Multilingual Sentiment Analysis: A Machine Learning Approach with a Focus on Malayalam

Shubham Ojha[1,2*] and Dr. Rajalakshmi R[2,3†]

[1*]Student, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, 600127, Tamil Nadu, India.
[2]Prof, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, 600127, Tamil Nadu, India.

*Corresponding author(s). E-mail(s): shubham.ojha2020@vitstudent.ac.in;
Contributing authors: rajalakshmi.r@vit.ac.in;
†These authors contributed equally to this work.

Abstract

Multilingual sentiment analysis poses a significant challenge in natural language processing, particularly in languages with limited resources such as Malayalam. This research addresses the problem of sentiment analysis in Malayalam by comparing various machine learning and deep learning models. Leveraging datasets from the CodaLab Fake News Detection in the Dravidian Languages competition, our study investigates the efficacy of Support Vector Machines (SVM), Random Forest (RF), Logistic Regression, Naive Bayes, and deep learning techniques including BERT classifiers. We apply these models to two distinct datasets: one comprising YouTube comments classified as original or fake, and the other categorizing news into five different categories false, mostly false, true, half true, and mostly true. Our experimentation involves three deep learning approaches: utilizing BERT classifiers directly on the Malayalam text, translating Malayalam to English using the Helsinki-NLP pipeline before employing BERT, and employing an IndicBERT tokenizer for enhanced tokenization. Recognizing the issue of fake news and the influence of social media comments on public opinion, our research seeks to address the challenges of identifying misinformation in online content. Our analysis reveals promising results, with the Naive Bayes classifier achieving an accuracy of 78.8% in detecting original YouTube comments, while Random Forest attains 68.6% accuracy in identifying fake news. Apart from this, we get 3.6% more accuracy by applying traditional ML models on English-translated datasets. Furthermore, we provide comprehensive comparisons using accuracy metrics and confusion matrices, shedding light on the strengths and limitations of each model in the context of Malayalam sentiment analysis. This research contributes to the advancement of sentiment analysis in understudied languages, offering insights into effective modeling techniques and highlighting avenues for future research in multilingual natural language processing, particularly in combating fake news and understanding social media discourse.

Keywords: NLP, BERT, IndicBERT

1 Introduction

Multilingual sentiment analysis is increasingly vital in computational linguistics, facilitating the comprehension of emotions across diverse linguistic landscapes and offering insights into public sentiment and social media trends on a global scale. Within this realm, our focus lies on harnessing the potential of multilingual sentiment analysis within the context of Malayalam, a significant South Indian language. This research endeavors to provide a comprehensive analysis of two distinct Malayalam datasets: one aimed at distinguishing between original and fabricated YouTube comments, and the other dedicated to categorizing news content into five distinct veracity levels.

Motivated by the need to unravel the intricate nuances embedded within Malayalam textual data, our study embarks on a thorough exploration of both traditional machine learning and cutting-edge deep learning methodologies. Traditional classifiers such as Support Vector Machines (SVM), Random Forest (RF), Logistic Regression, and Naive Bayes are rigorously evaluated alongside state-of-the-art techniques employing BERT classifiers and the Indic BERT tokenizer. By juxtaposing these diverse approaches, our aim is not only to enhance the accuracy of sentiment analysis in Malayalam but also to demonstrate the effectiveness of integrating traditional and advanced methodologies for multilingual sentiment analysis.

This interdisciplinary approach underscores the significance of bridging conventional techniques with advanced computational models, thereby paving the way for more robust and nuanced sentiment analysis across languages. Through this research, we aim to contribute to the ongoing discourse on multilingual sentiment analysis, shedding light on the challenges and opportunities inherent in analyzing emotions within diverse linguistic contexts.

The main contributions of this paper are:

- Implementing and evaluating a range of traditional machine learning algorithms, including Support Vector Machines (SVM), Random Forest (RF), Logistic Regression, and Naive Bayes, for sentiment classification on Malayalam datasets. Specifically, our focus lies in discerning between original and fake sentiments expressed in YouTube comments and categorizing news articles into five distinct veracity levels: false, mainly false, true, half true, and mostly true. Through this approach, we seek to assess the performance and efficacy of these conventional algorithms in capturing the nuanced emotions prevalent within Malayalam text.

- Adapting and fine-tuning BERT-based deep learning models to advance sentiment analysis of Malayalam YouTube comments and news articles. This entails leveraging the powerful capabilities of BERT (Bidirectional Encoder Representations from Transformers) to extract intricate emotional nuances embedded within the Malayalam language. Moreover, we explore different methodologies for incorporating BERT into our analysis, including direct BERT classification, translation of Malayalam text via the Helsinki-NLP pipeline, and utilization of the Indic BERT tokenizer for language-specific tokenization.
- Comprehensively comparing the effectiveness of these approaches in sentiment analysis, evaluating metrics such as accuracy, precision, recall, and F1-score on validation datasets. By rigorously assessing the performance of each model, we aim to ascertain which method yields the most reliable sentiment analysis results for Malayalam text. Furthermore, we delve into the efficacy of machine learning versus deep learning approaches in detecting fake comments and classifying news articles into nuanced truthfulness categories, providing insights into the strengths and limitations of each approach.

Ultimately, this research contributes to the broader field of multilingual sentiment analysis by addressing the unique challenges associated with language-specific sentiment assessment. Through comparative analyses of different algorithmic approaches, we seek to advance our understanding of the complexities inherent in multilingual sentiment analysis, thereby enhancing the accuracy and effectiveness of sentiment classification in diverse linguistic contexts.

2 Related Work

Salvador Contreras Hernández, María Patricia Tzili Cruz, and José Martín Espínola explored sentiment analysis of COVID-19-related tweets in Mexico using BERT-based models. Their research, focusing on semi-supervised learning with Spanish language models, demonstrated superior precision over multilingual BERT and traditional classifiers. This underscores the effectiveness of language-specific models in capturing public sentiment, offering significant implications for public health decision-making during the pandemic. [1]

George Manias, Argyro Mavrogiorgou, and Athanasios Kiourtis investigated multilingual sentiment analysis on Twitter, emphasizing the importance of language- and domain-agnostic approaches. Their study assessed the efficacy of four BERT-based classifiers against a zero-shot classification method. The findings suggest that while BERT-based classifiers are highly effective, zero-shot classification stands out as an innovative and scalable strategy, even though it may not reach the fine-tuned accuracy of its counterparts. [2]

Amina Amara, Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha spearheaded a study on COVID-19 trend analysis through the lens of Facebook data across seven languages using Latent Dirichlet Allocation (LDA). The research stands out by leveraging an underexplored platform for multilingual topic modelling, using graph visualization to trace the progression of public interest in the pandemic. The outcomes present unique insights into global sentiment and conversational trends on Facebook, mapping the chronological growth of discourse surrounding COVID-19. [3]

Rami Mohawesh, Sumbal Maqsood, and Althebyan's research offers a novel semantic solution to multilingual fake news detection by utilizing capsule neural networks. Their framework incorporates word embeddings and Qutaibah n-gram features, significantly enhancing fake news detection across languages. The results show a marked improvement over existing methods, highlighting the capability of capsule neural networks to adeptly manage the intricacies of multilingual text analysis. [4]

Anjum and Rahul Katarya have developed HateDetector, an advanced technique tailored for detecting hate speech on social media in various languages. Incorporating an improved seagull optimization algorithm and a hybrid diagonal-gated recurrent neural network, their method shows notable enhancements in accuracy, precision, recall, and F-measure. These results position HateDetector as a promising tool for effectively monitoring and curbing hate speech across multiple languages and social media platforms. [5]

Caio Mello, Gullal S. Cheema, and Gaurish Thakkar examine the construction of Olympic legacy narratives through multilingual sentiment analysis of news articles on the London 2012 and Rio 2016 Olympics. Their methodology intertwines four sentiment analysis (SA) algorithms with explainable AI techniques to scrutinize methodological constraints. While recognizing SA's value in content analysis, the research reveals complexities inherent in multilingual and specialized domains. A blend of leading classifiers paired with clear AI practices promises improvements, and an intriguing utopian versus dystopian narrative dichotomy in Olympic legacy portrayal is disclosed. [6]

Purbani Kar and Swapan Debbarma tackle the challenge of detecting hate speech and analyzing sentiments in multilingual code-mixed social media texts. They present an enhanced seagull optimization algorithm coupled with a novel hybrid diagonal gated recurrent neural network. The proposed methodology has shown considerable improvement in precision, recall, and F-measure over traditional approaches, asserting its effectiveness as a powerful tool for hate speech and sentiment analysis in diverse linguistic settings. [7]

Simran Sidhu, Surinder S. Khurana, Munish Kumar, and Parvinder Singh provide a thorough review of sentiment analysis techniques in Hindi, focusing on negation handling and the development of Hindi SentiWordNet. They explore a range of methodologies, including both semantic and machine learning approaches, and assess tools such as lexicons, stemmers, and morphological analyzers. The paper concludes with a call for more advanced sentiment analysis research for Hindi, highlighting its vast native-speaking community and expanding online footprint, while also pointing out potential areas for future investigation. [8]

Christian E. Lopez and Caleb Gallemore introduce a sizable multilingual Twitter dataset designed to support research on COVID-19 social discourse. With over 2.2 billion tweets enriched by sentiment analysis and named entity recognition, this resource permits an in-depth examination of discussions surrounding the pandemic. Their conclusion affirms the dataset's importance as a tool for tracking the progression of public sentiment and conversational patterns about COVID-19, enabling diverse analyses of social media data. [9]

Siroos Rahmani Zardak, Amir Hossein Rasekh, and Mohammad Sadegh Bashkari address the gap in sentiment analysis tools for Persian text by customizing the BERT algorithm for this context. Their work benchmarks the BERT algorithm's performance against former approaches across various datasets, concluding that BERT excels in analysing Persian sentiment, evidenced by superior accuracy and F1 scores. This breakthrough underscores BERT's adaptability and potency in processing Persian language datasets for sentiment analysis. [10]

3 Proposed Work

Our project embarks on advancing multilingual sentiment analysis with a concentrated focus on the Malayalam language, engaging with a pair of distinct datasets sourced from the CodaLab Fake News Detection in Dravidian Languages competition (Dravidian-LangTech@EACL 2024). [11] The first dataset comprises YouTube comments, each meticulously categorized as original or fake, serving as a testament to the intricacies of digital discourse.[11] The second dataset encompasses a diverse collection of news items, each painstakingly classified into one of five truthfulness categories: false, mainly false, true, half true, and mostly true.[11] This granular classification scheme presents a nuanced spectrum of information authenticity, critical for the discerning algorithms we deploy. The analytical journey of the project begins with the application of four classical machine learning classifiers; Support Vector Machine (SVM), Random Forest (RF), Logistic Regression, and Naive Bayes then executed on both datasets. The objective is to compare and contrast their validation performance rigorously, thereby determining the most efficacious model. The model that prevails in validation is then subjected to the crucible of testing within the datasets to ascertain its generalizability and robustness. Venturing into the realm of deep learning, the project explores three distinct methodologies. Initially, we implement a BERT classifier to delve into the sentiment analysis directly. Subsequently, we enhance our linguistic reach by employing the Helsinki-NLP pipeline to translate the Malayalam text into English, thereafter applying the BERT classifier to this translated corpus.[12] The final stride in our deep learning endeavor employs the Indic BERT tokenizer, specially designed to improve the tokenization of the Malayalam script, thus tailoring the BERT classifier to the linguistic nuances of Dravidian syntax and semantics. Through these varied approaches, we seek to construct a comprehensive analysis mechanism that stands at the vanguard of sentiment analysis for Malayalam language data, aiming for the highest echelons of accuracy and interoperability.

3.1 Traditional ML Models Steps

3.1.1 Data Preparation: Import the necessary libraries, including sci-kit-learn and pandas. Load the training dataset for model training.

3.1.2 Vectorization of Text: To convert the textual data into numerical vectors, use the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer.

3.1.3 The Logistic Regression Model: To find the ideal regularisation parameter (C), use cross-validation and Logistic Regression using a hyperparameter grid search. Utilize the determined ideal hyperparameters to train the classifier.

3.1.4 The Random Forest Model: To determine the optimal set of hyperparameters, such as the number of estimators, maximum depth, minimum samples split, and minimum samples leaf, use Random Forest in conjunction with a grid search. Utilize the optimized hyperparameters to train the Random Forest model.

3.1.5 The Support Vector Model (SVM): Use the Support Vector Machine (SVM) technique for categorization. Utilizing the TF-IDF vectorized training data, train the SVM model.

3.1.6 The Naive Bayes Model: Multinomial Naive Bayes should be used for categorization. Utilizing the TF-IDF vectorized training data, train the Naive Bayes model.

3.1.7 Verification Assessment of the dataset: To evaluate the model, load the validation dataset. Take care of missing values and vectorize the text column. Evaluate each classifier's accuracy score on the validation set.

3.1.8 Testing: Load the testing dataset. Utilizing the TF-IDF vectorizer, vectorize the text column of testing data. To predict labels for the test dataset, use the best classifier available.

3.1.9 Conclusion: Conclude the study by summarizing the chosen classifier's performance on the testing dataset.

3.2 Deep Learning Models Steps

3.2.1 Data Preparation: Begin by importing essential libraries such as transformers for the BERT model, torch for the deep learning framework, and pandas for data manipulation. Load the training dataset into a DataFrame for further processing.

3.2.2 Tokenization: Utilize the BERT tokenizer to convert the text into tokens that are understandable by the model. This step will include converting the Malayalam sentences into a format with token IDs, attention masks, and segment IDs suitable for BERT.

3.2.3 Model Configuration: Choose a pre-trained BERT model that is optimized for the Malayalam language or the multilingual version that includes Malayalam. Configure the BERT model with appropriate parameters, paying attention to the number of epochs, learning rate, and batch size for training.

3.2.4 Fine-Tuning: Using the tokenized text data, fine-tune the BERT model on the Malayalam sentiment analysis task. This involves training the model on the dataset, adjusting weights, and ensuring the model learns the context of the dataset effectively.

3.2.5 Validation Assessment: After the model has been fine-tuned, evaluate its performance on a separate validation set to gauge its effectiveness. Process the validation dataset similarly to the training set, with tokenization followed by the creation of DataLoader objects for the BERT model.

3.2.6 Testing: Load the testing dataset and process it through the BERT models using the same tokenization and DataLoader creation steps as the validation set. Apply the selected model to obtain predictions for sentiment classification.

3.2.7 Conclusion: Conclude the process by summarizing the performance of the BERT model on the testing dataset. Discuss the effectiveness of the model, its ability to generalize to unseen data, and any observations regarding its performance on multilingual sentiment analysis.

4 Results

The evaluation matrix used here are accuracy, training time and Confusion matrices.

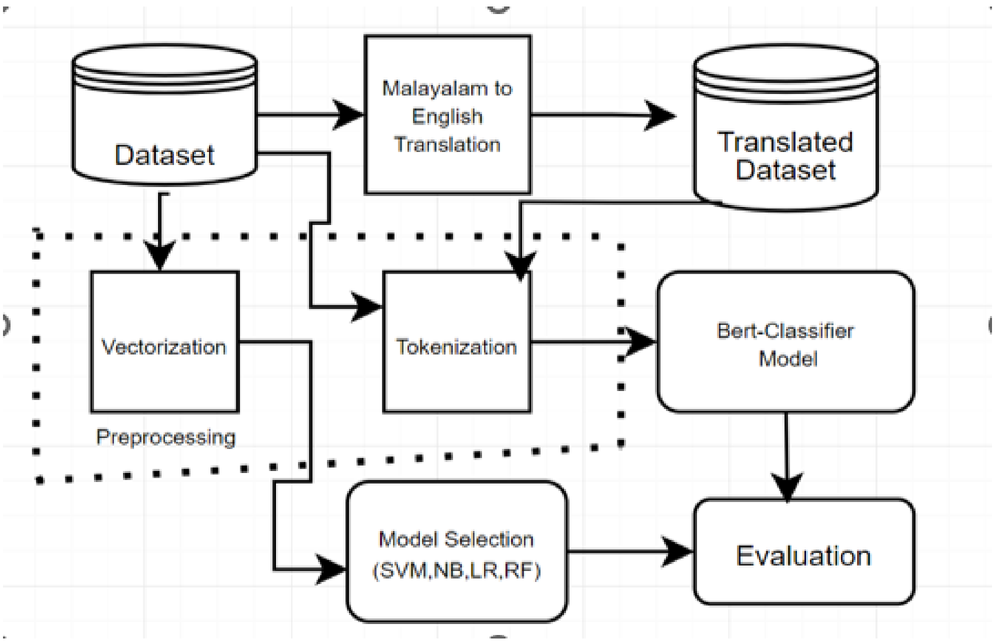

Fig. 1 Flowchart Diagram of Proposed System

Table 1 YouTube Comment Originality Detection Accuracy

Model	Accuracy (%)	Training Time (HH:MM:SS)
Naïve Bayes	78.8	00:00:51
SVM	50.1	00:00:27
Random Forest	75.5	00:12:31
Logistic Regression	78.6	00:01:57
BERT	75.2	01:30:22
IndicBERT	75.1	01:29:27

4.1 Youtube Comment Originality Detection

Tables 1-4 showcase the accuracy, F1 score, and training times for various machine learning and deep learning models applied to YouTube comment originality detection on datasets of different languages.

In Tables 1 and 3, models trained on the Malayalam language dataset exhibit varying levels of accuracy and F1 scores, with SVM and Naïve Bayes models demonstrating competitive performance. However, the Random Forest model achieves the highest accuracy at 75.5% but with a longer training time of 00:12:31.

In Table 2 and 4, evaluating models on an English-translated dataset reveals different trends. Here, the SVM model surpasses others with an accuracy of 58.3% and the highest F1 score of 0.3713, implying robust performance in the English context.

Table 2 YouTube Comment Originality Detection Accuracy on English Translated Dataset

Model	Accuracy (%)	Training Time (HH:MM:SS)
Naïve Bayes	76.8	00:00:47
SVM	58.3	00:00:25
Random Forest	78.5	00:10:41
Logistic Regression	71.6	00:01:37
BERT	74.0	01:16:20

Table 3 F1 Score for YouTube Comment Originality Detection on Malayalam Language Dataset

Model	Precision	Recall	F1 Score
SVM	0.4137	0.3156	0.3651
Random Forest	0.4324	0.2573	0.3204
Naïve Bayes	0.4209	0.2823	0.3388
Logistic Regression	0.3828	0.3027	0.3049
BERT	0.4277	0.2663	0.3291
IndicBERT	0.4194	0.2762	0.3344

Table 4 F1 Score for YouTube Comment Originality Detection on English Translated Dataset

Model	Precision	Recall	F1 Score
SVM	0.4025	0.3331	0.3713
Random Forest	0.4280	0.2485	0.3088
Naïve Bayes	0.4030	0.2548	0.3119
Logistic Regression	0.3670	0.3338	0.3543
BERT	0.4272	0.2612	0.3237

Notably, BERT demonstrates competitive performance with an F1 score of 0.3237, despite its longer training time of 01:16:20.

4.2 Fake News Detection

These four tables provide insights into the performance of various models in fake news detection tasks across different datasets and languages. Tables 5 and 6 highlight the accuracy and training times of models on Malayalam and English-translated datasets, respectively. Meanwhile, Tables 7 and 8 showcase the F1 scores for models on the same datasets. Overall, the tables shed light on the effectiveness of different models in detecting fake news under various linguistic contexts.

We evaluated the performance of various machine learning models on two distinct datasets: one composed of Malayalam YouTube comments aimed at distinguishing between original and fake content, and the other focusing on Malayalam news articles for fake news detection. Initially, we observed that across both datasets, models such as Naive Bayes, Random Forest, Logistic Regression, BERT, and IndicBERT exhibited competitive accuracies. However, there was a notable discrepancy when comparing the

Table 5 Fake News Detection Accuracy

Model	Accuracy (%)	Training Time (HH:MM:SS)
Naïve Bayes	63.5	00:00:59
SVM	65.8	00:00:37
Random Forest	68.6	00:20:76
Logistic Regression	61.7	00:12:07
BERT	60.52	01:56:12
IndicBERT	62.31	01:50:18

Table 6 Fake News Detection Accuracy on English Translated Dataset

Model	Accuracy (%)	Training Time (HH:MM:SS)
Naïve Bayes	60.5	00:00:56
SVM	67.7	00:00:28
Random Forest	72.2	00:14:76
Logistic Regression	64.7	00:11:72
BERT	65.7	01:33:12

Table 7 F1 Score for Fake News Detection on Malayalam Language Dataset

Model	Precision	Recall	F1 Score
SVM	0.5095	0.4442	0.4750
Random Forest	0.5280	0.3411	0.4199
Naïve Bayes	0.4700	0.3562	0.4036
Logistic Regression	0.4929	0.4092	0.4440
BERT	0.5682	0.3616	0.4461
IndicBERT	0.5542	0.2989	0.3832

performance on the original Malayalam dataset versus the English-translated dataset. Interestingly, upon applying the models to the English-translated dataset, we found a notable increase in accuracy for certain models, notably Naive Bayes and Random Forest, in both the YouTube comment originality and fake news detection tasks. Additionally, the training time for the translated dataset was generally reduced compared to the original Malayalam dataset. This observation suggests that while the translation process introduces linguistic variations, it may also enhance the model's ability to discern patterns in the data. Moreover, the reduced training time for the translated dataset implies that the models may require less computational resources to achieve comparable performance, potentially enhancing their scalability and efficiency in real-world applications. These findings underscore the importance of considering language translation techniques in multilingual text analysis tasks, offering insights into optimizing model performance and resource utilization.

Table 8 F1 Score for Fake News Detection on English Translated Dataset

Model	Precision	Recall	F1 Score
SVM	0.5588	0.4975	0.5222
Random Forest	0.5245	0.3008	0.3872
Naïve Bayes	0.4942	0.3556	0.4146
Logistic Regression	0.5202	0.4145	0.4606
BERT	0.5693	0.3465	0.4294

5 Conclusion and Future Work

The comparative analysis of sentiment analysis models on Malayalam datasets underscores the dominance of Naive Bayes for YouTube comment classification and Random Forest for fake news detection in terms of accuracy. Despite longer training times, BERT-based models offer deep linguistic insights, albeit not surpassing traditional models in accuracy. A critical factor influencing performance disparities is dataset size. Deep learning models like BERT thrive on vast data, whereas the modest datasets of 3,200 comments for YouTube and 1,700 articles for news may limit their potential. Traditional models exhibit notable efficiency, possibly due to their compatibility with smaller datasets and less complex feature spaces. The restricted dataset size emerges as a pivotal variable, potentially constraining deep learning algorithms' effectiveness. Expanding datasets or employing data augmentation strategies could enhance deep learning models' performance, leveraging their ability to grasp complex language patterns. This research illuminates current sentiment analysis model capabilities for Malayalam texts while emphasizing the necessity for larger datasets to maximize deep learning techniques' potential. Future endeavours should focus on dataset expansion or data augmentation strategies to enhance deep learning models' performance, striking a balance between model choice, dataset size, and computational efficiency in multilingual sentiment analysis research.

References

[1] Contreras Hernández, S., Tzili Cruz, M. P., Espínola Sánchez, J. M., & Pérez Tzili, A. (2023). Deep learning model for covid-19 sentiment analysis on Twitter. *New Generation Computing*, *41*(2), 189-212.

[2] Manias, G., Mavrogiorgou, A., Kiourtis, A., Symvoulidis, C., & Kyriazis, D. (2023). Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying Twitter data. *Neural Computing and Applications*, *35*(29), 21415-21431.

[3] Amara, A., Hadj Taieb, M. A., & Ben Aouicha, M. (2021). Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis. *Applied Intelligence*, *51*, 3052-3073.

[4] Mohawesh, R., Maqsood, S., & Althebyan, Q. (2023). Multilingual deep learning framework for fake news detection using capsule neural network. *Journal of Intelligent Information Systems*, *60*(3), 655-671.

[5] Anjum, & Katarya, R. (2023). HateDetector: Multilingual technique for the analysis and detection of online hate speech in social networks. *Multimedia Tools and Applications*, 1-28.

[6] Mello, C., Cheema, G. S., & Thakkar, G. (2023). Combining sentiment analysis classifiers to explore multilingual news articles covering London 2012 and Rio 2016 Olympics. *International Journal of Digital Humanities*, *5*(2), 131-157.

[7] Kar, P., & Debbarma, S. (2023). Multilingual hate speech detection sentimental analysis on social media platforms using optimal feature extraction and hybrid diagonal gated recurrent neural network. *The Journal of Supercomputing*, *79*(17), 19515-19546.

[8] Sidhu, S., Khurana, S. S., Kumar, M., Singh, P., & Bamber, S. S. (2023). Sentiment analysis of Hindi language text: a critical review. *Multimedia Tools and Applications*, 1-30.

[9] Lopez, C. E., & Gallemore, C. (2021). An augmented multilingual Twitter dataset for studying the COVID-19 infodemic. *Social Network Analysis and Mining*, *11*(1), 102.

[10] Zardak, S. R., Rasekh, A. H., & Bashkari, M. S. (2023). Persian Text Sentiment Analysis Based on BERT and Neural Networks. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, *47*(4), 1623-1634.

[11] Subramanian, M., Chakravarthi, B. R., Shanmugavadivel, K., Pandiyan, S., Kumaresan, P. K., Palani, B., Singh, M., Raja, S., Vanaja, & S, Mithunajha. (2023). Overview of the Shared Task on Fake News Detection from Social Media Text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages* (pp. 16-23). Varna, Bulgaria: Recent Advances in Natural Language Processing.

[12] Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020, November). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4948-4961).

[13] Devika, K., Haripriya, B., Vigneshwar, E., Premjith, B., & Chakravarthi, B. R. (2024, March). From Dataset to Detection: A Comprehensive Approach to Combating Malayalam Fake News. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages* (pp. 16-23).