DLRG@TamilNLP-ACL2022: Offensive Span Identification in Tamil using BiLSTM-CRF approach

Ratnavel Rajalakshmi [*], **Mohit Madhukar More, Bhamatipati Naga Shrikriti, Gitansh Saharan, Samyuktha Hanchate, Sayantan Nandy**
Vellore Institute of Technology, Chennai, India
rajalakshmi.r@vit.ac.in, mohitmadhukar.more2019@vitstudent.ac.in, shrikriti4@gmail.com, gitansh18saharan@gmail.com, samyukthahanchate@gmail.com, sayantann11@gmail.com

Abstract

Identifying offensive speech is an exciting and essential area of research, with ample traction in recent times. This paper presents our system submission to the subtask 1, focusing on using supervised approaches for extracting Offensive spans from code-mixed Tamil-English comments. To identify offensive spans, we developed the Bidirectional Long Short-Term Memory (BiLSTM) model with Glove Embedding. With this method, the developed system achieved an overall F1 of 0.1728. Additionally, for comments with less than 30 characters, the developed system shows an F1 of 0.3890, competitive with other submissions.

1 Introduction

Offensive speech, in general, is defined as the speech that causes an individual/group to feel displeased, upset, angry, or annoyed (Pavlopoulos et al., 2019). Often offensive speech is intended to vilify, humiliate, or incite hatred against a group or a class of persons based on race, religion, skin color, sexual identity, gender identity, ethnicity, disability, or national origin (Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021). Predominantly with social media outreach, this is more prevalent. Accordingly, pinpointing such offensive speech is vital to encourage healthy conversation across users. Moreover, such systems are essential in automatic content moderation, with minimal human involvement (Priyadharshini et al., 2021; Kumaresan et al., 2021).

Code-Mixing is yet another social media phenomenon that has crept into daily speech across all languages, including Tamil (B and A, 2021b,a). Often, we see the usage of more than one language like Tamil-English, Kannada-English, etc., which adds a layer of complexity in identifying offensive contents (Ghanghor et al., 2021a,b; Yasaswini et al., 2021). Code-mixing and Code-borrowing have become common among the multi-lingual people (Rajalakshmi and Agrawal, 2017). Even though offensive content classification on Code-mixed language has been studied by few researchers by applying machine learning (Ratnavel Rajalakshmi, 2020) and deep learning algorithms (Rajalakshmi et al., 2021), the span identification of offensive contents are not explored much. Dictionary learning approaches were proposed for short text classification and URL based classification applying machine learning techniques (R. and Aravindan, 2018; Rajalakshmi, 2014) , but the research work in Tamil is limited.

Tamil is a member of the southern branch of the Dravidian languages, a group of about 26 languages indigenous to the Indian subcontinent (Anita and Subalalitha, 2019b,a; Subalalitha and Poovammal, 2018; Subalalitha, 2019). It is also classed as a member of the Tamil language family, which contains the languages of around 35 ethno-linguistic groups, including the Irula and Yerukula languages (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019, 2020a,b, 2021). Malayalam is Tamil's closest significant cousin; the two began splitting during the 9th century AD. Although several variations between Tamil and Malayalam indicate a pre-historic break of the western dialect, the process of separating into a different language, Malayalam, did not occur until the 13th or 14th century.

This work, the shared task on offensive span identification handles the code-mixed Tamil-English comments and focuses on identification of character offsets of the offensive parts (**?**Ravikiran et al., 2022; Chakravarthi et al., 2022; Bharathi et al., 2022; Priyadharshini et al., 2022). There are multiple approaches for extracting spans. In this work, we treat the task of removing offensive span as an approach to token labeling. In this regard, we

[*]Corresponding Author

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 248 - 253
May 26, 2022

evaluated Bi-LSTM + CRF-based token labeling system for extracting offensive spans.

The rest of the paper is organized as follows. First, section 2 briefly discusses the literature on offensive span identification-related works. Then, in section 3, our system is described in detail, followed by Section 4, in which the experiments and results are presented. Finally, we conclude with possible implications for future work.

2 Related works

Offensive span can be solved in multiple ways ranging from token labeling to extracting spans using interpretability approaches. Unfortunately, the overall work is still developing for English and code-mixed languages, with very few well-established data sets and methods. (Pavlopoulos et al., 2021; Ravikiran and Annamalai, 2021). Interesting works related to offensive spans include Zhu et al. (2021) that employs token labeling using language models with a mixture of Conditional Random Fields (CRF). Usually, token labeling systems use BIO encoding of the text corresponding to offensive spans. Lexicon-based models (Burtenshaw and Kestemont, 2021) and statistical analysis (Palomino et al., 2021) are also widely explored. Finally, a few strategies utilize custom loss functions tailored explicitly for managing wrong spans. For code-mixed Tamil-English to date, we find there is only by Ravikiran and Annamalai (2021) that again uses token level labeling with language models.

3 Problem and System Description

An example of offensive span identification is shown in Figure 1. Given the input sentence, the task is to extract the range of spans corresponding to offensive content. In the above example, the word `Poramboku` contributes to offensiveness which corresponds to character offset of 47-56. A dataset with offensive span annotations details was released as part of the shared task on Toxic Span identification (Ravikiran et al., 2022). The description of this dataset is presented in Section 3.1.

3.1 Dataset Description

The released shared task dataset consists of two files with span annotations. The training dataset having 4816 samples with offensive spans and testing dataset with 876 samples without annotation. Additionally, the organizers released a stripped down version of train set which consists of span annotations for one or more words, but not the entire sentence. This was used for validation and hyper-parameter tuning.

3.2 Development Pipeline

The overall development pipeline used in this work is depicted in Figure 2. Our pipeline could be broken into three modules namely (a) Pre-processing Module (b) Encoding Module and (c) Bi-LSTM module respectively. Each of which is as described.

3.2.1 Preprocessing Module

In the preprocessing module, we extracted all the offensive parts of the comments from the given dataset and created individual parts it into list of tokens. These tokens are then converted to sequences using Tweet Tokenizer that is available as part of the nltk pipeline. Additionally, all the converted tokens are BIO encoded.

3.2.2 Encoding Module

In the encoding stage we use glove embedding pretrained on twitter data as initializer. We based this approach on the Vector Initialization (VI) alignment method, where after training embedding for one feature space, using it on related domain data will improve existing word embedding catering two new domain of data (code-mixed). We downloaded the Glove embedding which has 400K vocabulary size and each word corresponds to a 100-dimensional embedding vector. To use this embedding, we simply replace the one hot encoding word representation with its corresponding 100-dimensional vector.

3.2.3 Bi-LSTM Module

We follow Bi-LSTM + CRF architecture of Huang et al. (2015). The details of architecture is as shown in Figure 3 and consists of the following components.

- Input layer that accepts the input comments from which the span is to be identified.
- Embedding layer uses Glove embedding to create vectors suitable for training Bi-LSTM.
- The Bi-LSTM layer is more efficient in using the past features (via forward states) and future features (via backward states) for a specific time frame.
- CRF layer, that connects inputs to tags directly in turn identifying the offensive parts of the contents.

Example:
Input Sentence:
2017 Speech Super Star. 1996 endru solavakiran Poramboku Director
Output Spans:
"[47, 48, 49, 50, 51, 52, 53, 54, 55, 56]"

Figure 1: Example of offensive span identification used in the shared task.

Parameter	Value
Dropout	0.1
Recurrent Dropout	0.1
Max Sequence Length	128
Activation	ReLU

Table 1: Hyper-parameters

	F1	F1@30	F1@50	F1@>50
Bi-LSTM + CRF (Ours)	0.1728	0.3890	0.2523	0.1608
Random Baseline (Ravikiran et al., 2022)	0.3975	-	-	-

Table 2: Results obtained by our BiLSTM-CRF method

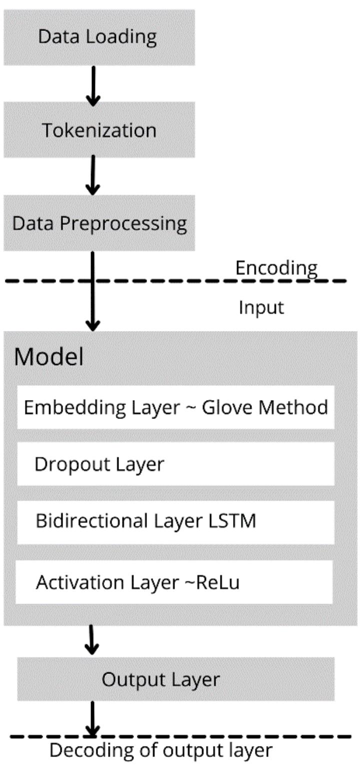

Figure 2: Overall pipeline used in this work

```
Layer (type)                 Output Shape              Param #
================================================================
input_1 (InputLayer)         (None, 128)               0
________________________________________________________________
embedding_1 (Embedding)      (None, 128, 100)          119351400
________________________________________________________________
dropout_1 (Dropout)          (None, 128, 100)          0
________________________________________________________________
bidirectional_1 (Bidirection (None, 128, 256)          234496
________________________________________________________________
time_distributed_1 (TimeDist (None, 128, 128)          32896
________________________________________________________________
crf_1 (CRF)                  (None, 128, 2)            266
================================================================
Total params: 119,619,058
Trainable params: 119,619,058
Non-trainable params: 0
```

Figure 3: Overall architecture of Bi-LSTM + CRF used in this work.

Finally the spans corresponding to words mapped as offensive are extracted. The hyper-parameters details are presented in Table 1.

4 Experiments and Results

We have conducted various experiments to study the performance of the model and submitted the best performing version of our model. The results obtained are as shown in Table 2. We can see that our model obtained an F_1 score of 0.1728 which is significantly lower than random baselines used by the organizers. To analyse the performance, we briefly studied the effects of our system on various sizes of text. We found that our model performs well for shorter comments sequences with an F_1 of 0.3890. We believe that, this may be because of lack of LSTM's ability to exploit long range sequences, especially with only one single layer. Accordingly, we plan to revisit this problem with deeper architectures and language models.

5 Conclusion

Offensive Span Identification is still a challenging task with multiple challenges including the need of learning less data and long range contexts. In this work, we studied Bi-LSTM + CRF model to predict offensive spans from code-mixed Tamil-English comments. Accordingly our system obtained the overall F_1 of 0.1728 which is significantly lower. However we found that the developed method is suitable for shorter sequences where we can see higher results. In the future we plan to revisit the architecture in detail with a study on effect of embeddings types, number of layers and advanced architectures.

Acknowledgment

We would like to thank Manikandan Ravikiran and the organizers for all the guidance on the shared tasks. We would also like to thank the management of Vellore Institute of Technology, Chennai, where this research work is carried out.

References

R Anita and CN Subalalitha. 2019a. An approach to cluster Tamil literatures using discourse connectives. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–4. IEEE.

R Anita and CN Subalalitha. 2019b. Building discourse parser for Thirukkural. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 18–25.

Bharathi B and Agnusimmaculate Silvia A. 2021a. SSNCSE_NLP@DravidianLangTech-EACL2021: Meme classification for Tamil using machine learning approach. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 336–339, Kyiv. Association for Computational Linguistics.

Bharathi B and Agnusimmaculate Silvia A. 2021b. 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*, pages 313–318, Kyiv. Association for Computational Linguistics.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, N Sripriya, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Ben Burtenshaw and Mike Kestemont. 2021. UAntwerp at SemEval-2021 task 5: Spans are spans, stacking a binary word level approach to toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 898–903, Online. Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2020. 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 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*, pages 61–72, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Phillip McCrae, Paul Buitaleer, Prasanna Kumar Kumaresan, and Rahul Ponnusamy. 2022. Findings of the shared task on Homophobia Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Nikhil 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*, pages 222–229, Kyiv. Association for Computational Linguistics.

Nikhil 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*, pages 197–203, Kyiv. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Prasanna Kumar Kumaresan, Ratnasingam Sakuntharaj, Sajeetha Thavareesan, Subalalitha Navaneethakrishnan, Anand Kumar Madasamy, Bharathi Raja Chakravarthi, and John P McCrae. 2021. Findings of shared task on offensive language identification in Tamil and Malayalam. In *Forum for Information Retrieval Evaluation*, pages 16–18.

Marco Palomino, Dawid Grad, and James Bedwell. 2021. GoldenWind at SemEval-2021 task 5: Orthrus - an ensemble approach to identify toxicity. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 860–864, Online. Association for Computational Linguistics.

John Pavlopoulos, Léo Laugier, Jeffrey Sorensen, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection (to appear). In *Proceedings of the 15th International Workshop on Semantic Evaluation*.

John Pavlopoulos, Nithum Thain, Lucas Dixon, and Ion Androutsopoulos. 2019. ConvAI at SemEval-2019 task 6: Offensive language identification and categorization with perspective and BERT. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 571–576, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on Abusive Comment Detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, and Rahul Ponnusamy. 2021. Overview of the DravidianCodeMix 2021 shared task on sentiment detection in Tamil, Malayalam, and Kannada. In *Forum for Information Retrieval Evaluation*, pages 4–6.

Rajalakshmi R. and Chandrabose Aravindan. 2018. An effective and discriminative feature learning for url based web page classification. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1374–1379.

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.

Ratnavel Rajalakshmi. 2014. Supervised term weighting methods for url classification. *J. Comput. Sci.*, 10:1969–1976.

Ratnavel Rajalakshmi, Yashwant Reddy, and Lokesh Kumar. 2021. DLRG@DravidianLangTech-EACL2021: Transformer based approach for offensive language identification on code-mixed Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 357–362, Kyiv. Association for Computational Linguistics.

Yashwanth Reddy B Ratnavel Rajalakshmi. 2020. DLRG@HASOC 2020: A hybrid approach for hate and offensive content identification in multilingual tweets. In *"Proceedings of FIRE '20, Forum for Information Retrieval Evaluation"*, pages 1–7.

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 Offensive Span Identification in code-mixed Tamil-English comments. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2016. A novel hybrid approach to detect and correct spelling in Tamil text. In *2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 1–6.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2017. Use of a novel hash-table for speeding-up suggestions for misspelt tamil words. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–5.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2021. Missing word detection and correction based on context of Tamil sentences using n-grams. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 42–47.

C. N. Subalalitha. 2019. Information extraction framework for Kurunthogai. *Sādhanā*, 44(7):156.

CN Subalalitha and E Poovammal. 2018. Automatic bilingual dictionary construction for Tirukural. *Applied Artificial Intelligence*, 32(6):558–567.

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.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2021. Sentiment analysis in Tamil texts using k-means and k-nearest neighbour. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 48–53.

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*, pages 187–194, Kyiv. Association for Computational Linguistics.

Qinglin Zhu, Zijie Lin, Yice Zhang, Jingyi Sun, Xiang Li, Qihui Lin, Yixue Dang, and Ruifeng Xu. 2021. HITSZ-HLT at SemEval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 521–526, Online. Association for Computational Linguistics.