BPHC@DravidianLangTech-ACL2022-A comparative analysis of classical and pre-trained models for troll meme classification in Tamil

Achyuta Krishna V **Mithun Kumar S R** **Aruna Malapati** **Lov Kumar**
BITS Pilani, Hyderabad Campus
{f20180165,p20190503,arunam,lovkumar}@hyderabad.bits-pilani.ac.in

Abstract

Trolling refers to any user behavior on the internet to intentionally provoke or instigate conflict, predominantly on social media. This paper aims to classify troll meme captions in Tamil-English code-mixed form. Embeddings are obtained for raw code-mixed text, and the translated and transliterated version of the text and their relative performances are compared. Furthermore, this paper compares the performances of 11 different classification algorithms using Accuracy and F1- Score. We conclude that we were able to achieve a weighted F1 score of 0.74 through MuRIL pretrained model.

1 Introduction

Technology is ingrained in every aspect of our lives. We require it to communicate with others and thrive in the modern world. We increasingly rely on text-based mediums to interact with technology every day. Hence there is a need for machines to understand natural human languages. However, it is challenging for computers to understand natural languages because of the inherent ambiguity in both the syntax and the semantics of natural language (Priyadharshini et al., 2021; Kumaresan et al., 2021).

With the ease of accessing the internet and the surge in the number of social media platforms, social media has become an essential and influential aspect of everyone's life (Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021). It has also brought about a change in the way regional languages are expressed. Native script of the regional language is not used for exchanges on social media platforms (Chakravarthi et al., 2021). Instead, native speakers use Roman script combined with English words or phrases through code-mixing to express their ideas. The text generated by users in social media contains a high amount of spelling mistakes, phonetic typing, wordplay characters, and modern internet slang (Sampath et al., 2022; Ravikiran et al., 2022; Chakravarthi et al., 2022; Bharathi et al., 2022; Priyadharshini et al., 2022). This is mainly due to the limitation of the English keyboard and the speed at which the modern world moves. Thus, the study of text expressed in code-mixed form is essential (Priyadharshini et al., 2020).

In this paper, we explain our submission to DravidianLangTech-ACL2022 for the task of Troll-meme classification in Tamil. Tamil is a member of the southern branch of the Dravidian languages, a group of about 26 languages indigenous to the Indian subcontinent (Subalalitha, 2019; Srinivasan and Subalalitha, 2019; Narasimhan et al., 2018). The earliest Old Tamil documents are small inscriptions in Adichanallur dating from 905 BC to 696 BC (Anita and Subalalitha, 2019b,a; Subalalitha and Poovammal, 2018). Tamil, one of the 22 scheduled languages in the Indian Constitution, was the first to be designated as a classical language of India (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019, 2020a,b, 2021). We brief on all the embedding techniques like TF-IDF, m-BERT, MuRIL, IndicFT, etc., and various classification algorithms, including Logistic Regression, Decision tree, SVM, etc., that were implemented in the process. The rest of the paper is organized as follows. Section 2 details the related works done in the field, and Section 3 describes the dataset used. Section 4 expands on the methodology and experimental setup, Section 5 discusses the results obtained, and Section 6 elucidates the conclusions.

2 Related work

There has been a rapid rise in the number of interesting studies performed in the domain of Dravidian code-mixed text analysis in the last few years (Chakravarthi et al., 2021, 2020; Ghanghor et al., 2021a,b; Yasaswini et al., 2021).

Sub-word level or morpheme level embedding

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 151 - 157

technique obtained using a 1D convolution layer with ReLU activation was proposed by Joshi et al. (2016). After getting a morpheme-level feature map, a 1-D maximum pooling layer is used to obtain its most prominent features. LSTMs are used to obtain the connections between each of these features due to their ability to process sequences and retain information.

Bharathi et al. (2021) proposed using TF-IDF and m-BERT embeddings coupled with classification models like Random Forest, Naive Bayes, and Multi-Layer Perceptron for the task of classifying English text code-mixed with Dravidian languages as offensive or not-offensive.

Selective translation and transliteration was performed by Sai and Sharma (2021), to convert the whole text to Tamil text in native script. XLM-RoBERTa multilingual model was used to obtain embeddings. Multiple classification algorithms were used to classify code-mixed text as offensive and not-offensive, and logistic regression was found to perform the best.

An ensembling of multiple classification algorithms applied on TF-IDF embedding was experimented in Kumar et al. (2021). It was found to perform well for shorter Dravidian code-mixed sentences like YouTube comments, which can be considered to be similar to meme captions.

We hypothesized translation and transliteration should improve the overall classification accuracy of the text when the troll script is code mixed. In addition, using language-specific embeddings would yield an improvement.

3 Data

3.1 Data description

The dataset used is the official dataset released in DravidianLangTech-ACL2022, which comprises captions from memes, and each caption is labelled as either troll or not troll. The data is represented in the Tamil-English code-mixed form, with the sentences comprising Tamil and English words but written in Roman script. (Suryawanshi et al., 2020)(Suryawanshi and Chakravarthi, 2021)(Suryawanshi et al., 2022)

Examples of the meme captions and its labels in Figure 1

3.2 Data distribution

The training data contains 2300 captions, each labelled as either troll or not troll, with a distribution of 1282 captions labelled as troll and 1018 captions labelled as not troll. The test dataset contains 667 captions, with 395 labelled as troll and 272 labelled as not troll.

Dataset	Troll	Not troll	Total
Train	1282	1018	2300
Test	395	272	667

Table 1: Distribution of the dataset

4 Methodology

4.1 Data pre-processing

The raw dataset was initially pre-processed to convert the text to lower case and remove URLs, special characters, extra spaces, and emoticons. Apostrophe abbreviated words like “they’re”, “it’s”, “I’m” etc., were converted to the long-form “they are”, “it is”, and “I am”, respectively. The words were lemmatized and stop words were removed using NLTK[1]. Named entity recognition was performed using the spaCy[2] library.

4.2 Experimental setup

4.2.1 Raw Tamil-English code-mixed text

The first set of techniques obtains embeddings from the dataset in the Tamil-English code-mixed form itself. The techniques under this category include TF-IDF, sub-word LSTM and m-BERT (Devlin et al. (2018)). TF-IDF was implemented using an n-gram range of (1,5), which was proven to perform better for the required use case (Bharathi et al., 2021). The hyperparameters for the sub-word LSTM model were tuned in accordance with the suggestions proposed by Joshi et al. (2016).

4.2.2 Translated and transliterated text

The second set of embedding techniques acts on the translated and transliterated version of the preprocessed dataset. The English words in the dataset were translated to Tamil using the deep translator API[3], and the Tamil words written in Roman script were transliterated to Tamil script using Indic transliteration API[4]. Embeddings for the resulting dataset were obtained using TF-IDF, IndicFT (Kakwani et al. (2020)), MuRIL (Khanuja et al. (2021)), and m-BERT.

[1]https://www.nltk.org/
[2]https://spacy.io/
[3]https://pypi.org/project/deep-translator/
[4]https://pypi.org/project/indic-transliteration/

	imagename	captions
2072	troll_238.jpg	Thangachi pasathula bamalaya minchiruva polaya...
1271	troll_1225.png	BUS'IL SINGLE'AAGA PAYANAM SAIBAVARGALUKKU MAT...
1430	troll_137.png	WHATSAPP GROUP CONVERSATIONS AFTER CREATING TH...
1567	troll_1539.jpg	*Beardless boys : Aiyoo perumale enna PET sir ...
733	Not_troll_742.jpg	idhe velaiya dhaan alaierangala kaalailairundhu..

Figure 1: Troll meme data

4.3 Classifier models

Eleven different classification algorithms, Logistic Regression (LR), MultinomialNaive Bayes (NB), Support Vector Machine - Linear kernel (L-SVM), Support Vector Machine - RBF kernel (R-SVM), Support Vector Machine - Polynomial kernel (P-SVM), Random Forest (RF), k-Nearnest Neighbours classifier (k-NN), Extra-tree classifier (ExT), AdaBoost classifier (AdB), XGBoost classifier (XgB), Multilayer Perceptron (MLP) and a voting ensemble of all the eleven classifiers were applied on the embeddings obtained from each of the above techniques.

Method	A	F1-m	F1-w
LR	0.62	0.53	0.57
NB	0.62	0.53	0.57
L-SVM	0.61	0.55	0.58
R-SVM	0.60	0.52	0.56
P-SVM	0.61	0.55	0.58
RF	0.61	0.57	0.60
k-NN	0.45	0.41	0.38
ExT	0.60	0.55	0.57
AdB	0.60	0.47	0.52
XgB	0.62	0.52	0.56
MLP	0.60	0.53	0.57
Voting	0.61	0.54	0.58

Table 2: TF-IDF - raw text

5 Results and Discussions

Tables 2, 3 and 4 depict the results obtained by performing experiments on raw code-mixed text. Tables 5, 6, 7 and 8 depict the results obtained by performing experiments on translated and transliterated text.

A weighted F1 score of 0.74 was achieved with MuRIL, beating our own previously published competition result of 0.60 obtained by random forest classifier, which held the top position in the rankings. This performs best due to the nature of input text which is code mixed and predominantly in Tamil and written either in Roman script or Tamil

Method	A	F1-m	F1-w
LR	0.57	0.47	0.51
NB	0.57	0.49	0.53
L-SVM	0.57	0.47	0.51
R-SVM	0.56	0.46	0.50
P-SVM	0.56	0.46	0.51
RF	0.57	0.49	0.53
k-NN	0.57	0.49	0.52
ExT	0.58	0.49	0.53
AdB	0.56	0.47	0.51
XgB	0.57	0.47	0.52
MLP	0.56	0.48	0.52
Voting	0.57	0.48	0.52

Table 3: Sub-word LSTM - raw text

Method	A	F1-m	F1-w
m-BERT	0.60	0.49	0.53

Table 4: m-BERT - raw text

Method	A	F1-m	F1-w
LR	0.59	0.51	0.55
NB	0.61	0.52	0.56
L-SVM	0.57	0.51	0.54
R-SVM	0.58	0.51	0.55
P-SVM	0.57	0.51	0.54
RF	0.56	0.51	0.54
k-NN	0.43	0.38	0.35
ExT	0.56	0.51	0.54
AdB	0.54	0.53	0.54
XgB	0.55	0.50	0.53
MLP	0.57	0.52	0.55
Voting	0.58	0.52	0.55

Table 5: TF-IDF - Translated, transliterated text

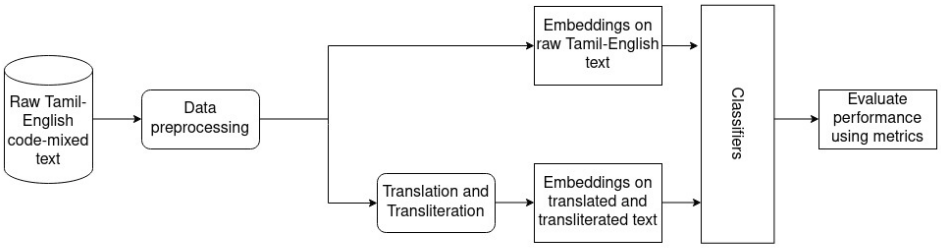

Figure 2: Experimental framework

Method	A	F1-m	F1-w
LR	0.59	0.37	0.74
NB	0.54	0.52	0.54
L-SVM	0.59	0.37	0.74
R-SVM	0.59	0.37	0.74
P-SVM	0.59	0.37	0.74
RF	0.53	0.48	0.55
k-NN	0.52	0.51	0.52
ExT	0.58	0.48	0.52
AdB	0.57	0.49	0.53
XgB	0.56	0.48	0.60
MLP	0.56	0.46	0.61
Voting	0.58	0.46	0.50

Table 6: MuRIL - Translated, transliterated text

Method	A	F1-m	F1-w
LR	0.60	0.49	0.66
NB	0.57	0.51	0.61
L-SVM	0.59	0.48	0.53
R-SVM	0.58	0.41	0.70
P-SVM	0.59	0.46	0.67
RF	0.55	0.50	0.56
k-NN	0.59	0.44	0.68
ExT	0.54	0.45	0.49
AdB	0.56	0.49	0.52
XgB	0.58	0.48	0.52
MLP	0.59	0.48	0.65
Voting	0.58	0.45	0.50

Table 7: IndicFT - Translated, transliterated text

Method	A	F1-m	F1-w
m-BERT	0.58	0.50	0.55

Table 8: m-BERT - Translated, transliterated text

script. With a pre-trained model like MuRIL, which is specifically trained on Tamil, we see a higher accuracy. IndicFT, a fastText model that is also trained on Indian languages, achieves a weighted F1 of 0.70. Another word embedding technique, m-BERT, performed slightly better with translated-transliterated text with an F1 score of 0.55 compared to 0.53 without. Translation and transliteration on TF-IDF, contrary to our hypothesis, performed poorly relative to direct usage of tokens in their raw form on all three metrics; Accuracy, macro F1 and weighted F1 scores. This could be attributed to the lower level of confidence in the translation and transliteration capabilities in code-mixed text.

6 Conclusion and Future Work

We compared the weighted F1 score of the various classifiers between the raw text and translated-transliterated text in Tamil-English code mixed troll memes. Language-specific word embedding techniques significantly improve the classification metrics like accuracy, macro F1 and weighted F1 scores. With a comparison between the various pre-trained models, MuRIL performs best with an F1 score of 0.74 relative to others. We have only considered text captions and not the images. We hypothesize that a unified framework to combine image with text will yield even better results. In the future, we would want to integrate a deep learning model that could take both the caption text and the images together.

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.

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.

B Bharathi et al. 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.

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, 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.

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.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021. Dataset for identification of homophobia and transophobia in multilingual YouTube comments. *arXiv preprint arXiv:2109.00227*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. 2016. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

SR Mithun Kumar, Nihal Reddy, Aruna Malapati, and Lov Kumar. 2021. An ensemble model for sentiment classification on code-mixed data in dravidian languages. Technical report, EasyChair.

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.

Anitha Narasimhan, Aarthy Anandan, Madhan Karky, and CN Subalalitha. 2018. Porul: Option generation and selection and scoring algorithms for a tamil flash card game. *International Journal of Cognitive and Language Sciences*, 12(2):225–228.

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.

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.

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.

Siva Sai and Yashvardhan Sharma. 2021. Towards offensive language identification for dravidian languages. In *Proceedings of the first workshop on speech and language technologies for Dravidian languages*, pages 18–27.

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.

Anbukkarasi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, Kogilavani Shanmugavadivel, Sajeetha Thavareesan, Sathiyaraj Thangasamy, Parameswari Krishnamurthy, Adeep Hande, Sean Benhur, Kishor Kumar Ponnusamy, and Santhiya Pandiyan. 2022. Findings of the shared task on Emotion Analysis in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

R Srinivasan and CN Subalalitha. 2019. Automated named entity recognition from tamil documents. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–5. IEEE.

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.

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.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, Susan Levy, Paul Buitaleer, Prasanna Kumar Kumaresan, Rahul Ponnusamy, and Adeep Hande. 2022. Findings of the second shared task on Troll Meme Classification in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020. A dataset for troll classification of tamilmemes. In *Proceedings of the WILDRE5–5th workshop on indian language data: resources and evaluation*, pages 7–13.

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.