Offensive Language Identification using Machine Learning and Deep Learning Techniques

C Jerin Mahibha[1], Sampath Kayalvizhi[2], Durairaj Thenmozhi [3] and Sundar Arunima [4]

[1]*Meenakshi Sundararajan Engineering College, Chennai*
[2]*Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam*
[3]*Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam*
[4]*Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam*

Abstract

Offensive Language is the use of abusive, rude or insulting language that upsets or embarrasses people towards whom it is been spoken. This paper aims to identify whether the given text is offensive or not using deep learning models. The proposed model uses a bidirectional dual-encoder with Additive Margin Softmax to perform the classification task. The performance of the model is also compared with a machine learning model and a recurrent model. When a cross lingual sentence embedding space transformer model was applied on the Tamil dataset provided by HASOC @ FIRE 2021 for Task 1, it was able to classify the offensive and non offensive data with a F1 score of 0.865 which brought our team to the top of the leaderboard score.

Keywords

Offensive, Classification, Transformer, Recurrent Network, Encoding, Deep Learning, Code-mixed data

1. Introduction

In this internet era with wide access to social media sites, online expression of opinions, through different social media is a common practice. Expressing opinions with the use of hate and offensive language are nowadays very common in the society via various social media like Facebook, Twitter, YouTube etc. and this might attack or offend the users for a variety of reasons. These may also be intended towards different communities, gender, ethnicity, language, caste etc. and may lead to problems in the society and hence identifying the information which acts as a source of spreading hate or offensive language has become necessary. This can also be considered as a step to reduce the crime rate in the society. Due to these reasons, detection of hate speech and offensive language has been considered as an emerging application in numerous research problems associated with the domain of Natural Language Processing. Even though there is a distinctive amount of work based on this in languages like English, Dravidian languages lack this and it has become important to extend the research associated with hate and offensive language identification for Dravidian languages[1] like Tamil, Malayalam etc. Various machine learning algorithms and deep learning models could be employed to detect the

FIRE 2021 : Forum for Information Retrieval Evaluation, December 13-17, 2021, India

jerinmahibha@gmail.com (C. J. Mahibha); kayalvizhis@ssn.edu.in (S. Kayalvizhi); d_theni@ssn.edu.in (D. T.); arunima1706@cse.ssn.edu.in (S. A.)

CEUR Workshop Proceedings (CEUR-WS.org)

hate / offensive information [2]. The deep learning models had been extended to identify hate speech in low resourced languages as well [3].

We cannot expect the performance of a particular classification model to act the same over a variety of datasets. Different classifiers have to be used to train the dataset and the classifier that produces the best performance has to be identified. When the same classifier is applied over different data set, it does not produce the same performance, hence researchers try to apply different classification algorithms over the same data set and try to identify the model that best suits the given data set. Classical machine learning approaches can also be used for classifying offensive and non offensive instances [4].

As individual models has their own disadvantages, cross validated ensemble models [5] are implemented by combining individual models and they can be used for offensive language identification. BERT based models show a great impact on the predictions when using an unbalanced dataset [6], [7] when hate speech has to be identified from multi-platform data like YouTube, Reddit, Wikipedia and Twitter.

We participated in the shared task: Dravidian CodeMix- HASOC@FIRE-2021[1] as a team SSN_NLP and focused on identifying Offensive Language from Code-Mixed Text in the Dravidian Language Tamil. The aim of the shared task was to do comment/post level binary classification on the given data setwith two labels representing that the language used is offensive or not.

2. Related Work

Hate speech and offensive language detection in languages like Tamil and Malayalam, where the social media posts are mostly code mixed, [8] had been implemented using 1 to 6-gram character TF-IDF features with models like Naive Bayes, Logistic Regression, and Neural Network [9]. This provided better performance measures when compared to more popular transfer learning models such as BERT and ULMFiT and hybrid deep models. Also it had been found that TFIDF and count vectorizer features provide better performance when compared to sentence embeddings [10] on the data set provided by DravidianLangTech-EACL2021.

The performance of eight different machine learning classifiers namely NB, SVM, KNN, DT, RF, AdaBoost, MLP and LR had been applied over publicly available dataset for hate speech detection and had shown that the support vector machine algorithm [11] when combined with bigram features out performed the other models with an accuracy of 79%.

The Naive Bayes classifier which is a rule based neural network approach had been used for a fine-grained classification of tweets [5] as offensive or not as a part of GermEval task 2018.

Various neural network models like CNN, GRU and CNN + GRU was used to identify offensive language in the language with few computational resources like Arabic [3] and the performance

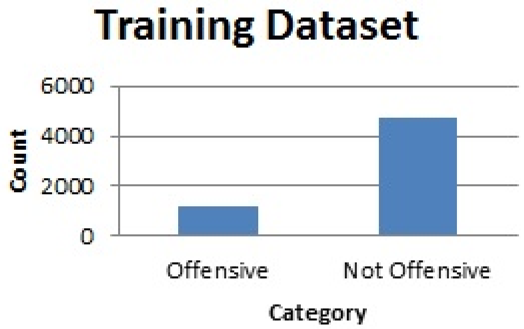


Figure 1: Distribution of data in the data set

were compared. The use of BERT for the problem of hate speech identification for Arabic language had also been carried out.

Offensive language detection in Dravidian languages like Malayalam, Tamil and Kannada had reaped better results when a pre-trained, fine-tuned, and ensembled versions of XLM-RoBERTa was used for the classification process [12]. Considering low resourced Indo-Aryan and Dravidian languages, multilingual transformer models are found to outperform monolingual models in the task of offensive language identification and the cross-lingual transformers show strong zero-shot and few-shot performance across languages.

As from the study it was evident that both machine learning models and deep learning models could be used for identifying offensive text, we started the task using a machine learning model which is logistic regression, then we moved on with a recurrent unit based deep learning model and finally we implemented the task of identifying offensive text using a transformer based model, by which we were able to improve the F1 score of the model.

3. Data set Description

The training and the testing data set used for the shared task was provided by HASOC@FIRE-2021. The comment/post had been collected from YouTube and few of it contained more than one sentence but one sentence was the corpus average sentence length. Each instance was annotated at the comment/post level using two labels OFF and NOT representing Offensive language and Non offensive language respectively. The data set had a total of 5880 instances with 1153 instances under the category OFF and 4727 instances under the non offensive category which shows the imbalanced nature of the data in the data set.

The Figure 1 shows the distribution of the data in the dataset provided by HASOC@FIRE-2021 for the Task1. It was evident that the dataset was an imbalanced one with more instances under the category Not- Offensive and less number of instances under the Offensive category.

As the validation data-set was not explicitly provided, 20% of the training data-set under both the offensive and the non offensive category was separated and used for validation purposes. The remaining 80% of the data-set was used for the training purpose. A separate data set was provided for testing for which the labels had to be predicted using the proposed model based on which the F1 score of the model was computed.

4. Proposed methodology

The shared task which we participated in the Dravidian Code mix HASOC@FIRE-2021 was the Task 1 which aimed at classifying the given code mixed Tamil data into offensive and not offensive instances. We tried the classification problem with Deep learning models and then performance was compared with Machine language model.

The machine learning model that we used for classification was logistic regression [13] which is a statistical machine learning algorithm that can be utilized for binary classification problems as in identifying offensive language which is represented by 1 and non offensive language which is represented by 0. The given data from the training dataset is preprocessed by removing all duplicates. The data from the data set was converted into a matrix of TF-IDF features. As the data set was an imbalanced data set, over-sampling of the vectorised data was done. The random over sampling mechanism was used which adds instances to the data set by repeating the instances which has fewer occurrence in the category and thus balancing the number of samples between classes in a dataset. This up-sampled vectorised data was used for training the model which used the concept of cross validation of the data set [1]. Then the label associated with the given test data set was predicted using the trained model. The values of the metrics that we obtained on the training data set using this model were: accuracy of 78.37%, precision of 51.62% and F1 score: 52.43%.

Two deep learning models were used over the given HASOC dataset. The dataset for validation of the model was generated from the training set by considering 20% of both the offensive and non offensive data that were provided with the training data set. The first model identified the offensive language using a sequential model with recurrent units [14], [15]. The task of classification was carried out using this model by transforming sequences from one form to another using recurrent neural networks. The recurrent neural network used is a deep multi-layer RNN which is unidirectional and uses LSTM as a recurrent unit[16]. The embedding layer generates the source and target embedding using the vocabulary generated from the training set. The generated embedding is fed to the encoder which is a multi layer RNN which is then fed to a decoder which also has access to the hidden layer of the encoder, based on which the predictions are carried out[2]. The performance of the system was improved by using scaled_luong as the attention mechanism. The batch size that we used for our implementation was 128 with a 20% dropout layer. The implementation used 4 encoder and 4 decoder layers. We trained the model

[1] https://towardsdatascience.com/yet-another-twitter-sentiment-analysis-part-1-tackling-class-imbalance-4d7a7f717d44

[2] https://github.com/tensorflow/nmt

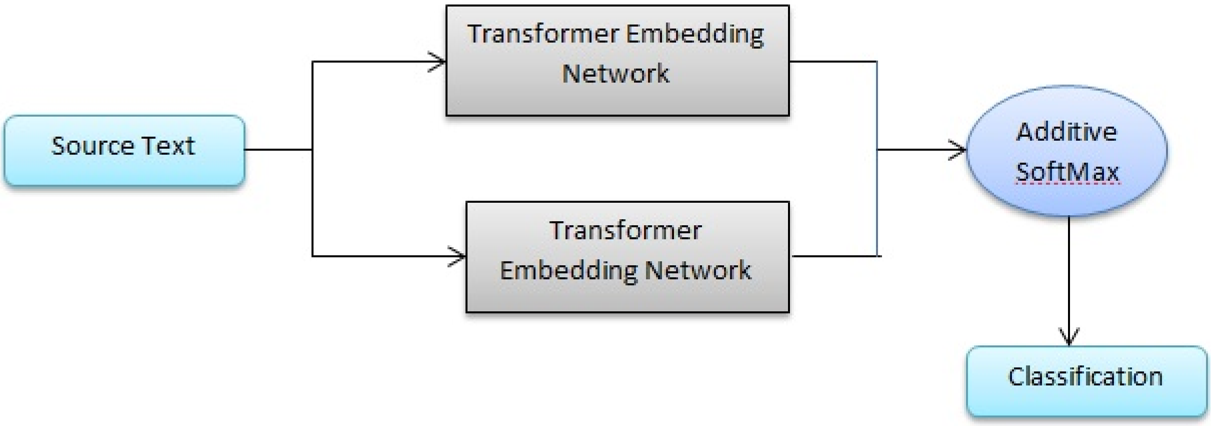


Figure 2: Model Architecture

with 16 epochs and the resultant vocabulary size generated by our model was 23867.The best accuracy obtained using the validation process was 79.6%.

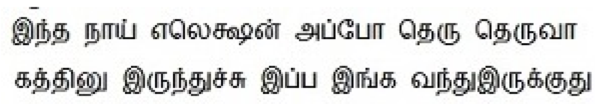
The other deep learning model that we used for offensive language classification was implemented using the transformers that made use of a cross lingual sentence embedding space. It is a multilingual model which boost the performance by implementing a combination of pre-training and fine-tuning strategies and it spans for over 109 languages. In this method we used the sentence based transformer model to map the input sentences into a dense vector space and based on this the classification problem was solved[17]. The model includes a dual-encoder framework for implementing a cross-lingual sentence embedding where the source and target sentences are encoded separately using a shared BERT based encoder. The similarity of these encoding are computed using the concept of cosine similarity based on which the classification is done[3]. As the classification boundaries are narrow in vector spaces, seperation of vectors becomes a difficult process. Additive Margin Softmax(AMS) is used for the classification which is a variation of the softmax function with an additional parameter which helps to increase the separability between the vectors [18]. Figure 2 shows the architecture of the proposed model.

The BERT based transformer model uses 12 layers with a global batch size of 2048 and hyper parameters like AdamW optimizer, initial learning rate of 1e-5 and linear weight decay mechanism. When the model was trained for five epochs by using 80-20% cross validation we obtained the MCC metric value as 0.45 and the F1 score computed was 0.5496. But when this

[3] https://huggingface.co/setu4993/LaBSE

Table 1
Submission score

Parameters	Score
Precision	0.856
Recall	0.864
F1 Score	0.859



இந்த நாய் எலெக்ஷன் அப்போ தெரு தெருவா
கத்தினு இருந்துச்சு இப்ப இங்க வந்துஇருக்குது

Figure 3: Sample data 1

was applied to the test data provided by HASOC for Task1 we were able to top the leader board with a F1 score of 0.85.

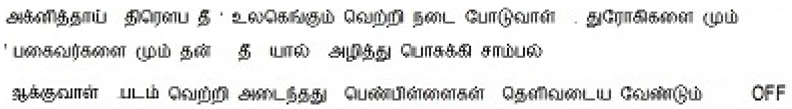
The identification of offensive language code mixed Tamil text was done using both machine learning and deep learning models. But it was found that the deep learning models outperformed the traditional machine learning algorithm that is Logistic regression.

The details given above show that the different metric values obtained for different models using the same data set. The deep learning model that used the cross lingual sentence embedding space for the classification of the given data was found to provide more accuracy than the Recurrent Network model. The result of the HASOC@FIRE-2021 was evaluated based on the weighted average F1 score generated for the prediction of the class labels for the test data provided. Our model was evaluated to generate the prediction with F1 score of 0.859.

The table 1 shows the value obtained by our model for various parameters based on the predictions done on the provided test data set. The F1 score generated by our model was greater than all other submissions and we obtained Rank 1 under task 1 of HASOC@FIRE-2021.

5. Error Analysis

As the text provided for Task 1 of HASOC@FIRE-2021 is basically posts from social media, there was every opportunity for the data not to follow grammatical rules. It may also have creative spellings, symbols and large usage of non standard abbreviations. Before providing the training data to the model pre-processing has to be done which will help the model to handle the data better during the training phase which could have improved the performance measure while predicting the class of the test data set.



அக்னித்தாய் திரௌப தீ ' உலகெங்கும் வெற்றி நடை போடுவாள் . துரோகிகளை யும்
' பகைவர்களை யும் தன் தீ யால் அழித்து பொசுக்கி சாம்பல்
ஆக்குவாள் படம் வெற்றி அடைந்தது பெண்பிள்ளைகள் தெளிவடைய வேண்டும் OFF

Figure 4: Sample data 2

We consider two statements given in the test dataset which has not been predicted correctly. The sample sentence from the test data set given in Figure 3 was predicted as the category ”NOT”, which represents that the text is not offensive, but considering this sentence the usage of words shows that it is an offensive text.

Figure 4 shows a sample sentence from the test data which had been predicted as an Offensive one. But while analysing it could be found that, even though the sentence use few words which are associated with offensive category, the overall structure and associated meaning of the sentence is not offensive. Both these statements have been wrongly captured by our model.By fine tuning the proposed model, these errors could be avoided which in turn would improve the performance of the model.

The data set provided was not balanced. The data set could be converted to a balanced one by carefully performing the process of up-sampling without losing the required features. Training the proposed model using this transformed data set can also be expected to improve the performance measures of the model to an extent.

6. Conclusion

From the output obtained from the different models it could be inferred that deep learning models outperform the machine learning models considering the offensive language classification problem for the data set provided by HASOC@FIRE-2021 for Task 1 associated with code mixed Tamil. Among the deep learning model transformer based models has done the more accurate predictions compared to recurrent models, hence more scope for transformer based models could be identified for research based on Dravidian languages and in specific Hate and Offensive language based researches.

References

[1] B. R. Chakravarthi, P. K. Kumaresan, R. Sakuntharaj, A. K. Madasamy, S. Thavareesan, P. B, S. Chinnaudayar Navaneethakrishnan, J. P. McCrae, T. Mandl, Overview of the HASOC-DravidianCodeMix Shared Task on Offensive Language Detection in Tamil and Malayalam, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.
]

[2] Z. Zhang, L. Luo, Hate speech detection: A solved problem? the challenging case of long tail on twitter, 2018. arXiv:1803.03662.

[3] R. Alshaalan, H. Al-Khalifa, Hate speech detection in saudi twittersphere: A deep learning approach, in: Proceedings of the Fifth Arabic Natural Language Processing Workshop, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 12–23. URL: https://aclanthology.org/2020.wanlp-1.2.

[4] H. A. Nayel, H. Shashirekha, DEEP at hasoc2019: A machine learning framework for hate speech and offensive language detection., 2019.

[5] J. Risch, E. Krebs, A. Löser, A. Riese, R. Krestel, Fine-grained classification of offensive language, Proceedings of GermEval (co-located with KONVENS) (2018) 38–44.

[6] J. Salminen, M. Hopf, S. A. Chowdhury, S.-g. Jung, H. Almerekhi, B. J. Jansen, Developing an online hate classifier for multiple social media platforms, Human-centric Computing and Information Sciences 10 (2020) 1–34.

[7] A. Kalaivani, D. Thenmozhi, SSN_NLP_MLRG@HASOC-FIRE2020: multilingual hate speech and offensive content detection in indo-european languages using ALBERT, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 188–194. URL: http://ceur-ws.org/Vol-2826/T2-12.pdf.

[8] B. R. Chakravarthi, R. Priyadharshini, N. Jose, A. Kumar M, T. Mandl, P. K. Kumaresan, R. Ponnusamy, H. R L, J. P. McCrae, E. Sherly, 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, Kyiv, 2021, pp. 133–145. URL: https://aclanthology.org/2021.dravidianlangtech-1.17.

[9] S. Saumya, A. Kumar, J. Singh, Offensive language identification in Dravidian code mixed social media text, 2021.

[10] B. Bharathi, et al., SSNCSE_NLP@ DravidianLangTech-EACL2021: offensive language identification on multilingual code mixing text, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 313–318.

[11] S. Abro, S. Shaikh, Z. H. Khand, Z. Ali, S. Khan, G. Mujtaba, Automatic hate speech detection using machine learning: A comparative study, International Journal of Advanced Computer Science and Applications 11 (2020). URL: http://dx.doi.org/10.14569/IJACSA.2020.0110861. doi:10.14569/IJACSA.2020.0110861.

[12] S. Sai, Y. Sharma, Towards offensive language identification for Dravidian languages, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 18–27. URL: https://aclanthology.org/2021.dravidianlangtech-1.3.

[13] K. Chopra, C. Srimathi, Logistic regression and convolutional neural networks performance analysis based on size of dataset, Complex Intell. Syst 6 (2018).

[14] T. Gowda, J. May, Neural machine translation with imbalanced classes, arXiv e-prints (2020) arXiv–2004.

[15] S. Yang, Y. Wang, X. Chu, A survey of deep learning techniques for neural machine translation, arXiv preprint arXiv:2002.07526 (2020).

[16] M. Luong, E. Brevdo, R. Zhao, Neural machine translation (seq2seq) tutorial, https://github.com/tensorflow/nmt (2017).

[17] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, arXiv preprint arXiv:2007.01852 (2020).

[18] Y. Yang, G. Hernandez Abrego, S. Yuan, M. Guo, Q. Shen, D. Cer, Y.-h. Sung, B. Strope, R. Kurzweil, Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 5370–5378. URL: https://doi.org/10.24963/ijcai.2019/746. doi:10.24963/ijcai.2019/746.