GJG@TamilNLP-ACL2022: Emotion Analysis and Classification in Tamil using Transformers

Janvi Prasad*
Vellore Institute of Technology
Vellore, India
janvi.prasad@gmail.com

Gaurang Prasad*
wikiHow Inc.
Palo Alto, CA, USA
gaurang@wikihow.com

Gunavathi Chellamuthu
Vellore Institute of Technology
Vellore, India
gunavathi.cm@vit.ac.in

Abstract

This paper describes the systems built by our team for the "Emotion Analysis in Tamil" shared task at the Second Workshop on Speech and Language Technologies for Dravidian Languages at ACL 2022. There were two multi-class classification sub-tasks as a part of this shared task. The dataset for sub-task A contained 11 types of emotions while sub-task B was more fine-grained with 31 emotions. We fine-tuned an XLM-RoBERTa and DeBERTA base model for each sub-task. For sub-task A, the XLM-RoBERTa model achieved an accuracy of 0.46 and the DeBERTa model achieved an accuracy of 0.45. We had the best classification performance out of 11 teams for sub-task A. For sub-task B, the XLM-RoBERTa model's accuracy was 0.33 and the DeBERTa model had an accuracy of 0.26. We ranked 2nd out of 7 teams for sub-task B.

1 Introduction

Emotions are a fundamental component of any language that are used to express how people feel about different things. Emotion detection and classification has becomes an important task in the field of Natural Language Processing (NLP) (Chakravarthi et al., 2021). Emotion analysis enables the improved understanding of user-generated text and has applications in understanding public opinions, healthcare, development of voice and language-based assistants, recommendation engines, etc (Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021a).

Over the past two decades, the internet has become the central avenue for communication. With the advent of web-based services and digital publication platforms, the volume of text-based content across all languages have sky rocketed (B and A, 2021b,a). This not only includes articles, blog posts, and scientific publications, but also user-generated opinions and comments in social networks (Priyadharshini et al., 2021; Kumaresan et al., 2021). People who feel apprehensive about in-person conversations and physical interactions also rely on social media to express their thoughts (Sampath et al., 2022; Ravikiran et al., 2022; Chakravarthi et al., 2022; Bharathi et al., 2022; Priyadharshini et al., 2022). Due to this, social media has become a modern channel of public expression for the people irrespective of the socio-economic boundaries (Priyadharshini et al., 2020). These mediums are not only used to express constructive and positive emotions but also a lot of negativity and hatred (Ghanghor et al., 2021a,b; Yasaswini et al., 2021). A lot of communities express these emotions in their native language. Identifying all these different kinds of emotions is extremely important for the development and improvement of software systems, NLP models, and Human-Computer Interaction.

India is a vast, multi-cultural, and multi-lingual country. A substantial amount of research work has been done for text classification tasks in global languages like English, Spanish, and Mandarin. There has also been NLP-research for Indian languages like Hindi and Urdu (Anita and Subalalitha, 2019b,a; Subalalitha and Poovammal, 2018). However, very little work has been done for Dravidian languages. Dravidian languages are a big part of the Indian culture. Even outside India, they are used in multiple regions for digital and in-person communication and publication (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019, 2020a,b, 2021).

The lack of research in Dravidian Language NLP tasks is largely due to the lack of annotated

*These authors contributed equally to this work

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 86 - 92

datasets. This task provides two datasets for researchers to work with - one coarse-grained and one fine-grained. The availability and publication of such datasets and shared tasks invites multiple approaches to solve downsteam NLP tasks for a Dravidian language like Tamil (Subalalitha, 2019; Srinivasan and Subalalitha, 2019; Narasimhan et al., 2018). In this shared task, we participated in both the sub-tasks: the coarse-grained classification sub-task A with 11 classes, and the fine-grained sub-task B with 31 output classes. The goal of our work is to demonstrate the performance of fine-tuning large pre-trained transformer-based models for a text-classification task in Tamil. We train an XLM-RoBERTa and DeBERTa model, both of which are pre-trained models, for each sub-task on the given train splits, optimize parameters, and evaluate their performance on the respective test splits (Conneau et al., 2019; He et al., 2021).

The rest of our paper is organized as follows: we discuss related work in Tamil emotion recognition, describe the datasets, our methodology, and conclude with the results and performance metrics.

We provide a link to our models and evaluations to provide reproducability, and empower future research in this space[1]. We hope to build on the learnings from this shared task to architect and build models specifically for downstream Tamil NLP tasks.

2 Related Work

There has been a lot of work in emotion analysis and classification for high-resource languages. Even for a low-resource language like Tamil, there have been multiple published works. Renjith and Manju (2017) used Cepstral Coefficients (LPCC) along with Neural Networks to detect emotions. They demonstrated higher accuracy with Hurst parameters as compared to LPCC, when considering individual features for a language like Tamil. Ram and Ponnusamy (2014) used Support Vector Machine (SVM) for emotion recognition in Tamil. They used Cepstral Coefficients for training their model. Sowmya and Rajeswari (2019) extracted features from Tamil audio signals and trained an SVM classifier. They demonstrated a classification accuracy of 85.4%. Saste and Jagdale (2017) also trained an SVM classifier using a feature vector formed by fusion of MFCCs and DWT. Poorna et al. (2018) demonstrated a weight-based emotion recognition system using audio signals for three South Indian languages. They used K-Nearest Neighbor, SVM, and a Neural Network as their classification models. Srikanth et al. (2017) proposed a Deep Belief Network (DBN) over Gaussian Mixture model (GMM) for Tamil emotion recognition. Fernandes and Mannepalli (2021) trained four LSTM-based models for emotion recognition in Tamil speech. They found that Deep Hierarchical LSTM and BiLSTM (DHLB) achieves the highest precision of about 84%. All of the aforementioned research has been focused on emotion detection and recognition using speech signals or features extracted from Tamil speech signals.

There has also been some work in emotion and sentiment analysis based on Tamil text. Raveendirarasa and Amalraj (2020) used sub-word level LSTM to build a behavioural profile for Facebook users, to be able to detect sentiment from Facebook comments. Priyadharshini et al. (2021) presented the findings of the shared task on sentiment analysis in Tamil, Malayalam, and Kannada. Chakravarthi and Muralidaran (2021b) presented the findings of a shared task on hope speech detection. These also focus on text classification tasks in Tamil, but emphasize other emotional and sentimental classes.

There have also been multiple published work that fine-tunes XLM-RoBERTa for text classification tasks. Zhao and Tao (2021) proposed a system using XLM-RoBERTa and DPCNN for detecting offensive text in Dravidian languages. Qu et al. (2021) used TextCNN and XLM-RoBERTa from emotion classification in Spanish. Ou and Li (2020) also demonstrated using XLM-RoBERTa for a hate speech identification classification task.

The number of published works using DeBERTa is fewer than that of XLM-RoBERTa. There have been some studies that use DeBERTa for entity extraction and text-classification tasks. (Martin and Pedersen, 2021; Khan et al., 2022)

3 Data

The annotated training and development datasets, for both the sub-tasks, were provided by the workshop organizers. The testing dataset, without labels, was released a few days prior to the run submission deadline for the teams to run their models on. Once the results were announced, the organizers released the labeled test dataset for

[1] https://tinyurl.com/GJGEmotionAnalysis

Label	Count	Label	Count
Neutral	4841	Anticipation	828
Joy	2134	Sadness	695
Ambiguous	1689	Love	675
Trust	1254	Surprise	248
Disgust	910	Fear	100
Anger	834		

Table 1: Classification labels for sub-task A and the number of rows under each label in the training set.

validation and verification purposes.

The datasets for both the sub-tasks consisted of Tamil sentences obtained from social media comments. A post/ row within the corpus may contain one or more sentences. However, the organizers ensured that the average sentence length of the corpora was 1. The annotations in the corpus were made at a comment / post level (Sampath et al., 2022). The posts could also contain extended words, emojis, and other special characters. The grammatical and lexical accuracy of the sentences were unchanged, in order to be representative of user-generated social media comments.

3.1 Sub-Task A

Sub-task A, the coarse-grained classification task, had a total of 11 output classes/ labels. The training dataset had a total of 14,208 rows while the development dataset had 3,552 rows. The test dataset had 4,440 rows. The classification labels along with the total count in the train set are represented in Table 1. The entire dataset was annotated with English labels, as compared to the sentences - which were in Tamil.

3.2 Sub-Task B

Sub-task B was significantly more fine-grained as compared to the sub-task A and contained a total of 31 output classes/ labels. Unlike sub-task A, the class labels for this sub-task were in Tamil and not English. The training dataset had a total of 30,179 rows, making it much larger than the training split for the coarse-grained classification task. The development dataset had 4,269 rows and the test dataset had 4,269 rows. Table 2 represents all the class labels in the train split (translated to English) along with the total number of rows in each label.

Label	Count	Label	Count
Admiration	4760	Caring	497
Realization	3499	Embarrass	484
Anticipation	2191	Sadness	470
Teasing	2128	Love	453
Approval	1853	Disappoint	422
Anger	1738	Disapproval	421
Annoyance	1277	Disgust	343
Joy	1276	Optimism	292
Neutral	1232	Fear	288
Pride	963	Grief	259
Gratitude	880	Nervous	255
Curiosity	782	Relief	238
Trust	713	Remorse	235
Confusion	709	Surprise	201
Amusement	625	Desire	147
Excitement	548		

Table 2: Translated classification labels for sub-task B and the number of rows under each in the training set.

4 Methodology

For each classification sub-task, we fine-tune an XLM-RoBERTa and DeBERTa base model on sentences from the training splits to create a classification model. We do not remove any stop words, special characters, or emojis from the test splits, in order to preserve the context of the comment. Extended train of special characters (examples: !!!, ..., etc.) and emojis provide useful context, especially for an emotion analysis task.

XLM-RoBERTa is a multilingual version of RoBERTa, which in itself was an improvement over BERT to achieve state-of-the-art results in multiple NLP tasks. XLM-RoBERTa is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages (Conneau et al., 2019). DeBERTa uses disentangled attention and enhanced mask decoder to enhance RoBERTa and outperform it in a majority of NLP tasks (He et al., 2021).

Table 3 represents the parameters used to fine-tune the XLM-RoBERTa and DeBERTa base models for both the sub-tasks.

5 Results

We use accuracy and the weighted averages of precision, recall, and F1-score as performance metrics to evaluate our classification models. While the shared task results were based on the macro average F1-score, we calculate all four evaluation metrics to get a better sense of the performance.

	Parameter Value			
	Sub-Task A		**Sub-Task B**	
	XLM-RoBERTa	DeBERTa	XLM-RoBERTa	DeBERTa
Batch Size	20	8	32	8
Max. Sequence Length	256	256	256	256
Number of Epochs	6	10	6	6
Learning Rate	1e-5	1e-5	1e-5	1e-5
Weight Decay	0	0	0	0
Use Class Weights	False	False	False	False

Table 3: Fine-tuning parameters of XLM-RoBERTa and DeBERTA models for both sub-tasks

Task	**Model**	**Accuracy**	**F1-score**	**Precision**	**Recall**
Sub-Task A	XLM-RoBERTa	0.46	0.44	0.44	0.46
	DeBERTa	0.45	0.38	0.38	0.45
Sub-Task B	XLM-RoBERTa	0.33	0.26	0.25	0.33
	DeBERTa	0.26	0.2	0.18	0.26

Table 4: Performance metrics for both sub-tasks.

For sub-task A, we find that the XLM-RoBERTa outperforms the DeBERTa in all evaluation metrics. There is a difference of 0.06 in the weighted F1-score and Precision between the two models. Despite the DeBERTa model using a smaller batch size and being trained for a higher number of epochs, the better performance of the XLM-RoBERTa is evident from the metrics.

The overall classification performance for sub-task 2 was lower than that for sub-task A. The fine-grained nature of the task made it a significantly more complex challenge. However, we still find that XLM-RoBERTa model easily outperforms the DeBERTa model (Table 4).

6 Conclusion

This paper presents the fine-tuning of a pre-trained XLM-RoBERTa and DeBERTa models for two multi-class text classification tasks in Tamil. The objective of the shared task was to classify a Tamil text into an emotion class. There were two sub-tasks: the coarse-grained sub-task A with 11 output classes and the fine-grained sub-task B with 31 classes. The dataset, including the training and validation splits, for both the sub-tasks were released by the organizers. The dataset consisted of Tamil text extracted from social media comments. The training split for sub-task A had a total of 14,208 rows and used English classification labels. The training split for sub-task B had 30,179 rows with Tamil classification labels.

We propose the fine-tuning of pre-trained transformer-based models for classifying Tamil text into emotion classes. We trained an XLM-RoBERTa and DeBERTa model for each sub-task while using the training split as-is. For sub-task A, the XLM-RoBERTa achieved a classification accuracy of 46% with a weighted F-1 of 0.44, precision of 0.44, and a recall value of 0.46. The DeBERTa model achieved an accuracy of 45% with weighted F-1 of 0.38, precision of 0.38, and 0.45 recall. For sub-task B, the XLM-RoBERTa achieved a classification accuracy of 33% with a weighted F-1 of 0.26, precision of 0.25, and a recall value of 0.33. The DeBERTa model achieved an accuracy of 26% with weighted F-1 of 0.2, precision of 0.18, and 0.26 recall.

We show that the XLM-RoBERTa model outperforms DeBERTa for both the sub-tasks. By using the training split as-is, we retain the information provided by special characters like emojis and extended punctuations. The XLM-RoBERTa model had the best classification performance out of 11 teams for the first sub-task and was the second-best in sub-task B out of 7 teams. We have open-sourced the code used in this study in a public GitHub repository.

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.

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. 2021a. 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 and Vigneshwaran Muralidaran. 2021b. 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.

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*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Bennilo Fernandes and Kasiprasad Mannepalli. 2021. Speech emotion recognition using deep learning lstm for tamil language. *Pertanika Journal of Science & Technology*, 29(3).

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Pervaiz Iqbal Khan, Imran Razzak, Andreas Dengel, and Sheraz Ahmed. 2022. Performance comparison of transformer-based models on twitter health mention classification. *IEEE Transactions on Computational Social Systems*.

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.

Anna Martin and Ted Pedersen. 2021. Duluth at semeval-2021 task 11: Applying deberta to contributing sentence selection and dependency parsing for entity extraction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 490–501.

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.

Xiaozhi Ou and Hongling Li. 2020. Ynu_oxz@ haspeede 2 and ami: Xlm-roberta with ordered neurons lstm for classification task at evalita 2020. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 2765:102–109.

SS Poorna, K Anuraj, and GJ Nair. 2018. A weight based approach for emotion recognition from speech: an analysis using south indian languages. In *International Conference on Soft Computing Systems*, pages 14–24. Springer.

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.

S Qu, Y Yang, and Q Que. 2021. Emotion classification for spanish with xlm-roberta and textcnn. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain*.

C Sunitha Ram and R Ponnusamy. 2014. An effective automatic speech emotion recognition for tamil language using support vector machine. In *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, pages 19–23. IEEE.

Vidyapiratha Raveendirarasa and CRJ Amalraj. 2020. Sentiment analysis of tamil-english code-switched text on social media using sub-word level lstm. In *2020 5th International Conference on Information Technology Research (ICITR)*, pages 1–5. 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.

S Renjith and KG Manju. 2017. Speech based emotion recognition in tamil and telugu using lpcc and hurst parameters—a comparitive study using knn and ann classifiers. In *2017 International conference on circuit, power and computing technologies (ICCPCT)*, pages 1–6. IEEE.

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, and Santhiya Ponnusamy, Kishor Kumar 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.

Sonali T Saste and SM Jagdale. 2017. Emotion recognition from speech using mfcc and dwt for security system. In *2017 international conference of electronics, communication and aerospace technology*, volume 1, pages 701–704. IEEE.

V Sowmya and A Rajeswari. 2019. Speech emotion recognition for tamil language speakers. In *International Conference on Machine Intelligence and Signal Processing*, pages 125–136. Springer.

M Srikanth, D Pravena, and D Govind. 2017. Tamil speech emotion recognition using deep belief network (dbn). In *International Symposium on Signal Processing and Intelligent Recognition Systems*, pages 328–336. Springer.

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.

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.

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.