www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 17(2), 2026, 206-214. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2026.v17i2.1280

Machine Learning Approach to Multiclass Sentiment Analysis in Dravidian Language Code-Mixed Text

Augustine Ekene Obiadoh [1], *Tolulope Olalekan Abiola* [2], *Grigori Sidorov**[2], *Liliana Chanona-Hernández* [3]

[1] Covenant University, Nigeria.
[2] Instituto Politécnico Nacional, Center for Computing Research, Mexico.
[3] Instituto Politécnico Nacional, ESIMEZ, Mexico.
obiadohaustine@gmail.com, tabiola2025@cic.ipn.mx, sidorov@cic.ipn.mx, lchanona@gmail.com
* Corresponding author

Abstract. In this study, we focused on sentiment analysis of code-mixed text in Dravidian languages, specifically Tamil-English and Tulu-English, using advanced natural language processing (NLP) techniques. We address the challenges posed by linguistic diversity, the unique nature of code-mixing, and the limited availability of annotated datasets with the application of the Logistic Regression machine learning model enhanced by feature engineering. Our preprocessing approach involves normalizing input text, handling class imbalances, and applying sub word tokenization to accommodate the agglutinative and mixed nature of Dravidian languages. The model was trained for a multi-class sentiment classification task, distinguishing between the classes. Our experiment achieved an F1 score of 0.4409 on Tamil-English and 0.4509 on Tulu-English, this demonstrates the model's ability and limit to process Dravidian scripts and capture sentiments in code-mixed data on a multiclass dataset. Our methodology approach can be extended and improved to tackle other low-resource language challenges in future work. The code and dataset used are available at https://github.com/ABITCONSULT/Sentiment-Analysis-in-Dravidian-Language.

Article Info
Received Dec 2, 2025
Accepted Jan 4, 2026

1 Introduction

Sentiment analysis is an important classification task in NLP for understanding the subjective opinions and emotional responses about an instance through text. It has varied uses for industry-based applications to perform reputation management, customer support, and monitoring of content over social media, among others (Wilson et al., 2005; Thavareesan & Mahesan, 2019, 2020). It summarizes the sentiments and interests of humans into textual feedback or total polarity comments, (Thavareesan & Mahesan, 2020). Identification of offensive language also is one other main task in the NLP area which intends to reduce harmful or unacceptable content on different online platforms. Over the years, there has been considerable attention on sentiment analysis and offence language identification for practical applications in content moderation.

Social media and online review platforms enable users to share their thoughts and opinions in informal and flexible environments. To enhance user experience, these platforms encourage communication in users' native languages or a mix of languages (Vyas et al., 2014). However, a significant challenge in analyzing this user-generated content arises because most NLP systems are trained on formal, grammatically correct data. Consequently, these systems often struggle with the informal, unstructured nature of social media content (Chanda et al., 2016; Pratapa et al., 2018). Moreover, most advancements in sentiment analysis and offensive

language detection have focused on monolingual datasets for high-resource languages, leaving mixed-language and under-resourced contexts underexplored (Winata et al., 2019; Jose et al., 2020).

Code-mixing also refers to the practice of alternating between two or more languages within a single document, paragraph, sentence, or even word. It is a common feature of communication in bilingual and multilingual communities (Barman et al., 2014) and is driven by structural, pragmatic, and sociolinguistic factors (Sridhar, 1978). Social media comments often exhibit code-mixing, particularly in multilingual societies. However, resources for analyzing code-mixed content remain limited, as most sentiment analysis and offensive language identification datasets are designed for monolingual text.

This paper presents sentiment analysis in code-mixed text, especially for Dravidian languages such as Tamil-English and Tulu-English. These languages are predominantly used on social media platforms and pose certain challenges because of code-switching and the non-native usage of most of the script. Addressing these challenges makes this work go toward more inclusive and effective development of NLP systems capable of managing diversity in multiple linguistic environments.

2 Literature Review

A variety of traditional machine learning methods (Reyes-Cocoletzi et al. 2025; Ojo et al., 2021; Ojo et al., 2020; Sidorov et al., 2013; Nath et al. 2025), and transformer models (Masmoudi et al. 2025; Abiola et al., 2025a; Adebanji et al., 2022; Abiola et al., 2025b; Mukhim et al. 2025) among others have been applied in the last few years for different text classification in NLP.

Sentiment analysis plays a vital role in understanding public opinions on various content, such as comments, tweets, videos, and events like elections or social movements (Krishna et al., 2013; Musto et al., 2017). By detecting sentiment polarity (positive, negative, neutral), sentiment analysis helps industries summarize audience perceptions and improve applications like recommendation systems and hate speech detection (Gitari et al., 2015). Social media has been a good source of sentimental data over the last two decades.

Most of the research done on sentiment analysis has focused on monolingual corpora, such as English by (Hu & Liu, 2004; Wiebe et al., 2005), and other languages like Russian and German (Cieliebak et al., 2017). Initially, techniques like the extraction of features using n-grams were widely used for sentiment classification, as reviewed (Kouloumpis et al., 2011). More recently, with large social media datasets, traditional methods have been replaced by deep learning models discussed (Patwa et al., 2020).

However, despite progress in SA, Dravidian languages, such as Tamil-English, Kannada-English, and Malayalam-English remain relatively under-explored in sentiment analysis research. The challenges associated with code-mixed languages, which interweave multiple languages within a single text, make the task of sentiment classification more challenging. To bridge this gap, Chakravarthi et al. (2020) created the Dravidian Code-mix corpus that includes datasets of code-mixed Tamil-English, Kannada-English, and Malayalam-English texts for sentiment analysis and offensive language identification. This is a manually annotated dataset and has become one of the most useful resources for the training and testing of sentiment analysis models for Dravidian languages.

The rapid growth in the volume of social media content also increases the usage of aggressive or offensive language online, compelling the development of automatic moderation systems. These systems would effectively detect harmful speech when trained on appropriate datasets and reduce the volume of offensive content on public platforms. Most research has so far been directed at the identification of offensive language in English texts (Zampieri et al., 2019), while efforts are expanding to languages like Arabic, Danish, Greek, and Turkish (Zampieri et al., 2020). The new resources for the identification of offensive language in Dravidian languages will help in extending the reach of NLP research in under-resourced languages.

Code-mixed native language interactions have increased with internet access and smartphone penetration in multilingual countries like India. The majority of these are spoken by speakers of the Dravidian language family, who, for most, English is the second language. Availability of code-mixed data is low for the Dravidian languages (Jose et al., 2020; Chanda et al., 2016). One of the pioneering works on Kannada-English code-mixed datasets by Sowmya Lakshmi and Shambhavi (2017) and Kannada-English sentiment analysis by Shalini et al. (2018) paved the way for further research in this direction. Later, these datasets were used for sentiment analysis using neural networks and other machine learning techniques.

While progress is being made, open large-scale code-mixed datasets for Dravidian languages are still rare. To stimulate research in this direction, Chakravarthi (2020) and Mandl et al. (2020) organized shared tasks which provided code-mixed Tamil-English,

Kannada-English, and Malayalam-English datasets for sentiment analysis and identification of offensive language. It is expected that these shared tasks will foster the creation of improved models for these resource-poor languages.

Current research in sentiment analysis and the detection of offensive language has focused on high-resource languages that provide a mass dataset on social media. Given the bilingual nature of social media users, systems should be capable of handling such under-resourced code-mixed languages effectively. This dearth of publicly available datasets of reasonable size is precisely the reason why resources like Dravidian Codemix provide a great amount of essential data to work on the training of code-mixed content in Dravidian languages.

3 Methodology

Sentiment analysis in code-mixed text is performed for the following paper in the Dravidian languages, namely, Tamil-English and Tulu-English. Such sets of languages present unique challenges due to their linguistic diversity, prevalence of code-mixing, and limited availability of annotated datasets. To address these challenges, we applied Logistic Regression (LR) model with feature engineering for Dravidian language-specific sentiment classification tasks.

Dataset Analysis
The dataset used in this study comprises multi-labelled comments and posts from social media platforms, specifically YouTube, in code-mixed Tamil-English and Tulu-English. These comments span various domains such as entertainment, politics, technology, and social issues. Each entry is annotated with sentiment labels indicating whether the comment expresses positive, negative, neutral, or mixed emotions.

Class Labels
The dataset follows a message-level sentiment classification task with four sentiment classes: Positive', 'unknown state', 'Mixed feelings', 'Negative for the Tamil-English dataset and positive, not Tulu, neutral, and mixed for the Tulu-English Dataset with 12458 rows of train dataset, 1254 rows for development and 1479 rows for testing the model while the Tamil-English dataset contains 31122, 3843, and 3459 respectively. This classification structure of the dataset enables us to identify how sentiments are expressed in mixed-language scenarios on social media.

Data Cleaning and Preprocessing
Due to the complexity of Dravidian languages and their distinct scripts, text preprocessing involved multiple steps. Code-mixed texts present additional difficulties, as they often contain English mixed with the native language, informal language, and abbreviations. To clean the data, we first removed non-textual elements such as URLs, web links, and special characters using regular expressions. Since our initial preview of the dataset shows so many non-word tokens which could be due to the extraction of the dataset from the internet.

To enrich the model's ability to recognize important linguistic features, we also extracted top bigrams using count vectorizer from scikit-learn library. These top bigrams were appended to the original text entries to guide the model in identifying key linguistic patterns that could influence sentiment classification and this is the core of our research as we discovered during development that varying the number of top bigrams extracted affects the model performance which we discovered with the development dataset that we reach the maximum likely result with the LR model when we extract 95 top bigrams and spend then to the rows of text where they originally existed.

Class Imbalance
The dataset revealed a slight class imbalance, with a higher representation of positive sentiment on Tamil-English Data and Not Tulu on Tulu-English Dataset categories compared to negative or mixed sentiments.

Model Architecture
We employed a Logistic Regression (LR) model imported from scikit-learn for sentiment classification. LR is a widely used linear model for classification tasks especially when computational resources are of concern, capable of handling high-dimensional sparse feature representations commonly found in text-based tasks. To extract meaningful features, after we append top 95 bigrams, we used Term Frequency-Inverse Document Frequency (TF-IDF) vectorization from scikit-learn, which helps capture the importance of words in the dataset while reducing noise from less informative terms. The input to the LR model consisted of TF-IDF-weighted n-grams, including bigrams was appended as a token, to enhance the model's understanding of sentiment-related phrases.

Experimental Setup
The experiments were conducted on a machine equipped with, Intel dual core processor, and 8GB of DDR4 RAM. The datasets were divided into three in the form of training, development and testing for both the Tamil-English and Tulu-English dataset. We trained the LR model and performance was evaluated based on metrics obtained from the test dataset.

Predictions on Unseen Data
We evaluated the trained LR model on a separate test dataset provided by the shared task organizers. This dataset consisted of unseen code-mixed comments in Tamil-English and Tulu-English. The model demonstrated a good performance across both languages, successfully classifying comments into the four sentiment categories with some inadequacy with the low-sourced class due to data imbalance, thereby showing its ability and limitation to handle the complexities of code-mixed, and imbalanced social media texts.

4 Results

Quantitative Evaluation
Our results highlight the effectiveness and limitation of LR in handling code-mixed texts in Dravidian languages for sentiment classification, demonstrating good performance even with limited annotated and imbalanced data. The LR model on the Tamil test dataset has an F1 score of 0.13 for Mixed Feelings class, 0.41 for Negative class, 0.78 for Positive class and 0.42 for unknown class with more details of the classification report on figure 3. We obtained 0.17 for Mixed class, 0.65 for Neutral class, 0.79 for Not Tulu and 0.69 for Positive class on the Tulu-English dataset as we present more classification report on figure 4.

Performance Analysis
Figure 1 and 2 and Table 1 and 2 presents the confusion matrix and classification report on Test Data for both the Tamil-English and Tulu-English results from the test dataset respectively, illustrating the model's classification performance across the four sentiment classes. The model exhibited strong performance in identifying both positive and Not-Tulu sentiments with an F1 score of 0.78 an 0.79 on the Tamil-English and Tulu English dataset respectively, for the positive class. However, some false positives were observed, particularly with other sentiment classes being misclassified

Figure 1: Confusion Matrix Tamil-English

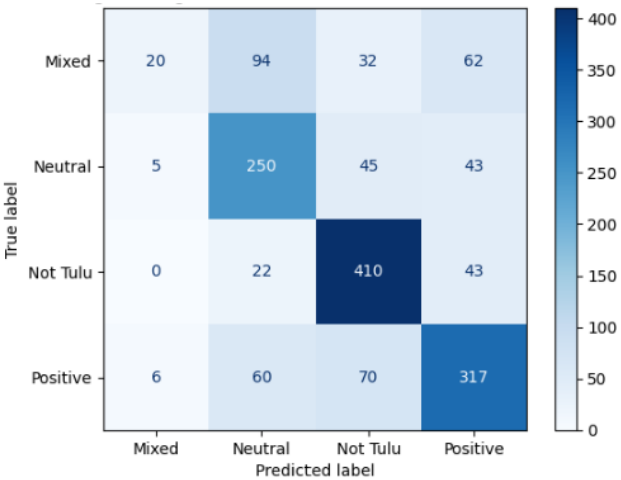


Figure 2: Confusion Matrix Tulu-English

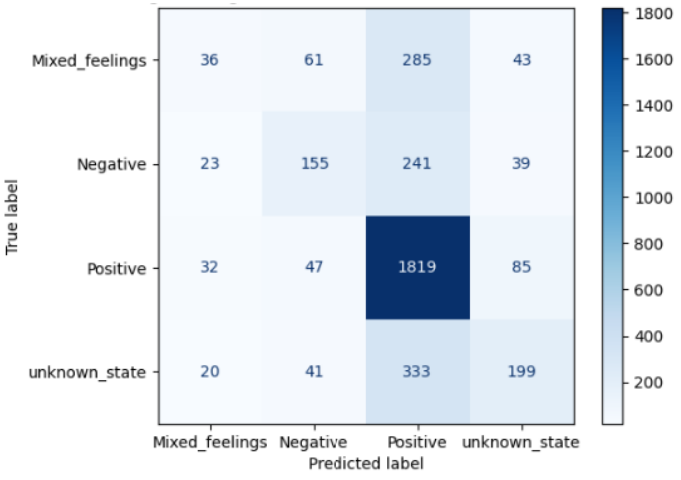


Table 1: Classification Report (Tulu)

Class	Precision	Recall	F1-score	Support
Mixed	0.65	0.10	0.17	208
Neutral	0.59	0.73	0.65	343
Not Tulu	0.74	0.86	0.79	475
Positive	0.68	0.70	0.69	453
accuracy			0.67	1479
macro avg	0.66	0.60	0.58	1479
weighted avg	0.67	0.67	0.64	1479

Table 2: Classification Report (Tamil)

Class	Precision	Recall	F1-score	Support
Mixed feelings	0.32	0.08	0.13	425
Negative	0.51	0.34	0.41	458
Positive	0.68	0.92	0.78	1983
Unknown state	0.54	0.34	0.42	593
accuracy			0.64	3459
macro avg	0.51	0.42	0.43	3459
weighted avg	0.59	0.64	0.59	3459

Linguistic Complexity

The model's ability to handle the complex linguistic features of Dravidian languages was thoroughly tested. TF-IDF vectorization effectively captured key lexical features present in Tamil-English and Tulu-English code-mixed texts. Additionally, the inclusion of bigrams helped in identifying sentiment-laden phrases, improving classification accuracy.

Comparative Strengths

To further compare the classes prediction on the test data we printed the ROC-curve to visualize the classification for both tasks. The ROC curve in figure 3 and 4 for visual performance analysis of the tasks.

Figure 3: ROC Curve Tamil-English

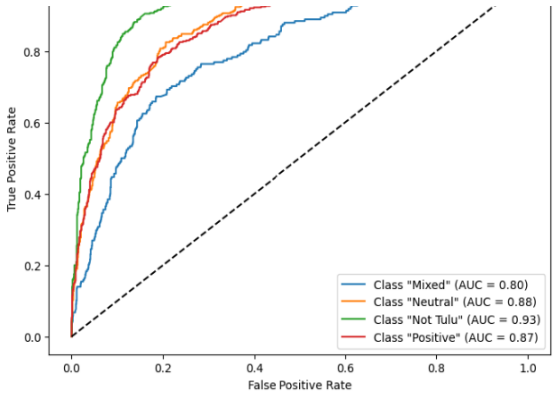
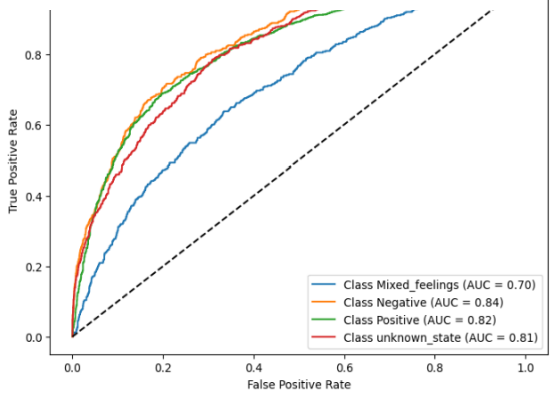


Figure 4: ROC Curve Tulu-English



When compared to other traditional machine learning models like Random Forest and Support Vector Machines (SVMs), LR performed competitively in terms of precision, recall, and F1 score. This can be attributed to LR's ability to handle high-dimensional feature spaces efficiently while maintaining interpretability and computational efficiency.

5 Limitations

The major limiting factor of this research is the low computational resources at disposal as at the time of performing this experiment as we believe some LLM would have performed better on this task, also availability of a large dataset evenly distributed across the classes would surely improve the model performance.

References

Adebanji, O. O., Gelbukh, A., Calvo, H., & Ojo, O. E. (2022). Sequential models for sentiment analysis: A comparative study. In *Advances in Computational Intelligence: 21st Mexican International Conference on Artificial Intelligence (MICAI 2022), Proceedings, Part II* (Lecture Notes in Computer Science). Springer.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)* (pp. 30–38). Association for Computational Linguistics.

Agrawal, R., Chenthil Kumar, V., Muralidharan, V., & Sharma, D. (2018). No more beating about the bush: A step towards idiom handling for Indian language NLP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association.

Andronov, M. S. (1970). *Dravidian languages*. Nauka Publishing House.

Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014). "I am borrowing ya mixing?" An analysis of English-Hindi code mixing in Facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching* (pp. 116–126). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3914

Barman, U., Das, A., Wagner, J., & Foster, J. (2014). Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching* (pp. 13–23). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3902

Blackburn, S. H. (2006). *Print, folklore, and nationalism in colonial South India*. Orient Blackswan.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)* (pp. 161–168). ACM. https://doi.org/10.1145/1143844.1143865

Chakravarthi, B. R. (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 Emotions in Social Media* (pp. 41–53). Association for Computational Linguistics.

Chakravarthi, B. R., & Muralidaran, V. (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* (pp. 61–72). Association for Computational Linguistics.

Chakravarthi, B. R., Anand Kumar, M., McCrae, J. P., Premjith, B., Soman, K., & Mandl, T. (2020). Overview of the track on HASOC–offensive language identification–DravidianCodeMix. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020)*. CEUR Workshop Proceedings.

Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2020). A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop of SLTU and CCURL*. European Language Resources Association.

Chakravarthi, B. R., Muralidaran, V., Priyadharshini, R., & McCrae, J. P. (2020). Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop of SLTU and CCURL*. European Language Resources Association.

Chakravarthi, B. R., Priyadharshini, R., Jose, N., Kumar, M. A., Mandl, T., Kumaresan, P. K., Ponnusamy, R., R. L. H., McCrae, J. P., & Sherly, E. (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* (pp. 133–145). Association for Computational Linguistics.

Chakravarthi, B. R., Priyadharshini, R., Muralidaran, V., Suryawanshi, S., Jose, N., Sherly, E., & McCrae, J. P. (2020). Overview of the track on sentiment analysis for Dravidian languages in code-mixed text. In *Forum for Information Retrieval Evaluation* (pp. 21–24). ACM. https://doi.org/10.1145/3441501.3441515

Chanda, A., Das, D., & Mazumdar, C. (2016). Unraveling the English-Bengali code-mixing phenomenon. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching* (pp. 80–89). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-5810

Cieliebak, M., Deriu, J. M., Egger, D., & Uzdilli, F. (2017). A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 45–51). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1106

Clarke, I., & Grieve, J. (2017). Dimensions of abusive language on Twitter. In *Proceedings of the First Workshop on Abusive Language Online* (pp. 1–10). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3001

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

de Gispert, A., Iglesias, G., & Byrne, B. (2015). Fast and accurate preordering for SMT using neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1012–1017). Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1105

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

El Boukkouri, H., Ferret, O., Lavergne, T., Noji, H., Zweigenbaum, P., & Tsujii, J. (2020). CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6903–6915). International Committee on Computational Linguistics.

Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics, 49*(3), 291–304. https://doi.org/10.1198/004017007000000245

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)* (pp. 168–177). Association for Computing Machinery. https://doi.org/10.1145/1014052.1014073

Jiang, Q., Chen, L., Xu, R., Ao, X., & Yang, M. (2019). A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 6279–6284). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1654

Jin, S., & Pedersen, T. (2018). Duluth UROP at SemEval-2018 Task 2: Multilingual emoji prediction with ensemble learning and oversampling. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)* (pp. 482–485). Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1077

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30*(1), 61–70. https://doi.org/10.1177/001316447003000105

Krishnamurti, B. (2003). *The Dravidian languages*. Cambridge University Press.

Lample, G., & Conneau, A. (2019). *Cross-lingual language model pretraining* (arXiv:1901.07291). arXiv. https://arxiv.org/abs/1901.07291

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). *ALBERT: A lite BERT for self-supervised learning of language representations* (arXiv:1909.11942). arXiv. https://arxiv.org/abs/1909.11942

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach* (arXiv:1907.11692). arXiv. https://arxiv.org/abs/1907.11692

Mahadevan, I. (2003). *Early Tamil epigraphy: From the earliest times to the sixth century AD*. Harvard University Press.

Mandl, T., Modha, S., Kumar, M. A., & Chakravarthi, B. R. (2020). Overview of the HASOC track at FIRE 2020: Hate speech and offensive language identification in Tamil, Malayalam, Hindi, English, and German. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE 2020)* (pp. 29–32). Association for Computing Machinery. https://doi.org/10.1145/3441501.3441517

Masmoudi, A., Aridhi, N., & Hadrich Belguith, L. (2025). Pre-trained model sentiment analysis of Tunisian telecommunications operators' comments on social media. *Computación y Sistemas, 29*(3), 1283–1305. https://doi.org/10.13053/CyS-29-3-5918

Mukhim, B., Maji, A. K., & Das, S. (2025). Sentiment analysis with Khasi low-resource language through generation of sentiment words using machine learning. *Computación y Sistemas, 29*(3), 1225–1236. https://doi.org/10.13053/CyS-29-3-5915

Nath, B., Sarkar, S., & Mukhopadhyay, S. (2025). AstraMT: Instruction-tuned few-shot Assamese–English translation with context-aware prompting and reranking. *Computación y Sistemas, 29*(3), 1167–1178. https://doi.org/10.13053/CyS-29-3-5886

Ojo, O. E., Gelbukh, A., Calvo, H., & Adebanji, O. O. (2021). Performance study of n-grams in the analysis of sentiments. *Journal of the Nigerian Society of Physical Sciences, 3*(4), 477–483. https://doi.org/10.46481/jnsps.2021.201

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)* (pp. 271–278). Association for Computational Linguistics. https://doi.org/10.3115/1218955.1218990

Park, H. (2013). An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing, 43*(2), 154–164.

Patwa, P., Aguilar, G., Kar, S., Pandey, S., Pykl, S., Gambäck, B., Chakraborty, T., Solorio, T., & Das, A. (2020). SemEval-2020 Task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*. Association for Computational Linguistics.

Pratapa, A., Bhat, G., Choudhury, M., Sitaram, S., Dandapat, S., & Bali, K. (2018). Language modeling for code-mixing: The role of linguistic theory-based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers) (pp. 1543–1553). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1143

Reyes-Cocoletzi, L., Aldama-Ramos, J. A., Elias-Zapata, A., Betancourt-González, J., & Rojas-Hernández, J. (2025). Detection of tendency to depression through text analysis. *Computación y Sistemas, 29*(3). https://doi.org/10.13053/CyS-29-3-5887

Sakuntharaj, R., & Mahesan, S. (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)* (pp. 1–6). IEEE.

Sakuntharaj, R., & Mahesan, S. (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)* (pp. 1–5). IEEE.

Salomon, R. (1998). *Indian epigraphy: A guide to the study of inscriptions in Sanskrit, Prakrit, and the other Indo-Aryan languages*. Oxford University Press.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (arXiv:1910.01108). arXiv. https://arxiv.org/abs/1910.01108

Sekhar, A. C. (1951). Evolution of Malayalam. *Bulletin of the Deccan College Research Institute, 12*(1–2), 1–216.

Severyn, A., Moschitti, A., Uryupina, O., Plank, B., & Filippova, K. (2014). Opinion mining on YouTube. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1252–1261). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1118

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest, and KNN models for text classification. *Augmented Human Research, 5*(1), 1–16.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2013). Empirical study of machine learning based approach for opinion mining in tweets. In *Advances in Artificial Intelligence (MICAI 2012)* (Lecture Notes in Computer Science, Vol. 7629, pp. 1–14). Springer. https://doi.org/10.1007/978-3-642-37807-2_1

Sowmya Lakshmi, B. S., & Shambhavi, B. R. (2017). An automatic language identification system for code-mixed English-Kannada social media text. In *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)* (pp. 1–5). IEEE. https://doi.org/10.1109/CSITSS.2017.8447784

Sridhar, S. N. (1978). On the functions of code-mixing in Kannada. *International Journal of the Sociology of Language, 16*, 109–118.

Sridhar, S. N., & Sridhar, K. K. (1980). The syntax and psycholinguistics of bilingual code mixing. *Canadian Journal of Psychology, 34*(4), 407–416.

Takahashi, T. (1995). *Tamil love poetry and poetics* (Vol. 9). Brill.

Thamburaj, K. P., & Rengganathan, V. (2015). A critical study of SPM Tamil literature exam paper. *Asian Journal of Assessment in Teaching and Learning, 5*, 13–24.

Thamburaj, K. P., Arumugum, L., & Samuel, S. J. (2015). An analysis on keyboard writing skills in online learning. In *2015 International Symposium on Technology Management and Emerging Technologies (ISTMET)* (pp. 373–377). IEEE.

Thavareesan, S., & Mahesan, S. (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)* (pp. 320–325). IEEE. https://doi.org/10.1109/ICIIS47346.2019.9063341

Thavareesan, S., & Mahesan, S. (2020a). Sentiment lexicon expansion using Word2Vec and fastText for sentiment prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)* (pp. 272–276). IEEE. https://doi.org/10.1109/MERCon50084.2020.9185369

Thavareesan, S., & Mahesan, S. (2020b). Word embedding-based part of speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)* (pp. 478–482). IEEE. https://doi.org/10.1109/ICIIS51140.2020.9342640

Thenmozhi, D., & Aravindan, C. (2018). Ontology-based Tamil-English cross-lingual information retrieval system. *Sādhanā, 43*(10).

Tian, Y., Galery, T., Dulcinati, G., Molimpakis, E., & Sun, C. (2017). Facebook sentiment: Reactions and emojis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 11–16). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1102

Vyas, Y., Gella, S., Sharma, J., Bali, K., & Choudhury, M. (2014). POS tagging of English-Hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 974–979). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1105

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation, 39*(2), 165–210. https://doi.org/10.1007/s10579-005-7880-9

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (pp. 347–354). Association for Computational Linguistics.

Winata, G. I., Lin, Z., & Fung, P. (2019). Learning multilingual meta-embeddings for code-switching named entity recognition. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP 2019)* (pp. 181–186). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4320

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1)* (pp. 1415–1420). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1144

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç. (2020). SemEval-2020 Task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*. Association for Computational Linguistics.

Zvelebil, K. V. (1991). Comments on the Tolkappiyam theory of literature. *Archív Orientální, 59*, 345–359.