A Novel LSTM Based Model for Sentiment Detection in Hindi-English Code-Switched Texts

Namitha Shambhu Bhat[1][a], Kuldeep Sambrekar[1][b] and Shridhar Allagi[2][c]
[1]*Computer Science and Engineering, KLS Gogte Institute of Technology, Belagavi, India*
[2]*Computer Science and Engineering, KLE Institute of Technology, Hubli, India*

Keywords: Sentiment Analysis, Code-Switching, Language Identification, LSTM, Switch Point Detection.

Abstract: Sentiment analysis in multilingual conversations is laborious yet important in understanding emotions and opinions expressed across multiple languages and cultures. Code-switching, a prevalent technique poses challenges due to linguistic diversity, cultural nuances, and contextual dependencies. The research in this article provides an LSTM-based framework for sentiment analysis in Hindi-English code-switched text, addressing the challenges of multilingual content in social media. The methodology adopted in this research incorporates three key components: language-specific encoders to obtain linguistic patterns, a switcher module for understanding language transitions, and a sentiment analysis module for extracting sentiment within a multilingual text. A Hindi-English dataset containing 4,954 samples with positive, neutral, and negative sentiments is used for training and evaluation. The model achieves an overall accuracy of 89.9%, and an F1-score of 0.9 across all sentiments investigated. This work contributes substantially to multilingual sentiment analysis, eliminating the shortcomings of conventional approaches and offering a robust method for analyzing complex code-switched text.

1 INTRODUCTION

Multilingual Code-Switching (MCS) is a prevalent phenomenon in global societies, where individuals switch between two or more languages within a single conversation(Myers-Scotton 1993). This linguistic phenomenon is widespread in multicultural communities, social media, and online forums (Plaza-del-Arco et al. 2021) (AlGhamdi et al. 2016). Sentiment analysis (SA) in MCS texts is crucial for understanding public opinions, emotions, and intentions (Pang and Lee 2008). However, MCS poses significant challenges for SA due to its complex linguistic patterns, cultural nuances, and contextual dependencies. The increasing volume of MCS text data necessitates effective SA tools to extract valuable insights. Traditional SA approaches focus on monolingual text analysis, neglecting the complexities of MCS (Jamatia et al. 2020) (Zhu et al. 2022) (Tan, Lee, and Lim 2023). Recent studies have addressed MCS SA using rule-based and machine-learning approaches (Kim n.d.; P and Mahender 2024), (Ullah et al., 2022). However, these methods have limitations, such as relying on hand-crafted features or requiring large annotated datasets.

Deep learning techniques have shown promise in SA (Dutta, Agrawal, and Kumar Roy 2021) (Santos and Gatti 2014). However, their application to MCS SA is still developing. This research aims to bridge this gap by proposing a novel deep-learning approach for MCS SA. The complexity of MCS texts arises from various factors, including Language identification (S. D. Das et al., 2019)(D. Das & Petrov, 2011), contextual understanding (Code-switching n.d.) (Gardner-Chloros 2009), and cultural influences (Albahoth et al., 2024; Hofstede, 2001). Current approaches to MCS SA mainly focus on Rule-based methods that use hand-crafted rules and linguistic features (Agüero-Torales et al., 2021), and Machine learning leveraging supervised learning techniques with annotated datasets (Chakravarthi et

[a] https://orcid.org/0009-0007-8945-4197
[b] https://orcid.org/0000-0002-3542-1716
[c] https://orcid.org/0000-0001-7502-0741

Bhat, N. S., Sambrekar, K. and Allagi, S.
A Novel LSTM-Based Model for Sentiment Detection in Hindi- English Code-Switched Texts.
DOI: 10.5220/0013593600004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 2*, pages 422-428
ISBN: 978-989-758-763-4

al., 2020). However, these approaches have limitations wherein Hand-crafted features may not capture complex linguistic patterns., Annotated datasets are scarce, especially for low-resource languages, Cultural and contextual nuances are often overlooked.

This research is further organised as follows: We have presented the related prior research in section 2, and the methodology for assimilating the gaps and required solutions is formulated and presented in section 3. Section 4 showcases the experimental results obtained using the datasets and the processing techniques mentioned in the method with elaborate discussions. In section 5, we provide compelling concluding remarks highlighting the challenges of sentiment analysis and outline significant opportunities for future research and improvement.

2 RELATED STUDIES

Research on computational models for code-mixing is limited due to the scarcity of this phenomenon in conventional text corpora, making data-greedy approaches difficult to apply. Recent studies have highlighted the complexity and importance of sentiment analysis in multilingual and code-switched contexts. For instance, Sharma et al. (Sharma et al., 2023a) developed a sentiment analysis system for code-switched data using a late fusion approach combining two transformers, which showed promising results on English-Hindi and English-Spanish datasets (Sharma et al., 2023b). Similarly, Vilares et al. (Vilares et al., 2016) explored sentiment analysis on monolingual, multilingual, and code-switching Twitter corpora, emphasizing the unique challenges posed by code-switching. Additionally, Mamta and Ekbal (Mamta and Ekbal 2025) proposed a transformer-based multilingual joint learning framework for code-mixed and English sentiment analysis, demonstrating improved performance through shared-private, multi-task learning.

Furthermore, a study by Yekrangi and Abdolvand (Yekrangi and Abdolvand 2021) they were focused on augmenting sentiment prediction capabilities for code-mixed languages, including English and Roman Urdu, using robust transformer-based algorithms. These studies underscore the growing recognition of code-switching in sentiment analysis and the need for advanced models to handle the linguistic diversity present in social media content. Another study by Kumar et al. (Kumari and Kumar 2021) revealed the impact of code-switching on sentiment analysis, highlighting the need for specialized models to handle the linguistic diversity in social media content. Additionally, a recent study by Patwa et al. (Patwa et al., 2020) examined the effectiveness of various machine learning models for sentiment analysis in code-mixed social media text, finding that transformer-based models outperformed traditional approaches. An exhaustive study of the prior research is presented in Table 1.

These studies underscore the growing recognition of code-switching in sentiment analysis and the need for advanced models to handle the linguistic diversity present in social media content. In the research presented herein, we have used multiple techniques to address the issues related to sentiment analysis and achieve an optimum result that bridges the gap mentioned in the related research. This research proposes a novel deep learning approach for MCS SA, incorporating a combination of three techniques namely,

- Language-specific encoders to capture linguistic patterns.
- A switcher module to detect language switching.
- A sentiment analysis module to capture contextual nuances.

This approach effectively utilizes deep learning techniques, aiming to improve accuracy in MCS SA, reduce reliance on hand-crafted features, and enable more effective sentiment analysis in multilingual settings.

3 METHODOLOGY

This research proposes a novel deep-learning approach for sentiment analysis in multilingual code-switching text shown in Figure 1.

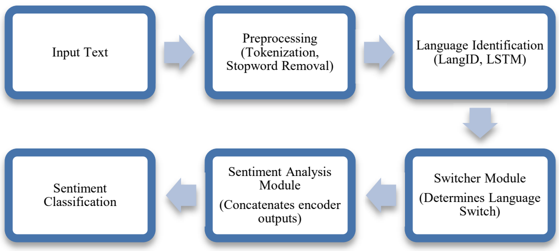

Figure 1: Proposed Methodology for Multilingual Sentiment Analysis.

In the dataset preparation and preprocessing a Hindi-English mixed dataset samples are collected from social media. Sentiment labels such as Positive, Neutral, and Negative are annotated to the dataset. A total of 4954 samples are used, consisting of a mix of 1850 Positive, 1700 Neutral, and 1404 Negative samples. Tokenization was executed to split the input text into individual words for further analysis.

Table 1: Previous Research on Multilingual Sentiment Analysis.

Study	Methodology	Language(s)	Dataset	Accuracy/F-score	Limitations	Year
(Kasmuri et al. 2020)	Rule Based	Malay-English	Social media	0.87 (F1-score)	Limited to specific language pairs, relies on hand-crafted rules	2020
(Younas et al. 2020)	State-of-the-art Deep learning models (mBERT, XLM-R)	Roman Urdu-English	Social media	mBERT: 69% (Accuracy) 0.69 (F1-score) XLM-R: 71% (Accuracy) 0.71 (F1-score)	Lack of exploration into the contextual nuances and cultural references in the code-mixed text that could influence sentiment interpretation.	2020
(Sharma, Chinmay, and Sharma 2023b)	mBERT-BERT Late fusion models	English-Hindi	Social media	0.6124 (F1-score)	Focuses on widely spoken languages. Low resource languages need to be explored	2023
(Adel et al. 2013)	RNN Language Model	English-German	Speech Data	-	Focuses on language modeling, may not generalize well to sentiment analysis	2013
(Santos and Gatti 2014)	CNN Sentiment Analysis	English-Portuguese	Short Texts	0.925 (F1-score)	May not perform well with longer texts or different language pairs	2014
(Klementiev, Titov, and Bhattarai 2012)	Cross-lingual Sentiment	English-Spanish-	Multilingual	0.859 (Accuracy)	Assumes parallel corpora availability, may not handle code-switching within sentences	2012
(Glorot, Bordes, and Bengio 2011)	Domain Adaptation	English-French	Multilingual	0.885 (Accuracy)	Requires target domain labeled data, may not adapt well to unseen domains	2011
(You, Jin, and Luo 2017)	Visual Sentiment Analysis	English	Image Data	0.921 (Accuracy)	Limited to visual features, may not capture textual sentiment	2016
(Zadeh et al. 2017)	Tensor Fusion Network for Multimodal Sentiment	English	Multimodal	0.942 (F1-score)	Requires multimodal data, may face challenges in integrating multiple modalities	2017

Stopwords like “is”, “and”,” or”, “the”, “aur”, “mein” etc. which did not carry any significant meanings were removed from the tokenized words. This step reduces the dimensionality whilst retaining the quality of the data. Special characters such as emojis, hashtags and punctuations are omitted to prevent trivial features from affecting the performance of the model.

LangID, a prebuilt library is used to tag each token with its language English or Hindi. All English words are converted to lowercase and the Hindi words to Unicode for a uniformity in formats. The dataset is later split into training and testing sets at an 80:20 ratio. The Architecture of the model was built using:

3.1 Language Identification Module

The Pre-trained LangID embeddings are used to detect language-specific patterns and switching between Hindi and English. The Language-Specific Encoders contain separate input layers which are defined for Hindi and English and embedding layers are created to convert integer-encoded words into dense vectors. The dimension of embedding vectors in this layer is set to 100 to convert input sequences into embedded representations. A single layer of LSTM with 64 units is implemented to capture sequential dependencies. This step helps in understanding where and why a language switch occurs. A dropout layer with a dropout rate of 0.2 is created for regularization. It helps prevent overfitting by randomly dropping 20% of the units during training.

3.2 Switcher Module

This module is applied to identify where the language changes from Hindi to English or vice versa within a sentence. The execution is carried out in two steps namely Sequence Labelling with LSTM-based Network and Training with Annotated Switch Points. The former step involves assigning a label to each token in the sentence and the latter step is used to train the model using a labelled dataset where each token was annotated with its respective language. For the switcher module, labels denote the language of the token i.e. English or Hindi.

3.3 Sentiment Analysis Module

The output from the language-specific encoders is fed to the sentiment analysis module to predict the sentiment of the input text. This module integrates information from both the Hindi and English parts of the sentence. The outputs from the Hindi and English language-specific encoders are concatenated. These outputs are high-dimensional vectors (embeddings) representing the semantic meaning of the text in both languages. A fully connected dense layer with 64 hidden units is applied to the concatenated vector. Rectified Linear Unit (ReLU) introduces non-linearity, allowing the model to learn complex patterns. A final Dense Layer with Softmax Activation predicts the sentiment class:

Positive (e.g., "I love this movie!"),
Neutral (e.g., "The day was okay."),
Negative (e.g., "I hated the food.").

Softmax activation converts raw scores into probabilities, ensuring they sum to 1. The class with the highest probability is chosen as the sentiment prediction.

3.4 Model Compilation and Training

Adam optimizer with a learning rate of 0.001 is used to compile the model. Sparse Categorical Cross entropy is applied as the loss function to understand the difference between the true class label (as an integer) and the predicted probability distribution over the classes output by the model. The model is trained using 3964 samples and the remaining 990 samples are used for validation with batch size 32 for 20 epochs.

3.5 Evaluation Metrics

Evaluation Metrics used to measure the model's performance are accuracy, precision, recall and F1-score. Accuracy measures the percentage of appropriately classified samples out of the total samples.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

Precision measures the proportion of correctly predicted positive samples out of all samples predicted as positive.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

Recall measures the proportion of actual positive samples that were correctly predicted as positive.

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

F1-score is the harmonic mean of Precision and Recall, providing a balance between the two.

$$F1 = 2\times\frac{Precision\ X\ Recall}{Precision+Recall} \tag{4}$$

The results of these metrics are discussed and presented in the next section.

4 EXPERIMENTAL RESULTS

The results of the proposed approach to the sentiment analysis in the English-Hindi code-switching dataset are shown using the confusion matrix in Table 2. The positive and negative classes attain marginally higher performance metrics, implying the model's confidence in determining these sentiments. The neutral class, exhibits slightly lower recall, leading to incidental misclassifications into positive or negative categories.

Table 2: Confusion Matrix of the Proposed Approach.

Class	Predicted Positive	Predicted Neutral	Predicted Negative
True Positive	1700	100	50
True Neutral	120	1500	80
True Negative	70	76	1258

It is observed that the model has achieved strong bilingual performance with an overall accuracy of 89.9% with better recall, precision and an average F1 score of 0.9 calculated using equations 1 to 4 respectively for positive, neutral, and negative classes shown in Table 3.

Table 3: Precision, Recall, And F1-Score for Each Class.

Class	Precision	Recall	F1-Score
Positive	0.899	0.919	0.909
Neutral	0.895	0.882	0.889
Negative	0.906	0.896	0.901
Overall Model Performance	0.9	0.899	0.899

The ROC curve represented in Figure 2 reveals that both the positive and negative classes possess high AUC values of 0.93, indicating outstanding separability. The neutral class showcases a robust ability to distinguish its predictions with an inconsiderable lower AUC of 0.91. The curves rise steeply, implying that the model achieves high True Positive Rates (TPR) with minimal False Positive Rates (FPR). All curves observed in Figure 2 overlie the diagonal chance line, confirming the model's effectiveness over random assumptions. In principle, the model performs reliably, with balanced classification capabilities across all sentiment categories.

5 CONCLUSION AND FUTURE IMPLICATIONS

This research demonstrates a notable advancement in sentiment analysis for Hindi-English code-switched texts through a novel LSTM-based model that excels in both accuracy and F1-score, key metrics for evaluating classification models. Achieving an overall accuracy of 89.9%, the model correctly classifies a high proportion of sentiments across positive, neutral, and negative categories, showcasing its reliability in real-world applications. The F1-score, a harmonic mean of precision and recall, underscores the model's balanced performance by ensuring both correctness and completeness in predictions. With an F1-score of 0.9, the study highlights the model's proficiency in capturing sentiments even in linguistically complex and contextually varied code-switched text, reflecting its ability to minimize false positives and negatives effectively. This balance is particularly important in multilingual sentiment analysis, where nuances in language transitions can pose significant challenges. By achieving such metrics, the study not only validates the robustness of its methodology but also establishes a strong foundation for addressing broader applications, including low-resource language pairs and more complex multilingual contexts, while encouraging further refinements in accuracy and sentiment detection capabilities.

Further research can be extended by exploring additional low-resource language pairs, improving neutral sentiment detection, and involving advanced deep learning architectures like transformers to enhance performance and adaptability to distinct multilingual contexts.

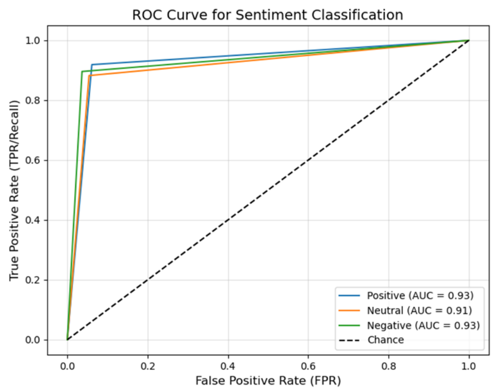

Figure 2: ROC-AUC Curve.

REFERENCES

Adel, H., Vu, N. T., Kraus, F., Schlippe, T., Li, H., & Schultz, T. (2013). Recurrent neural network language modeling for code switching conversational speech. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 8411–8415. https://doi.org/10.1109/ICASSP.2013.6639306

Agüero-Torales, M. M., Abreu Salas, J. I., & López-Herrera, A. G. (2021). Deep learning and multilingual sentiment analysis on social media data: An overview. Applied Soft Computing, 107, 107373. https://doi.org/10.1016/J.ASOC.2021.107373

Albahoth, Z. M., Jabar, M. A. bin A., & Jalis, F. M. B. M. (2024). A Systematic Review of the Literature on Code-Switching and a Discussion on Future Directions. International Journal of Academic Research in Business and Social Sciences, 14(2). https://doi.org/10.6007/ijarbss/v14-i2/20452

AlGhamdi, F., Molina, G., Diab, M., Solorio, T., Hawwari, A., Soto, V., & Hirschberg, J. (2016). Part of Speech Tagging for Code Switched Data. EMNLP 2016 - 2nd Workshop on Computational Approaches to Code Switching, CS 2016 - Proceedings of the Workshop, 98–107. https://doi.org/10.18653/V1/W16-5812

Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., & Mccrae, J. P. (2020). A Sentiment Analysis Dataset for Code-Mixed Malayalam-English. https://pypi.org/project/langdetect/

Code-switching. (n.d.). Retrieved November 11, 2024, from https://www.cambridge.org/core/books/codeswitching/11E359BFC45F331519170EE425117736

Das, D., & Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1, 600–609.

Das, S. D., Mandal, S., & Das, D. (2019). Language identification of Bengali-English code-mixed data using character & phonetic based LSTM models. ACM International Conference Proceeding Series, 60–64. https://doi.org/10.1145/3368567.3368578

Dutta, S., Agrawal, H., & Kumar Roy, P. (2021). Sentiment Analysis on Multilingual Code-Mixed Kannada Language. https://dravidian-codemix.github.io/2021/index.html

Gardner-Chloros, P. (2009). Code-switching. Code-Switching, 1–242. https://doi.org/10.1017/CBO9780511609787

Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. Proceedings of the 28th International Conference on Machine Learning, ICML 2011, 1, 513–520.

Hofstede, G. (2001). Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations. Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations, 41(7), 861–862. https://doi.org/https://doi.org/10.1016/S0005-7967(02)00184-5

Jamatia, A., Swamy, S. D., Gambäck, B., Das, A., & Debbarma, S. (2020). Deep Learning Based Sentiment Analysis in a Code-Mixed English-Hindi and English-Bengali Social Media Corpus. International Journal on Artificial Intelligence Tools, 29(5). https://doi.org/10.1142/S0218213020500141

Kasmuri, E., Fakulti, H. B., Maklumat, T., Komunikasi, D., & Maklumat, F. T. (2020). Segregation of Code-Switching Sentences using Rule-Based Technique. Int. J. Advance Soft Compu. Appl, 12(1).

Kim, H. (n.d.). 1 Semi Annual Edition.

Klementiev, A., Titov, I., & Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. 24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers, December, 1459–1474.

Kumari, J., & Kumar, A. (2021). Offensive Language Identification on Multilingual Code Mixing Text. CEUR Workshop Proceedings, 3159, 643–650. https://github.com/Abhinavkmr/Dravidian-hate-speech.git

Mamta, & Ekbal, A. (2025). Quality achhi hai (is good), satisfied! Towards aspect based sentiment analysis in code-mixed language. Computer Speech and Language, 89. https://doi.org/10.1016/j.csl.2024.101668

Myers-Scotton, C. (1993). Common and Uncommon Ground: Social and Structural Factors in Codeswitching. Language in Society, 22(4), 475–503. https://doi.org/10.1017/S0047404500017449

P, K. V., & Mahender, C. N. (2024). A Named Entity Recognition System for the Marathi Language. JOURNAL OF ADVANCED APPLIED SCIENTIFIC RESEARCH, 6(3). https://doi.org/10.46947/JOAASR632024937

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Found Trends Inf Retr, 2(1–2), 1–135. https://doi.org/10.1561/1500000011

Patwa, P., Aguilar, G., Kar, S., Pandey, S., Srinivas, P. Y. K. L., Gambäck, B., Chakraborty, T., Solorio, T., & Das, A. (2020). SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets. 14th International Workshops on Semantic Evaluation, SemEval 2020 - Co-Located 28th International Conference on Computational Linguistics, COLING 2020, Proceedings, 774–790. https://doi.org/10.18653/v1/2020.semeval-1.100

Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). No Title. 166(114), 120. https://doi.org/10.1016/j.eswa.2020.114120

Santos, C. dos, & Gatti, M. (2014). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts (pp. 69–78). https://aclanthology.org/C14-1008

Sharma, G., Chinmay, R., & Sharma, R. (2023a). Late Fusion of Transformers for Sentiment Analysis of Code-Switched Data. Findings of the Association for Computational Linguistics: EMNLP 2023, 6485–6490.

https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.430
Sharma, G., Chinmay, R., & Sharma, R. (2023b). Late Fusion of Transformers for Sentiment Analysis of Code-Switched Data. Findings of the Association for Computational Linguistics: EMNLP 2023, 6485–6490. https://doi.org/10.18653/v1/2023.findings-emnlp.430
Ullah, A., Khan, S. N., & Nawi, N. M. (2022). Review on sentiment analysis for text classification techniques from 2010 to 2021. Multimedia Tools and Applications 2022 82:6, 82(6), 8137–8193. https://doi.org/10.1007/S11042-022-14112-3
Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2016). En-ES-CS: An English-Spanish code-switching twitter corpus for multilingual sentiment analysis. Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, 4149–4153. http://grupolys.org/
Yekrangi, M., & Abdolvand, N. (2021). Financial markets sentiment analysis: developing a specialized Lexicon. Journal of Intelligent Information Systems, 57(1), 127–146. https://doi.org/10.1007/S10844-020-00630-9/TABLES/2
You, Q., Jin, H., & Luo, J. (2017). Visual Sentiment Analysis by Attending on Local Image Regions. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1), 123–125. https://doi.org/10.1609/aaai.v31i1.10501
Younas, A., Nasim, R., Ali, S., Wang, G., & Qi, F. (2020). Sentiment analysis of code-mixed roman Urdu-English social media text using deep learning approaches. Proceedings - 2020 IEEE 23rd International Conference on Computational Science and Engineering, CSE 2020, 66–71. https://doi.org/10.1109/CSE50738.2020.00017
Zadeh, A., Chen, M., Cambria, E., Poria, S., & Morency, L. P. (2017). Tensor Fusion Network for Multimodal Sentiment Analysis. EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings, 1103–1114. https://doi.org/10.18653/v1/d17-1115