Amrita_CEN_NLP@DravidianLangTech-EACL2021: Deep Learning-based Offensive Language Identification in Malayalam, Tamil and Kannada

Sreelakshmi K, Premjith B and Soman K.P
Center for Computational Engineering and Networking (CEN)
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
b_premjith@cb.amrita.edu

Abstract

This paper describes the submission of the team Amrita_CEN_NLP to the shared task on Offensive Language Identification in Dravidian Languages at EACL 2021. We implemented three deep neural network architectures such as a hybrid network with a Convolutional layer, a Bidirectional Long Short-Term Memory network (Bi-LSTM) layer and a hidden layer, a network containing a Bi-LSTM and another with a Bidirectional Recurrent Neural Network (Bi-RNN). In addition to that, we incorporated a cost-sensitive learning approach to deal with the problem of class imbalance in the training data. Among the three models, the hybrid network exhibited better training performance, and we submitted the predictions based on the same.

1 Introduction

In recent years, people from all walks of life use social platforms like Twitter, Instagram, Facebook. So, it is demanding to monitor their behavior to avoid violence, hateful and offensive content (Thavareesan and Mahesan, 2019, 2020a,b). Offensive content is any non-verbal or oral communication expressing disparity against a group or person based on their religion, age, sexual orientation, race, gender, nationality, and ethnicity (Chakravarthi and Muralidaran, 2021; Suryawanshi and Chakravarthi, 2021).

A substantial amount of work was done to identify offensive content in English, but much work is not done in Dravidian languages (Chakravarthi et al., 2018, 2019; Chakravarthi, 2020). The Dravidian languages were first documented in Tamili script engraved on cave walls in Tamil Nadu's Madurai and Tirunelveli districts in the 6th century BCE. India being a multilingual country, a lot of people use regional languages along with English. The usage of two languages to communicate is called code-mixing. It is even more challenging to identify hateful content from code-mixed language owing to the non-standard grammar and spelling (Sreelakshmi et al., 2020), (Sreelakshmi et al., 2019), (Sasidhar et al., 2020).

DravidianLangTech-EACL2021 is a task to identify offensive content from code-mixed Tamil-English (Tam-Eng), Malayalam-English (Mal-Eng), and Kannada-English (Kan-Eng). In this task, we came up with a Deep learning model to identify offensive content from Malayalam-English, Tamil-English, and Kannada-English datasets. We employed three different deep learning models for solving the classification problem. A hybrid model that includes a convolutional layer followed by a Bi-LSTM (Graves et al., 2013), (Premjith et al., 2018) and a fully connected network attained the maximum scores while training. Therefore, the labels for the test data were predicted using the aforementioned model.

The rest of the contents are explained in the following sections: Section 2 presents the literature review. Dataset details are provided in Section 3. Section 4 explains the system description, and section 5 relates to experimental details and results. Finally, the work is concluded in Section 6.

2 Literature Review

Different abusive and offense language identification problems and shared tasks have been explored in the literature ranging from aggression to cyberbullying, hate speech, toxic comments, and offensive language. Below we discuss each of them briefly.

In 2018, Adithya et.al (Bohra et al., 2018) evolved a dataset consisting of 4500 hate and non-hate code-mixed Hindi-English tweets. The dataset was congregated using Twitter API and annotated by two linguists. Machine learning models like

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 249–254

Random Forest and SVM and handcrafted features like character N-Grams, punctuation count, emoticon count, negation words, word N-Grams were used for classification.

SemEval (Zampieri et al., 2019) conducted three tasks in 2019, of which one task is on offensive and non-offensive comments detection from English tweets. The dataset (OLID) used for the task has 13240 tweets for training and 860 tweets for testing. Several models like Convolutional Neural Networks (CNN), Bidirectional Encoder Representations from Transformers (BERT), Long Short Term Memory (LSTM), LSTM with attention, Embeddings from Language Models (ELMo) were used by various teams. Even basic machine learning models like SVM was a part of the assorted models used.

SemEval 2020 conducted a task on offensive language identification in multilingual languages (OffensEval) such as English, Arabic, Danish, Greek, and Turkish. The same task was also conducted for Indo-European languages in FIRE 2019 (Mandl et al., 2019).

In 2020, FIRE conducted a shared task called Hate Speech and Offensive Content Identification from code-mixed posts in Dravidian languages (Malayalam-English and Tamil-English) (Chakravarthi et al., 2020d; Mandl et al., 2020; Chakravarthi et al., 2020b). Different teams came up with diversified approaches of which include, the work by Gaurav Arora (Arora, 2020). He came up with an approach based on a pretraining ULMFiT on code-mixed data, which are generated synthetically. The code-mixed data was modeled as a Markov process using Markov chains.

3 Dataset Description

The dataset (Chakravarthi et al., 2021), (Chakravarthi et al., 2020a), (Chakravarthi et al., 2020c), (Hande et al., 2020) consists of sentences from three code-mixed languages namely Tamil-English, Malayalam-English and Kannada-English. The Kannada-English and Tamil-English dataset have sentences labeled to six classes and, the Malayalam-English dataset has five labels. The labels for each language are given in Table 1 and the dataset statistics is given in Table 2.

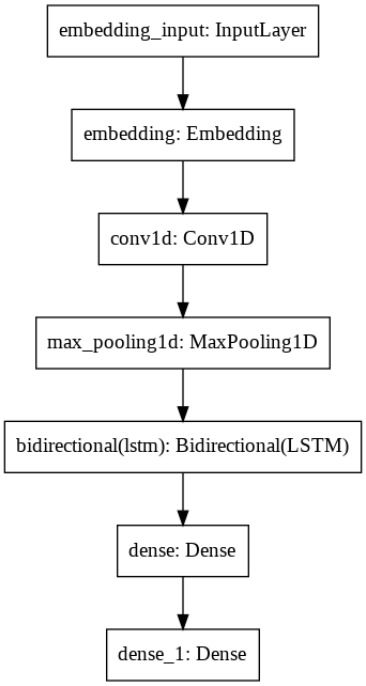

Figure 1: An illustration of the deep learning model submitted to the shared task

4 System Description

This section describes the details of the model submitted to the shared task. We have experimented with various deep neural networks for identifying the underlying patterns in the text required for classification.

4.1 Preprocessing

The dataset provided for the shared tasks contains words in both native languages (Malayalam, Tamil, and Kannada) and English. The dataset comprises social media texts and hence includes user names, hashtags, and URLs. Since these entities do not contribute much to the classification task, we employed a preprocessing step to remove such entities from the text. In addition to that, the preprocessing step involved steps to remove the punctuation and to lower-case the English characters.

4.2 Models

We experimented with different deep learning models for classifying the social media text into different categories. The model which obtained the highest accuracy when tested with the validation data is a hybrid of a 1-D convolution layer, a 1-D max-pooling layer, a Bidirectional-LSTM, and a fully connected network along with another fully connected layer dedicated for classification, and

Langauge	Label
Malayalam-English	Offensive_Targeted_Insult_Group, Offensive_Targeted_Insult_Individual, Offensive_Untargetede, Not_offensive, not-Malayalam
Tamil-English	Not_offensive, Offensive_Targeted_Insult_Group, Offensive_Targeted_Insult_Individual, Offensive_Targeted_Insult_Other, Offensive_Untargetede, not-Tamil
Kannada-English	Not_offensive, Offensive_Targeted_Insult_Group, Offensive_Targeted_Insult_Individual, Offensive_Targeted_Insult_Other, Offensive_Untargetede, Not_Kannada

Table 1: Details of the class labels available in the dataset.

Language	Train set	Valid set	Test set
Mal-Eng	16010	1999	2001
Tam-Eng	35139	4388	4392
Kan-Eng	6217	777	778

Table 2: Statistics of the dataset used in the shared task.

is illustrated in Figure 1. The same model is used for all the classification tasks. The other models considered are a network containing one Bi-LSTM layer and another with a Bi-RNN layer.

The cleaned text is fed into the model after another sequence of preprocessing steps which involve the following,

- Tokenization: An "<OOV>" token is used to mark the Out-of-Vocabulary (OOV) words in the test data.
- Translation of words into indexes.
- Padding the sequences with zeros to nake the sequence length equal: Here, maximum sequence length is set to the length of the lengthiest sentence in the dataset. Here, the zeros are padded at the end of the sequences. This padded sequences are fed into the model.

The dataset used for this task is highly imbalanced. To reduce the bias towards the majority class, we applied a cost-sensitive learning approach. This approach computes weights for each class so that the majority class gets minimum weight, and the minority class gets maximum weight. Equation 1 is used for computing the class weights.

$$cw = \frac{N}{N_c} \tag{1}$$

Where cw is the class weight, N is the total number of data points in the corpus, and N_c is the number of sentences in the class c.

Hyperparameter	Value
Embedding dimension	100
Convolution filter size	128
Convolution kernel size	5
Activation function at Conv1D layer	ReLU
Padding at Conv1D layer	Same
Pool size	5
Padding at the pooling layer	Same
No. of neurons in the Bi-LSTM	32
No. of neurons in the fully connected layer	32
Activation function at the fully connected layer	ReLU
Activation function at the output layer	Softmax
Loss	Categorical crossentropy
Optimizer	Adam
Learning rate	0.01

Table 3: Set of hyperparameters used in building the model.

4.3 Hyperparameter Tuning

Hyperparameter tuning is a crucial step in building a deep learning model. The performance of a deep learning model heavily relies on the optimal selection of the hyperparameters. In this model, we chose the hyperparameters from a set of values based on the metrics considered for evaluating the model. The metrics used in this model are accuracy, precision, and recall, and AUC. The hyperparameters were chosen based on the performance of the model on validation data. A grid search method was used to find the optimal hyperparameters from a set of values.

Model	Accuracy	Precision	Recall	AUC
Model-1	0.9677	0.9241	0.9135	0.9847
Model-2	0.9494	0.9420	0.7959	0.9255
Model-3	0.9432	0.8613	0.8539	0.9635

Table 4: Training performance of various models experimented for the Malayalam-English data.

Model	Accuracy	Precision	Recall	AUC
Model-1	0.8932	0.6936	0.6438	0.8872
Model-2	0.8680	0.6371	0.4829	0.8456
Model-3	0.8364	0.9100	0.0207	0.8801

Table 5: Training performance of various models experimented for the Tamil-English data.

Model	Accuracy	Precision	Recall	AUC
Model-1	0.8795	0.6951	0.4929	0.8396
Model-2	0.8567	0.5767	0.5277	0.8469
Model-3	0.8368	0.5113	0.4672	0.7891

Table 6: Training performance of various models experimented for the Kannada-English data.

Dataset	Precision	Recall	F1-score
Mal-Eng	0.90	0.82	0.85
Tam-Eng	0.64	0.62	0.62
Kan-Eng	0.65	0.54	0.58

Table 7: Performance of the model over the test data.

The set of optimal hyperparameters for this model are shown in Table 3. We used the same model for all the tasks and hence didn't change the hyperparameters for individual tasks.

5 Results and Discussion

We experimented with three deep learning models for three subtasks in the shared task. The first model, Model-1 is a hybrid of CNN, Bi-LSTM, and a fully connected layer apart from the output layer, the second model, Model-2, has a Bi-LSTM layer, and the third model, Model-3, is made up of a Bi-RNN layer. We used validation data to evaluate the models to identify the best performing model. The performance was measured using metrics such as accuracy, precision, recall, and AUC. Among the three models, the network containing CNN+Bi-LSTM+Dense layers achieved the best scores. Even though all the models exhibited comparable accuracy, the decisive factor was the recall score. The hybrid model performed substantially better than the other two models in terms of recall and precision. Besides that, it is also evident that the hybrid model could use the class weights effectively. This trend is visible in all three tasks. Tables 4, 5, and 6 shows the training performance of the Malayalam-English dataset, Tamil-English dataset and Kannada-English dataset, respectively. We submitted the predictions obtained by Model-1 based on the training performance.

The performance of the submitted model over the testing data is given in Table 7.

6 Conclusion

This paper presents the submission of Amrita_CEN_NLP to the shared task at EACL 2021 on Offensive Language Identification from three Dravidian Languages, namely Tamil-English (Tam-Eng), Malayalam-English (Mal-Eng), and Kannada-English (Kan-Eng). Three Deep Learning architectures, such as a hybrid network with a Convolutional layer, a Bidirectional Long Short-Term Memory network (Bi-LSTM) layer, and a fully connected network, a network containing a Bi-LSTM, and another with a BidirectionalRecurrent Neural Network (Bi-RNN) were implemented. The class imbalance problem was solved using the cost-sensitive learning approach. The hybrid of CNN, Bi-LSTM, and a fully-connected layer model gave the highest result of 90% accuracy for Mal-Eng, 64% accuracy for Tam-Eng, and 65% accuracy for kan-Eng.

References

Gaurav Arora. 2020. Gauravarora@ HASOC-Dravidian-CodeMix-FIRE2020: Pre-training ULMFiT on Synthetically Generated Code-Mixed Data for Hate Speech Detection. *arXiv preprint arXiv:2010.02094*.

Aditya Bohra, Deepanshu Vijay, Vinay Singh,

Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 36–41.

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.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee.

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.

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*, pages 14–17.

B Premjith, KP Soman, and Prabaharan Poornachandran. 2018. A deep learning based Part-of-Speech (POS) tagger for Sanskrit language by embedding character level features. In *FIRE*, pages 56–60.

T Tulasi Sasidhar, B Premjith, and KP Soman. 2020. Emotion Detection in Hinglish (Hindi+ English) Code-Mixed Social Media Text. *Procedia Computer Science*, 171:1346–1352.

K Sreelakshmi, B Premjith, and KP Soman. 2019. Amrita CEN at HASOC 2019: Hate Speech Detection in Roman and Devanagiri Scripted Text. In *FIRE (Working Notes)*, pages 366–369.

K Sreelakshmi, B Premjith, and KP Soman. 2020. Detection of Hate Speech Text in Hindi-English Code-mixed Data. *Procedia Computer Science*, 171:737–744.

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.

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). *arXiv preprint arXiv:1903.08983*.