CUET-NLP@DravidianLangTech-ACL2022: Exploiting Textual Features to Classify Sentiment of Multimodal Movie Reviews

Nasehatul Mustakim[Ψ], **Nusratul Jannat**[Ψ], **Md. Maruf Hasan**[Ψ], **Eftekhar Hossain**[$], **Omar Sharif**[$Ψ] and **Mohammed Moshiul Hoque**[Ψ]
[Ψ]Department of Computer Science and Engineering
[$]Department of Electronics and Telecommunication Engineering
[$Ψ]Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1604109, u1604115, u1604089}@student.cuet.ac.bd
{eftekhar.hossain, omar.sharif, moshiul_240}@cuet.ac.bd

Abstract

With the proliferation of internet usage, a massive growth of consumer-generated content on social media has been witnessed in recent years that provide people's opinions on diverse issues. Through social media, users can convey their emotions and thoughts in distinctive forms such as text, image, audio, video, and emoji, which leads to the advancement of the multimodality of the content users on social networking sites. This paper presents a technique for classifying multimodal sentiment using the text modality into five categories: highly positive, positive, neutral, negative, and highly negative. A shared task was organized to develop models that can identify the sentiments expressed by the videos of movie reviewers in both Malayalam and Tamil languages. This work applied several machine learning (LR, DT, MNB, SVM) and deep learning (BiLSTM, CNN+BiLSTM) techniques to accomplish the task. Results demonstrate that the proposed model with the decision tree (DT) outperformed the other methods and won the competition by acquiring the highest macro f_1-score of 0.24.

1 Introduction

Over the years, sentiment analysis has grown to an influential research domain with widespread commercial applications in the enterprise. To date, a significant number of applications have already been used for classifying or analyzing textual sentiment, including customer feedback (Pankaj et al., 2019; Hossain et al., 2021a), recommendation systems (Preethi et al., 2017), medicine analysis (Rajput, 2019), marketing, financial strategies (Jangid et al., 2018) and so on (Sampath et al., 2022; Ravikiran et al., 2022; Chakravarthi et al., 2022; Priyadharshini et al., 2022). Usually, people express their opinions, emotions, and ideas through text over the internet (Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021). However, the mode of communication is gradually shifting from unimodal to multimodal due to the rapid growth of all sorts of media content, including massive collections of videos (e.g., YouTube, Facebook, TikTok), audio clips, and images (Chakravarthi et al., 2020b; Bharathi et al., 2022). Classification of sentiment utilizing multiple modalities is becoming increasingly important and an exciting research topic. Multimodal sentiment analysis can analyze public opinions based on the speaker's language, facial gestures and acoustic behaviours, and voice's intensity (Ghanghor et al., 2021a,b; Yasaswini et al., 2021).

In recent years, a few studies have been performed on unimodal sentiment analysis concerning low-resource languages (e.g., Tamil, Malayalam, Bengali) (Priyadharshini et al., 2020, 2021; Kumaresan et al., 2021; Chakravarthi et al., 2021a, 2020a). The most challenging task in categorizing movie reviews is the interpretation of the words as most of the time, words are anticipated to the elements of a movie, not the opinion of the reviewer (Wöllmer et al., 2013; Mamun et al., 2022). Moreover, most language processing works mainly concentrate on high-resource languages like English, Arabic, and other European languages, where standard datasets are not available for low-resource languages. This work addresses the multimodal sentiment analysis from movie reviews 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. 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). Tamil is an official language of Tamil Nadu, Sri Lanka, Singapore, and the Union Territory of Puducherry in India. 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

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 191 - 198

Andaman and Nicobar Islands. 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 (Subalalitha, 2019; Srinivasan and Subalalitha, 2019; Narasimhan et al., 2018). Tamil is also the native language of Sri Lankan Moors. The term "Old Tamil" refers to the time of the Tamil language from the 10th century BC to the 8th century AD. The earliest Old Tamil documents are small inscriptions in Adichanallur dating from 905 BC to 696 BC. These inscriptions are written in Tamil-Brahmi, a variation of the Brahmi script. The Tolkppiyam, an early work on Tamil grammar and poetics, is the first extended book in Old Tamil, with layers dating back to the late 6th century BC (Anita and Subalalitha, 2019b,a; Subalalitha and Poovammal, 2018).

The significant contributions of this work illustrate as follows,

- Developed various machine learning (ML) and deep learning (DL) based techniques to classify the sentiments into five classes (i.e., highly positive, positive, neutral, negative, and highly negative) for the Tamil language.
- Investigated the performance of the developed models with careful experimentation and error analysis.

2 Related Work

With the rapid popularization of social media, people's eagerness to express their views or opinions on these mediums increases sharply. However, sentiment analysis in low-resource languages is still rudimentary due to the scarcity of standard corpora and limited language processing tools. Few ML-based methods such as support vector machine (SVM), logistic regression (LR), naive Bayes (NB) have been used to analyze the textual sentiment in Bengali (Naeem et al., 2020; Sharif et al., 2019). Thavareesan and Mahesan (2019) performed sentiment analysis on five different Tamil text corpora using various ML techniques with BoW and TF-IDF features, which obtained the highest accuracy of 79% with Extreme Gradient Boosting with FastText. Singla et al. (2017) experimented with NB, DT, and SVM with the 10-fold cross-validation achieving 81.75% accuracy with SVM. Phani et al. (2016) used the SAIL corpus to assess the sentiment of tweets. They achieved the best performance in Tamil with NB and Hindi, Bengali with LR. The performance of these models is not very impressive as they were unable to capture semantic and contextual information in the text. The major obstacles are the inherent ambiguity of the language, the computational complexity of exploring large amounts of content, resource-poor language problems, and the contextual understanding of natural language (Zhou et al., 2021; Hossain et al., 2021b).

Different DL models were applied to Malayalam tweets to classify them into positive and negative where Gated Recurrent Unit (GRU) achieved the highest accuracy (Soumya and Pramod, 2019). Several approaches, including lexicon, supervised ML, hybrid, were experimented on Tamil texts (Thavareesan and Mahesan, 2019; Phani et al., 2016; Prasad et al., 2016). Abid et al. (2019) proposed a joint structure that combines CNN and RNN layers along with GloVE embeddings for capturing long-term dependencies of Twitter data. In another similar work, the sentiment lexicon is used to enhance the sentiment features, and then CNN-GRU networks are combined to analyze the sentiment of product reviews (Yang et al., 2020). Pranesh and Shekhar (2020) presented 'MemeSem' where VGG19 is used for visual and BERT for textual modality to analyze the sentiment of memes. MemeSem outperformed all the unimodal and multimodal baseline by 10.69% and 3.41% respectively. Recently, the CNN + Bi-LSTM model (Xuanyuan et al., 2021) has been employed to classify the sentiment of Twitter data and gained the highest accuracy of 90.2% for binary classification (positive and negative).

3 Dataset Description

The dataset we have used for this task is provided by the shared task organizers[1]. It is a collection of videos, audios, and text accumulated from YouTube and manually annotated. The dataset is divided into three sets (i.e., train, validation, and test) and annotated into five classes: highly positive, positive, neutral, negative, and highly negative. The dataset consists of a total of 134 videos, out of which 70 are Malayalam videos and the remaining 64 are Tamil videos (Chakravarthi et al., 2021b; Premjith et al., 2022). The length of the videos is between 1 minute to 3 minutes. Table 1 presents the distribution of the dataset. Table 2 shows the

[1]https://competitions.codalab.org/competitions/36406

number of samples in each category. This work dealt with the Tamil language dataset only.

This work employed textual features to address the assigned task. Participants have the freedom to utilize unimodal data (i.e., video, audio, or text) or multimodal (i.e., a combination of any two or three modalities) features to perform the classification task. Each text is provided in the *.docx file format. Therefore, we extracted all the texts from documents before starting the experimentation and evaluation (Section 4 provides a detailed description).

4 Methodology

The objective of the task is to classify the underlying sentiments from movie reviews using video, audio, and text modalities. However, we have used only textual data to attain this goal. Initially, texts were taken from the *.docx files and preprocessed for further use. Subsequently, feature extraction techniques are applied to get the features. Finally, the extracted features are utilized to develop ML and DL models to perform the classification task. Figure 1 illustrates an abstract view of the sentiment analysis model.

4.1 Feature Extraction

TF-IDF technique has been used to extract the unigram textual features for developing the ML models. On the other hand, Word2vec and FastText (Grave et al., 2018) embeddings are used to train the DL models. This work used pre-trained word vectors which were trained on Common Crawl and Wikipedia texts with a dimension of 300. In case of Word2Vec embeddings, we employed the Keras embedding layer to generate the vectors of length 260.

4.2 Classifiers

Four popular ML models (LR, DT, SVM, and MNB) have been developed to address the task using the 'scikit-learn' library. We devised the LR technique with the regularization parameter (C=5) and 'lbfgs' optimizer. The smoothing parameter (α) settled to 1.0 in the case of MNB. Meanwhile linear kernel with balanced class weight and C = '2' was used for SVM. However, the DT model parameters are: class weight = 'balanced' and criterion = 'gini'.

The task investigates two DL models (CNN and BiLSTM) and their combination (CNN+BiLSTM). For BiLSTM, we utilize the features extracted by the Word2Vec with an embedding dimension of 100. The BiLSTM consists of 128 units, and the dropout rate is set to 0.2 to reduce the overfitting. Finally, features are flattened and passed to the softmax layer for prediction. The model is trained for 30 epochs with a batch size of 32. For the CNN+BiLSTM based approach, we have used pre-trained FastText embedding. The output of Conv1D having 128 filters was fed to the max-pooling layer to downsample the features. These features were propagated to a bidirectional LSTM layer with 128 units. The model was also trained with a batch size of 32 for 30 epochs. For both models, the learning rate settled to 0.001, and 'sparse categorical crossentropy' is used to evaluate the loss. Keras callback function is utilized to save the best model during training used for the final evaluation. Table 3 shows the summary of hyper-parameters used in the experiment.

5 Results and Analysis

Table 4 illustrates the performance of the models in terms of precision, recall and $f1$-score measures. The $f1$-score is used to decide the superiority of the model.

The results demonstrate that the DT model outperformed the other ML and DL models. The DT models showed 86.6% of increased performance compared to the ML models and improved by 140% than the best DL model (CNN+BiLSTM). The other ML models, such as LR, SVM, and MNB, were classified all test instances as the positive class. The BiLSTM with Word2Vec features predicts maximum reviews as neutral ones having a macro f_1 score of 0.07. However, after using the pre-trained word embedding with combined CNN and BiLSTM model, the macro f_1-score has grown to 0.10. Thus, the model shows an increase in performance using pre-trained word embedding. However, it cannot beat the DT model developed based on the TF-IDF features.

Table 5 shows the class-wise $f1$-score of the models. The PS class obtained the maximum f_1-score (0.80) in DT model because this class contained the highest number of instances in the dataset. In contrast, the HPS and HNE classes showed the lowest f_1-score (0.0). That means any model cannot predict any sample of HPS and HNE classes due to the minimal number of samples in the dataset. In particular, HPS class contained only

	Tamil			Malayalam			
Dataset	Train	Test	Validation	Train	Test	Validation	Size(MB)
Video	44	10	10	50	10	10	1111.8
Audio	44	10	10	50	10	10	162.2
Text	44	10	10	50	10	10	1.003
Total	132	30	30	150	30	30	1275.003

Table 1: Statistics of dataset

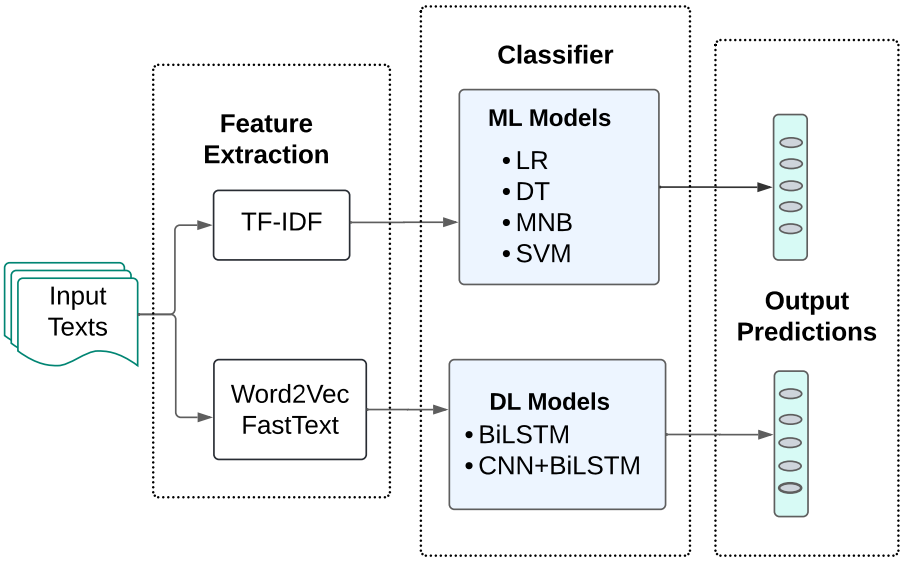

Figure 1: An overview of the sentiment analysis model

Class Label	Tamil	Malayalam
HPS	9	8
PS	39	38
NT	8	8
NE	12	5
HNE	2	5
Total	64	70

Table 2: Class-wise data sample distribution for each language. Here HPS, PS, NT, NE, and HNE indicate highly positive, positive, neutral, negative, and highly negative, respectively

Hyperparameters	Values
Dropout rate	0.2
Optimizer	'adam'
Learning rate	0.001
Epoch	30
Batch size	32

Table 3: Summary of tuned hyperparameters

9 samples whereas, HNE consisting only 2.

5.1 Error Analysis

Table 4 confirmed that the DT model is the best for the assigned task. The model's performance is further investigated using the confusion matrix (Figure 2) with detailed error analysis.

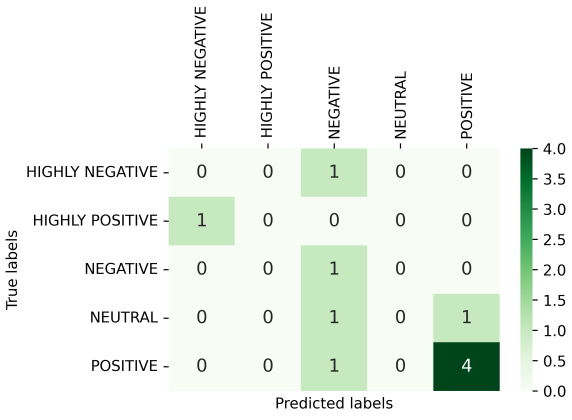

Figure 2: Confusion matrix of the best model (DT)

The model can genuinely predict 4 positive re-

Approach	Classifier	Precision	Recall	$f1$-score
ML models	LR	0.20	0.10	0.13
	DT	**0.21**	**0.36**	**0.24**
	MNB	0.20	0.10	0.13
	SVM	0.20	0.10	0.13
DL models	BiLSTM	0.20	0.04	0.07
	CNN+BiLSTM	0.12	0.09	0.10

Table 4: Performance comparison of various models on the test set

Classifier	HPS	PS	NT	NE	HNE
LR	0.0	0.67	0.0	0.0	0.0
DT	0.0	0.80	0.0	0.40	0.0
MNB	0.0	0.67	0.0	0.0	0.0
SVM	0.0	0.67	0.0	0.0	0.0
BiLSTM	0.0	0.0	0.33	0.0	0.0
CNN+BiLSTM	0.0	0.50	0.0	0.0	0.0

Table 5: Class-wise f_1-score of classifiers

views among 5 reviews. It miss-classifies only one neutral review as the positive class. The model predicted the negative class correctly but miss-classified the highly negative, neutral, and positive class as a negative one. The DT model is failed to predict the highly positive and the neutral classes. The model's low performance can be due to the lack of training data samples. Since this work considered the text modality only, it might miss some essential features associated with video and audio samples. The use of multimodal features might improve the performance of the system.

6 Conclusion

This paper investigated several ML and DL techniques to address the sentiment analysis task on a multimodal dataset in Tamil. Although the provided dataset included text, audio, and video modalities, this work considered the text modality. Results indicate that the DT model outperformed the other ML and DL models obtaining the maximum macro f_1-score (0.24). Surprisingly, DL models showed poor performance compared to their ML counterparts. Since the dataset is too small and crooked, data oversampling techniques or any open source large corpora can be used to create synthetic data to improve performance. The scarcity of training samples might cause lower scores. Moreover, excluding the video and audio features might also hurt the model's performance. We aim to incorporate multimodal features (video, audio, text) and address the task with the recent transformer-based models (i.e., IndicBERT, mBERT, XML-R, MuRIL) in the future.

Acknowledgements

This work supported by the ICT Innovation Fund, ICT Division, Ministry of Posts, Telecommunications and Information Technology, Bangladesh.

References

Fazeel Abid, Muhammad Alam, Muhammad Naveed Yasir, and Chen Li. 2019. Sentiment analysis through recurrent variants latterly on convolutional neural network of twitter. *Future Gener. Comput. Syst.*, 95:292–308.

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.

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. 2020a. 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, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020b. 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, KP Soman, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kingston Pal Thamburaj, John P McCrae, et al. 2021b. Dravidian-multimodality: A dataset for multi-modal sentiment analysis in Tamil and Malayalam. *arXiv preprint arXiv:2106.04853*.

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.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H. Sarker. 2021a. Sentilstm: A deep learning approach for sentiment analysis of restaurant reviews. In *Hybrid Intelligent Systems*, pages 193–203, Cham. Springer International Publishing.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2021b. Sentiment polarity detection on bengali book reviews using multinomial naïve bayes. In *Progress in Advanced Computing and Intelligent Engineering*, pages 281–292, Singapore. Springer Singapore.

Hitkul Jangid, Shivangi Singhal, Rajiv Ratn Shah, and Roger Zimmermann. 2018. Aspect-based financial sentiment analysis using deep learning. *Companion Proceedings of the The Web Conference 2018*.

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.

Md Mashiur Rahman Mamun, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Classification of textual sentiment using ensemble technique. *SN Computer Science*, 3(1):1–13.

Saud Naeem, Doina Logofătu, and Fitore Muharemi. 2020. Sentiment analysis by using supervised machine learning and deep learning approaches. pages 481–491.

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.

Pankaj, Prashant Pandey, Muskan, and Nitasha Soni. 2019. Sentiment analysis on customer feedback data: Amazon product reviews. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 320–322.

Shanta Phani, Shibamouli Lahiri, and Arindam Biswas. 2016. Sentiment analysis of tweets in three Indian languages. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 93–102, Osaka, Japan. The COLING 2016 Organizing Committee.

Raj Ratn Pranesh and Ambesh Shekhar. 2020. Memesem:a multi-modal framework for sentimental analysis of meme via transfer learning.

Sudha Shanker Prasad, Jitendra Kumar, Dinesh Kumar Prabhakar, and Sachin Tripathi. 2016. Sentiment mining: An approach for Bengali and Tamil tweets. pages 1–4.

G. Preethi, P. Venkata Krishna, Mohammad S. Obaidat, V. Saritha, and Sumanth Yenduri. 2017. Application of deep learning to sentiment analysis for recommender system on cloud. In *2017 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 93–97.

B Premjith, Bharathi Raja Chakravarthi, B Bharathi, Malliga Subramanian, K.P Soman, Dhanalakshmi Vadivel, K Sreelakshmi, Arunaggiri Pandian, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on Multimodal Sentiment Analysis in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

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.

Adil E. Rajput. 2019. Natural language processing, sentiment analysis and clinical analytics. *ArXiv*, abs/1902.00679.

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.

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.

Omar Sharif, M. Hoque, and E. Hossain. 2019. Sentiment analysis of bengali texts on online restaurant reviews using multinomial naïve bayes. *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–6.

Zeenia Singla, Sukhchandan Randhawa, and Sushma Jain. 2017. Sentiment analysis of customer product reviews using machine learning. pages 1–5.

S Soumya and K V Pramod. 2019. Sentiment analysis of Malayalam tweets using different deep neural network models-case study. pages 163–168.

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.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

Minzheng Xuanyuan, Le Xiao, and Mengshi Duan. 2021. Sentiment classification algorithm based on multi-modal social media text information. *IEEE Access*, 9:33410–33418.

Li Yang, Ying Li, Jin Wang, and R. Simon Sherratt. 2020. Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning. *IEEE Access*, 8:23522–23530.

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.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2021. Challenges in automated debiasing for toxic language detection.