

Hate and Offensive content identification from Dravidian social media posts: A deep learning approach

Anu Priya¹, Abhinav Kumar²

¹Central University of Punjab, Bathinda, Punjab, India

²Department of Computer Science & Engineering, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, India

Abstract

Identifying hate and offensive content in social media posts is one of the most challenging tasks for Natural Language Processing. The usage of non-standard acronyms, misspellings, poor grammar, and multilingualism in social media posts makes detecting hate and offensive language much more difficult. This work proposes a deep neural network-based model for the identification of offensive social media posts from Tamil script-mixed, Tamil code-mixed, and Malayalam code-mixed messages. The combination of one to six-gram character-level Term-Frequency Inverse Document Frequency (TF-IDF) features with a four-layered deep neural network model performed better than the other combinations of character-level n-gram TF-IDF features. For Tamil script-mixed, Tamil code-mixed, and Malayalam code-mixed social media postings, the suggested model attained weighted $F1$ -scores of 0.84, 0.65, and 0.71, respectively. The code for the proposed models is available at <https://github.com/Abhinavkmr/Hate-Speech-Identification-Dravidian-Language.git>

Keywords

Hate speech, Social media, Deep learning, Offensive language

1. Introduction


Along with population growth during the previous ten years, the number of people using social networking has increased dramatically. People may express and share their opinions globally on social media sites like Twitter and Facebook, which has resulted in a flood of textual information on these platforms [1, 2]. These platforms were developed with the purpose of connecting people from all around the globe together. Currently, social platforms are used for a variety of objectives, including allowing governments to engage citizens, allowing consumers to make educated decisions, disaster management, and so on [3, 4]. Social media have a dark side as well [5, 6]. The widespread use of such platforms has led to the spread of abusive and controversial content, which has resulted in cyberstalking. The lack of social media monitoring norms contributes to the unhealthy use of these networks. The number of negative comments on social media grew rapidly over time. This has created the enormous interest of many experts who are trying to figure out how to filter out these hate and offensive social media contents

FIRE 2021: Forum for Information Retrieval Evaluation, December 13-17, 2021, India

✉ apriyarani3@gmail.com (A. Priya); abhinavanand05@gmail.com (A. Kumar)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

[7, 8, 9]. The identification of hate and offensive social media contents creates another level of complexity when the posted messages are in the form of code-mixed and script-mixed [10, 11].

In the non-native English-speaking country, a huge fraction of multi-lingual social media contents (mainly code-mixed and script-mixed) are posted by the users. Several works [12, 13, 14, 15] have been proposed for the detection of hate and offensive contents from English, Hindi, and German social media posts. Raj et al.[14] proposed Convolutional Neural Networks (CNN), Bi-directional Long Short-Term Memory (BiLSTM), and hybrid models (CNN+BiLSTM) model. Ray and Garain [15] uses TF-IDF, Word2Vec, and other textual features to train Random Forest and RoBERTa model whereas, Ou and Li [13] proposed XLM-RoBERTa and Ordered Neurons LSTM (ON-LSTM)-based model. Recently, a few works [16, 17, 18] have been reported by different researchers to identify hate and offensive language from Dravidian social media posts. Sai and Sharma [17] performed translation and transliteration of the posts and combined several transformer-based models to identify hate and offensive contents from Dravidian social media posts. Tula et al. [6] proposed an ensemble-based model by combining several popular BERT-based models. Kumar et al. [10] explored the usability of different deep learning models such as attention-based Long Short Term Memory (LSTM), Convolution Neural Network (CNN), and conventional machine learning models such as support vector machine, Logistic regression, Random forest, and Naive Bayes in the identification of hate contents. In their experiment, they found the use of character-level TF-IDF features with conventional machine learning models achieved state-of-the-art performance. Further, Saumya et al. [18] explore several popular models such as BERT, ULMFiT, hybrid deep learning models, and several conventional machine learning models. They also found that the use of character-level features with conventional machine learning models outperformed several complex deep learning models for the hate content identification from Dravidian social media posts.

In line with these studies, this work proposes a deep neural network-based model that uses character n-gram TF-IDF features to classify 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 [19]. 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 rest of the sections are organized as follows: section 2 discusses the proposed methodology in detail, section 3 list the finding of the proposed system. Finally, section 4 concludes the paper.

2. Methodology

The systematic diagram for the proposed dense neural network-based model can be seen in Figure 1. The proposed system is validated with the dataset provided in HASOC-Dravidian-CodeMix-FIRE2021 challenge [19]. The overall data statistic can be seen in Table 1.

The extensive experimentation was performed to find the best-suited features for the dense neural network. In the extensive experiments, we found the first 30,000 character one to six-gram TF-IDF features performed best for Tamil script-mixed and Tamil code-mixed dataset whereas the first 20,000 character one to six-gram TF-IDF features for Malayalam code-mixed

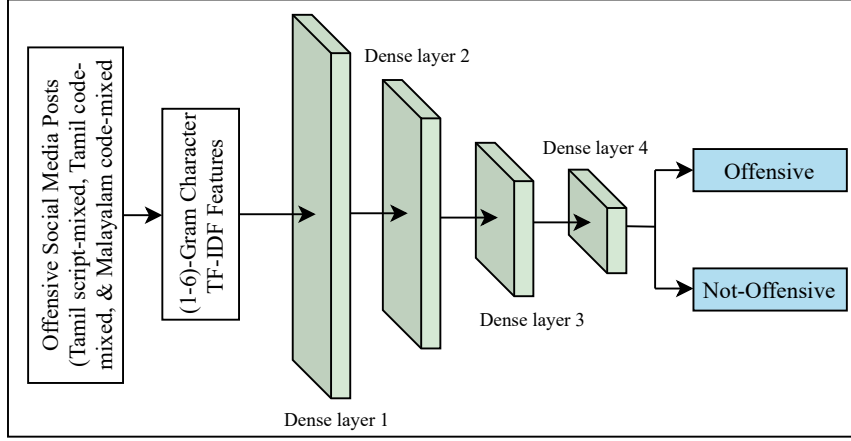


Figure 1: Proposed Dense Neural Network (CNN)-based model for the identification of offensive social media Dravidian posts

Table 1

Overall data statistic for Tamil script-mixed, Tamil code-mixed, and Malayalam code-mixed datasets

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

dataset performed best. The proposed system for Tamil script-mixed and Tamil code-mixed datasets have four dense layers containing 4,096, 512, 64, and 2-neurons, respectively. At the output layer, a softmax activation function is used to calculate the probabilities to decide the final class. Similarly, for the Malayalam code-mixed dataset, the proposed dense neural network has four layers containing 2,048, 512, 64, and 2-neurons, respectively. A softmax layer is then used to calculate the final class for a given social media post. As the deep learning models are very sensitive to the chosen hyper-parameters, we performed a sensitivity analysis of the model by varying learning rate, batch size, epochs, optimizer, and loss function. The best-suited hyper-parameters for the proposed system can be seen in Table 2.

3. Results

The performance of the proposed system is measured in terms of precision, recall, F_1 -score, AUC-ROC curve, and confusion matrix. Along with these metrics, weighted precision, weighted

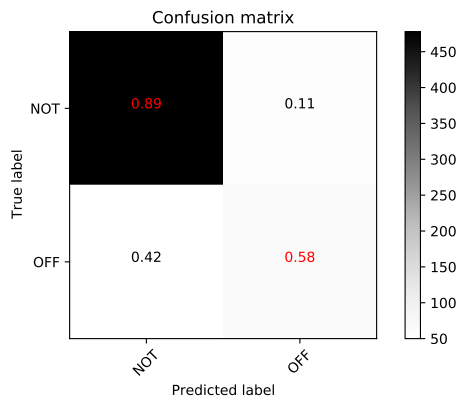


Figure 2: Confusion matrix (Tamil script-mixed)

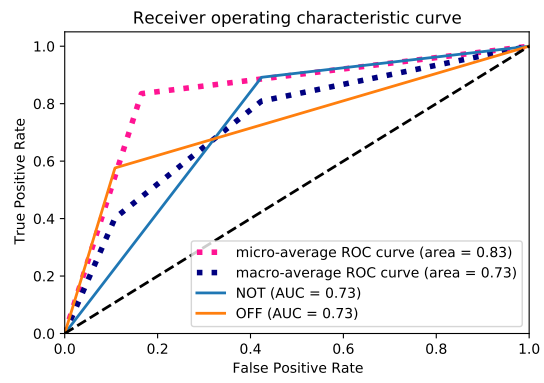


Figure 3: ROC curve (Tamil script-mixed)

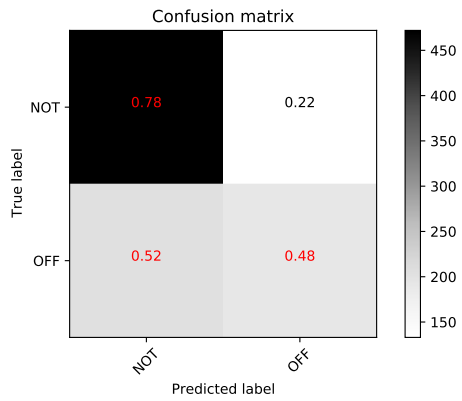


Figure 4: Confusion matrix (Tamil code-mixed)

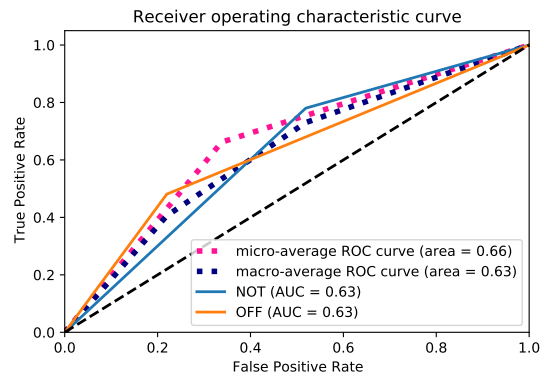


Figure 5: ROC curve (Tamil code-mixed)

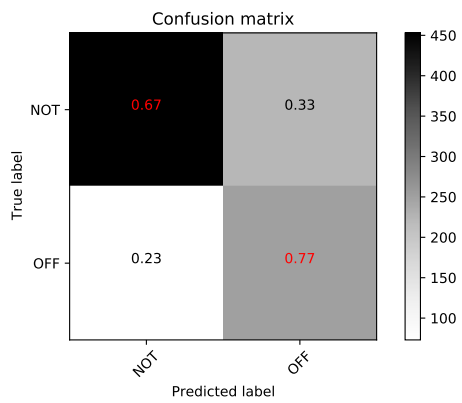


Figure 6: Confusion matrix (Malayalam code-mixed)

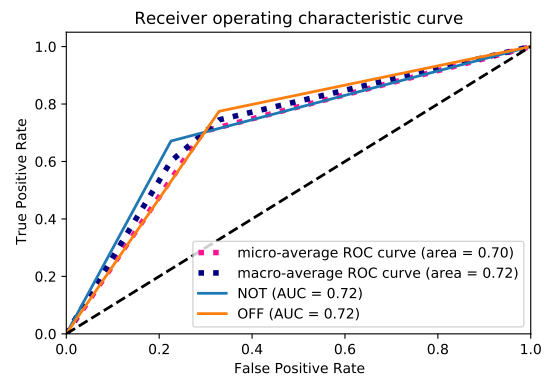


Figure 7: ROC curve (Malayalam code-mixed)

Table 2

Best suited hyper-parameters for the proposed dense neural network-based model

Hyper-parameters	Tamil script-mixed	Tamil code-mixed	Malayalam code-mixed
Dense layers	4	4	4
Number of neurons at each layer	4,096, 512, 64, 2	4,096, 512, 64, 2	2,048, 512, 64, 2
Dropout	0.2	0.2	0.2
Activation function	ReLU, Softmax	ReLU, Softmax	ReLU, Softmax
Optimizer	Adam	Adam	Adam
Loss	Binary cross-entropy	Binary cross-entropy	Binary cross-entropy
Learning rate	0.001	0.001	0.001
Batch size	20	20	20
Epochs	50	50	50

Table 3

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	0.91	0.89	0.90
	Offensive	0.54	0.58	0.56
	Weighted Avg.	0.84	0.83	0.84
Tamil code-mixed (Task-2)	Not-offensive	0.70	0.78	0.74
	Offensive	0.59	0.48	0.53
	Weighted Avg.	0.65	0.66	0.65
Malayalam code-mixed (Task-2)	Not-offensive	0.86	0.67	0.75
	Offensive	0.53	0.77	0.63
	Weighted Avg.	0.75	0.70	0.71

recall, and weighted F_1 -score are also calculated. The performance of the proposed model for Tamil script-mixed, Tamil code-mixed, and Malayalam code-mixed datasets are listed in Table 3.

For Tamil script-mixed dataset, the proposed system achieved a weighted precision, recall, and F_1 -score of 0.84, 0.83, and 0.84, respectively. The confusion matrix and ROC curve for the Tamil script dataset can be seen in Figures 2 and 3, respectively. For Tamil code-mixed dataset, the proposed system achieved a weighted precision, recall, and F_1 -score of 0.65, 0.66, and 0.65, respectively. The confusion matrix and ROC curve for the Tamil code-mixed dataset can be seen in Figures 4 and 5, respectively. For Malayalam code-mixed dataset, the proposed dense neural network-based model achieved a weighted precision, recall, and F_1 -score of 0.75, 0.70, and 0.71, respectively. The confusion matrix and ROC curve for the Tamil code-mixed dataset can be seen in Figures 6 and 7, respectively.

4. Conclusion

The identification of hate and offensive content from script-mixed and code-mixed social media posts is one of the hot research topics in recent times. This work proposes a dense neural

network-based model that uses character-level TF-IDF features to identify offensive messages from Tamil script-mixed, Tamil code-mixed, and Malayalam code-mixed datasets. The proposed model achieved a weighted F_1 -score of 0.84, 0.65, and 0.71 for Tamil script-mixed, Tamil code-mixed, and Malayalam code-mixed social media posts, respectively. As the character-level features for Dravidian social media posts are giving the promising performance, therefore the character-level features can be explored further in the future. An ensemble-based model can also be made in feature to achieve better performance.

References

- [1] J. P. Singh, Y. K. Dwivedi, N. P. Rana, A. Kumar, K. K. Kapoor, Event classification and location prediction from tweets during disasters, *Annals of Operations Research* 283 (2019) 737–757.
- [2] A. Kumar, J. P. Singh, Location reference identification from tweets during emergencies: A deep learning approach, *International journal of disaster risk reduction* 33 (2019) 365–375.
- [3] A. Kumar, J. P. Singh, Y. K. Dwivedi, N. P. Rana, A deep multi-modal neural network for informative twitter content classification during emergencies, *Annals of Operations Research* (2020) 1–32.
- [4] A. Kumar, J. P. Singh, Disaster severity prediction from twitter images, in: *Intelligence Enabled Research*, Springer, 2021, pp. 65–73.
- [5] K. Sreelakshmi, B. Premjith, K. Soman, Detection of hate speech text in hindi-english code-mixed data, *Procedia Computer Science* 171 (2020) 737–744.
- [6] D. Tula, P. Potluri, S. Ms, S. Doddapaneni, P. Sahu, R. Sukumaran, P. Patwa, Bitions@ DravidianLangTech-EACL2021: Ensemble of multilingual language models with pseudo labeling for offence detection in Dravidian languages, in: *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, 2021, pp. 291–299.
- [7] A. K. Mishraa, S. Saumyab, A. Kumara, IIIT_DWD@ HASOC 2020: Identifying offensive content in indo-european languages (2020).
- [8] B. R. Chakravarthi, R. Priyadharshini, N. Jose, A. Kumar M, T. Mandl, P. K. Kumaresan, R. Ponnusamy, H. R L, J. P. McCrae, E. Sherly, 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, Kyiv, 2021, pp. 133–145. URL: <https://aclanthology.org/2021.dravidianlangtech-1.17>.
- [9] S. Banerjee, B. Raja Chakravarthi, J. P. McCrae, Comparison of pretrained embeddings to identify hate speech in indian code-mixed text, in: *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2020, pp. 21–25. doi:10.1109/ICACCCN51052.2020.9362731.
- [10] 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.
- [11] 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.
- [12] T. Mandl, S. Modha, A. Kumar M, B. R. Chakravarthi, Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: FIRE, 2020, pp. 29–32.
 - [13] X. Ou, H. Li, Ynu_oxz at hasoc 2020: Multilingual hate speech and offensive content identification based on xlm-roberta., in: FIRE (Working Notes), 2020, pp. 121–127.
 - [14] R. Raj, S. Srivastava, S. Saumya, NSIT & IIITDWD@ HASOC 2020: Deep learning model for hate-speech identification in indo-european languages., in: FIRE (Working Notes), 2020, pp. 161–167.
 - [15] B. Ray, A. Garain, JU at HASOC 2020: Deep learning with RoBERTa and random forest for hate speech and offensive content identification in Indo-European languages., in: FIRE (Working Notes), 2020, pp. 168–174.
 - [16] B. R. Chakravarthi, A. K. M, J. P. McCrae, B. Premjith, K. Soman, T. Mandl, Overview of the track on HASOC-offensive language identification-DravidianCodeMix., in: FIRE (Working Notes), 2020, pp. 112–120.
 - [17] S. Sai, Y. Sharma, Siva@ HASOC-Dravidian-CodeMix-FIRE-2020: Multilingual offensive speech detection in code-mixed and romanized text., in: FIRE (Working Notes), 2020, pp. 336–343.
 - [18] 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.
 - [19] 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.