CUSATNLP@DravidianLangTech-EACL2021:Language Agnostic Classification of Offensive Content in Tweets

Sara Renjit
Research Scholar
Department of Computer Science
CUSAT
Kerala, India
sararenjit.g@gmail.com

Sumam Mary Idicula
Professor
Department of Computer Science
CUSAT
Kerala, India
sumam@cusat.ac.in

Abstract

Identifying offensive information from tweets is a vital language processing task. This task concentrated more on English and other foreign languages these days. In this shared task on Offensive Language Identification in Dravidian Languages, in the First Workshop of Speech and Language Technologies for Dravidian Languages in EACL 2021, the aim is to identify offensive content from code mixed Dravidian Languages Kannada, Malayalam, and Tamil. Our team used language-agnostic BERT (Bidirectional Encoder Representation from Transformers) for sentence embedding and a Softmax classifier. The language-agnostic representation based classification helped obtain good performance for all the three languages, out of which results for the Malayalam language are good enough to obtain a third position among the participating teams.

1 Introduction

These days, social media platforms support almost all forms of languages for people to communicate. Internet use has increased through the support of language technologies, and code-mixing is common nowadays (Thavareesan and Mahesan, 2019, 2020a,b). With more and more language support, free form communication is possible among people, and as such, comment filtering or evaluating comment content based on its offensiveness to an individual, group, or other forms is an essential task (Jose et al., 2020; Priyadharshini et al., 2020). These platforms also support languages from the Dravidian family (Chakravarthi et al., 2020c; Mandl et al., 2020). As a result, it is necessary to identify offensive content from Tweets in English and all other languages, including code mixed tweets used in these platforms (Hande et al., 2020; Ghanghor et al., 2021b,a).

Dravidian languages are spoken and used in written form by the people in South India and Sri Lanka. This language is also used by a small portion of people in many parts of the world. Dravidian languages are not new, it is almost 4500 years old (Kolipakam et al., 2018), but it remains an under-resourced language (Chakravarthi, 2020). The commonly used languages of this type include Kannada, Malayalam, and Tamil. These are free word order languages and have high agglutinating nature.

Tweet classification based on its offensive content has been challenging over these years. Initially, the focus was on the English language, which is the prime language spoken and used worldwide. Nowadays, its focus is on the low resource languages. The tweets classification is based on different classes, and it can be a binary classification into offensive or not-offensive class. In an alternative way, we can classify tweets based on offensive content, their targeted audience (individual, group, or other), language, and form of insult.

This working note paper discusses the methodology used to identify offensive content from tweets in Kannada, Malayalam, and Tamil. This paper's subsequent sections are as follows: Section 2 discusses the related works based on offensive content identification. Details about the task and the dataset are in Section 3. The methods used for this classification task is in Section 4. We project the results and discuss the observations in Section 5. The last section, Section 6, concludes the working note.

2 Related Works

We discuss the works done so far for offensive content classification in this section. It is one of the very relevant topics of concern these days as communication is mostly through online platforms, especially in a pandemic situation like COVID-19 that we face now. Offensive language identifica-

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 236–242

tion has started with English and other foreign languages, and it has gained its importance concerning other languages, including Dravidian and low resource languages.

Apart from the different forms of offensive content, such as hate speech, abusive language, cyberbullying, cyber-aggression, there are different kinds of offensive content. Zampieri et al. (2019) discuss the different kinds of offensive content classes and a basic definition of the different classes. They proposed a hierarchical model of offensive content in different levels, namely, Level A: Offensive language detection, Level B: Categorization of offensive language, Level C: Offensive language target identification.

The first level (Level A), Offensive language detection, consists of two major classes, namely **Not Offensive**, defined as posts/comments without any offense or profanity and **Offensive** defined as posts with profanity (non-acceptable language) or offense, including insults, threats, and swear words.

The second level (Level B) consists of categorizing offensive language into targeted insult and untargeted class. The **Targeted_Insult** class consists of posts that contain insult to an individual, group, or others. The **Untargeted** class contains posts that are not targeted but have non-acceptable language or swearing.

The third level (Level C) classifies the targets of insults or threats. It is of three types: Individual, Group, and Other. The **Individual (IND)** class has posts that target an individual, may or may not be a named person, also called cyberbullying, whereas the **Group (GRP)** class targets people of the same religion, affiliation, or gender (also referred to as Hate speech). The **Other (OTH)** category targets organizations, situations, events, or issues that do not fall in the above categories. Zampieri et al. (2019) also classified the Offensive Language Identification (OLID) dataset based tweets based on the defined classes using SVM, CNN, BiLSTM, out of which CNN performed better for the three levels of classification.

Offensive language identification has taken place in shared tasks these years. We discuss a few of the notable works here. Apart from the English language, this problem is also being solved in other languages. They predicted offensive content in Bengali, Hindi, and Spanish languages. Pretrained BERT models combined with CNN were used for Arabic, Greek, and Turkish offensive language identification (Safaya et al., 2020).

The significant contributions of offensive language identification from different teams that participated in SemEval, 2020 are discussed in Zampieri et al. (2020). Three sub-tasks corresponding to three levels of the OLID dataset hierarchy were offered for Arabic, Danish, English, Greek, and Turkish languages. German BERT model based on offensive language identification was experimented with for the German language by Risch et al. (2019) as part of GermEval,2019.

3 Task Details and Dataset

The shared task on Offensive Language Identification in Dravidian languages aims to identify offensive content from Kannada-English, Malayalam-English, and Tamil-English code-mixed tweet datasets (Chakravarthi et al., 2021). It is an imbalanced multi-class classification at the comment level. The dataset for this task consists of classes, namely,

1. Not_Offensive (NO)
2. Offensive_Targeted_Insult_Group (OTIG)
3. Offensive_Targeted_Insult_Individual (OTII)
4. Offensive_Targeted_Insult_Other (OTIO)
5. Offensive_Untargetede (OU)
6. Not_in_indented_language (not-Malayalam (NM), not-Kannanda (NK), not-Tamil (NT))

Language	Train	Dev	Test	Total
Kannada	6217	777	778	7772
Malayalam	16010	1999	2001	20010
Tamil	35139	4388	4392	43919

Table 1: Data statistics for Kannada, Malayalam and Tamil

The dataset statistics for the three languages are shown in Table 1. The average length of each post at sentence level is 1. The dataset includes tweets that are code mixed with the intended language (Kannada, Malayalam, or Tamil) and the English language.

3.1 Kannada Dataset

Hande et al. (2020) presented the Kannada-English code mixed dataset collected from YouTube and

Classes	Kannada		Malayalam		Tamil	
	Train	Dev	Train	Dev	Train	Dev
Not_Offensive	3544	426	14153	1779	25425	3193
Not_in_indented_language	1522	191	1287	163	1454	172
Offensive_Targeted _Insult_Group	329	45	140	13	2557	295
Offensive_Targeted _Insult_Individual	487	66	239	24	2343	307
Offensive_Targeted _Insult_Other	123	16	0	0	454	65
Offensive_Untargetede	212	33	191	20	2906	356

Table 2: Training and Development Dataset for Kannada, Malayalam and Tamil languages

annotated for sentiment analysis and offensive language detection. The dataset statistics for the Kannada language is in Table 2. This dataset has 6217 tweets for training and 777 tweets for development.

3.2 Malayalam Dataset

According to Chakravarthi et al. (2020a), systems trained on monolingual data sometimes fail with multilingual data written in non-native scripts. Hence a Malayalam-English code mixed dataset is created, which stands as a gold standard for offensive data identification in Dravidian languages. It helps moderate offensive content in these languages in social media comments or posts. These dataset details are in Table 2, and it has 16010 tweets as training dataset and 1999 tweets as development dataset.

3.3 Tamil Dataset

Tamil-English code mixed corpus with sentiment annotations were also developed, and polarities are assigned with an inter-annotator agreement (Chakravarthi et al., 2020b). The Tamil dataset has 35139 tweets for training, 4388 tweets for the development set and its data statistics are in Table 2.

4 Methods

Our classification method uses a good sentence representation that is purely language-agnostic in nature. The system design consists of a representation module and a classifier module.

Representation module : Efficient sentence embedding for the comments or tweets is derived using the language-agnostic BERT (LaBSE). The comments are in the form of code mixed language scripts written in the non-native script, English, and transliterated form. We have not used any preprocessing technique to make the method more generic and add to the language-agnostic nature.

The language-agnostic BERT uses a masked language model and a translation language model based pretrained encoder. The dual encoder architecture helps to encode source and target sentences separately and feed a combinational function. We obtain the individual embeddings from the [CLS] token's embeddings using a shared BERT encoder. The source and target sentence embeddings comparison using cosine similarity and additive margin softmax loss learn useful cross-lingual embeddings. The architecture of language-agnostic BERT is in Figure 1.

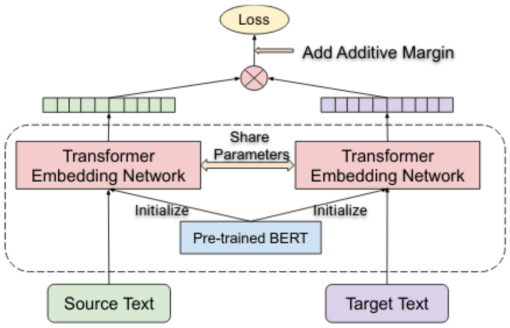

Figure 1: LaBSE model

This approach performed equally well for all the resource-rich languages and resource-poor languages. In zero-shot transfer, it showed above-average performance. This method was evaluated on different corpus like BUCC shared task for parallel sentence extraction, United Nations parallel corpus, Tatoeba corpus for 109 languages, and zero-shot transfer for languages without training data (Feng et al., 2020).

Class	P	R	F1	S
Not_offensive	0.69	0.78	0.73	427
Offensive_Targeted _Insult_Group	0.30	0.23	0.26	44
Offensive_Targeted _Insult_Individual	0.56	0.52	0.54	75
Offensive_Targeted _Insult_Other	0.50	0.07	0.12	14
Offensive _Untargetede	0.00	0.00	0.00	33
not-Kannada	0.66	0.56	0.61	185
accuracy			0.63	778
macro avg	0.45	0.36	0.38	778
weighted avg	**0.62**	**0.63**	**0.62**	**778**

Table 3: Results for Kannada language in terms of Precision (P), Recall (R), F1-score (F1) and Support (S)

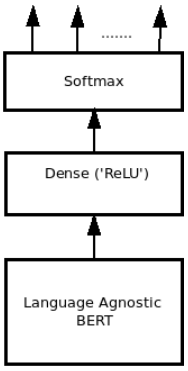

Figure 2: System design

Classification module : The embedded representation of sentences are passed through a perceptron based classifier. It is a two-layer classifier with Dense(100) as an input layer with ReLU activation and Dense(5,6) as an output layer with a softmax activation function for multiclass classification. The classifier is trained for 20 epochs with validation data used from the development dataset provided. The same approach is followed for all three languages. The overall system design is shown in Figure 2.

5 Results and Observations

This classification task's experimental results are measured in terms of weighted average precision, recall, and F1-score using Scikit-Learn classification report. Weighted average averages the support weighted mean per label [1] and are used when the datasets are highly imbalanced.

5.1 Kannada

The classification report for the Kannada language offensive content identification into six classes is in Table 3. We observe that the system is good only at classifying not-offensive classes and comments that do not belong to the Kannada language.

The misclassified instances can be clearly seen through the confusion matrix shown in Table 4. Samples from Offensive_Targeted_Insult_Group, Offensive_Targeted_Insult_Other, Offensive_Untargeted classes are the most misclassified instances.

5.2 Malayalam

The results for Malayalam code mixed tweet classification is in Table 7. From the table, we find an above-average performance for the classification of all the classes, resulting in an F1-score of 0.95. This model's better classification F1-score for the Malayalam dataset implies that LaBSE based sentence representations are suitable for Malayalam code mixed texts.

The confusion matrix for Malayalam dataset classification is shown in Table 5. Here we observe that majority of samples from each class are classified correctly which lead to an F1-score of 0.95.

5.3 Tamil

For the Tamil language, the classification report for offensive content classification is in Table 8, and we observe above average results only for the not-offensive class and an average value for tweets that do not belong to the Tamil language.

The confusion matrix for Tamil dataset classification is shown in Table 6, from which it is found that Not_offensive and not-Tamil are the only classes classified adequately, which resulted in lesser performance.

5.4 Comparison with MBERT

The tweets are encoded with Multilingual BERT (Devlin et al., 2018), bert-base-multilingual-cased, a pretrained encoder with 12 layer, 768 hidden layer size, 12 heads and 110M parameters trained for 104 languages. The same classification module is used. LaBSE based results are compared with the multilingual BERT based classification results in Table 9 in terms of its weighted F1 score. We

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

Actual Classes \ Predicted Classes	NO	OTIG	OTII	OTIO	OU	NK
NO	**333**	11	18	0	20	45
OTIG	23	**10**	4	1	3	3
OTII	23	4	**39**	0	5	4
OTIO	6	6	0	**1**	1	0
OU	25	2	4	0	**0**	2
NK	72	0	5	0	4	**104**

Table 4: Confusion Matrix for Kannada dataset

Actual Classes \ Predicted Classes	NO	OTIG	OTII	OU	NM
NO	**1726**	3	9	8	19
OTIG	10	**12**	1	0	0
OTII	12	0	**14**	0	1
OU	11	0	0	**18**	0
NM	29	0	0	0	**128**

Table 5: Confusion Matrix for Malayalam dataset

Actual Classes \ Predicted Classes	NO	OTIG	OTII	OTIO	OU	NT
NO	**2823**	116	79	8	117	47
OTIG	172	**60**	23	1	29	3
OTII	177	28	**60**	7	26	17
OTIO	46	8	6	**1**	10	0
OU	203	37	23	4	**93**	8
NT	65	6	4	0	4	**81**

Table 6: Confusion matrix for Tamil dataset

find that the performance of both the models are almost similar with respect to this task.

6 Conclusions

In this paper, we used language-agnostic BERT for embedding comments/tweets of any language and identified the different categories of offensive content using a softmax classifier. This model achieved comparatively good results without applying any pre-processing techniques. The quality of the embedding model helps to encode information more effectively. Comparing LaBSE for Kannada, Malayalam, and Tamil language, this embedded representation is more efficient for the Malayalam language and helped improve classification results. Our team was ranked 3rd for the Malayalam language, 11th for the Kannada language, and 9th for the Tamil language in DravidianLangTech 2021, part of EACL, 2021.

References

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

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

Class	P	R	F1	S
Not_offensive	0.97	0.98	0.97	1765
Offensive_Targeted _Insult_Group	0.80	0.52	0.63	23
Offensive_Targeted _Insult_Individual	0.58	0.52	0.55	27
Offensive _Untargetede	0.69	0.62	0.65	29
not-malayalam	0.86	0.82	0.84	157
accuracy			0.95	2001
macro avg	0.78	0.69	0.73	2001
weighted avg	**0.95**	**0.95**	**0.95**	**2001**

Table 7: Results for Malayalam language in terms of Precision(P), Recall(R), F1-score(F1),and Support (S)

Class	P	R	F1	S
Not_offensive	0.81	0.88	0.85	3190
Offensive_Targ _Insult_Group	0.24	0.21	0.22	288
Offensive_Targ _Insult_Individual	0.31	0.19	0.24	315
Offensive_Targ _Insult_Other	0.05	0.01	0.02	71
Offensive _Untargetede	0.33	0.25	0.29	368
not-Tamil	0.52	0.51	0.51	160
accuracy			0.71	4392
macro avg	0.38	0.34	0.35	4392
weighted avg	**0.67**	**0.71**	**0.69**	**4392**

Table 8: Results for the Tamil language in terms of Precision(P), Recall(R), F1-score(F1),and Support (S)

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. 2020c. 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.

Language	LaBSE	MBERT
Kannada	0.62	0.68
Malayalam	0.95	0.93
Tamil	0.69	0.68

Table 9: Comparison of LaBSE and MBERT encoders in terms of weighted F1-score.

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*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

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.

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.

Vishnupriya Kolipakam, Fiona M Jordan, Michael Dunn, Simon J Greenhill, Remco Bouckaert, Russell D Gray, and Annemarie Verkerk. 2018. A bayesian phylogenetic study of the dravidian language family. *Royal Society open science*, 5(3):171504.

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.

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.

Julian Risch, Anke Stoll, Marc Ziegele, and Ralf Krestel. 2019. hpidedis at germeval 2019: Offensive language identification using a german bert model.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059.

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.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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 1415–1420.

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.