DLRG@DravidianLangTech-EACL2021: Transformer based approach for Offensive Language Identification on Code-Mixed Tamil

Ratnavel Rajalakshmi, B. Yashwant Reddy, Lokesh Kumar
School of Computer Science and Engineering,
Vellore Institute of Technology, Chennai
rajalakshmi.r@vit.ac.in

Abstract

Internet advancements have made a huge impact on the communication pattern of people and their life style. People express their opinion on products, politics, movies etc. in social media. Even though, English is predominantly used, nowadays many people prefer to tweet in their native language and sometimes by combining it with English. Sentiment analysis on such code-mixed tweets is challenging, due to large vocabulary, grammar and colloquial usage of many words. In this paper, the transformer based language model is applied to analyse the sentiment on Tanglish tweets, which is a combination of Tamil and English. This work has been submitted to the the shared task on DravidianLangTech-EACL2021. From the experimental results, it is shown that an F_1 score of 64% was achieved in detecting the hate speech in code-mixed Tamil-English tweets using bidirectional transformer model.

1 Introduction

Recent advancements in the Internet technologies and the usage of smart phones changed the behaviour of communication pattern among the people of all walks of life. Many people use the social media to post their opinion about various domains, including politics, movies, sports etc. This has led to a sharp increase in offensive posts and hate speech targeted towards an individual or a group or women. Recently, the trend of using Code-Mixed language for tweets has increased, in which the words/phrases from more than one Language is used to express their feelings (Chakravarthi et al., 2020a; Chakravarthi, 2020a; Hande et al., 2020). Many research works have been reported for identifying the Hate Speech in English / German tweets using mono-lingual models (Rajalakshmi and Reddy, 2019), but limited works are performed for under-resourced languages like Tamil (Thavareesan and Mahesan, 2019, 2020a,b). The Third Tamil Sangam Period is the period of history from c. 6th century BCE[1] to c. 3rd century CE of ancient Tamil Nadu, Kerala and parts of Sri Lanka (then known as Tamilakam) (Sivanantham and Seran, 2019). The Tamilakam (Puranuru 168. 18) referred to the whole ancient Tamil-speaking territory in the Old Tamil language, approximately equivalent to the area known today as southern India, consisting of the regions of the present-day Indian states of Tamil Nadu, Kerala, parts of Andhra Pradesh, parts of Karnataka, and also known as Eelam in northern Sri Lanka. Tamil language is the oldest language in India, all the Dravidian languages evolved from Tamil language [2] (Chakravarthi and Muralidaran, 2021; Suryawanshi and Chakravarthi, 2021; Chakravarthi et al., 2021a,b).

The machine learning algorithms with handcrafted features are not sufficient to capture the semantic information from these multi-lingual tweets. This has made the researchers to move towards the Sequence Models which have proven capability to capture the semantic Information from the data using the Transformer Architecture. Among the Sequence Models, Bi-Directional Encoder Representation from Transformer(BERT) has gained more popularity because of its ability to capture the semantic relationship from both the directions of text and suitable method for understanding context-heavy sentences (Ghanghor et al., 2021b,a; Puranik et al., 2021; Hegde et al., 2021; Yasaswini et al., 2021). In this research work, we have studied the effectiveness of BERT to identify hate speech in the code-mixed tweets. This work is submitted for the shared task on DravidianLangTech, the first workshop on Speech and Language Technologies for

[1] Iron Age - Early Historic Transition in South Indian Appraisal

[2] Tamil-language

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 357–362

Dravidian Languages at EACL 2021. This paper is organized as follows: Related works are presented in Section 2 followed by the proposed methodology in Section 3. Experimental details are discussed in Section 4 followed by the Conclusion in Section 5.

2 Related Works

Sentiment analysis on social media tweets is an important problem being studied for various reasons like identifying the opinion of people about a product, movie or sports etc. Extracting the sentiments from the tweets is a challenging task, as the users give their comments explicitly or implicitly. (Soubraylu and Rajalakshmi, 2020) performed sentiment analysis on movie reviews and proposed a method to determine the explicit opinions by combining the advantages of Convolution Neural Networks with the Bidirectional Long Short Memory. The task of identifying the implicit opinions from the tweets is more challenging and many works are reported to address the same. A detailed survey is presented in (Ganganwar and Rajalakshmi, 2019).

Many research works are reported in the literature for hate speech detection in social media tweets. (Corazza et al., 2020) studied this problem in multi-lingual context and evaluated their deep learning and machine leaning approaches on three different languages viz., English, German and Italian. In a study by (Rani et al., 2020) on code-mixed Hindi tweets, the performance of CNN is found to be better than the linear classifiers such as SVM. (Rajalakshmi and Reddy, 2019) proposed an ensemble based approach for detecting hate speech in Hindi and German languages. To process the multi-lingual queries quickly, differentiating Code-Mixing and Code borrowing is important (Chakravarthi et al., 2018, 2019b,a, 2020c; Chakravarthi, 2020b). For this, a relevance based metric is proposed by (Rajalakshmi and Agrawal, 2017) to rank the borrowing likeliness of the words in Hindi-English tweets. Code-mixing is common among many bi-lingual speakers (Priyadharshini et al., 2020; Jose et al., 2020) and to identify hate speech in Tamil-English tweets, a corpus is created by (Chakravarthi et al., 2020a). The overview of the different approaches to address this issue is presented in (Chakravarthi et al., 2020b; Mandl et al., 2020). A novel methodology has been proposed by (Sainik Kumar Mahata and Bandyopadhyay, 2020) by combining BLSTM with language tags to identify the hate speech in Tanglish tweets. (Devlin et al., 2018) proposed a new language model Bidirectional Encoder Representations from Transformers (BERT) that is suitable for many NLP tasks. BERT is simple and can be fine-tuned for various downsteam task in NLP. In this work, we have applied BERT for the task of identifying the hate speech in code-mixed Tamil-English tweets.

3 Methodology

The objective of this shared task is to identify the hate speech in the code-mixed Tamil-English tweets that contain any one of the following 6 category labels viz., Not-Offensive, Offensive-TargetedGroup, Offensive-Targeted-Individual, Offensive-TargetedOther, Offensive-Untargeted and not-Tamil. The data distribution among the categories is not uniform and 72% of them are not-offensive tweets with the remaining tweets split among the other categories with 7%, 6%, 1%, 8% and 4% respectively. As part of this task, the training and validation set were released with 35,139 and 4388 labelled tweets that followed the same distribution as mentioned above.

The code-mixed Tanglish tweets contain both Tamil and English words and phrases. So, we have converted the Tamil words into English terms by using Tamil to English Mapping Corpus. We have utilized NLTK(Natural Language Tool Kit) package in python to clean the data. For any classification task, the pre-processing steps are important which helps to improve the classifier performance. We have performed some of the cleaning steps like stop word removal,lemmatization and removed the special characters. For example, after pre-processing, the tweet

> "@Bala sundar ayyo sorry...antha line ah clarify pannama vittutu irukan[:drowsy]ok na solran(en appavum indha grant work ku vanthurukkaru,neenga en appava paakala pola....en appavukku munnadiye ipdi enna affront panra maathri kevi kettu asinga paduthuringa nu solraaru[:yeah][:yeah] ' chiiiii karumam podinnngggg... asingama vaila vanthurum...."

is converted into

> "bala sundar ayyo sorry antha line ah clarify pannama vittutu irukandrowsyok na solran en appavum indha grant

work ku vanthurukkaruneenga en appava paakala pola en appavukku munnadiye ipdi enna affront panra maathri kevi kettu asinga paduthuringa nu solraaruyeahyeah chi karumam poding asingama vaila vanthurum."

For any classification task, suitable vector representation need to be chosen that can capture the relationship between the terms in multiple aspects. After the pre-processing step, we have used (Pennington et al., 2014) Glove for obtaining vector representation of the words. The reason to choose Glove embedding is that, it has the ability to capture the linguistic similarity among the words. In this method, the cosine similarity metric has been employed to find the nearest neighbour, which helps in revealing the rare and relevant terms from the corpus. By this method, we could obtain the interesting patterns among the words and their relationship in a better way.

The data set has 6 categories of tweets viz., Not Offensive (25425), Offensive Targeted Insult Group (2557) , Offensive Targeted Insult Individual (2343), Offensive Targeted Insult Other (454), Offensive Untargeted (2906) and not-Tamil (1454) with imbalanced distribution. As this data set is highly imbalanced, we have applied SOUP(Similarity-based Oversampling and Undersampling Pre-processing) (Janicka et al., 2019). In this technique, the number of minority class samples are increased and the number of majority class samples are decreased to make the data set a balanced one. This is performed by removing the most unsafe examples until a desired class cordinality is obtained. The calculation of the safe level is done by using the Heterogeneous Value difference metric (HVDM). By this method, the multi-class imbalance problem is solved and then we used this balanced data for performing classification task. After applying SOUP the samples of all the classes are balanced.

To determine the sentiment expressed in the multi-lingual tweet, we have also applied the BERT model (Devlin et al., 2018). BERT is the language representation model that extracts the context from the input sentence from both the directions. To capture semantic and linguistic features from the given sentence, a bidirectional encoder representation is applied. In general, tweets may have one or more sentences. BERT has the ability to consider these input sentences into a single sequence that can unambiguously result in a better input representation. BERT embeddings are effective compared to other language models, as it combines the token embedding, segment embedding and positional embedding. Another advantage is that, pre-training of BERT combines both Masked Language Model (MLM) and Next Sequence Prediction (NSP). The pre-trained BERT can be fine-tuned to suit the downstream tasks. By adding a classification layer, BERT can be be used for this task of Hate speech identification in code-mixed Tamil-English tweets.

Unlike GloVe, that deals with word-like tokens, BERT considers the word-pieces which are segmented input units. The out-of-vocabulary tokens are also avoided when we use BERT language model. Compared to other language models, we have chosen BERT as it is possible to capture the context in a bidirectional way and efficient than the single directional models. The multi-head attention is applied to identify the key terms that are more important in determining the sentiment of the sentence expressed in Tamil. Considering the baseline models, such as LSTM, BLSTM, the BERT based model performed well on the validation data set. The fine-tuned model was used to predict the sentiment on the released test set. The details of experiments and the results are reported in the next section.

4 Results and Discussion

To study the performance of the proposed method, we have conducted the experiments on the released Code-mixed Tamil-English tweets. All the experiments were carried on a workstation with Intel Xeon Quad Core Processor, 32 GB RAM, NVIDIA Quadro P4000 GPU 8GB. For implementation, we have used Python 3, scikit-learn with NLTK library. For the base line experiments, we have tried with different vector representatiosn like TF-IDF and Glove with machine learning techniques. In order to capture the linguistic terms that contribute more to identify the hate speech, we have tried deep learning technique (BERT) with attention mechanism. We have applied BERT and fine-tuned the parameters. The following parameters were fixed, based on its better performance on the validation set. The experiments were conducted with a learning rate of $3e-5$, batch size of 4 with 5 epochs. We have obtained the validation accuracy of 65% for these parameters. We have applied the obtained best model, on the test set and we could achieve the

Label	Precision	Recall	F-1
Not Offensive	0.79	0.78	0.78
Offensive Targeted Insult Group	0.14	0.14	0.14
Offensive Targeted Insult Individual	0.00	0.00	0.00
Offensive Targeted Insult Other	0.00	0.00	0.00
Offensive Untargeted	0.21	0.38	0.27
not-Tamil	0.99	0.5	0.66
Accuracy	0.63		
Macro-Average	0.36	0.30	0.31
Weighted-Average	0.65	0.63	0.64

Table 1: Performance of the proposed approach

same 63% accuracy. The obtained results for different classes are presented in Table 1. With the BERT based model, the weighted average precision, recall and F_1 values of 0.65, 0.63 and 0.64 have been achieved. It could be observed that, the classifier is able to differentiate the non-offensive and non-Tamil classes alone and could not correctly identify the hate-speech tweets in code-mixed Tamil-English text. As mentioned in Section 3, the data set is highly skewed with more than 75% samples from non-offensive category, the model could not generalize well for the other categories. We are trying to address this issue of multi-class imbalance in the future work by combining other techniques to improve the performance.

5 Conclusion

The social media plays an important role in reflecting people's opinion on different issues. This work focused on identifying the hate speech posted in social media that contains code-mixed Tamil-English tweets, and is submitted to the DravidianLangTech-EACL2021 shared task. To capture the linguistic terms with higher contextual awareness, we have used BERT language model for this task. From the experiments, it is shown that, an F_1 score of 64% has been achieved using this bidirectional transformer based model. The data set is highly skewed, hence suitable techniques for addressing the multi-class imbalance will be explored in future.

6 Acknowledgement

The authors would like to thank the management of Vellore Institute of Technology, Chennai for providing the support to carry out this work. Also, the authors thank the Science and Engineering Research Board, Govt. of India for their financial support (ECR/2016/000484).

References

Bharathi Raja Chakravarthi. 2020a. 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. 2020b. *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. 2019a. Comparison of different orthographies for machine translation of under-resourced dravidian languages. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019b. 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 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. 2020a. 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, Shubhanker Banerjee, Richard Saldhana, John Philip McCrae, Anand Kumar M, Parameswari Krishnamurthy, and Melvin Johnson. 2021a. Findings of the shared task on Machine Translation in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

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. 2021b. 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. 2020b. Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text. *Forum for Information Retrieval Evaluation*.

Bharathi Raja Chakravarthi, Navaneethan Rajasekaran, Mihael Arcan, Kevin McGuinness, Noel E. O'Connor, and John P. McCrae. 2020c. Bilingual lexicon induction across orthographically-distinct under-resourced Dravidian languages. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 57–69, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology*, pages 1–24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Vaishali Ganganwar and R. Rajalakshmi. 2019. Implicit aspect extraction for sentiment analysis: A survey of recent approaches. *Procedia Computer Science*, 165:485–491.

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.

Małgorzata Janicka, Mateusz Lango, and Jerzy Stefanowski. 2019. Using information on class interrelations to improve classification of multiclass imbalanced data: A new resampling algorithm. *International Journal of Applied Mathematics and Computer Science*, 29:769–781.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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.

R. Rajalakshmi and Rohan Agrawal. 2017. Borrowing likeliness ranking based on relevance factor. In *Proceedings of the Fourth ACM IKDD Conferences on Data Sciences*, CODS '17, New York, NY, USA. Association for Computing Machinery.

R Rajalakshmi and B Yashwant Reddy. 2019. Dlrg@ hasoc 2019: An enhanced ensemble classifier for hate and offensive content identification. In *FIRE (Working Notes)*, pages 370–379.

Priya Rani, Shardul Suryawanshi, Koustava Goswami, Bharathi Raja Chakravarthi, Theodorus Fransen, and John Philip McCrae. 2020. A Comparative Study of Different State-of-the-Art Hate Speech Detection Methods in Hindi-English Code-Mixed Data. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 42–48, Marseille, France. European Language Resources Association (ELRA).

Dipankar Das Sainik Kumar Mahata and Sivaji Bandyopadhyay. 2020. JUNLP@Dravidian-CodeMix-FIRE2020: Sentiment Classification of Code-Mixed Tweets using Bi-Directional RNN and Language Tags. *Forum for Information Retrieval Evaluation*.

R Sivanantham and M Seran. 2019. Keeladi: An Urban Settlement of Sangam Age on the Banks of River Vaigai. *India: Department of Archaeology, Government of Tamil Nadu, Chennai*.

Sivakumar Soubraylu and Ratnavel Rajalakshmi. 2020. Hybrid convolutional bidirectional recurrent neural network based sentiment analysis on movie reviews. *Computational Intelligence*, pages 1–23.

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.