Offensive Language Identification on Multilingual Code Mixing Text

Jyoti Kumari[1], Abhinav Kumar[2]

[1]Department of Computer Science & Engineering, National Institute of Technology Patna, Patna, India
[2]Department of Computer Science & Engineering, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, India

Abstract

Hate and offensive language identification from social media platforms have been an active area of research for the researchers. As the user-generated social media posts contain several grammatical errors, spelling mistakes, and non-standard abbreviations, the identification of hate and offensive posts have become a challenging task. In non-native English-speaking countries, social media texts are often code mixed or script mixed/switched, making it considerably more difficult. This work proposes ensemble-based models for the identification of offensive language from Tamil script-mixed, Tamil code-mixed, and Malayalam code-mixed social media posts. The use of character n-gram TF-IDF features with the ensemble-based model have shown promising results with weighted F_1-scores of 0.83 for Tamil script-mixed, 0.67 for Tamil code-mixed, and 0.77 for Malayalam code-mixed social media posts. The code for the proposed models is available at https://github.com/Abhinavkmr/Dravidian-hate-speech.git

Keywords

Hate speech, Dravidian language, Code-mixed, Social media

1. Introduction

The technology advancement aimed to ease the people life has attracted much users towards digitization specially the young generation. Today, the life of a person is incomplete without social media [1]. Online social media platforms like Facebook, Twitter etc. allow users to connect with their friends, make friends, share their thoughts, pictures, videos, etc.[2]. The users are also increasing day by day. Along with huge data generation [3, 4], the use of offensive language or terminologies are also increasing at a rapid pace[1]. This is generating a serious issue to the sustainable society [5].

The offensive language broadly comprises of hate speeches including race, age, sexual orientation, disability, religion, and racism against violence or hate promoting contents [2]. These contents impact a user's mental health terribly leading to depression, sleeplessness, and even suicide. Few countries have already adopted strict rules or policies against such activities caused due to freedom of expression or freedom to write. [6].

FIRE 2021: Forum for Information Retrieval Evaluation, December 13-17, 2021, India
j2kumari@gmail.com (J. Kumari); abhinavanand05@gmail.com (A. Kumar)

[1]https://ucr.fbi.gov/hate-crime/
[2]https://support.google.com/youtube/answer/2801939?hl=en

The manual identification of hate speech is impossible due to various reasons like huge amount of data, different policies, various types of hate speeches etc. Rather it should be done automatically [6, 7]. Few researchers have tried to build such models [8, 9, 10]. Agarwal and Sureka [11] extracted linguistic, semantic, and sentimental features and learned an ensemble classifier to detect racist contents. Kapil et al. [6] proposed LSTM and CNN based model to identify the hate speech in social media posts whereas, Badjatiya et al. [12] learned semantic word embedding to classify each tweet as racist, sexist, or neither. Kumari and Singh [13] presented a deep learning model to detect hate speech for English text. A considerable amount of research work is present for English language in the literature. The major challenges arises for the code-mixed and script-mixed sentences due to the unavailability of a sufficient datasets.

The purpose of this study is to recognize the hate speech from Tamil script-mixed, Tamil code-mixed, and Malayalam code-mixed social media posts into offensive and not-offensive classes. The proposed model is validated with the datasets provided by HASOC-Dravidian-CodeMix-FIRE2021 challenge [14]. Two different tasks were given by the organizer: (i) Task-1: classification of YouTube Tamil comments into offensive and not-offensive classes, (ii) Task-2: classification of code-mixed Tamil and Malayalam tweets into offensive and not-offensive classes. The current paper explores the usability of character-level features with ensemble-based model to classify Tamil script-mixed, Tamil code-mixed, and Malayalam code-mixed social media posts into offensive and not-offensive classes.

The rest of the article is organized as follows; The proposed methodology is explained in Section 2. The experiment setting and obtained results are discussed in Section 3. Finally, the paper is concluded in Section 4.

2. Methodology

This section discusses the proposed methodology for the identification of offensive social media posts. The proposed model is validated with three datasets [14]: (i) Tamil script-mixed, (ii) Tamil code-mixed, and (iii) Malayalam code-mixed social media posts. The overall data statistic used in this study can be seen in Table 1. Two different ensemble-based methods are proposed: (i) Ensemble of Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) for the Tamil code-mixed and Malayalam code-mixed social media posts (see Figure 1, (ii) Ensemble of AdaBoost classifier trained on three different validation split (see Figure 2).

2.1. Ensemble-based model for Tamil and Malayalam code-mixed dataset

The systematic diagram for the proposed ensemble-based model for the identification of offensive Tamil and Malayalam code-mixed social media posts can be seen in Figure 1. Character N-gram TF-IDF (Term-Frequency Inverse-Document-Frequency) features were given to SVM, LR, and RF classifiers. The predicted probabilities from each of the classifiers for offensive and not-offensive classes is then averaged to get the final probability values for each of the classes. The higher probability gets the final class label (as can be seen in Figure 1). The experiment has been performed with different combinations of character (1-gram to 6-gram) TF-IDF features. In this extensive experiment, it is observed that the first 30,000 one to six-gram character TF-IDF features have performed best. The results of the proposed model are listed in section 3.

Table 1
Overall data statistic for the Tamil script-mixed, Tamil code-mixed, and Malayalam code-mixed dataset

Language	Class	Train	Validation	Test
Tamil (Script-mixed)	Offensive	1,153	-	118
	Not-offensive	4,724	-	536
	Total	5,977	-	654
Tamil (Code-mixed)	Offensive	1,980	475	395
	Not-offensive	2,019	465	605
	Total	3,999	940	1,000
Malayalam (Code-mixed)	Offensive	1,952	478	324
	Not-offensive	2,047	473	675
	Total	3,999	951	999

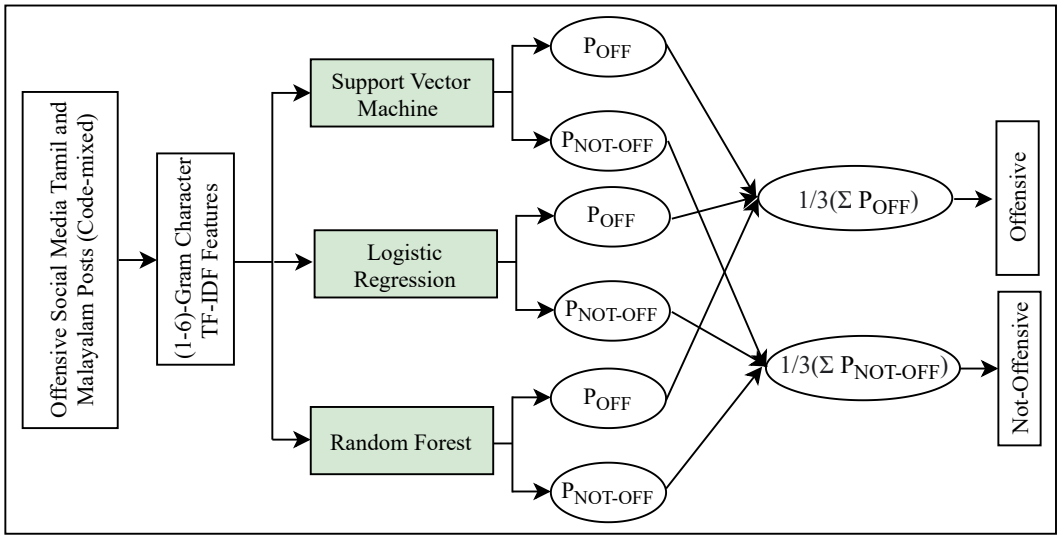

Figure 1: Ensemble-based model diagram for code-mixed social media posts

2.2. Ensemble-based model for Tamil script-mixed dataset

The systematic diagram for the proposed ensemble-based model for the identification of offensive Tamil script-mixed social media posts can be seen in Figure 2. Similar to the previous model (Figure 1), character n-gram TF-IDF features are input to AdaBoost classifier with three different validation splits. Three different random seeds 10, 20, and 42 are used to select the data samples into training and validation sets. The predicted probabilities of offensive and not-offensive classes from all the three AdaBoost model are then averaged to get the final classification probability. In this extensive experiment, it is observed that the first 50,000 one to six-gram character TF-IDF features performed best. The results of the proposed model are listed in section 3.

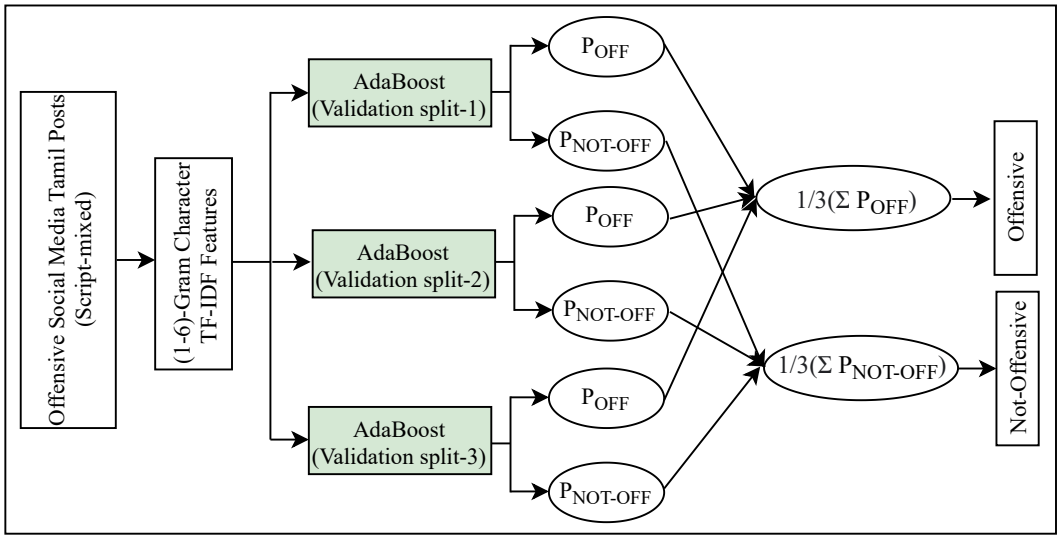

Figure 2: Ensemble-based model diagram for script-mixed Tamil social media posts

Table 2
Results of the proposed model for the identification of offensive posts in Tamil script-mixed, Tamil code-mixed, and Malayalam code-mixed tasks

Dataset	Class	Precision	Recall	F_1-score
Tamil script-mixed (Task-1)	Not-offensive	0.87	0.95	0.91
	Offensive	0.61	0.36	0.45
	Weighted Avg.	0.82	0.84	0.83
Tamil code-mixed (Task-2)	Not-offensive	0.73	0.73	0.73
	Offensive	0.58	0.58	0.58
	Weighted Avg.	0.67	0.67	0.67
Malayalam code-mixed (Task-2)	Not-offensive	0.85	0.78	0.82
	Offensive	0.61	0.72	0.66
	Weighted Avg.	0.78	0.76	0.77

3. Results

The performance of the proposed models are measured in terms of precision, recall, and F_1-score. Along with these, the confusion matrix and AUC-ROC curve are also plotted. The results for the Tamil script-mixed, Tamil code-mixed, and Malayalam code-mixed dataset is listed in Table 2. The proposed ensemble-based model has achieved a weighted precision of 0.82, weighted recall of 0.84, and weighted F_1-score of 0.83 for the Tamil script-mixed dataset. The confusion matrix and ROC curve for the Tamil script-mixed dataset are illustrated in Figures 3, and 4, respectively.

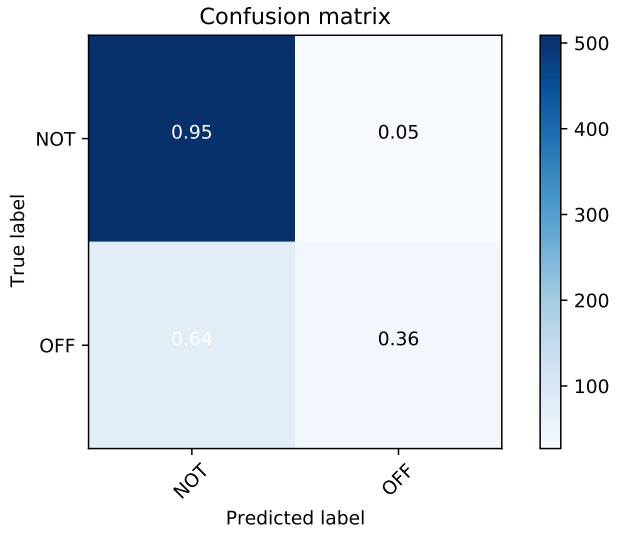

Figure 3: Confusion matrix for script-mixed Tamil social media posts

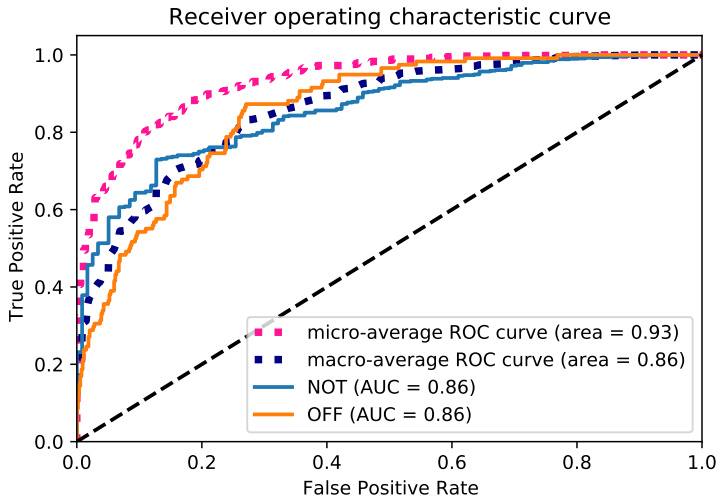

Figure 4: ROC curve for script-mixed Tamil social media posts

Similarly, the proposed ensemble-based model for Tamil code-mixed dataset has achieved weighted precision, reacll, and F_1-score of 0.67. Whereas, the proposed ensemble-based model has achieved weighted precision of 0.78, weighted recall of 0.76, and weighted F_1-score of 0.77. The confusion matrix and ROC curve for the Tamil code-mixed and Malayalam code-mixed datasets can be seen in Figures 5 and 6, 7 and 8, respectively.

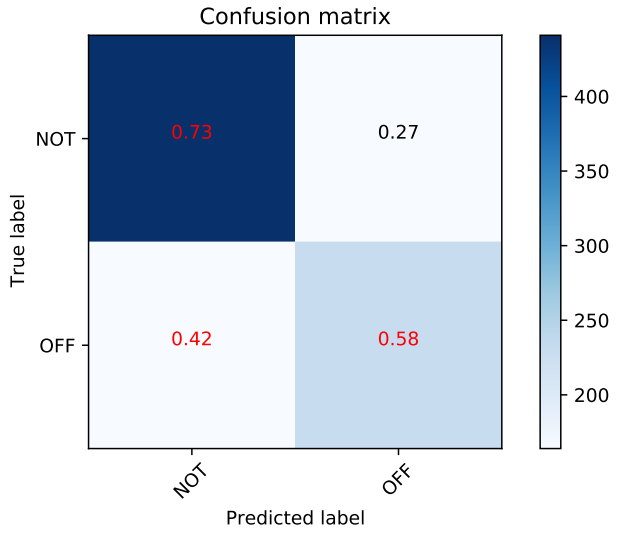

Figure 5: Confusion matrix for code-mixed Tamil social media posts

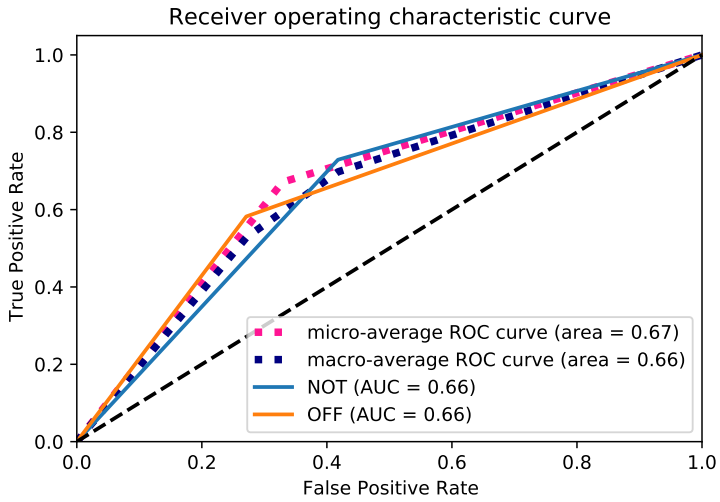

Figure 6: ROC curve for code-mixed Tamil social media posts

4. Conclusion

Hate and abusive language detection from code-mixed and script-mixed Dravidian social media postings are one of the most challenging tasks for natural language processing. Two different ensemble-based models have been developed, one for Tamil and Malayalam code-mixed and another one for Tamil script-mixed social media posts. The proposed model has achieved weighted F_1-scores of 0.83 for Tamil script-mixed, 0.67 for Tamil code-mixed, and 0.77 for Malayalam code-mixed social media posts. As the character-level features are giving promising

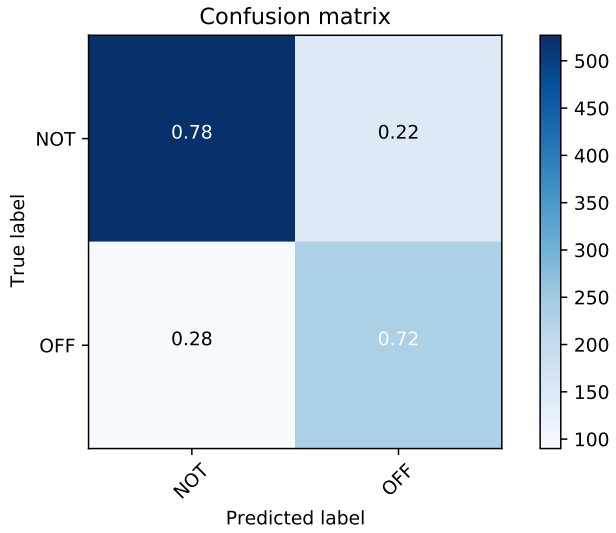

Figure 7: Confusion matrix for code-mixed Malayalam social media posts

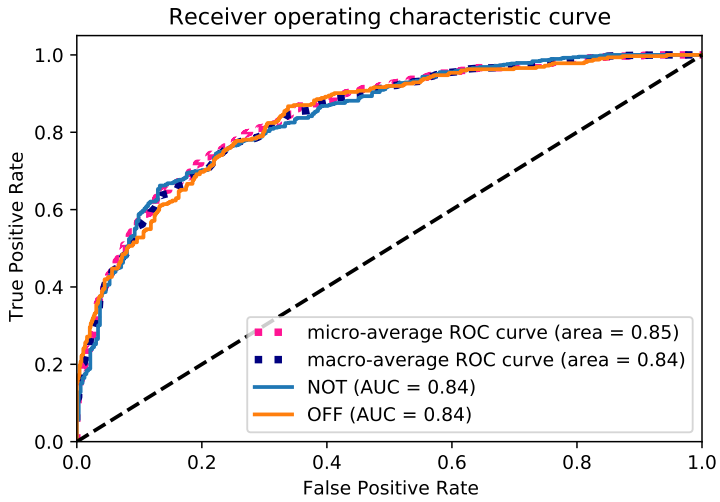

Figure 8: ROC curve for code-mixed Malayalam social media posts

results for code-mixed and script-mixed social media posts, it can be explored further for developing a robust system in the future.

References

[1] P. Kumar, Y. Dasari, S. Nath, A. Sinha, Controlling and mitigating targeted socio-economic attacks, in: Conference on e-Business, e-Services and e-Society, Springer, 2016, pp. 471–476.

[2] K. Gaurav, A. Sinha, J. P. Singh, P. Kumar, Facebook like: Past, present and future, in: Data Engineering and Intelligent Computing, Springer, 2018, pp. 617–625.
[3] A. Kumar, J. P. Singh, S. Saumya, A comparative analysis of machine learning techniques for disaster-related tweet classification, in: 2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129), IEEE, 2019, pp. 222–227.
[4] A. Kumar, N. C. Rathore, Relationship strength based access control in online social networks, in: Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2, Springer, 2016, pp. 197–206.
[5] S. Saumya, J. P. Singh, Detection of spam reviews: A sentiment analysis approach, Csi Transactions on ICT 6 (2018) 137–148.
[6] P. Kapil, A. Ekbal, D. Das, Investigating deep learning approaches for hate speech detection in social media, arXiv preprint arXiv:2005.14690 (2020).
[7] A. Kumar, S. Saumya, J. P. Singh, NITP-AI-NLP@ HASOC-FIRE2020: Fine tuned bert for the hate speech and offensive content identification from social media., in: FIRE (Working Notes), 2020, pp. 266–273.
[8] A. Kumar, S. Saumya, J. P. Singh, NITP-AI-NLP@ HASOC-Dravidian-CodeMix-FIRE2020: A machine learning approach to identify offensive languages from Dravidian code-mixed text., in: FIRE (Working Notes), 2020, pp. 384–390.
[9] A. K. Mishraa, S. Saumyab, A. Kumara, Iiit_dwd@ hasoc 2020: Identifying offensive content in indo-european languages (2020).
[10] S. Saumya, A. Kumar, J. P. Singh, Offensive language identification in Dravidian code mixed social media text, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 36–45.
[11] S. Agarwal, A. Sureka, Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website, arXiv preprint arXiv:1701.04931 (2017).
[12] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th International Conference on WWW Companion, 2017, pp. 759–760.
[13] K. Kumari, J. P. Singh, Ai_ml_nit patna at hasoc 2019: Deep learning approach for identification of abusive content, in: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019), 2019, pp. 328–335.
[14] B. R. Chakravarthi, P. K. Kumaresan, R. Sakuntharaj, A. K. Madasamy, S. Thavareesan, P. B, S. Chinnaudayar Navaneethakrishnan, J. P. McCrae, T. Mandl, Overview of the HASOC-DravidianCodeMix Shared Task on Offensive Language Detection in Tamil and Malayalam, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.