CUET_Binary_Hackers@DravidianLangTech-EACL 2024: Sentiment Analysis using Transformer-Based Models in Code-Mixed and Transliterated Tamil and Tulu

Asrarul Hoque Eusha, Salman Farsi, Ariful Islam
Jawad Hossain, Shawly Ahsan and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{asrar2860, salman.cuet.cse, arif.cse18cuet}@gmail.com
{u1704039, u1704057}@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

Textual Sentiment Analysis (TSA) delves into people's opinions, intuitions, and emotions regarding any entity. Natural Language Processing (NLP) serves as a technique to extract subjective knowledge, determining whether an idea or comment leans positive, negative, neutral, or a mix thereof toward an entity. In recent years, it has garnered substantial attention from NLP researchers due to the vast availability of online comments and opinions. Despite extensive studies in this domain, sentiment analysis in low-resourced languages such as Tamil and Tulu needs help handling code-mixed and transliterated content. To address these challenges, this work focuses on sentiment analysis of code-mixed and transliterated Tamil and Tulu social media comments. It explored four machine learning (ML) approaches (LR, SVM, XGBoost, Ensemble), four deep learning (DL) methods (BiLSTM and CNN with fastText and Word2Vec), and four transformer-based models (m-BERT, MuRIL, L3Cube-IndicSBERT, and Distilm-BERT) for both languages. For Tamil, L3Cube-IndicSBERT and ensemble approaches outperformed others, while m-BERT demonstrated superior performance among the models for Tulu. The presented models achieved the 3^{rd} and 1^{st} ranks by attaining macro F1-scores of 0.23 and 0.58 in Tamil and Tulu, respectively.

1 Introduction

TSA plays a crucial role in production and content creation, offering insights into how consumers perceive offerings and providing immediate feedback. Utilizing the internet and social media, studies focus on sentiment analysis in monolingual comments, achieving high accuracy levels (Wankhade et al., 2022). While research addresses multilingual, code-mixed, and code-switched text, extensive exploration focuses on well-resourced languages like English and Chinese (Xu et al., 2022). In contrast, low-resourced languages such as Tamil and Tulu need more exploration, particularly in code-mixed and code-switched contexts. The challenge arises from comments written in English letters, like Romanized Tamil or Tulu, attracting recent attention from academia (S. K. et al., 2024a).

ML and DL approaches like LSTM and BiLSTM, and transformer-based models like BERT, m-BERT, XLMR, and Distilm-BERT have been extensively studied for monolingual and multilingual text, encompassing code-mixed, code-switched, and Romanized formats in low-resource languages (Sharif et al., 2019; Kalaivani and Thenmozhi, 2021). Researchers focus on enhancing accuracy, particularly in Tamil-English and Tulu-English. Transformer-based models exhibit proficiency in handling sequence dependencies, motivating deeper exploration in these languages for improved contextual understanding.

The critical contributions of this research work are outlined below:

- Investigate several ML, DL, and transformer-based models with fine-tuning to classify sentiment in Tamil and Tulu languages into four classes: Positive, Neutral, Mixed, and Negative.
- Explored the suitable model for identifying textual sentiment from Tamil and Telegu texts on the available dataset.

2 Related Work

Understanding audience feedback is critical for social media content creators, fostering self-improvement and broader outreach. Similarly, grasping user sentiment in the restaurant industry is vital for improving services and cuisine quality (Sharif et al., 2019). The study examines several

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 205–211

ML models, such as DT, RF, and MNB models, for classifying user reviews, where MNB achieved the highest accuracy (80.48%). SA on Bengali book reviews using a MNB attained an accuracy of 84% (Hossain et al., 2021). Moreover, a study on TSA in Tamil and Tulu code-mixed texts, utilizing SVM and ensemble models with fastText and TF-IDF, obtained F1 scores of 0.14 and 0.204, respectively (Rachana et al., 2023). Numerous DL methods have been explored for TSA across various high-resourced languages. For instance, an Arabic aspect-based sentiment analysis employed bidirectional GRU, achieving F1 scores of 70.76% and 83.98% for aspect-based sentiment and sentiment polarity classification, respectively (Abdelgwad et al., 2022). Additionally, a fusion-based deep learning model analyzed sentiment in COVID-19 tweets, outperforming individual models like CNN, BiGRU, and DistilBERT (Basiri et al., 2021).

Recent studies have explored transformer-based models for sentiment analysis. For instance, a proposed aspect-category sentiment analysis based on RoBERTa integrated 1D CNN, cross-attention, document attention, and fully connected layers for classification (Liao et al., 2021). Another study introduced a BERT-based sentiment analysis model focusing on software engineering, fine-tuning BERT, ALBERT, and RoBERTa models, and employing an ensemble of these models (Batra et al., 2021). Additionally, a hybrid model combining RoBERTa and LSTM layers demonstrated effectiveness in sentiment analysis (Tan et al., 2022). In multilingual sentiment analysis, a study utilizing multilingual BERT achieved notable F1 scores in Tamil, Malayalam, and Kannada, including English code-mixed text (Kalaivani and Thenmozhi, 2021). Similarly, sentiment analysis on a code-mixed Tamil-English dataset using transformer-based models revealed the superiority of XLM-RoBERTa over BERT and RoBERTa models (Sangeetha and Nimala, 2023). Furthermore, an investigation into sentiment analysis in Tamil and Tulu code-mixed text highlighted the efficacy of fine-tuned transformer-based models across various scenarios (Hegde et al., 2023).

3 Task and Dataset Descriptions

In the shared task on 'Sentiment Analysis in Tamil and Tulu' (S. K. et al., 2024b), participants were tasked with exploring distinct models for each language. Using the provided datasets, we conducted multi-class classification to discern whether a given comment falls into categories such as 'Positive,' 'Neutral,' 'Negative,' or 'Mixed' within the Tamil-English code-mixed dataset developed by Chakravarthi et al. (2020). Similarly, Hegde et al. (2022) developed Tulu-English code-mixed dataset SA containing the same classes as Tamil. These gold standard datasets encompass code-mixed Tamil-English, Romanized Tamil, code-mixed Tulu-English, and Romanized Tulu texts. Each corpus consists of distinct training, validation, and test sets.

Table 1 displays Tamil's dataset details for sentiment analysis. In this task, the training set contains 90704 unique words, and the test set contains 4832 unique words, with 2330 out-of-vocabulary words. The average lengths of samples are 10, 10, and 13 in the training, validation, and test sets, respectively.

Data	Class	SC	UWC	OOV	AL
Train	Positive	20,070	90,704		10
	Neutral	5,628			
	Mixed	4,020			
	Negative	4,271			
Validation	Positive	2,257	16,111	2,330	10
	Neutral	611			
	Mixed	438			
	Negative	480			
Test	Positive	73	4,832		13
	Neutral	137			
	Mixed	101			
	Negative	338			

Table 1: Detailed dataset statistics of sentiment analysis in Tamil. The acronyms SC, UWC, OOV, and AL denote sample count, unique word count, out-of-vocabulary words, and average sample length, respectively.

Table 2 presents the dataset details for the sentiment analysis task in Tulu. Here, the majority of training samples belong to the positive class. The unique word counts in the training, validation, and test sets are 18056, 2004, and 2145, respectively. These statistics indicate that out of 2145 unique words in the test samples, 1094 are out-of-vocabulary words unseen by the models during training. The average sample lengths are 7, 6, and 7 in the training, validation, and test sets.

Notably, in both the Tamil and Tulu test sets, the 'Positive' class accounted for nearly two-thirds of the samples in the training set, resulting in highly imbalanced datasets.

Data	Class	SC	UWC	OOV	AL
Train	Positive	3,352	18,056		7
	Neutral	1,854			
	Mixed	1,041			
	Negative	698			
Validation	Positive	231	2,004	1,094	6
	Neutral	124			
	Mixed	90			
	Negative	55			
Test	Positive	248	2,145		7
	Neutral	140			
	Mixed	70			
	Negative	43			

Table 2: Detailed dataset statistics of sentiment analysis in Tulu.

4 Methodology

The developed SA method starts with the text undergoing preprocessing to eliminate unwanted special characters, spaces, line breaks, and emojis. For ML, we employed TF-IDF (Takenobu, 1994), while pre-trained fastText (Bojanowski et al., 2017) and Word2Vec (Mikolov et al., 2013) word embeddings were utilized for DL. Figure 1 shows an abstract process and employed models for TSA.

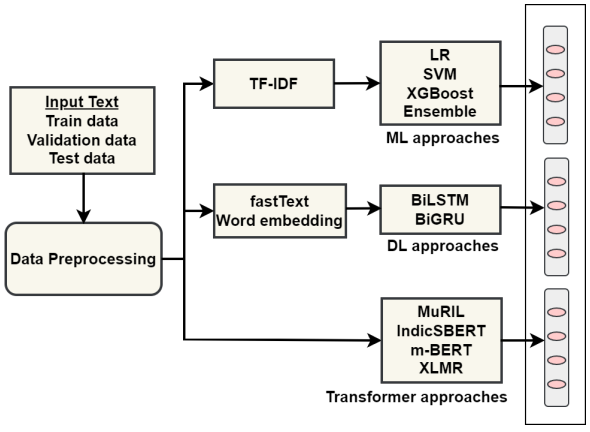

Figure 1: Schematic process of sentiment analysis in Tamil and Tulu

TF-IDF, a statistical method, is an information retrieval technique for words within text commonly used in NLP tasks. In this work, we used word-level n-grams for feature extraction using TF-IDF. On the other hand, word embeddings transform words into numerical representations, enabling the capture of semantic meaning and relationships within a continuous vector space. This work utilized the 300 dimensions for both fastText and Word2Vec.

4.1 Classifiers

This work explored several ML, DL, and transformer-based approaches to classify sentiment from Tamil and Tule texts.

ML Approaches: Logistic Regression (LR) is developed using a 'balanced' class weight approach to handle imbalanced datasets. It employs an 'l2' penalty for L2 regularization and the 'lbfgs' solver, with a maximum iteration set to 200 and a regularization parameter C kept at 1.0. Support Vector Machine (SVM) was used with a 'linear' kernel and applied 'balanced' class weights. The ensemble method combined multiple ML-based classifiers to generate a new classifier. Its superior performance in classification tasks over individual ML models has already been established (Roy et al., 2018). We constructed an ensemble method using DT, RF, SVM, and LR, implementing majority voting for prediction. We set the class weight to 'balanced' and utilized the 'gini' criterion for RF and DT. For RF, we used a value of 100 for 'n_estimators.' The parameters for SVM and LR remain consistent with their previous settings. XGBoost was employed with the 'multi:softmax' objective, employing 'n_estimators' of 200, a learning rate of 0.3, and a maximum depth of 6.

DL Approaches: This work developed the BiLSTM model (Hameed and Garcia-Zapirain, 2020) using Word2Vec and fastText. They consist of a single BiLSTM layer featuring 100 units. For classification within the output layer, we applied softmax activation. During training, we set a learning rate of $3e^{-3}$, a batch size of 32, utilized 15 epochs, and introduced a dropout of 0.2 to prevent overfitting. We employed the CNN model (O'Shea and Nash, 2015) utilizing Word2Vec and fastText. Our approach involved a single layer of CNN, comprising 128 units with max-pooling. All other parameters were configured identically to those of the BiLSTM.

Transformer-based Appraches: We selected several pre-trained transformer models available through HuggingFace[1] (Wolf et al., 2019), including m-BERT, Distil-mBERT, L3Cube-IndicSBERT, and MuRIL. The task dataset contains code-mixed multilingual text, so these models proved particularly suitable. We fine-tuned these four models, adjusting hyperparameters to attain optimal results, utilizing a maximum length of 50, a batch size of 16, and a learning rate of $5e^{-6}$. Also utilized the

[1]https://huggingface.co/

number of epochs 10 and 15 for Tamil and Tulu, respectively.

MuRIL is a pre-trained BERT model on 17 major Indian languages, including their transliterated counterparts (Khanuja et al., 2021). L3Cube-IndicSBERT (Deode et al., 2023) utilizes the MuRIL approach, trained on NLI datasets encompassing 10 primary Indian languages, Tamil and Tulu included. m-BERT (Devlin et al., 2018) is a pre-trained model trained on a vast multilingual corpus, covering 104 languages in a self-supervised manner. We employed 'bert-base-multilingual-cased.' We also employed Distil-mBERT, a multilingual variant of DistilBERT (Sanh et al., 2019). It serves as a smaller and faster iteration of m-BERT. We used 'distilbert-base-multilingual-cased.'

5 Results and Analysis

This section details the performance analysis of the proposed system, trained and evaluated on separate corpora. The best models were employed for test data predictions and evaluated using the macro-averaged F1-score. Table 3 shows the results of all ML, DL, and transformer-based models.

The ensemble approach in Tamil sentiment analysis outperformed most DL and transformer models, except L3Cube-IndicSBERT, achieving a precision (P) of 0.28, a recall (R) of 0.26, and a macro F1-score (F) of 0.23. Meanwhile, L3Cube-IndicSBERT achieved a similar macro F1-score with a precision of 0.24 and a recall of 0.28. XGBoost showed poor performance, possibly due to overfitting. In Tulu code-mixed sentiment analysis, m-BERT excelled with precision of 0.59, recall of 0.58, and a macro F1-score of 0.58, surpassing other models.

5.1 Error Analysis

The best-performed models, ensemble (for Tamil) and m-BERT (for Tulu), are further investigated using quantitative and qualitative analysis for more insights regarding their performance.

Quantitative Analysis: Figure 2 illustrates that the model misclassified a significant portion of the test samples in the TSA task in Tamil. This misclassification stems from the fact that while two-thirds of the training samples were 'Positive,' the test set comprised half as 'Negative,' a classless frequency in the training set. Consequently, the models predominantly predicted test samples as 'Positive,' leading to increased misclassification.

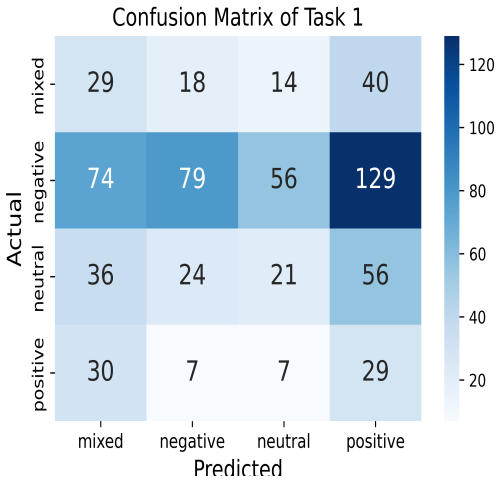

Figure 2: Confusion matrix for sentiment analysis in Tamil using an ensemble of ML techniques

The developed m-BERT model (for Tulu) shows promise, but there is room for improvement. Figure 3 shows that most 'Mixed' and 'Negative' test samples and a notable portion of 'Neutral' samples are misclassified. Imbalanced datasets during training could be the cause. Adjusting class weights could enhance results.

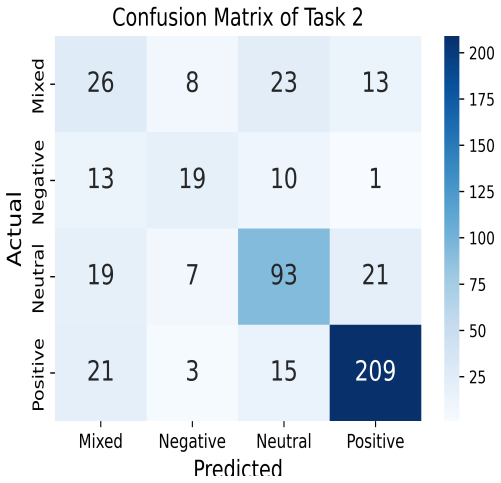

Figure 3: Confusion matrix for sentiment analysis in Tulu using m-BERT

Qualitative Analysis: Figure 4 displays predicted outcomes from the ensemble model for text samples 2 and 3, which align with the actual classes in the TSA task in Tamil. However, for samples 1, 4, and 5, misclassification of text occurs. Figure 5, concerning the TSA in Tulu, shows that m-BERT misclassified samples 2, 3, and 5. Whereas samples 1 and 4 were predicted to match the actual labels. It is noted that English translations of Tamil texts are

Methods	Classifiers	Tamil			Tulu		
		P	R	F	P	R	F
ML	LR	0.29	0.26	0.14	0.53	0.53	0.53
	SVM	0.27	0.24	0.12	0.53	0.54	0.53
	XGBoost	0.22	0.21	0.09	0.24	0.24	0.23
	Ensemble	0.28	0.27	**0.23**	0.53	0.54	0.53
DL	BiLSTM + WV	0.27	0.27	0.22	0.27	0.26	0.26
	BiLSTM + FT	0.27	0.24	0.19	0.25	0.24	0.21
	CNN + WV	0.26	0.25	0.21	0.26	0.26	0.26
	CNN + FT	0.25	0.25	0.19	0.24	0.23	0.21
Transformer	MuRIL	0.25	0.29	0.21	0.54	0.53	0.53
	Indic-SBERT	0.24	0.28	0.23	0.55	0.56	0.56
	m-BERT	0.29	0.26	0.20	0.59	0.58	**0.58**
	Distil-mBERT	0.29	0.26	0.18	0.53	0.53	0.52

Table 3: Evaluation results on the test set using various ML, DL, and transformer-based models. P, R, F1, WV, and FT represents precision, recall, macro F1-score, Word2Vec, and fastText respectively

Text Sample	Actual	Predicted
Sample 1. இப்போது இந்த 9தொல்லை அதிகமாக ஆயிருச்சு (Now this 9 trouble is more)	Negative	Mixed
Sample 2. இது புதுவகை கொள்ளை கூட்டம் (This is a new type of robbery)	Neutral	Neutral
Sample 3. என்று சொல்லாதீர்கள் திருநங்கை என்று கூறுங்கள் முதலில்.(Don't say that, say transgender first)	Positive	Positive
Sample 4. Romba thollai pannuthunga yaarume ketka matangala (No one bothers to listen to religions.)	Negative	Positive
Sample 5. காவல்துறை கண்டிப்பாக நடவடிக்கை எடுக்க வேண்டும். (Police must take action.)	Neutral	Positive

Figure 4: Some predicted samples in Tamil using ensemble

Text Sample	Actual	Predicted
Sample 1. ಥರ್ಡ್ ಕ್ಲಾಸ್ ಕಾಮಿಡಿ (Third Class Comedy)	Negative	Negative
Sample 2. ಈ ಪೊಣ್ಣಗ್ ಪನ್ಪಿನಗುಲ್ ಯೀ ಲ್ಲಾ ಇಜ್ಜಿರ (This girl is very beautiful.)	Negative	Mixed
Sample 3. ಕುಸಾಸ್ ಇತ್ತುಂಡು. ಪ್ರಾಕ್ಟಿಸ್ ಕಮ್ಮಿ. ಒಟ್ಟುಗ್ ಓಕೆ. (Having fun. Practice together. All okay.)	Mixed	Neutral
Sample 1. ಸರ್ ಮಸ್ತ್ ಧನ್ಯವಾದ ರಡ್ ಜನಕ್ಲೆಗ್ಲ ನಿಕ್ಲೆಗ್ ಬೊಕ್ಕ ಈರ್ನ ಟೀಮ್ಗುಲ ಕಾಫಿ ನಾಡುನ್ ಎಡ್ಡೆ ತೋಜದರ್ಯ (Sir, thank you very much, for those people who came out, and for the team.)	Positive	Positive
Sample 5. Vol ittar anna onji vaara.. (Can you give me a little time..)	Neutral	Positive

Figure 5: Few predictions in Tulu using m-BERT

accomplished with Google Translate, and ChatGPT does English translations of Tulu texts.

6 Limitations

The developed systems suffered some significant limitations.

- The DL and transformer-based models rely on extensive training data. Limited or biased datasets can notably impact results, especially when dealing with diverse or uncommon sentiment expressions.

- The system encountered difficulty effectively managing class imbalances in both tasks.

7 Conclusion and Future Work

This paper explored sentiment analysis on code-mixed Tamil and Tulu datasets, investigating various ML, DL, and transformer-based models. For improved performance, this work delved into experiments with transformer-based models such as m-BERT, L3Cube-IndicSBERT, Distilm-BERT, and MuRIL. In the case of Tamil, both the ensemble and L3Cube-IndicSBERT outperformed, achieving macro F1-scores of 0.23. Conversely, m-BERT exhibited superior performance for Tulu with a macro F1-score of 0.58.

Future research should explore adaptive learning rates, ensembles comprising different BERT models, and advanced word embedding techniques (ELMO, ULMFiT). Lastly, developing a fair model applicable across languages can improve accuracy.

References

Mohammed M Abdelgwad, Taysir Hassan A Soliman, Ahmed I Taloba, and Mohamed Fawzy Farghaly. 2022. Arabic aspect based sentiment analysis using bidirectional GRU based models. *Journal of King Saud University-Computer and Information Sciences*, 34(9):6652–6662.

Mohammad Ehsan Basiri, Shahla Nemati, Moloud Abdar, Somayeh Asadi, and U Rajendra Acharrya. 2021. A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets. *Knowledge-Based Systems*, 228:107242.

Himanshu Batra, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. 2021. BERT-based sentiment analysis: A software engineering perspective. In *Database and Expert Systems Applications: 32nd International Conference, DEXA 2021, Virtual Event, September 27–30, 2021, Proceedings, Part I 32*, pages 138–148. Springer.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. *arXiv preprint arXiv:2006.00206*.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3Cube-IndicSBERT: A simple approach for learning cross-lingual sentence representations using multilingual BERT. *arXiv preprint arXiv:2304.11434*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zabit Hameed and Begonya Garcia-Zapirain. 2020. Sentiment classification using a single-layered BiLSTM model. *Ieee Access*, 8:73992–74001.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed Tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40, Marseille, France. European Language Resources Association.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Cn, SK Lavanya, Durairaj Thenmozhi, Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023. Findings of the shared task on sentiment analysis in Tamil and Tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 64–71.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2021. Sentiment polarity detection on Bengali book reviews using multinomial naive bayes. In *Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2020*, pages 281–292. Springer.

Adaikkan Kalaivani and Durairaj Thenmozhi. 2021. Multilingual Sentiment Analysis in Tamil Malayalam and Kannada code-mixed social media posts using MBERT. In *FIRE (Working Notes)*, pages 1020–1028.

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*.

Wenxiong Liao, Bi Zeng, Xiuwen Yin, and Pengfei Wei. 2021. An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa. *Applied Intelligence*, 51:3522–3533.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Keiron O'Shea and Ryan Nash. 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

K Rachana, M Prajnashree, Asha Hegde, and HL Shashirekha. 2023. MUCS@ DravidianLangTech2023: Sentiment Analysis in Code-mixed Tamil and Tulu Texts using fastText. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 258–265.

Arjun Roy, Prashant Kapil, Kingshuk Basak, and Asif Ekbal. 2018. An ensemble approach for aggression identification in English and Hindi text. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 66–73.

Lavanya S. K., Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, and Rajkumar Charmathi Kumaresan, Prasanna Kumar. 2024a. Overview of Second Shared Task on Sentiment Analysis in Code-mixed Tamil and Tulu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Lavanya S. K., Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rajeswari Natarajan, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Thenmozhi Durairaj, and Rajkumar Charmathi Kumaresan, Prasanna Kumar. 2024b. Overview of Second Shared Task on Sentiment Analysis in Code-mixed Tamil and Tulu. In *Proceedings*

of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, Malta. European Chapter of the Association for Computational Linguistics.

M Sangeetha and K Nimala. 2023. Sentiment Analysis on Code-Mixed Tamil-English Corpus: A Comprehensive Study of Transformer-Based Models.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Association for Computational Linguistics.

Omar Sharif, Mohammed Moshiul Hoque, and Eftekhar Hossain. 2019. Sentiment analysis of Bengali texts on online restaurant reviews using multinomial Naïve Bayes. In *2019 1st international conference on advances in science, engineering and robotics technology (ICASERT)*, pages 1–6. IEEE.

Tokunaga Takenobu. 1994. Text categorization based on weighted inverse document frequency. *Information Processing Society of Japan, SIGNL*, 94(100):33–40.

Kian Long Tan, Chin Poo Lee, Kalaiarasi Sonai Muthu Anbananthen, and Kian Ming Lim. 2022. RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access*, 10:21517–21525.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yuemei Xu, Han Cao, Wanze Du, and Wenqing Wang. 2022. A survey of cross-lingual sentiment analysis: Methodologies, models and evaluations. *Data Science and Engineering*, 7(3):279–299.