Findings of the Second Shared Task on Offensive Span Identification from Code-Mixed Tamil-English Comments

Manikandan Ravikiran[†*], Ananth Ganesh, Anand Kumar Madasamy[⋆]
Ratnavel Rajalakshmi [⊕], Bharathi Raja Chakravarthi[‡]
[†]Georgia Institute of Technology, Atlanta, Georgia
[⋆]National Institute of Technology Karnataka Surathkal, India
[⊕]Vellore Institute of Technology, Chennai, India
[‡]School of Computer Science, Univeristy of Galway, Ireland
mravikiran3@gatech.edu, bharathi.raja@insight-centre.org

Abstract

Maintaining effective control over offensive content is essential on social media platforms to foster constructive online discussions. Yet, when it comes to code-mixed Dravidian languages, the current prevalence of offensive content moderation is restricted to categorizing entire comments, failing to identify specific portions that contribute to the offensiveness. Such limitation is primarily due to the lack of annotated data and open source systems for offensive spans. To alleviate this issue, in this shared task, we offer a collection of Tamil-English code-mixed social comments that include offensive comments. This paper provides an overview of the released dataset, the algorithms employed, and the outcomes achieved by the systems submitted for this task.

1 Introduction

Combating offensive content is crucial for different entities involved in content moderation, which includes social media companies as well as individuals (Subramanian et al., 2022; Chinnaudayar Navaneethakrishnan et al., 2023). To this end, moderation is often restrictive with either usage of human content moderators, who are expected to read through the content and flag the offensive mentions (Arsht and Etcovitch, 2018). Alternatively, there are semi-automated and automated tools that employ trivial algorithms and block lists (Jhaver et al., 2018). Though content moderation looks like a one-way street, where either it should be allowed or removed, such decision-making is fairly hard (Bharathi and Agnusimmaculate Silvia, 2021; Bharathi and Varsha, 2022; Swaminathan et al., 2022). This is more significant, especially on social media platforms, where the sheer volume of content is overwhelming for human moderators especially (Kumaresan et al., 2022; Chakravarthi, 2022b,a). With ever increasing offensive social media contents focusing on offensive comments and statements semi-automated and fully automated content moderation is favored (Ravikiran et al., 2022; Chakravarthi, 2023; Chakravarthi et al., 2023a).

Tamil is an classical ancient language (Subalalitha, 2019a; Anita and Subalalitha, 2019a; Thavareesan and Mahesan, 2019, 2020a,b) with a history dating back to 580 BCE (Sivanantham and Seran, 2019). It is primarily spoken in Tamil Nadu, India, and also in Sri Lanka, Malaysia, and Singapore. Tamil holds official language status in Tamil Nadu, Sri Lanka, Singapore, and the Union Territory of Puducherry (Subalalitha, 2019b; Sakuntharaj and Mahesan, 2016, 2017, 2021). Additionally, there are significant Tamil-speaking communities in Kerala, Karnataka, Andhra Pradesh, Telangana, and the Andaman and Nicobar Islands. The Tamil diaspora is spread across countries across the world and is recognized as a scheduled language in the Indian Constitution. It has a rich literary tradition dating back to the 6th century BCE (Anita and Subalalitha, 2019b), with rock edicts and "hero stones" serving as some of the earliest known written records. De-

[*] Corresponding Author, Work done during graduate school

Third Workshop on Speech and Language Technologies for Dravidian Languages, pages 52–58, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_007

spite its own script, with the advent of social media, code-switching has permeated into the Tamil language across informal contexts like forums and messaging outlets (Ravikiran et al., 2022). As a result, code-switched content is part and parcel of offensive conversations in social media.

Despite many recent NLP advancements, handling code-mixed offensive content is still a challenge in Dravidian Languages (Sitaram et al., 2019) including Tamil owing to limitations in data and tools. However, recently the research of offensive code-mixed texts in Dravidian languages has seen traction (Priyadharshini et al., 2020; Chakravarthi, 2020; Chakravarthi et al., 2023a,b). Yet, very few of these focus on identifying the spans that make a comment offensive (Ravikiran and Annamalai, 2021; Ravikiran et al., 2022). However, highlighting these specific spans segments can greatly assist content moderators and semi-automated tools that prioritize identifying and attributing offensive content. In line with this objective, we presented second iteration of code-mixed social media text in Tamil, including offensive spans, and invited participants to develop and submit systems under two distinct settings for this shared task. Our CodaLab website[1] will remain open to foster further research in this area.

2 Related Work

2.1 Offensive Span Identification

Existing literature on identifying offensive spans primarily finds its origins in SemEval Offensive Span Identification shared task, which predominantly centers around the English language (Pavlopoulos et al., 2021). More than 36 different systems have been developed using various approaches. For Dravidian Languages there are quite few works namely Ravikiran and Annamalai (2021); LekshmiAmmal et al. (2022); Rajalakshmi et al. (2022).

3 Task Description

Our task of offensive span identification required participants to identify offensive spans i.e, character offsets that were responsible for the offensive of the comments, when identifying such spans was possible. To this end, we created two subtasks each of which are as described.

[1] https://codalab.lisn.upsaclay.fr/competitions/11174

3.1 Subtask 1: Supervised Offensive Span Identification

Given comments and annotated offensive spans for training, here the systems were asked to identify the offensive spans in each of the comments in test data. This task could be approached as supervised sequence labeling, training on the provided posts with gold offensive spans. It could also be treated as rationale extraction using classifiers trained on other datasets of posts manually annotated for offensiveness classification, without any span annotations.

3.2 Subtask 2: Less data Offensive Span Identification

All the participants of subtask 1 are encouraged to also submit a *Less Data approach*, where the participants are expected to submit a model while using only parts (not fully) of training data of subtask 1. Participants were asked to develop systems to achieve competitive performance with limited data. To this end, participants were empowered to use creative ways to do this including data subset selection, coreset theory etc.

4 Dataset

For this shared task, we build upon dataset from earlier work of Ravikiran et al. (2022), which originally released 4786 code-mixed Tamil-English comments with 6202 offensive spans. We released this dataset to the participants during training phase for model development. Additionally, the test set from the same work with 1006 samples were released for development/validation purposes. Meanwhile for testing we extended this dataset with new additional annotated comments. To this end, we use dataset of Priyadharshini et al. (2022) that consist of abusive comments. From this we selected 366 comments for testing purpose.

Split	Train	Test
Number of Sentences	4786	361
Number of unique tokens	22096	2947
Number of annotated spans	6202	677
Average size of spans (# of characters)	21	21
Min size of spans (# of characters)	4	4
Max size of spans (# of characters)	82	58

Table 1: Dataset Statistics used in this shared task

Following previous research (Ravikiran et al., 2022), we created span-level annotations for 361 newly selected test comments. We followed the same process and guidelines for annotation, anonymity maintenance etc. Profanity in data was explained apriori with an option to withdraw from the annotation process if necessary. To ensure quality each annotation was verified by one or more annotation verifier, prior to merging and creating gold standard test set. The overall dataset statistics is given in the Table 1. Overall for the 361 comments we obtained Cohen's Kappa inter-annotator agreement of 0.64 inline with Ravikiran et al. (2022).

5 Competition Phases

5.1 Training Phase

In the training phase, the train split with 4786 comments, and their annotated spans were released for model development. Participants were given training data and offensive spans. Along with this development/validation set was also released. Participants were also emphasized on cross-validation by creating their splits for preliminary evaluations or hyperparameter tuning. In total, 48 participants registered for the task and downloaded the dataset.

5.2 Testing Phase

Test set comments without any span annotation were released in the testing phase. Each participating team was asked to submit their generated span predictions for evaluation. Predictions are submitted via Google form, which was used to evaluate the systems. Though CodaLab supports evaluation inherently, we used google form due to its simplicity. Finally, we assessed the submitted spans of the test set and were scored using character-based F1 (See section 7.2).

6 System Descriptions

Overall we received only a total of 3 submissions from three teams out of 48 registered participants. All these were only for subtask 1. No submissions were made for subtask 2. Each of their respective systems are as described.

6.1 The AJNS Submission

The best performing system from AJNS experimented with rationale extraction (Atharva et al., 2023) by training offensive language classifiers and employing model-agnostic rationale extraction mechanisms to produce toxic spans as explanations of the decisions of the classifier. Specifically to achieve accurate classification, it employed the Bidirectional and Autoregressive Transformers model (Lewis et al., 2020), which is based on zero-shot learning and effectively captures the semantic meaning and context of the input text. BART's ability to generalize from limited labeled data allows for higher accuracy despite using less data compared to traditional models. This initial classification step helps us narrow down the focus to offensive spans within the text. Once the offensive spans are identified, we further process them using the Bidirectional Encoder Representations from Transformers (Devlin et al., 2019) in conjunction with the Local Interpretable Model-Agnostic Explanations (Ribeiro et al., 2016). The BERT+LIME model extracts specific span words and their positions within the parent sentence. They obtain F1 score of 0.2858

6.2 The DLRG-R1 and DLRG-R2 submission

The DLRG team formulated the problem as a combination of token labeling and span extraction. Specifically, the team created word-level BIO tags i.e., words were labelled as B (beginning word of a offensive span), I (inside word of a offensive span), or O (outside of any offensive span). Following which character level embeddings is created and an LSTM model (DLRG-R1) is trained. This system produces F1 of 0.2254. The DLRG-R2 employed similar strategy like DLRG-R2 team except they used GRU instead of LSTM. This system produces F1 of 0.2134.

6.3 The DLRG-R2 submission

7 Evaluation

This section focuses on the evaluation framework of the task. First, the official measure that was used to evaluate the participating systems is described. Then, we discuss baseline models that were selected as benchmarks for comparison reasons. Finally, the results are presented.

7.1 Evaluation Measure

In line with work of Pavlopoulos et al. (2021) each system was evaluated F1 score computed on character offset. For each system, we computed the F1 score per comments, between the predicted and the ground truth character offsets. Following this we calculated macro-average score over all the 876 test comments. If in case both ground truth and predicted character offsets were empty we assigned a F1 of 1 other wise 0 and vice versa.

7.2 Benchmark

To establish fair comparison we first created following baseline benchmark system which are as described.

- `BENCHMARK` is a random baseline model which randomly labels 50% of characters in comments to belong to be offensive. To this end, we run this benchmark 10 times and average results are presented in Table 2.

Table 2: Official rank and F1 score (%) of the 3 participating teams that submitted systems. The baselines benchmarks are also shown.

RANK	TEAM	F1 (%)
BASELINE	BENCHMARK	37.24
1	AJNS	28.58
2	DLRG-R1	22.54
2	DLRG-R2	21.34

8 Analysis, Discussion and Remarks

In general, we were pleased to witness the level of engagement in this shared task, with numerous participants signing up, expressing interest in obtaining datasets, and seeking potential baseline codes for the project. Although only three teams ultimately submitted their systems, the variety of approaches taken to tackle the problem is quite promising. Nevertheless, we have included some of our observations below, which stem from our evaluation and the insights gained from the results.

Table 2 shows the scores and ranks of two teams that made their submission. NITK-IT_NLP (Section 6.1) was ranked first, followed by DLRG (Section 6.3) that scored 27% lower was ranked second. The median score was 31.08%, which is far below the top ranked team and the benchmark baseline models.

`BENCHMARK 1` achieves a considerably high score and, hence, is very highly ranked with character F1 of 37.24%. Combination of BART with LIME interpretability by model AJNS is behind `BENCHMARK 1` by 9%, indicating the language models ability to not so effectively rationalize and identify the spans. Meanwhile DLRG-R1 and DLRG-R2 has large gap compared to random baselines, indicating the proposed approaches by these teams are not suitable for practical use. To this end, these methods employ direct token labeling which is more surprising.

Table 3: Results of submitted systems across comments of different lengths.

	F1@30 (%)	F1@50 (%)	F1@>50 (%)
AJNS	41.01	41.61	22.48
DLRG-R1	38.03	34.39	17.03
DLRG-R2	31.03	30.33	16.06

8.1 General remarks on the approaches

Though neither of teams that made final submissions created any simple baselines, we could see that all the submissions use well established approaches in recent NLP focusing on pretrained language models. Meanwhile DLRG used well-grounded Non-Transformer based approach. Yet neither of teams used any ensembles, data augmentation strategies or modifications to loss functions that are seen for the task of span identification in the past across shared tasks.

8.2 Error Analysis

Table 2 shows maximum result of 37.24% for baseline model with AJNS showing highest result of 28.58% with DLRG failing significantly compared to random baseline. To this end, we wonder if potentially these approaches have any weaknesses or strengths. To understand this, first we study the character F1 results across sentences of different lengths. Specifically we analysis results of (a) comments with less than 30 characters (F1@30) (b) comments with 30-50 characters (F1@50) (c) comments with more than 50 characters (F1@>50). The results so obtained are as shown in Table 3.

Firstly from Table 3 we can see though AJNS shows high results overall for cases of comments with larger lengths the model fails significantly by 19%. Meanwhile for DLRG-R1 and DLRG-R2 the results are more mixed, especially we can see that for comments with less than 30 characters the model shows improvement in F1 by around 10%. Meanwhile for shorter comments, the results high indicating the methods are indeed useful. However these short sentences often contained only cuss words or clearly abusive words that are easily identifiable and often present in the train set, indicating the deficiency of the submitted systems.

9 Conclusion

In this work, we set up a second shared task that was centred on locating offensive language spans in code-mixed Tamil-English text. Compared to our earlier iteration, we had 6,153 social media comments that were tagged to identify abusive spans. Only three teams submitted their systems out of 48 registered participants. We described their strategies in this study and talked about the results they got. It's interesting that a strategy for reason extraction that combines BART and LIME was effective but was not able to beat random baseline. The LSTM/GRU model, on the other hand, performed noticeably worse than the random baseline and showed sensitivity to shorter sentences. We have made the baseline models and information available to the public in order to aid future research. Moving forward, we intend to redo the offensive span identification task under multitask setup with identification of different types of offensiveness alongside the offensive spans.

Acknowledgements

We thank our anonymous reviewers for their valuable feedback. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors only and does not reflect the view of their employing organization or graduate schools. The shared task was result of series projects done during CS7646-ML4T (Fall 2020), CS6460-Edtech Foundations (Spring 2020) and CS7643-Deep learning (Spring 2022) at Georgia Institute of Technology (OMSCS Program). Bharathi Raja Chakravarthi were supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 (Insight_2).

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.

Andrew Arsht and Daniel Etcovitch. 2018. The human cost of online content moderation. *Harvard Journal of Law & Technology*.

Joshi Atharva, Subramanian Narayanan, and Rajalakshmi Ratnavel. 2023. Team_ajns@dravidianlangtech: Offensive language identification on multilingual code mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

B Bharathi and A Agnusimmaculate Silvia. 2021. 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 and Josephine Varsha. 2022. SSNCSE NLP@TamilNLP-ACL2022: Transformer based approach for detection of abusive comment for Tamil language. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 158–164, Dublin, Ireland. 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. 2022a. Hope speech detection in youtube comments. *Social Network Analysis and Mining*, 12(1):75.

Bharathi Raja Chakravarthi. 2022b. Multilingual hope speech detection in english and dravidian languages. *International Journal of Data Science and Analytics*, 14(4):389–406.

Bharathi Raja Chakravarthi. 2023. Detection of homophobia and transphobia in youtube comments. *International Journal of Data Science and Analytics*, pages 1–20.

Bharathi Raja Chakravarthi, Manoj Balaji Jagadeeshan, Vasanth Palanikumar, and Ruba Priyadharshini. 2023a. Offensive language identification in dravidian languages using mpnet and cnn. *International Journal of Information Management Data Insights*, 3(1):100151.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023b. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Subalalitha Chinnaudayar Navaneethakrishnan, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Malliga Subramanian, Prasanna Kumar Kumaresan, Bharathi, Lavanya Sambath Kumar, and Rahul Ponnusamy. 2023. Findings of shared task on sentiment analysis and homophobia detection of youtube comments in code-mixed dravidian languages. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval*

Evaluation, FIRE '22, page 18–21, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Shagun Jhaver, Sucheta Ghoshal, Amy S. Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Trans. Comput. Hum. Interact.*, 25(2):12:1–12:33.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Elizabeth Sherly, Sangeetha Sivanesan, and Bharathi Raja Chakravarthi. 2022. Transformer based hope speech comment classification in code-mixed text. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 120–137. Springer.

Hariharan LekshmiAmmal, Manikandan Ravikiran, and Anand Kumar Madasamy. 2022. NITK-IT_NLP@TamilNLP-ACL2022: Transformer based model for toxic span identification in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 75–78, Dublin, Ireland. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 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*.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

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. IEEE.

Ratnavel Rajalakshmi, Mohit More, Bhamatipati Shrikriti, Gitansh Saharan, Hanchate Samyuktha, and Sayantan Nandy. 2022. DLRG@TamilNLP-ACL2022: Offensive span identification in Tamil usingBiLSTM-CRF approach. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 248–253, Dublin, Ireland. 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 S, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022. Findings of the shared task on offensive span identification fromCode-mixed Tamil-English comments. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 261–270, Dublin, Ireland. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the Demonstrations Session, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 97–101. The 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. IEEE.

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. IEEE.

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. IEEE.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and A. Black. 2019. A survey of code-switched speech and language processing. *ArXiv*, abs/1904.00784.

R Sivanantham and M Seran. 2019. Keeladi: An urban settlement of sangam age on the banks of river vaigai. *India: Department of Archaeology, Government of Tamil Nadu, Chennai*.

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

CN Subalalitha. 2019b. Information extraction framework for kurunthogai. *Sādhanā*, 44(7):156.

Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. Offensive language detection in tamil youtube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 76:101404.

Krithika Swaminathan, Bharathi B, Gayathri G L, and Hrishik Sampath. 2022. SSNCSE_NLP@LT-EDI-ACL2022: Homophobia/transphobia detection in multiple languages using SVM classifiers and BERT-based transformers. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 239–244, Dublin, Ireland. 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. IEEE.

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. IEEE.

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. IEEE.