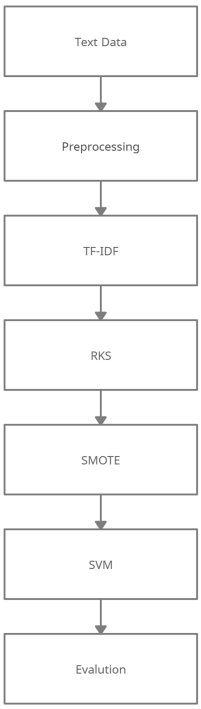

Figure 1: Steps involved in training our submitted Model

4.1 Preprocessing

The datasets used for this shared task contains comment with words in Tamil and English. The social media text contains noise such as URLs, Hashtags and other unwanted characters such as punctuation. The preprocessing step includes the removal of noise to make the data clean. In this step, we removed emojis, hashtags, URLs and non-alphabetical characters.

4.2 Text Representation and Classifier

Text Representation is one of the fundamental task in Natural Language Processing where the text is represented with array of numbers. We used TF-IDF with RKS for text representation. TF-IDF (Term Frequency - Inverse Document Frequency) is vector semantics text representation technique which uses the frequency of a word in a given document and the number of documents in which the particular word is present(Sammut and Webb, 2010). We used character character word bound n-grams based TF-IDF with RKS for increasing the dimension of the data. We used different max features for TF-IDF and different dimension size for RKS.

From Table 2 and Table 3 it is evident that the datasets are highly imbalanced. In order to solve this class imbalance problem, we used a oversampling technique called SMOTE (Synthetic Minority Over-sampling Technique) with k neighbors being 1. It uses the k-nearest neighbor algorithm by creating a plane based on the k neighbors and generates new samples from the plane(Chawla et al.). In our work, we used SMOTE by utilizing imblearn API.

RKS (Random Kitchen Sink) is an effective method for mapping features from their feature space to a higher dimensional space without explicit kernel mapping by using Fourier coefficients. The methodology is able to emulate the characteristics of the shift invariant kernel functions satisfactorily (S et al.).

SVM Classifier is used for classification due to its ability to perform well in case of higher dimensional data (Cortes and Vapnik, 1995). We used Polynomial Kernel with the regularization parameter set to 1. We used Scikit-learn API to do the classification task.

4.3 Hyperparameters

Hyperparameter tuning is an important step in building a model. The model performance is heavily dependent on hyperparameters. We selected the hyperparameters from a set of values and reported the models with hyperparameters that gave better result while valdating the model (trained and validated on Tamil Dataset) in terms of F1-score. Table

Class Name	Train Data	Val Data	Test Data
Counter-speech	149	36	47
Homophobia	35	8	8
Hope-Speech	86	11	26
Misandry	446	104	127
Misogyny	125	24	48
None-of-the-above	1296	346	416
Not-Tamil	2	0	0
Transphobic	6	2	2
Xenophobia	95	29	25

Table 1: Classwise distribuiton of Tamil dataset

Class Name	Train Data	Val Data	Test Data
Counter-speech	348	95	88
Homophobia	172	43	56
Hope-Speech	213	53	70
Misandry	830	218	292
Misogyny	211	50	57
None-of-the-above	3720	919	1143
Transphobic	157	40	58
Xenophobia	297	70	95

Table 2: Classwise distribuiton of Tamil-English dataset

Language	Train	Valid	Test
Tamil	2240	560	699
Tamil-English	5948	1488	1859

Table 3: Shared Task Dataset Statistics

Hyperparameter	Value
TFIDF ngram range	(1,5)
TFIDF Max-Features	2000
RKS Dimension	10*Max-Features
SVM Kernel	Poly
SVM C Parameter	100

Table 4: Hyperparameter used for building the models

4 shows the optimal hyperparameter used for building the models and we used the same parameters for both the datasets.

5 Result and Analysis

We experimented with four different machine learning classification models. All the four models initially uses TF-IDF with char-wb analyzer and max features being 2000 and SVM classifier with polynomial kernel and regularization parameter being 100. Model-1 uses only SVM and TF-IDF. Model-2 additionally uses SMOTE oversampling technique. Model-3 additionally uses RKS for increasing the size of text representation. Model-4 additionally uses RKS and SMOTE. The classification models' performance are measured in terms of macro average Precision, marco average Recall and marco average F1-Score across all the classes. Table 5 and Table 6 shows the performance of the models on validation dataset, Tamil and Tamil-English respectively. We used Model-4 in both cases due to its higher macro F1-score and secured rank 1 for Tamil and rank 7 for Tamil-English in the shared task (Priyadharshini et al., 2022). Table 7 contains the result obtained for Tamil and Tamil-English test datasets using model 4.

By comparing our predictions from the model-4 for Tamil dataset against the ground truth of Tamil test data, we found that None-of-the-above class has the highest individual f1-score of 0.83 and Transphobic class has the lowest individual f1-score of 0 since it has only two data points in the test data. In Tamil-English dataset, None-of-the-above class has the highest individual f1-score of 0.85 and Misogyny class has the lowest individual f1-score of 0.18. Table 8 contains the class-wise f1-score for both the datasets.

Model	Precision	Recall	F1-Score
Model-1	**0.51**	0.28	0.31
Model-2	0.41	0.31	0.33
Model-3	0.50	0.29	0.32
Model-4	0.43	**0.32**	**0.34**

Table 5: Results For Tamil Validation dataset

Model	Precision	Recall	F1-Score
Model-1	0.68	0.40	0.47
Model-2	0.64	0.45	0.51
Model-3	**0.70**	0.42	0.48
Model-4	0.67	**0.46**	**0.52**

Table 6: Results For Tamil-English Validation dataset

Dataset	Precision	Recall	F1-Score
Tamil	0.38	0.29	0.32
Tamil-English	0.30	0.23	0.25

Table 7: Results For Test datasets

Class Name	Tamil	Tamil-English
Counter-speech	0.35	0.38
Homophobia	0.67	0.37
Hope-Speech	0.26	0.23
Misandry	0.71	0.68
Misogyny	0.54	0.18
None-of-the-above	0.83	0.85
Transphobic	0.00	0.32
Xenophobia	0.18	0.65

Table 8: Classwise F1-Score Obtained Using Model-4

6 Conclusion and Future work

This paper briefs the submission of team CEN-Tamil to the shared task at ACL 2022 on Abusive Comments Detection in Tamil. We experimented character word bound n-grams based TF-IDF with and without RKS. The max features for the TF-IDF is taken to be 2000. The highly class imbalance problem was solved using SMOTE. We also reported the results obtained without using SMOTE. The SVM classifier with polynomial kernel and 100 as regularization with TF-IDF, RKS and SMOTE gave high macro F1-Score of 0.32 for Tamil and 0.25 for Tamil-English which secured first and seventh place in the shared task respectively.

We have not explored Transformers based approaches for abusive comment detection. As future works we like to experiment with different transformers like BERT and LaBSE along with deep learning architecture like LSTM and CNN to improve the results.

References

Muhammad Pervez Akhter, Zheng Jiangbin, Irfan Raza Naqvi, Mohammed AbdelMajeed, and Tehseen Zia. Abusive language detection from social media comments using conventional machine learning and deep learning approaches.

HB Barathi Ganesh, M Anand Kumar, and KP Soman. 2016. From vector space models to vector space models of semantics. In *Forum for Information Retrieval Evaluation*, pages 50–60. Springer.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

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*, pages 61–72, Kyiv. Association for Computational Linguistics.

Nitesh V. Chawla, Kevin W. Boyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Maliha Jahan, Istiak Ahamed, Md. Rayanuzzaman Bishwas, and Swakkhar Shatabda. Abusive comments detection in bangla-english code-mixed and transliterated text.

B Premjith, KP Soman, M Anand Kumar, and D Jyothi Ratnam. 2019. Embedding linguistic features in word embedding for preposition sense disambiguation in english—malayalam machine translation context. In *Recent Advances in Computational Intelligence*, pages 341–370. Springer.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on Abusive Comment Detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Athira S, Harikumar K, Sowmya V, and Soman K P. Parameter analysis of random kitchen sink algorithm.

Claude Sammut and Geoffrey I. Webb, editors. 2010. *TF-IDF*, pages 986–987. Springer US, Boston, MA.

Dhanya Sathyan, Kalpathy Balakrishnan Anand, Aravind Jaya Prakash, and Bhavukam Premjith. 2018. Modeling the fresh and hardened stage properties of self-compacting concrete using random kitchen sink algorithm. *International journal of concrete structures and materials*, 12(1):1–10.

KP Soman, R Loganathan, and V Ajay. 2009. *Machine learning with SVM and other kernel methods*. PHI Learning Pvt. Ltd.

Arunima Sundar, Akshay Ramakrishnan, Avantika Balaji, and Thenmozhi Durairaj. Hope speech detection for dravidian languages using cross-lingual embeddings with stacked encoder architecture.