NITK-IT_NLP@TamilNLP-ACL2022: Transformer based model for Offensive Span Identification in Tamil

Hariharan RamakrishnaIyer LekshmiAmmal[1], **Manikandan Ravikiran**[2],
Anand Kumar Madasamy[1]
[1]Department of Information Technology,
National Institute of Technology Karnataka, Surathkal
[2]Georgia Institute of technology, Atlanta, Georgia
hariharanrl.197it003@nitk.edu.in, mravikiran3@gatech.edu
m_anandkumar@nitk.edu.in

Abstract

Offensive Span identification in Tamil is a shared task that focuses on identifying harmful content, contributing to offensiveness. In this work, we have built a model that can efficiently identify the span of text contributing to offensive content. We have used various transformer-based models to develop the system, out of which the fine-tuned MuRIL model was able to achieve the best overall character F1-score of 0.4489.

1 Introduction

As far as social media and entities involved in content moderation are concerned, identifying offensive content is critical. However, most of these companies employ content moderators for determining and mitigating offensive content, but they are frequently swamped by their volume (Arsht and Etcovitch, 2018). Small firms cannot utilize human moderators because of the cost, and hence they turn off their comment sections fully.

Code-mixing is the mixing of various linguistic units from two or more languages in a conversation or even in a single utterance. When the Indian perspective is considered, English is primarily influenced by all Indian languages, including Dravidian languages like Tamil, Malayalam, and Kannada (Chakravarthi et al., 2020). Hence this has become a part of different conversations in social media. Many recent works address the whole comment classification as offensive or not but do not consider the span of text that makes it offensive. Identifying this span of text will further help moderators who deal with these contents.

Offensive Span identification is a shared task organised as a part of DravidianLangTech @ACL-2022[1]. They had two subtasks, Supervised Offensive Span Identification and Semi-Supervised Offensive Span Identification, where we were given annotated as well as non-annotated data. The task was to identify the offensive span of text content.

In this paper, we have used multilingual transformer-based models and Local Interpretable Model Agnostic Explanations (LIME) (Ribeiro et al., 2016) to identify the span of text. The paper is presented as follows; Section 2 explains the Dataset, Section 3 is about the Methodology used, Section 4 explains the Experiments and Results, which follows the Conclusions and Future Scope.

2 Related Works

Offensive language identification is one of the widely explored problems. Most of the work on offensive language identification tasks is of classification type rather than identifying the span of texts. Recent works (Kedia and Nandy, 2021; Sharif et al., 2021; Jayanthi and Gupta, 2021) have explored various transformer-based models and some (Saha et al., 2021; Zhao and Tao, 2021) have made an ensemble of different ones which are focused on classification task. Offensive Span identification is in its developing stage, (Pavlopoulos et al., 2021) was the first to introduced a shared task and Offensive Span dataset.

3 Dataset Description

Two subtasks were given on the codalab competition website[2] for Offensive Span identification in Tamil, namely Supervised Offensive Span Identification and Semi-Supervised Offensive Span Identification. The Dataset (Ravikiran and Annamalai, 2021; Ravikiran et al., 2022) had training and testing sets for both tasks, which are retrieved from YouTube, whose details are given in Table 1, which contained Code-mixed Tamil comments. The supervised task had annotated data for spans (some entire comments are annotated for full spans);

[1] https://dravidianlangtech.github.io/2022/

[2] https://competitions.codalab.org/competitions/36395

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 75 - 78
May 26, 2022 ©2022 Association for Computational Linguistics

along with that, we had partial annotated data. The test data had 876 comments for prediction. We have used the HASOC-2021 (Chakravarthi et al., 2021) shared task dataset for training, which had 4000 comments with an equal number of offensive and not-offensive labels.

Task	Comments
Supervised	4816

Table 1: Dataset Details

4 Methodology

We had two subtasks as a part of this shared task on Offensive Span Identification. The first task was to use a supervised method to identify offensive span of text in the data, and the second task was to use a semi-supervised method to do the same. We used the supervised method, which uses the transformer-based model to train the data for classifying offensive content. This trained model is used to predict whether the given comment is offensive or not. On top of this, we further examine the results on each class individually using input perturbation-based explanation method involving Local Interpretable Model Agnostic Explanations (LIME) (Ribeiro et al., 2016).

Here we are training the model for offensive content identification from comment text. We have used data from the HASOC-2021 shared task[3] (Chakravarthi et al., 2021), which contains comments extracted from YouTube annotated with labels 'offensive' and 'not-offensive.' The comments are Preprocessed, Tokenized, and fed to the pre-trained model and further fine-tuned to make predictions.

Hyperparameter	Model		
	M-BERT	ELECTRA	MuRIL
Learning rate	3e-5	3e-5	3e-5
Batch_size	32	32	32
Optimizer	ADAM	ADAM	ADAM
Epochs	10	10	10
Sequence length	160	160	160

Table 2: Hyperparameter Values

4.1 Text Preprocessing and Tokenization

We have used code-mixed data from HASOC-2021 for training our model. The data given was retrieved from YouTube comments which we cleaned using cleantext[4] library from python for removing unknown characters and ASCII conversions.

We use the tokenization[5] method, which corresponds to the pre-trained models[6] and expects tokens to be in some explicit format.

4.2 Model Description

We have used three pre-trained models named Multilingual BERT (M-BERT) (Kenton et al., 2018), MuRIL (Khanuja et al., 2021) and Electra (Clark et al., 2020) from Google. Among these, the M-BERT and MuRIL were trained on multilingual data. MuRIL was specially trained for the Indian context, with multilingual representations for Indian languages, and they have explicitly augmented monolingual text with translated and transliterated document pairs for training. As shown in Table 2 we used the recommended hyperparameters for all the models.

5 Experiments and Result

We have fine-tuned the pre-trained models for the HASOC 2021 data using the hyperparameters mentioned in Table 2, we have used Adam (Kingma and Ba, 2014) as the optimizer. The experiments were performed on Tesla P100 16GB GPU provided by Kaggle.

Model	Overall F1-Score	F1@30[a]	F1@50[b]	F1@100[c]
MuRIL	0.4489[1]	0.3726	0.2844	0.2968
DLRG-Run1	0.1727	0.3890	0.2522	0.1628

[a] This is character level F1 score calculated for sentences with less than 30 characters

[b] This is character level F1 score calculated for sentences with less than 50 characters but greater than 30 characters

[c] This is character level F1 score calculated for sentences with less than 50 characters but greater than 100 characters

Table 3: Final Results

We initially trained our model with M-BERT for the code-mixed offensive data. This model is used to predict the test data given for identifying offensive/harmful content. The final prediction is interpreted using LIME, which will give a score for each of the words contributing to the offensive and non-offensive contents. Those words contributing to the offensive texts are extracted and predicted for

[3] https://competitions.codalab.org/competitions/31146

[4] https://pypi.org/project/clean-text/

[5] https://huggingface.co/docs/tokenizers/python/latest

[6] http://huggingface.co/models

the given task and its span in the whole comment. We employed a similar procedure for the Electra and MuRIL model; because of the better representation for the Indian context, MuRIL was able to give a more reasonable prediction and we got first position for the same which is given in Table 3. Hence, it got the best score among the others. The Table 3 gives the final score of our MuRIL model and the next best score from participants, and we have not included the scores from M-BERT and Electra as they were not released. The figure 1,2 shows examples of LIME interpretations for given comment text along with contributing words for offensive and not-offensive.

Example Comment text [offensive] :

@USER Ungotha ku rate ellam illa ,free eh othukittu irukan da ungotha thevidiya mundaiyae.

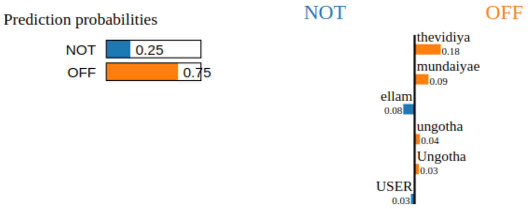

Figure 1: Example of LIME Interpretation for Offensive Class

Example Comment text [Not-offensive]:

*Ellarum Saptacha ..??? *** : Yen Sapadu Vangi kuduka Poriya ?? Unaku RT dhana Venum .... Straight ah Kelu !! Ama..!! Tag #TAG.*

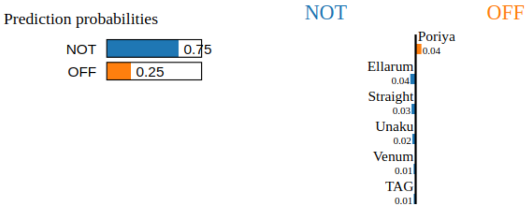

Figure 2: Example of LIME Interpretation for Not-Offensive Class

6 Conclusions and Future Scope

Social media is the primary source from which people use to get information. Hence, the companies that handle these need to moderate content so that the offensive and harmful content are not propagated. In this paper, we have explored the transformer-based model along with LIME interpretations to identify the span of harmful content in comments. Google's MuRIL achieved the best result from different models, which came first in the leaderboard for the shared task. In the future, we would like to explore more on the Code-mixed data and develop improved solutions to this problem.

References

Andrew Arsht and Daniel Etcovitch. 2018. The human cost of online content moderation.

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Ratnasingam Sakuntharaj, Anand Kumar Madasamy, Sajeetha Thavareesan, Premjith B, Subalalitha Chinnaudayar Navaneethakrishnan, John P. McCrae, and Thomas Mandl. 2021. 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.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. 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.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators.

Sai Muralidhar Jayanthi and Akshat Gupta. 2021. SJ_AJ@DravidianLangTech-EACL2021: Task-adaptive pre-training of multilingual BERT models for offensive language identification. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 307–312, Kyiv. Association for Computational Linguistics.

Kushal Kedia and Abhilash Nandy. 2021. indicnlp@kgp at DravidianLangTech-EACL2021: Offensive language identification in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 330–335, Kyiv. Association for Computational Linguistics.

Ming-wei Chang Kenton, Lee Kristina, and Jacob Devlin. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (Mlm).

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual Representations for Indian Languages.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Manikandan Ravikiran and Subbiah Annamalai. 2021. DOSA: Dravidian code-mixed offensive span identification dataset. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 10–17, Kyiv. Association for Computational Linguistics.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha Sivanesan, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022. Findings of the shared task on Toxic Span Identification in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Debjoy Saha, Naman Paharia, Debajit Chakraborty, Punyajoy Saha, and Animesh Mukherjee. 2021. Hate-alert@DravidianLangTech-EACL2021: Ensembling strategies for transformer-based offensive language detection. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 270–276, Kyiv. Association for Computational Linguistics.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. NLP-CUET@DravidianLangTech-EACL2021: Offensive language detection from multilingual code-mixed text using transformers. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 255–261, Kyiv. Association for Computational Linguistics.

Yingjia Zhao and Xin Tao. 2021. ZYJ123@DravidianLangTech-EACL2021: Offensive language identification based on XLM-RoBERTa with DPCNN. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 216–221, Kyiv. Association for Computational Linguistics.