SSNCSE_NLP@DravidianLangTech-EACL2021: Offensive Language Identification on Multilingual Code Mixing Text

B. Bharathi & Agnusimmaculate Silvia A
Department of CSE
Sri Siva subramaniya Nadar College of Engineering
Kalavakkam - 603110
bharathib@ssn.edu.in
agnusimmaculate18011@cse.ssn.edu.in

Abstract

Social networks made a huge impact in almost all fields in recent years. Text messaging through the Internet or cellular phones has become a major medium of personal and commercial communication. Everyday we have to deal with texts, emails or different types of messages in which there are a variety of attacks and abusive phrases. It is the moderator's decision which comments to remove from the platform because of violations and which ones to keep but an automatic software for detecting abusive languages would be useful in recent days. In this paper we describe an automatic offensive language identification from Dravidian languages with various machine learning algorithms. This is work is shared task in DravidanLangTech-EACL2021. The goal of this task is to identify offensive language content of the code-mixed dataset of comments/posts in Dravidian Languages ((Tamil-English, Malayalam-English, and Kannada-English)) collected from social media. This work explains the submissions made by SSNCSE_NLP in DravidanLangTech-EACL2021 Code-mix tasks for Offensive language detection. We achieve F1 scores of 0.95 for Malayalam, 0.7 for Kannada and 0.73 for task2-Tamil on the test-set.

1 Introduction

Social media has become an important tool to connect the people all over the world. This is because it allows its users to share the content they want quickly, efficiently and in real-time. However, user-created content shared on social media is not always organized by the rules. In fact, nowadays, content written in an offensive language has become widespread on social media (Chakravarthi and Muralidaran, 2021; Chakravarthi et al., 2021; Suryawanshi and Chakravarthi, 2021). Offensive language is defined as message which contains insulting or threatening expressions written by one person to another (Chakravarthi et al., 2020d; Mandl et al., 2020). The severity of this problem is increasing each day; consequently, it is very important to deal with this problem in terms of government policy, social media terms and policies and online community plans (Chakravarthi et al., 2020b). At this stage, there is a need for effective methods. Social media generates a large amount of data daily as mentioned above. Therefore, it is very difficult to manually determine offensive language on the social media even by an expert. Some words that have multiple meanings that could be offensive to some people from some places. There is an increasing demand for offensive language identification on social media texts which are largely code-mixed (Chakravarthi, 2020).

Code-mixing is a prevalent phenomenon in a multilingual community and the code-mixed texts are sometimes written in non-native scripts (Jose et al., 2020; Priyadharshini et al., 2020). Systems trained on monolingual data fail on code-mixed data due to the complexity of code-switching at different linguistic levels in the text (Chakravarthi et al., 2018, 2019). This shared task on Offensive Language Identification in Dravidian Languages-EACL 2021 presents a new gold standard corpus for offensive language identification of code-mixed text in Dravidian languages (Tamil-English, Malayalam-English, and Kannada-English). A recorded Tamil writing has been archived for more than 2600 years (Thavareesan and Mahesan, 2019, 2020a,b). The oldest time of Tamil writing, Sangam writing, is dated from ca. 600 BC – AD 300. It has the most established surviving writing among Dravidian languages. Over 55% of the epigraphical engravings (around 55,000) found by the Archeological Survey of India are in the Tamil language.

The remainder of the paper is organized as fol-

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 313–318
April 20, 2021

Task description	Class label	Train-set	Dev-set	Test-set
Tamil-code mixed	Not_offensive	25425	3193	3190
	Offensive_Untargeted	2906	356	368
	Offensive_Targeted_Insult_Group	2557	309	315
	Offensive_Targeted_Insult_Individual	2343	295	288
	not-Tamil	1454	172	160
	Offensive_Targeted_Insult_Other	454	65	71
Malayalam-code mixed	Not_offensive	14153	1779	1765
	not-malayalam	1287	163	157
	Offensive_Targeted_Insult_Individual	239	24	29
	Offensive_Untargeted	191	20	27
	Offensive_Targeted_Insult_Group	140	13	23
Kananda-code mixed	Not_offensive	3544	426	427
	not-Kannada	1522	191	185
	Offensive_Targeted_Insult_Individual	487	66	75
	Offensive_Targeted_Insult_Group	329	45	44
	Offensive_Untargetede	212	33	33
	Offensive_Targeted_Insult_Other	123	16	14

Table 1: Data Distribution

lows. Section 2 discusses the related work on offensive language identification task. The dataset about the shared task is described in Section . Section 4 outlines the features and machine learning algorithms used for this task. Results are discussed in Section 5. Section 6 concludes the paper.

2 Related work

Offensive language identification for Greek language is described in (Pitenis et al., 2020). This paper uses different machine learning and deep learning models to for offensive language identification task with Offensive Greek Tweet Dataset. Multilingual offensive language identification using cross lingual embeddings with transfer learning is carried out in (Puranik et al., 2021; Hegde et al., 2021; Yasaswini et al., 2021; Ghanghor et al., 2021b,a). In (Razavi et al., 2010), flame detection approach which extracts features at different conceptual levels and applies multilevel classification for flame detection was described. Arabic offensive language identification task was described in (Alakrot et al., 2018). The survey of offensive language identification task is explained in (Pradhan et al., 2020). Machine learning approach for detecting offensive language identification on Twitter data is carried out in (Gaydhani et al., 2018).

3 Data-set Analysis and Preprocessing

The goal of this task is to identify offensive language content of the code-mixed dataset of comments/posts in Dravidian Languages ((Tamil-English, Malayalam-English, and Kannada-English)) collected from social media(Chakravarthi et al., 2021, 2020a,c; Hande et al., 2020). The comment/post may contain more than one sentence but the average sentence length of the corpora is 1. Each comment/post is annotated at the comment/post level. This dataset also has class imbalance problems depicting real-world scenarios. The dataset containing YouTube comments with class labels such as Not-offensive, offensive-untargeted, offensive-targeted-individual, offensive-targeted-group, offensive-targeted-other, or Not-in-indented-language. The train-set, dev-set and test-set distribution with class-wise distribution is shown in Table 1. There is a clear imbalance in the data-set distribution. This could cause a bias towards a particular class and the model trained on this data-set would be more inclined towards the dominant class.

4 Experimental setup and features

For feature extraction, the n-gram model and BERT embedding model are experimented upon. As the content of the comments is a mix of Dravidian language grammar in Roman lexicons along with English grammar, it becomes challenging to find

Features	Classifier	Precision	Recall	F1-score
Tfidf	Random forest	0.77	0.75	0.66
Tfidf	K-nearest	0.69	0.75	0.70
Tfidf	Adaboost	0.62	0.74	0.65
Tfidf	Decision tree classifier	0.66	0.66	0.67
Count vec	Random forest	0.74	0.75	0.66
Count vec	MLP classifier	0.74	0.76	0.75
BERT	MLP classifier	0.71	0.72	0.69

Table 2: Performance of the proposed approach of Tamil-English code mixed text using dev data

Features	Classifier	Precision	Recall	F1-score
Tfidf	k-nearest	0.91	0.91	0.91
Tfidf	MLP	0.97	0.97	0.97
Tfidf	SVM	0.87	0.75	0.81
Countvec	k-nearest	0.90	0.91	0.91
Countvec	MLP	0.97	0.97	0.97
Countvec	SVM	0.96	0.96	0.96
BERT	MLP classifier	0.95	0.87	0.85

Table 3: Performance of the proposed approach of Malayalam-English code mixed text using dev data

pre-trained models for this context. So a simple n-gram approach is considered. Also, the advancements done by the transformer model for pre-training and the availability of multilingual trained models encourage to experiment with BERT pre-trained embeddings. In the proposed approach, TFIDF, Count vectorizer and BERT embeddings were extracted from the input text. Then the extracted features were trained with different machine learning models such as K-nearest neighbour, MLP classifier, random forest classifier, Ada boost classifier, decision tree classifier and voting classifier etc. The experiments were conducted for Tamil-English, Malayalam-English, and Kannada-English data sets and best models obtained for these tasks were used to generate the scores for the test-set.

5 Observations

5.1 Tamil-English dataset

The features such as TFIDF, Count vectorizer and BERT embeddings were extracted from the youtube comments specified in Tamil-English code mixed text. The extracted features were trained with different machine learning models and the models are evaluated using the development data and results are tabulated in Table 2.

From the Table 2, count vectorizer feature with MLP classifier is giving the F1-score of 0.75.

5.2 Malayalam-English dataset

The features such as TFIDF, Count vectorizer and BERT embeddings were extracted from the youtube comments specified in Malayalam-English code mixed text. The extracted features were trained with different machine learning models and the models are evaluated using the development data and results are tabulated in Table 3.

From the Table 3, count vectorizer feature with MLP classifier is giving the F1-score of 0.97.

5.3 Kannada-English dataset

The features such as TFIDF, Count vectorizer and BERT embeddings were extracted from the youtube comments specified in Kannada-English code mixed text. The extracted features were trained with different machine learning models and the models are evaluated using the development data and results are tabulated in Table 4.

From the Table 3, count vectorizer feature with MLP classifier is giving the F1-score of 0.69.

Three runs were submitted using the Tamil-English, Malayalam-English, and Kannada-English code mixed text. The results of the runs were tabulated in Table in 5.

From Table 5, it has been noted that the performance of the proposed system is better for

Features	Classifier	Precision	Recall	F1-score
Tfidf	k-nearest	0.63	0.65	0.63
Tfidf	MLP	0.68	0.70	0.68
Tfidf	SVM	0.68	0.73	0.69
Countvectorizer	k-nearest	0.62	0.65	0.62
Countvectorizer	MLP	0.69	0.71	0.69
Countvectorizer	SVM	0.67	0.69	0.68
SentenceTransformer	MLP classifier	0.67	0.65	0.64

Table 4: Performance of the proposed approach of Kannada-English code mixed text using dev data

Dataset	Precision	Recall	F1-score	Rank
Tam-Eng	0.74	0.73	0.73	6
Mal-Eng	0.95	0.96	0.95	3
Kan-Eng	0.71	0.74	0.70	5

Table 5: Performance of the proposed approach using test data

Malayalam-English code mixed text.

6 Conclusion

There is an increasing demand for offensive language identification on social media texts which are largely code-mixed. The goal of this task is to identify offensive language content of the code-mixed dataset of comments/posts in Dravidian Languages ((Tamil-English, Malayalam-English, and Kannada-English)) collected from social media. In the proposed work, the basic TFIDF and count vectorizer features are giving better performance when compared to sentence embeddings. The examples are not sufficient to train the deep learning models. The machine learning models are giving better performance than the deep learning models.

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. Arabic Computational Linguistics.

Bharathi Raja Chakravarthi. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2018. Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global Wordnet Conference*, pages 77–86, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019. WordNet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7, Dublin, Ireland. European Association for Machine Translation.

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, M Anand Kumar, John Philip McCrae, Premjith B, Soman KP, and Thomas Mandl. 2020b. Overview of the track on HASOC-Offensive Language Identification-DravidianCodeMix. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020). CEUR Workshop Proceedings. In: CEUR-WS. org, Hyderabad, India*.

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. 2020c. 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. 2020d. 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.

Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. 2018. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach.

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.

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.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Rahul Pradhan, Ankur Chaturvedi, Aprna Tripathi, and Dilip Sharma. 2020. *A Review on Offensive Language Detection*, pages 433–439.

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.

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Advances in Artificial Intelligence*, pages 16–27.

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.

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.