Findings of the Shared Task on Multi-task Learning in Dravidian Languages

Bharathi Raja Chakravarthi[1], **Ruba Priyadharshini**[2], **CN Subalalitha**[3]
Sangeetha Sivanesan[4], **Malliga Subramanian**[5], **Kogilavani Shanmugavadivel**[5],
Parameswari Krishnamurthy[6], **Adeep Hande**[7], **Siddhanth U Hegde**[8]
Roshan Nayak[9], **Swetha Valli**[10]
[1] National University of Ireland Galway, [2]Madurai Kamaraj University
[3]SRM Institute Of Science And Technology, [4]NIT Tiruchirappalli,[5]Kongu Engineering College
[6]University of Hyderabad, [7]Indian Institute of Information Technology Tiruchirappalli
[8]University Visvesvaraya College of Engineering, [9]BMS College of Engineering
[10]Thiyagarajar college of engineering
bharathi.raja@insight-centre.org

Abstract

We present our findings from the first shared task on Multi-task Learning in Dravidian Languages at the second Workshop on Speech and Language Technologies for Dravidian Languages. In this task, a sentence in any of three Dravidian Languages is required to be classified into two closely related tasks namely *Sentiment Analyis* (**SA**) and *Offensive Language Identification* (**OLI**). The task spans over three Dravidian Languages, namely, Kannada, Malayalam, and Tamil. It is one of the first shared tasks that focuses on Multi-task Learning for closely related tasks, especially for a very low-resourced language family such as the Dravidian language family. In total, 55 people signed up to participate in the task, and due to the intricate nature of the task, especially in its first iteration, 3 submissions have been received.

1 Introduction

The term "Social media" provides a channel through which people engage in interactive communities and networks by creating, sharing, and exchanging thoughts and information. It has received users from almost all generations and all around the world (Chakravarthi et al., 2020a). Users can interact and connect with others and form communities through social media. It allows users to share their ideas, views and information openly on various topics. This gives license to the users to write hateful and offensive comments sometimes. People come from a variety of racial backgrounds and hold a diversity of belief systems. This can often cause conflict of opinions during their interactions on social media platforms (Chakravarthi et al., 2021a,b, 2020b; Priyadharshini et al., 2020; Chakravarthi, 2020). Due to the COVID 19 pandemic, the internet community has become more popular than it has ever been. The amount of false narratives and derogatory remarks shared on online platforms has risen exponentially. A large number of social media users share malicious posts despite understanding that they are infringing on their rights to free expression (Bhardwaj et al., 2020). Sentiment analysis is a text mining task that identifies and extracts personal information from source material, allowing a company/researcher to better understand the social sentiment of its brand, product, or service while monitoring online conversations.

Multi-task learning (MTL) is a practical approach to improving system performance by utilising shared characteristics of tasks (Caruana, 1997). The goal of MTL is to use learning multiple tasks at the same time to improve system performance (Martínez Alonso and Plank, 2017). Because SA and OLI are essentially sequence classification tasks, we were motivated to conduct the shared task, due to the recent developments in large language modeling. Kannada and Malayalam are Dravidian languages that are widely spoken in South India and are also official languages in the states of Karnataka and Kerala (Reddy and Sharoff, 2011; Chakravarthi et al., 2020a, 2019, 2018; Ghanghor et al., 2021a,b). Tamil is an official language in Tamil Nadu, India, as well as Sri Lanka, Singapore, Malaysia, and other parts of the world. Dravidian languages are morphologically rich; with code-mixing, processing these languages becomes even more difficult, and they are under-resourced (Priyadharshini et al., 2021; Kumaresan et al., 2021;

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 286 - 291
May 26, 2022 ©2022 Association for Computational Linguistics

Chakravarthi and Muralidaran, 2021; Chakravarthi et al., 2020c; Sampath et al., 2022; Ravikiran et al., 2022; Chakravarthi et al., 2022a; Bharathi et al., 2022; Priyadharshini et al., 2022). Significant minority speak Tamil in the four other South Indian states of Kerala, Karnataka, Andhra Pradesh, and Telangana, as well as the Union Territory of the Andaman and Nicobar Islands (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019, 2020a,b, 2021). It is also spoken by the Tamil diaspora, which may be found in Malaysia, Myanmar, South Africa, the United Kingdom, the United States, Canada, Australia, and Mauritius. Tamil is also the native language of Sri Lankan Moors. Tamil, one of the 22 scheduled languages in the Indian Constitution, was the first to be designated as a classical language of India (Subalalitha, 2019; Srinivasan and Subalalitha, 2019; Narasimhan et al., 2018). Tamil is one of the world's longest-surviving classical languages. The earliest epigraphic documents discovered on rock edicts and "hero stones" date from the 6th century BC. Tamil has the oldest ancient non-Sanskritic Indian literature of any Indian language (Anita and Subalalitha, 2019b,a; Subalalitha and Poovammal, 2018).

The shared task on MTL in Dravidian Languages investigates whether it is beneficial to train models using MTL, as obtaining extensive annotated data for under resourced languages is difficult. Additionally, SA and OLI have discourse properties in common (Chakravarthi et al., 2022b). The lack of large labelled data for user-generated code-mixed datasets motivated the selection of these tasks. Past studies have shown us that the benefits to MTL are two folds, namely, reducing the space/time complexity, and the ability for the model to learn from each other tasks (Hande et al., 2021). Our dataset contains a wide range of code-mixing, from simple script mixing to morphological mixing. The task is to determine the polarity of sentiment and offensiveness in a code-mixed dataset of Tamil-English, Malayalam-English, and Kannada-English comments or posts. This paper presents an overview of the task description, dataset, description of the participating systems, analysis, and provide insights from the shared task.

2 Dataset

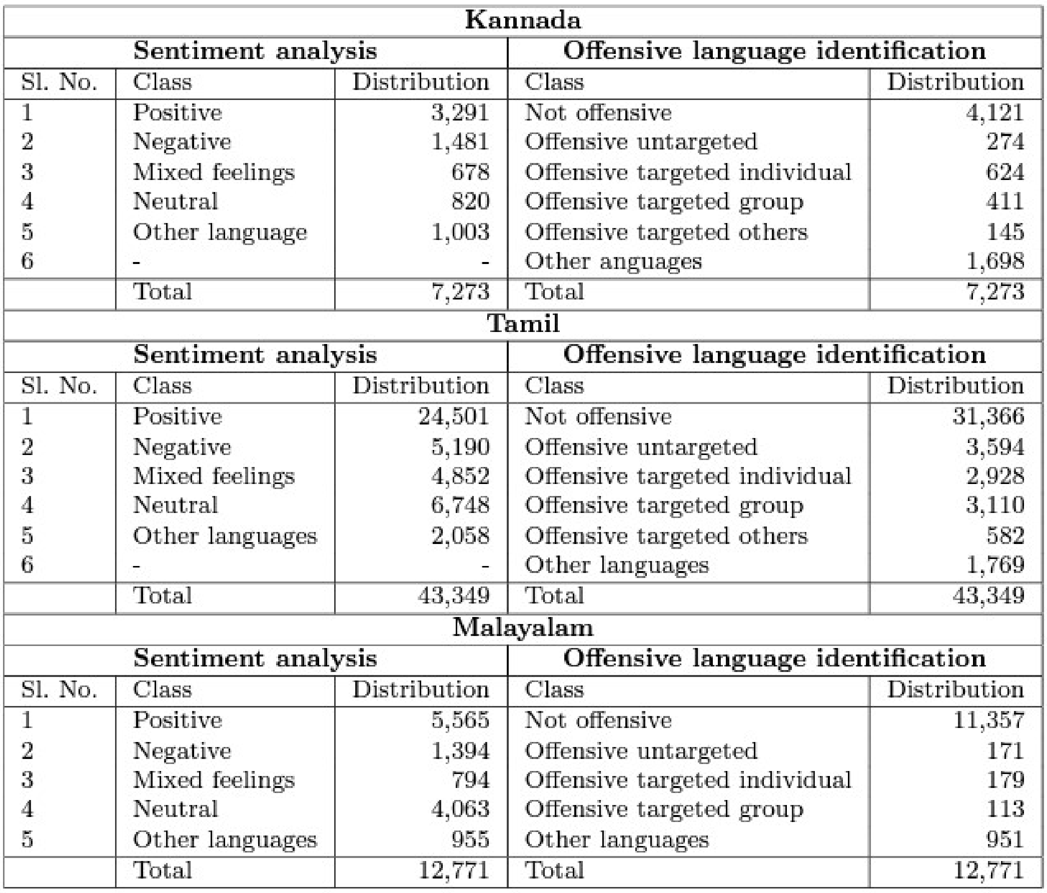
The DravidianCodeMix dataset (Chakravarthi et al., 2022b) is the primary resource of the shared task. It comprises of over 60,0000 manually annotated comments scraped from YouTube. Additionally, DravidianCodeMix spans three languages in the Dravidian language family, namely, Kannada, Malayalam, and Tamil. The Kannada code-mixed dataset has 7,273 comments, while the Malayalam and Tamil codemixed datasets have 12,711 and 43,349 comments, respectively. Following the removal of repetitive sentences, Figure 1 shows the class-wise distribution of the datasets which will be split into train, validation, and test sets.

Sentiment Analysis:

- **Positive state:** Comment contains an explicit or implicit clue in the text suggesting that the speaker is in a positive state.
- **Negative state:** Comment contains an explicit or implicit clue in the text suggesting that the speaker is in a negative state.
- **Mixed feelings:** Comment contains an explicit or implicit clue in both positive and negative feeling.
- **Neutral state:** Comment does not contain an explicit or implicit indicator of the speaker's emotional state.
- **Not in intended language:** For Kannada if the sentence does not contain Kannada script or Latin script then it is not Kannada.

Offensive Language Identification :

- **Not Offensive**: Comment does not contain offence or profanity.
- **Offensive Untargeted** : Comment contains offence or profanity without any target. These are comments which contain unacceptable language that does not target anyone.
- **Offensive Targeted Individual**: Comment contains offence or profanity which targets the individual.
- **Offensive Targeted Group**: Comment contains offence or profanity which targets the group.
- **Offensive Targeted Other**: Comment contains offence or profanity which does not belong to any of the previous two categories (e.g., a situation, an issue, an organization or an event).



Kannada				
	Sentiment analysis		Offensive language identification	
Sl. No.	Class	Distribution	Class	Distribution
1	Positive	3,291	Not offensive	4,121
2	Negative	1,481	Offensive untargeted	274
3	Mixed feelings	678	Offensive targeted individual	624
4	Neutral	820	Offensive targeted group	411
5	Other language	1,003	Offensive targeted others	145
6	-	-	Other anguages	1,698
	Total	7,273	Total	7,273
Tamil				
	Sentiment analysis		Offensive language identification	
Sl. No.	Class	Distribution	Class	Distribution
1	Positive	24,501	Not offensive	31,366
2	Negative	5,190	Offensive untargeted	3,594
3	Mixed feelings	4,852	Offensive targeted individual	2,928
4	Neutral	6,748	Offensive targeted group	3,110
5	Other languages	2,058	Offensive targeted others	582
6	-	-	Other languages	1,769
	Total	43,349	Total	43,349
Malayalam				
	Sentiment analysis		Offensive language identification	
Sl. No.	Class	Distribution	Class	Distribution
1	Positive	5,565	Not offensive	11,357
2	Negative	1,394	Offensive untargeted	171
3	Mixed feelings	794	Offensive targeted individual	179
4	Neutral	4,063	Offensive targeted group	113
5	Other languages	955	Other languages	951
	Total	12,771	Total	12,771

Figure 1: Classwise distribution of the datasets for Kannada, Malayalam, and Tamil

- **Not in indented language**: Comment not in the Kannada language.

In general, all languages have similar class types. Kannada and Tamil code-mixed datasets have six classes in OLI, while Malayalam has five classes. The Malayalam dataset lacks the Offensive Language Others (OTO) class.

2.1 Training Phase

In the first phase, data is made available for training and/or development of offensive language detection models. Participants were given training and validation datasets for preliminary evaluations or tuning of hyper-parameters. They were also given the option of performing cross-validation on the training data. In total, 57 people registered for the task and downloaded the data.

2.2 Evaluation Phase

In the second phase, test sets for all three languages are made available for evaluation. Each team that took part submitted their generated prediction for evaluation. Predictions have been submitted to the organising committee via Google form for evaluation. CodaLab is a well-known platform for organising collaborative tasks. However, due to issues with running the evaluation, we decided to evaluate manually. The macro average F1 score is the metric used for evaluation.

3 System Description

MUCIC (Gowda et al., 2022) - The authors submitted their predictions for all three languages. They treated this as a single task and fine-tuned the multilingual DistilBERT language model, and aggregated the outputs.
MUCS (Hegde and Coelho, 2022) - The authors submitted their predictions for all three languages. Similar to the other team, they treated it a single task. They used Dynamic Meta Embedding as a feature in training a DL-based LSTM model to predict test set labels.

4 Evaluation, Results and Discussion

The submissions were primarily evaluated using major classification metrics such as Macro Aver-

Team Name	Kannada		Rank
	Sentiment Analysis	Offensive Language Identification	
MUCS	0.201	0.221	1
MUCIC	0.177	0.199	2
Team Name	**Malayalam**		**Rank**
	Sentiment Analysis	**Offensive Language Identification**	
MUCIC	0.192	0.245	1
MUCIC	0.148	0.079	2
Team Name	**Tamil**		**Rank**
	Sentiment Analysis	**Offensive Language Identification**	
MUCS	0.296	0.176	1
MUCIC	0.255	0.171	2

Table 1: Macro Average F1-Score of the systems submitted for the MTL shared Task.

aged and Weighted Average Precision, Recall, and F1-Score. We predominantly used Macro Averaged F1 Score to rank the teams because it identifies the F1 score to every label and calculates their unweighted mean.

MTL in its essence is a very challenging problem, especially when we focus this aspect on low-resourced language family such as Dravidian Languages (Kannada, Malayalam, and Tamil). Table 1 represents the results of the teams **MUCS** (Hegde and Coelho, 2022) and **MUCIC** (Gowda et al., 2022) on the two tasks of the three languages.

5 Conclusion

In its first iteration, the shared task on MTL for Dravidian Languages opened up new avenues for research in low-resource Multi-task Learning. The task involved multiple languages, namely, Kannada, Malayalam, and Tamil. This overview article analyzed the systems that were submitted to the shared task. The main inference from the participants is that MTL is a very challenging problem, especially for morphologically rich languages and all participants performed Single Task Learning and aggregated the outputs.

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.

Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

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, Mihael Arcan, and John P. McCrae. 2018. Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global Wordnet Conference*, pages 77–86, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019. WordNet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7, Dublin, Ireland. European Association for Machine Translation.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. 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 177–184, Marseille, France. European Language Resources association.

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. 2020b. 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. 2022a. 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, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2022b. DravidianCodeMix: sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text. *Language Resources and Evaluation*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020c. Overview of the track on sentiment analysis for Dravidian languages in code-mixed text. In *Forum for Information Retrieval Evaluation*, pages 21–24.

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

Bharathi Raja Chakravarthi, Priya Rani, Mihael Arcan, and John P McCrae. 2021b. A survey of orthographic information in machine translation. *SN Computer Science*, 2(4):1–19.

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.

Anusha Gowda, Fazlourrahman Balouchzahi, HL Shashirekha, and G Sidorov. 2022. MUCIC@DravidianLangTech-ACL2022: multi-task learning for dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Adeep Hande, Siddhanth U. Hegde, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced dravidian languages. *CoRR*, abs/2108.03867.

Asha Hegde and Sharal Coelho. 2022. MUCS@DravidianLangTech@ACL 2022: Multi-task Learning for Sentiment Analysis and Offensive Language Identification in Dravidian Languages using Dynamic Meta Embedding. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

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.

Héctor Martínez Alonso and Barbara Plank. 2017. When is multitask learning effective? semantic sequence prediction under varying data conditions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 44–53, Valencia, Spain. Association for Computational Linguistics.

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 Reddy and Serge Sharoff. 2011. Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources. In *Proceedings of the Fifth International Workshop On Cross Lingual Information Access*, pages 11–19, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

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.

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.