IRNLP_DAIICT@DravidianLangTech-EACL2021: Offensive Language identification in Dravidian Languages using TF-IDF Char N-grams and MuRIL

Bhargav Dave
DA-IICT / Gandhinagar, India
bhargavdave1@gmail.com

Shripad Bhat
DA-IICT / Gandhinagar, India
shripadbhat30@gmail.com

Prasenjit Majumder
DA-IICT / Gandhinagar, India
prasenjit.majumder@gmail.com

Abstract

This paper presents the participation of the IRNLP_DAIICT team from Information Retrieval and Natural Language Processing lab at DA-IICT, India in DravidianLangTech-EACL2021 Offensive Language identification in Dravidian languages. The aim of this shared task is to identify Offensive Language from a code-mixed data-set of YouTube comments. The task is to classify comments into Not Offensive (NO), Offensive Untargetede (OU), Offensive Targeted Individual (OTI), Offensive Targeted Group (OTG), Offensive Targeted Others (OTO), Other Language (OL) for three Dravidian languages: Kannada, Malayalam and Tamil. We use TF-IDF character n-grams and pretrained MuRIL embeddings for text representation and Logistic Regression and Linear SVM for classification. Our best approach achieved Ninth, Third and Eighth with weighted F1 score of 0.64, 0.95 and 0.71 in Kannada, Malayalam and Tamil on test dataset respectively. Our code is publicly available here[1].

1 Introduction

The emergence of smartphones and widespread access to internet, social media platforms like Facebook, Twitter, YouTube, etc have become increasingly popular. People express their opinions on such platforms on various issues. Due to lack of moderation, a significant amount of offensive content is often posted on these platforms. Since social media is become a indispensable part of our life, the content posted on the social media platforms has great impact on the society. It has been seen recently that offensive and provoking content on these platforms can lead to riots. Hence the moderation of content in these platforms is a very important task. Since the amount of content generated is very high, there is a need to perform automated moderation of the content.

Natural language processing can used to perform automated analysis of text. By identifying and classifying the text content posted, offensive content can be removed and users could be warned. This can help in curbing offensive content online.

The goal of this shared task is to classify offensive language from a code-mixed data-set of comments of Dravidian languages collected from YouTube. Shared task was introduced as six class classification of the YouTube for three Dravidian languages: Kannada, Malayalam and Tamil (Chakravarthi et al., 2021)

Our approaches use TF-IDF character n-grams and MuRIL embeddings for the text representation and Logistic Regression, Linear SVM, Randam Forest for classification.

The rest of this paper is organized as follows. In section we discuss the related work followed by Section 3 which describes the shared task dataset and Methods are presented in Section 4. Results and Analysis is given in final Section 5 and Section 6 present Conclusion.

2 Related Work

The identification of hate and offensive Speech in social media is of great importance and receives much attention in the text classification community. Due to the lack of resources and morphological complexity there is a huge demand for research in code-mixed Dravidian languages.

Some of the shared tasks in the recent past are OffensEval (Zampieri et al., 2019, 2020), HateEval(Basile et al., 2019) and HASOC (Mandl et al., 2019, 2020). Among these shared tasks, HASOC 2020 shared task is based on Dravidian languages. In OffensEval (Zampieri et al., 2020), BERT (Devlin et al., 2019), ROBERTa (Liu et al., 2019) and

[1] https://github.com/bhargav25dave1996/IRNLP_DAIICT_DravidianLangTech-EACL2021

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 266–269

Label	Kannada			Malayalam			Tamil		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
NO	3544	426	427	14153	1779	1765	25425	3193	3190
OU	212	33	33	191	20	29	2906	356	368
OTI	487	66	75	239	24	27	2557	307	315
OTG	329	45	44	140	13	23	2343	295	288
OTO	123	16	14	-	-	-	2343	172	160
OL	1522	191	185	1287	163	157	1454	65	71
Total	6217	777	778	16010	1999	2001	35139	4388	4392

Table 1: Offensive Language Identification Detection shared task Dataset Statistics

ELMo (Peters et al., 2018) were used for offensive language identification. In HASOC (Mandl et al., 2020) Track multilingual transformer based methods like XLM-ROBERTa, mBERT, etc were employed and fine tuned for the task. Other popular techniques were TF-IDF along with Character n-grams combined with machine learning classifiers like Logistic Regression, SVM and XGboost.

3 Dataset

Offensive Language Identification shared task organizers provide datasets in three languages Kannada (Hande et al., 2020), Malayalam (Chakravarthi et al., 2020a) and Tamil (Chakravarthi et al., 2020b). Dataset has been curated from Youtube comments using the YouTube Comment Scraper [2]. The number of comments is 28451 in the Kannada dataset, 20198 in the Tamil dataset and 10705 in the Malayalam dataset. Full statistic of dataset given in Table 1

4 Methods

For all the YouTube comments we first preprocess the text and then create a text representation and finally classify the text using the machine learning classifiers. Figure 1 illustrates the set of steps used to classify the YouTube comments.

4.1 Pre-Processing

Since there is a lot of noise in the social media text we perform the following preprocessing operations. URL's, user mentions of the form @user, emojis, digits, punctuations are removed and the text is lowercased.

[2] https://github.com/philbot9/youtube-comment-scraper

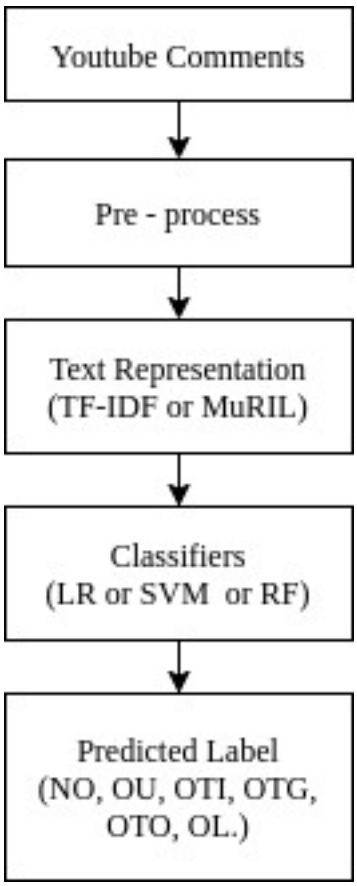

Figure 1: Steps involved in Offensive language identification

4.2 Text representation and Classifiers

Representation of the text is one of the fundamental tasks in Natural language processing. We explore two representation techniques: TF-IDF and MuRIL. TF-IDF is a very popular text representation technique which takes into account the frequency of the word in a given document and the number of documents in which a word is present. We employ Scikit-learn TF-IDF vectorizer API for obtaining the text representation. Instead of word based TF-IDF representation we make use of character

n-grams based TF-IDF representation so as to effectively capture morphological variations of the words.

MuRIL[3] (Multilingual Representations for Indian Languages) is a transformer based language model trained on 17 Indian languages on self-supervised masked language modeling task. MuRIL training consists of translation and transliteration segment pairs in addition to the standard training used in Multilingual BERT. Pretrained MuRIL model is used to obtain the text representation in the form vectors of 768 dimension.

Logistic Regression (LR), Random Forest (RF), Linear SVM classifiers have been used to perform the classification of the text. Scikit-learn API has been used to implement the classification task. So the six approaches we have used are TF-IDF (Char) + LR, TF-IDF(Char) + SVM, TF-IDF(Char) + RF, MuRIL + LR, MuRIL + SVM, MuRIL + RF for each language.

5 Results and Analysis

The submission evaluation on the test data of all the three languages is shown in table tables 2, 3, 4. Our methods achieve Ninth, Third and Eighth rank in Kannada, Malayalam and Tamil respectively. For the Kannada language task, TF-IDF (Char) + LR and TF-IDF (Char) + SVM achieves the best results with weighted F1 score of 0.65 but for ranking task organize consider TF-IDF(char) + RF with weighted F1 score of 0.64. For the Malayalam language task, MuRIL representation of the text with Random Forest classifier (MuRIL + RF) achieves the best weighted F1 score of 0.95. For Tamil language TF-IDF character n-grams representation of the text with Logistic regression classifier (TF-IDF (Char) + LR) achieves the best results with weighted F1 score of 0.71.

Model	**W-Avg F1-score**
TF-IDF (Char) + LR	**0.65**
TF-IDF(Char) + SVM	**0.65**
TF-IDF(Char) + RF	0.64
MuRIL + LR	0.39
MuRIL + SVM	0.39
MuRIL + RF	0.60

Table 2: Results for the Kannada on test dataset.

It can be observed that MuRIL is not performing well particularly in case of Kannada and Tamil language but it has help us achieve rank 3 with weighted F1 score of 0.95 in Malayalam. TF-IDF character n-grams representation performs better than MuRIL in Kannada and Tamil. So there is no clear winner in terms of text representation or the classifier across all the three languages. Also, data imbalance problem is prevalent in all three languages making the task more difficult.

[3] https://tfhub.dev/google/MuRIL/1

Model	**W-Avg F1-score**
TF-IDF (Char) + LR	0.89
TF-IDF(Char) + SVM	0.90
TF-IDF(Char) + RF	0.94
MuRIL + LR	0.83
MuRIL + SVM	0.83
MuRIL + RF	**0.95**

Table 3: Results for the Malayalam on test dataset.

Model	W-Avg F1-score
TF-IDF (Char) + LR	**0.71**
TF-IDF(Char) + SVM	0.70
TF-IDF(Char) + RF	0.66
MuRIL + LR	0.61
MuRIL + SVM	0.61
MuRIL + RF	0.63

Table 4: Results for the Tamil on test dataset.

6 Conclusion

In this paper the details regarding our submission in Offensive Language Identification shared task have been presented. We explored two text representation techniques: TF-IDF character n-grams and MuRIL and conclude that MuRIL works better for Malayalam while TF-IDF works better for other two languages. In future to mitigate the data imbalance problem text augmentation techniques like back translation could be applied to oversample the minority class. Also, other methods like CNN, RNN and fine tuning of Language models could be explored.

Acknowledgments

We would like to thank the Offensive language identification task organizers for giving us a opportunity to work on a new task. We also like to acknowledge the IRNLP lab at DAIICT, Gandhinagar for their contribution.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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, 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.