hate-alert@DravidianLangTech-ACL2022: Ensembling Multi-Modalities for Tamil TrollMeme Classification

Mithun Das, Somnath Banerjee, Animesh Mukherjee
Indian Institute of Technology, Kharagpur, India
mithundas@iitkgp.ac.in, som.iitkgpcse@kgpian.iitkgp.ac.in, animeshm@cse.iitkgp.ac.in

Abstract

Social media platforms often act as breeding grounds for various forms of trolling or malicious content targeting users or communities. One way of trolling users is by creating memes, which in most cases unites an image with a short piece of text embedded on top of it. The situation is more complex for multilingual(e.g., Tamil) memes due to the lack of benchmark datasets and models. We explore several models to detect Troll memes in Tamil based on the shared task, "Troll Meme Classification in DravidianLangTech2022" at ACL-2022. We observe while the text-based model MURIL performs better for Non-troll meme classification, the image-based model VGG16 performs better for Troll-meme classification. Further fusing these two modalities help us achieve stable outcomes in both classes. Our fusion model achieved a 0.561 weighted average F1 score and ranked **second** in this task.

1 Introduction

Over the past few years, social media platforms have been expanding rapidly. Users of the platform interact by sharing content to enrich their knowledge and social connections. Although most of the content on social media platforms that existed so far was textual, recently, a unique message was born: the *meme* (Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021). A meme is usually created by an image and a short piece of text on top of it, entrenched as part of the image. Memes are generally meant to be harmless and conceived to look humorous, but sometimes, bad actors use memes for threatening and abusing individuals or specific target communities (Ghanghor et al., 2021a,b; Yasaswini et al., 2021). Such memes are collectively known as Offensive/Troll memes in social media.

Trolling is the exercise of publicizing a message via social media that is planned to be abusive, inciting, or threatening to distract, which often has rambling or off-topic content to provoke the audience(Bishop, 2014; Suryawanshi et al., 2020a). In addition, such memes can be treacherous as they can easily harm the reputation of individuals, famous celebs, political entities, businesses, or social groups, e.g., minorities. Although various studies have been conducted to detect offensive posts using different natural language techniques, Troll meme classification has not yet been explored.

The situation for countries like India is more complicated due to the immense lanuage diverisy[1]. The meme in the Indian context, can be composed in English, local language (native or foreign script) or in combination of both language and script (Sampath et al., 2022; Ravikiran et al., 2022; Chakravarthi et al., 2022; Bharathi et al., 2022; Priyadharshini et al., 2022). This adds another challenge for the troll meme classification. Tamil is one of the world's longest-surviving classical languages (Anita and Subalalitha, 2019b,a; Subalalitha and Poovammal, 2018). 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). It is also classed as a member of the Tamil language family, which contains the languages of around 35 ethno-linguistic groups, including the Irula and Yerukula languages (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019, 2020a,b, 2021).

Recently, there has been a lot of effort to investigate the malicious side of memes, e.g., focusing on hate(Gomez et al., 2020), offensive(Suryawanshi et al., 2020a), and harmful(Pramanick et al., 2021) memes. However, the majority of the studies are centralized around the English language. Further several shared tasks like HASOC 2021(Modha et al., 2021), DravidianLangTech

---

[1] https://en.wikipedia.org/wiki/Languages_of_India

(a) An example of a Troll meme

(b) An example of a Non-troll meme

Figure 1: Examples of troll and not-troll meme

Split	Troll	Non-troll	Total
Train	1,282	1,018	2,300
Test	395	272	667
Total	1,677	1,290	2,967

Table 1: Dataset statistics

2021(Chakravarthi et al., 2021), have been organized on multiple languages for hostile content detection in the Indian context, but it is limited to textual classification. Extending those tasks further, the organizer of this shared task has organized a classification task to identify troll memes in Tamil by providing 2,967 memes. This paper illustrates the methodologies we used to identify Tamil troll memes, which helped us achieve **second** place in the final leader-board standings of shared tasks.

2 Related Work

This section discusses some of the text-based abusive content detection methods and briefly explains the multi-modal techniques used so far to detect malicious memes.

2.1 Text-based abusive content detection

Recently, a lot of work has been carried out to identify abusive speech using text from social media
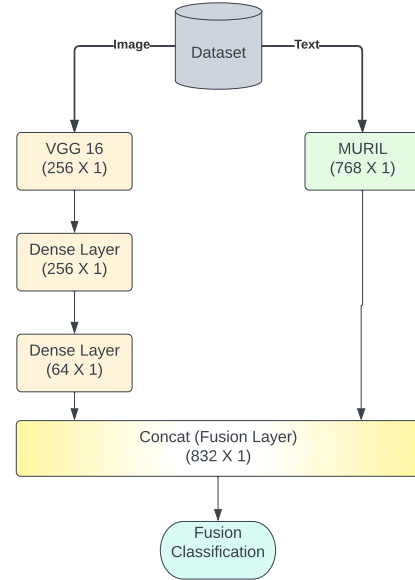

Figure 2: Our fusion model architecture with VGG16 and MURIL

posts (Das et al., 2020). In 2017, Davidson et al. (2017) made public a Twitter dataset in which thousands of tweets were labeled offensive, hate, and neither. The earlier efforts to create such classifiers used easy methods such as linguistic features, word n-grams, bag-of-words, etc (Davidson et al., 2017). With the availability of larger datasets, researchers have started utilizing complex models such as deep learning and graph embedding(Das et al., 2021b) strategies to improve the classifier performance of hate speech detection in social media posts. In 2018, Pitsilis et al. (2018) used deep learning-based models, such as the recurrent neural networks (RNNs), to detect the abusive tweets in the English language and witnessed that it was pretty effective in this task. In contrast, RNNs have been established to perform well with several language models. In addition, other neural network models, such as LSTM and CNN, have succeeded in detecting abusive speech (Goldberg, 2015; la Peña Sarracén et al., 2018). Recently, Transformer-based (Vaswani et al., 2017) language models such as BERT, (Devlin et al., 2019) are becoming quite prevalent in several downstream tasks, such as spam detection, classification(Das et al., 2021a; Banerjee et al., 2021), etc. Having observed the exceptional performance of these Transformer based models, we also utilize a Transformer based model, MURIL, which is pre-trained explicitly in Indian Languages.

Model	Accuracy	F1 Score(T)	F1 Score(w)	Precision(w)	Recall(w)
MURIL	0.556	0.637	<u>0.552</u>	<u>0.549</u>	0.556
VGG16	**0.587**	**0.736**	0.458	0.522	**0.587**
Fusion	<u>0.566</u>	<u>0.649</u>	**0.561**	**0.558**	<u>0.567</u>

Table 2: Performance Comparisons of Each Model. T: Troll Class. w: Weighted-Average. The best performance in each column is marked in **bold** and second best is <u>underlined</u>
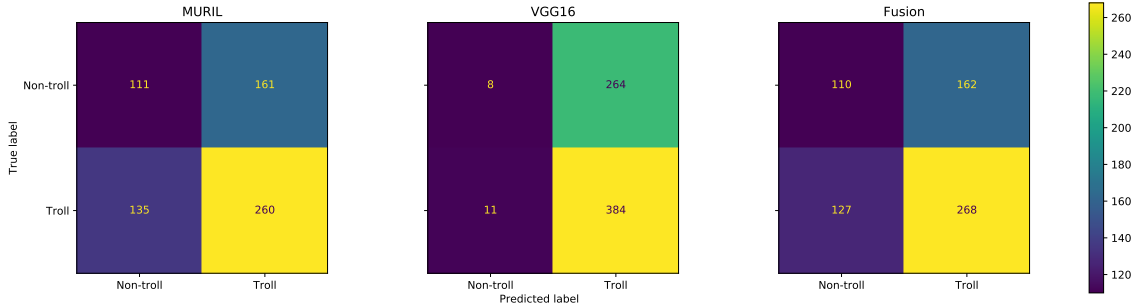

Figure 3: Confusion Matrix on Test Data for Each Model

2.2 Multi-modal abusive content detection

Lately, several datasets have been made public to the research community for abusive meme detection. Sabat et al.(2019) created a dataset of 5,020 memes for hate speech detection. The MMHS150K hate meme dataset developed by Gomez et al.(2020) is one of the enormous datasets collected from Twitter, consisting of 150K posts. Similarly, Facebook AI (Kiela et al., 2020) introduced another Hateful Meme dataset of 10K+ posts labeled hateful and non-hateful. As part of the hateful meme detection, an array of techniques with diverse architecture ranging from the text-based model, image-based model, and multi-modal models have been employed, including Glove embedding, FastText embedding, ResNet-152, VGG16, VisualBERT, UNITER, ViLBERT CC, V-BERT COCO(Pramanick et al., 2021; Chandra et al., 2021).

In this work, we use the VGG16 model, which is extensively used for several classification problems, to extract the features of all the memes and finally use it with the textual features to design our final model.

3 Dataset Description

The shared task on Troll Meme Classification in DravidianLangTech2022 (Suryawanshi et al., 2022) at ACL-2022 is based on a classification problem with the aim of moderating and minimizing the offensive/harmful content in social media. The objective of the shared task is to devise method-

ologies and vision-language models for troll meme detection in Tamil. We show the class distribution of the dataset(Suryawanshi et al., 2020b; Suryawanshi and Chakravarthi, 2021) in Table 1. The training set consisting of 2,300 memes (out of which 1,282 memes were labeled as troll meme) and the test set consisting of 667 memes. In addition, the latin transcribed texts were shared for all memes. We show example of both Troll and Non-troll memes in Figure 1.

4 Methodology

In this section, we discuss the different parts of the pipeline that we pursued for the detection of troll meme using the dataset.

4.1 Uni-modal Models

As part of our initial experiments, we created the following two uni-model models, one utilizing text features and the other using image-based features. **MURIL:** MURIL(Khanuja et al., 2021) is a transformer encoder having 12 layers with 12 attention heads and 768 dimensions. We used the pre-trained model which has been trained on 17 Indian languages and their transliterated counterparts using the MLM (masked language model) and the next sentence prediction (NSP) loss functions. The dataset used for pre-training is obtained by using the publicly available corpora from Wikipedia and Common Crawl. We pass all the texts associated with the meme via pre-trained MURIL [2] to get the

---

[2] https://huggingface.co/google/

53

768-dimensional feature vectors for each meme and then finally fed it to a output node for the final prediction.

VGG16: VGG16 (Simonyan and Zisserman, 2014) is a Convolutional Neural Network architecture, a variant of the VGG model which consists of 16 layers and is very appealing because of its very uniform architecture. We pass all the images(meme) via VGG16 and get the 256-dimensional feature vectors, then we pass it to the two dense layer of size 256 (with dropout of 0.5), 64 and finally fed it two the output node for the final prediction.

4.2 Fusion Model

The uni-modal models we used so far do not use the relation between the text and image present in the meme. To have better understanding between the text and image, we design a new MURIL+VGG16 fusion classifier, where we first concatenate the embedding from the both MURIL and VGG16 models discussed above, then we pass the concatenated embedding to a classification node for the final prediction. The detail of the pipeline is presented in Figure 2.

All the models are trained with binary cross-entropy loss functions and Adam optimizer for 20 epochs.

5 Results

Table 2 demonstrates the performance of each model. We observe among the uni-modal models, VGG16 has the highest Accuracy(MURIL: 0.556, VGG16: 0.587) and F1 score (MURIL: 0.637, VGG16: 0.736) for troll class. Though in terms of weighted F1 score(MURIL: 0.552, VGG16: 0.458), the text-based model MURIL performs better. When we fuse these two models, the fusion model achieves the highest weighted F1 score(0.561) among all the models. To further understand the model's weakness, we show the confusion matrix of each model in Figure 3. We observe that while the MURIL performs better on the Non-troll meme datapoints, VGG16 performs better on the troll meme datapoints. Whereas on the non-troll meme data points, VGG16 shows inferior performance. The fusion model brings the positive characteristics of both MURIL and VGG16 and performs the best by understanding better connections between the text and image of the memes.

`muril-base-cased`

6 Conclusion

In this shared task, we deal with a novel problem of detecting Tamil troll memes. We evaluated different uni-modal models and introduced a fusion model. We found that text-based model MURIL performs better on the Non-troll class, whereas VGG16 performs better on the Troll class. Ensembling these two models help us in gaining stable outcomes in both classes. We plan to explore further other vision-based models to improve classification performance as an immediate next step.

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.

Somnath Banerjee, Maulindu Sarkar, Nancy Agrawal, Punyajoy Saha, and Mithun Das. 2021. Exploring transformer based models to identify hate speech and offensive content in english and indo-aryan languages. *arXiv preprint arXiv:2111.13974*.

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.

Jonathan Bishop. 2014. Dealing with internet trolling in political online communities: Towards the this is why we can't have nice things scale. *International Journal of E-Politics (IJEP)*, 5(4):1–20.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 61–72, Kyiv. Association for Computational Linguistics.

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

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, RL Hariharan, John Philip McCrae, Elizabeth Sherly, et al. 2021. Findings of the shared task on offensive language identification in tamil, malayalam, and kannada. In *Proceedings of the first workshop on speech and language technologies for Dravidian languages*, pages 133–145.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "subverting the jewtocracy": Online antisemitism detection using multimodal deep learning. In *13th ACM Web Science Conference 2021*, pages 148–157.

Mithun Das, Somnath Banerjee, and Punyajoy Saha. 2021a. Abusive and threatening language detection in urdu using boosting based and bert based models: A comparative approach. *arXiv preprint arXiv:2111.14830*.

Mithun Das, Binny Mathew, Punyajoy Saha, Pawan Goyal, and Animesh Mukherjee. 2020. Hate speech in online social media. *ACM SIGWEB Newsletter*, (Autumn):1–8.

Mithun Das, Punyajoy Saha, Ritam Dutt, Pawan Goyal, Animesh Mukherjee, and Binny Mathew. 2021b. You too brutus! trapping hateful users in social media: Challenges, solutions & insights. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 79–89.

Thomas Davidson, Dana Warmsley, M. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

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.

Yoav Goldberg. 2015. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

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*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Gretel Liz De la Peña Sarracén, Reynaldo Gil Pons, C. E. Muñiz-Cuza, and P. Rosso. 2018. Hate speech detection using attention-based lstm. In *EVALITA@CLiC-it*.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Forum for Information Retrieval Evaluation*, pages 1–3.

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.

Georgios K. Pitsilis, H. Ramampiaro, and H. Langseth. 2018. Detecting offensive language in tweets using deep learning. *ArXiv*, abs/1801.04433.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455.

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.

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.

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *arXiv preprint arXiv:1910.02334*.

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.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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, and Paul Buitelaar. 2020a. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41.

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. 2020b. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

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.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

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.