Journal of Research High School

Vol. 2025 (12) 971 – 977
ISSN 2688-3651 [online]

Comparison of Different Models for Hindi-English Code-Switched Sentiment Analysis

Aditya Aiyer[1] *

[1] James Logan High School, Union City, CA, USA
*Corresponding Author: aditya12aiyer@gmail.com

Advisor: Arun Kumar Rajasekaran, rarunrd@gmail.com

Received December 27, 2024; Revised July 3, 2025; Accepted August 7, 2025

Abstract

Natural Language Processing (NLP) can aid in translating and making conclusions from data, through its ability to quickly scan and analyze text on a large scale. However, limited research has been put into developing models that can handle multiple languages, and only a few datasets involving code-switched sentences are easily accessible. Due to the increased use of code-switching, it has become important to improve models that help overcome language barriers. Especially in low-resource languages such as Hindi, not all models produce satisfactory results. This study aims to evaluate the effectiveness of the BERT-base-uncased model and the Google MuRIL model in combination with four different sequence classifiers for performing sentiment analysis on Hindi-English code-switched data. The models were tested on a Hugging Face dataset containing tweets of various sentiments. In particular, the performance of the models and respective classifiers was verified across different sentiment categories. The findings demonstrate varied performance across each model-classifier combination with the BERT-base-uncased model and BertForSequenceClassification resulting in the highest accuracy. The results also reflect the importance of sentiment biases within the dataset and how they influence the selection of models and classifiers. This work highlights the challenges for sentiment analysis in low-resource, code-switched languages. Furthermore, the paper examines the correlation of model performance concerning content sentiment, which is often overlooked.

Keywords: Code-switching, Natural Language Processing, Sentiment Analysis, Models, Sequence Classifiers

1. Introduction

Code-switching was initially documented as a linguistic phenomenon by sociologists and linguists, particularly John J. Gumperz. It is defined as "the juxtaposition of passages of speech belonging to two different grammatical systems or subsystems, within the same exchange." This alternation usually occurs when a speaker uses a second language to reiterate a message given in one language or to act as a reply (Gumperz, 1977). Yet in recent years the definition of code-switching has evolved past simply linguistics and more towards a cultural identity - code-switching alters a person's self-presentation to conform with social expectations and norms including race, ethnicity, gender, sexuality, age socioeconomic status, and disability status (Sharma, 2023). Moreover, code-switching may be implemented to add emphasis or emotion to a statement or idea, hence its connection to sentiment analysis.

Sentiment analysis is an approach to NLP that recognizes an emotional tone or purpose behind the text. Its usefulness manifests as a method for organizations to categorize reviews or opinions about a product, service, or idea. Sentiment analysis uses data mining, machine learning (ML), AI, and computational linguistics to mine text for subjective information (Gillis, 2024). The ability to detect emotion is a major advantage while reading and understanding data due to the inherent use of connotation while making a statement. Merriam-Webster defines connotation as the suggested meaning of a word apart from its explicit or literal definition. Connotation occurs because of biases and emotions, so being able to take control of these factors through sentiment analysis is highly valuable in

Vol. 2025 (12) 971 – 977
ISSN 2688-3651 [online]

NLP applications. As products are developed and the population grows, real-world applications of sentiment analysis will increase exponentially. Inevitably, cases such as product reviews will involve different languages from different geographical locations, many of which will be code-switched between multiple languages. Therefore, the importance of models that can read code-switched data and perform sentiment analysis at a large scale and high speeds will be incredibly useful.

To understand the performance levels of current models of sentiment analysis on code-mixed data, a dataset consisting of mixed code-switched tweets with varied sentiment values was used. This paper presents sentiment analysis on Hindi-English code-switched data. Considering that Hindi is the fourth most spoken language in the world with 345 million native speakers, it is a surprisingly low-resource language. Therefore, conducting a comprehensive performance analysis of various models in Hindi-English code-switched data will lay a solid foundation for future research in other low-resource languages and the broader field of AI linguistics. Previously, code-mixed datasets have been created such as the Twitter Corpus for Hindi-English Code-Mixed POS Tagging (Singh et al.) and the Hindi-English Code-Switching Corpus (Dey and Fung). However, a comparative analysis of existing models has not yet been performed to gain an understanding of the current landscape of Hindi-English code-switched NLP. Therefore, this study aims to contribute to that direction by comparing the accuracy results of two such models.

A Hindi-English code-switched dataset obtained from the works of Anjali et al. was utilized for this work. It contains code-mixed tweets and sentiment labels “negative”, “neutral”, and “positive”. To classify the data, multiple classifiers were used to understand how each classification method would impact the overall accuracy of the model. The two models that were tested were bert-base-uncased and google-muril, and for each model, four competitive sequence classifiers were used: BertForSequenceClassification, AutoModelForSequenceClassification, ElectraForSequenceClassification, and RobertaForSequenceClassification. These models and classifiers were chosen because they have already been trained in foreign languages, particularly Indian languages. In this study, different combinations will be used to assess the accuracy for each individual sentiment: negative, neutral, and positive as well as overall accuracy.

2. Research Methodology

The Google-MuRIL and BERT models were chosen for this work. MuRIL stands for Multilingual Representations for Indian Languages and is pre-trained in 17 Indian languages and transliterated counterparts (Khanuja, 2021). Therefore, overall high accuracy is anticipated from this model. Another chosen model was bert-base-uncased, which is mainly trained for English data so this would give an understanding of how well the English model would perform with code-switched data (Devlin, 2018). Because the dataset contains romanized letters, the results from the bert-base-uncased model would be worthy of analysis. To standardize the tweets, a preprocessing function was created which removed punctuation, converted text to lowercase, and stored the clean version prior to tokenization.

The sequence classifiers selected to be used with the chosen models are Bert, AutoModel, Electra, and Roberta. BertForSequenceClassification is a model configuration class with the parameters of the Bert model. AutoModelForSequenceClassification is a generic model class that is instantiated as one of the sequence classification model classes of the library. The Electra Model is a new pre-training approach that trains a generator and a discriminator, and the sequence classifier is a model configuration class with the parameters of the model. The RoBERTa model builds on the BERT model and modifies key hyperparameters, and the sequence classifier is a model transformer with a regression head on top, like Bert and RobertaForSequenceClassification.

The data from the works of Anjali et al. consists of 3100 training rows and 776 validation and testing rows each. Of the 776 testing rows, there are 114 negative, 392 neutral, and 270 positive lines. The learning rate was set at 1e-5, and 20 epochs were used for each test.

The questions that were resolved in this study that influenced the chosen methodology include determining which model was more robust, whether the selection of classifiers mattered, and whether dataset sentiment bias influenced the accuracy and selection of the model or classifier.

Vol. 2025 (12) 971 – 977
ISSN 2688-3651 [online]

Figure 1 displays the workflow of the study conducted. Two models, google-muril and bert-base-uncased, were paired with four different classifiers: Bert, AutoModel, Electra, and Roberta. Each model-classifier combination was evaluated on a dataset including positive, negative, and neutral sentiments, where each combination was assigned, an overall accuracy encompassing all three sentiments, as well as individual accuracy scores representing how well the combination performed on each sentiment.

3. Results

Each of the chosen models, along with their possible combinations of chosen classifiers, were benchmarked on the pre-processed Hindi-English dataset. Although sentiment classification remains the central focus of this study, the costs and benefits of models and classifiers when working on varying sentiments are of particular interest in this study. Keeping this in mind, the results of the classification performance are segregated into three sentiments - namely, positive, negative and neutral. The idea is to explore whether the performance of the models (as well as their corresponding classifiers) can be attributed to the original sentiment bias of the dataset.

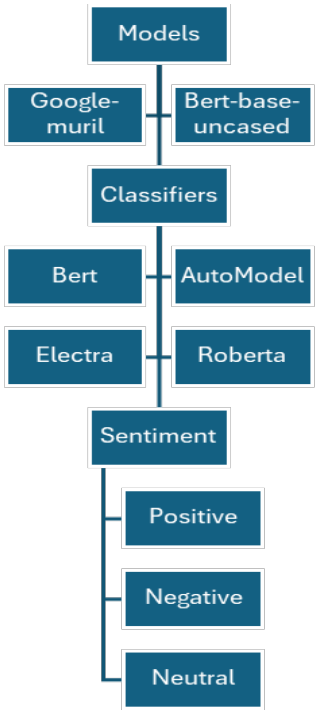

Figure 1. Flowchart of the models, classifiers, and sentiments within the dataset that were employed as part of the study. The reproducible source code with dataset and instructions can be found at https://github.com/Arunimad/Hindi-English_CodeSwitch_Analysis.

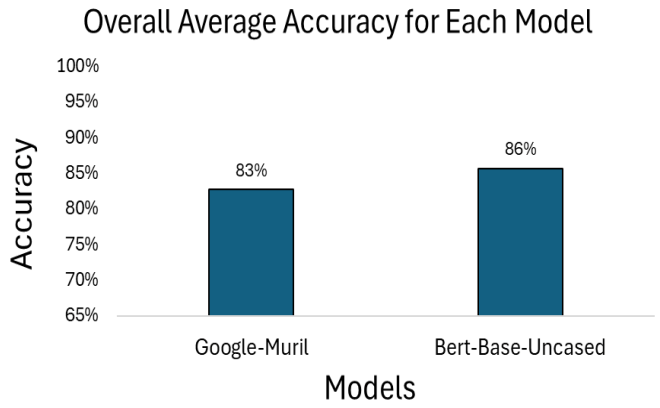

Figure 2. The average accuracy of both models used during the study. Each value was determined by averaging the individual accuracies of the model-classifier combinations

The average accuracy of both models is around 85%, with the bert-base-uncased resulting in a slightly higher average which could be the result of a margin of error. While both models seem to perform well on the selected dataset, a closer inspection was done on the individual model-classifier combinations to further examine whether the selected classifier affects the overall accuracy, see Figures 3 and 4.

Figure 3 displays the overall accuracies for the entire testing set specifically for the google-muril model with each classifier. The google-muril model performed the highest with the Roberta Sequence Classifier with an accuracy of 87%, while it performed the lowest with the Bert Sequence Classifier with an accuracy of 78%. The total range across classifiers was 9%, indicating that the choice of classification method likely did influence the resulting accuracy.

Figure 4 displays the overall accuracy for the entire testing set specifically for the bert-base-

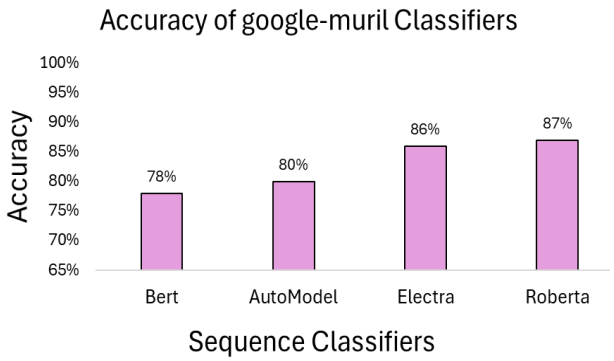

Figure 3. Overall accuracy of the google-muril model across all four sequence classifiers on the Hindi-English dataset

Vol. 2025 (12) 971 – 977
ISSN 2688-3651 [online]

uncased model with each classifier. The bert-base-uncased model performed the highest with the Bert Sequence Classifier with an accuracy of 89%, while it performed the lowest with the Electra Sequence Classifier with an accuracy of 82%. The total range across classifiers was 7%, indicating that the choice of classifier may have impacted the accuracy, but the model was generally robust across classifiers.

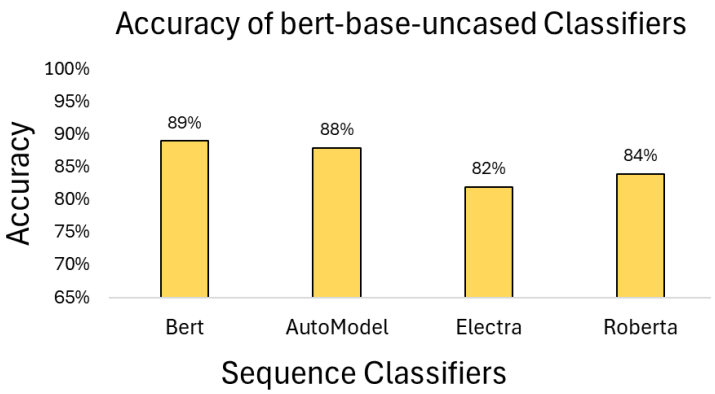

Figure 4. Overall accuracy of the bert-base-uncased model across all four sequence classifiers on the Hindi-English dataset

Based on this data, the bert-base-uncased model has outperformed google-muril both overall and concerning its robustness across classification methods in the selected Hindi-English code-switched dataset. However, a further investigation was conducted to identify changes in performance based on parts of the dataset that were sentiment-based. Each model-classifier combination was tested on the individual negative, neutral, and positive sentiments. The results are displayed in Figure 5.

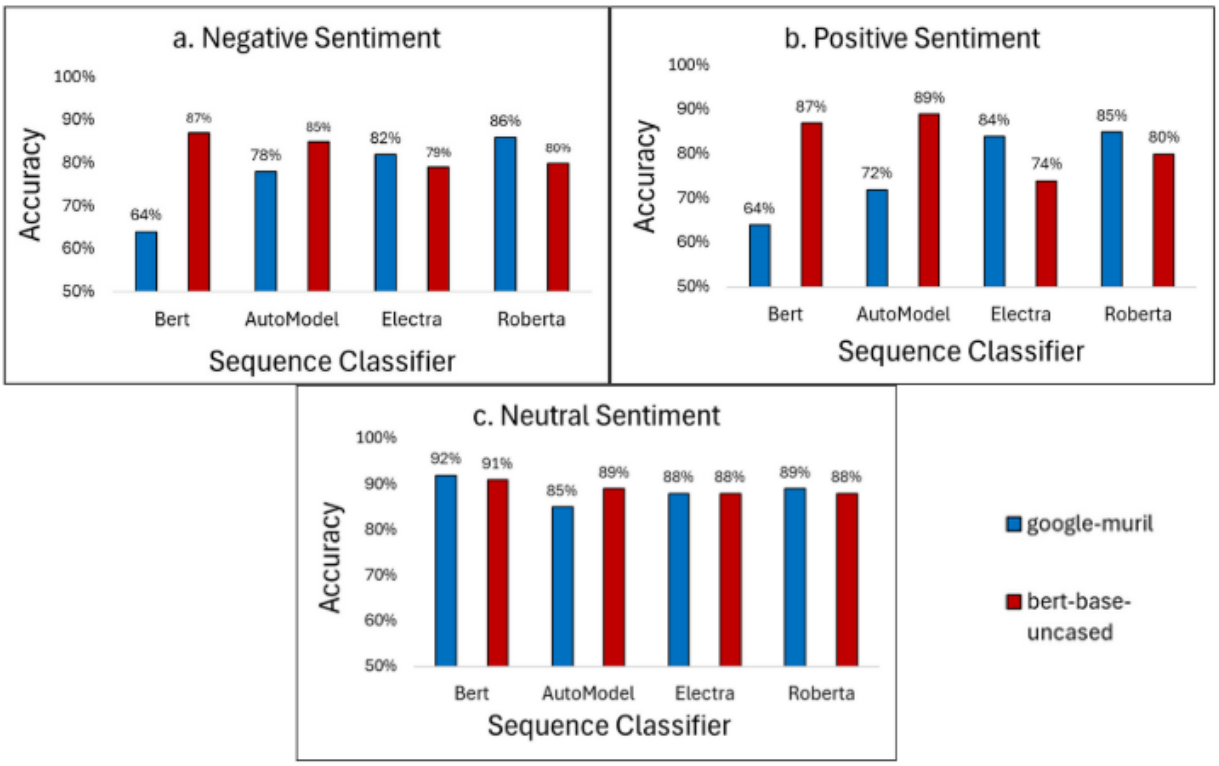

Figure 5. Accuracy of google-muril and bert-base-uncased across all four sequence classifiers in predicting individual sentiments on the Hindi-English dataset. The results are displayed on the graphs, where blue represents the results of google-muril and red represents the results of bert-base-uncased.

Figure 5a displays the results of each model-classifier combination for the negative sentiment specifically. The bert-base-uncased model with the Bert Sequence Classifier performed the highest with an accuracy of 87%, while the google-muril model with the Bert Sequence Classifier performed the lowest with an accuracy of 64%. For the negative sentiment, changes in the classification method resulted in a difference of 22% for the google-muril model, indicating that the choice of classifier had a notable influence on the accuracy of this model. However, the bert-base-uncased

Vol. 2025 (12) 971 – 977
ISSN 2688-3651 [online]

model reported a difference of 8%, with generally robust performances but still a possible influence from the classifier choices.

Figure 5b displays the results of each model-classifier combination for the positive sentiment specifically. The bert-base-uncased model with the AutoModel Sequence Classifier performed the highest with an accuracy of 89%, while the google-muril model with the Bert Sequence Classifier performed the lowest with an accuracy of 64%. The google-muril model had a difference of 21%, and the bert-base-uncased model had a difference of 15% across models, indicating that the chosen classifier influenced both models.

Figure 5c displays the results of each model-classifier combination for the neutral sentiment specifically. The google-muril model with the Bert Sequence Classifier performed the highest with an accuracy of 92%, while the google-muril model with the AutoModel Sequence Classifier performed the lowest with an accuracy of 85%. For the neutral sentiment, both models displayed a robust performance across all classifiers, indicating that the classifiers may not have influenced as much for this sentiment.

For negative and positive sentiments, the trend is that Bert and AutoModel work better with the bert-base-uncased model, while Electra and Roberta show higher accuracies with the google-muril model, as shown in Figures 5a and 5b. For the neutral sentiment, every test delivered a satisfactory result, with the lowest combination of google-muril with AutoModel still producing an accuracy of 85%, as shown in Figure 5c.

Among the overall accuracies, the model-classifier combination that returned the lowest result was google-muril with BertForSequenceClassification at 78% (Figure 3). The highest result was bert-base-uncased with BertForSequenceClassification at 89% (Figure 4).

4. Discussion

The results reveal key insights into the model behavior of google-muril and BERT-base-uncased on code-switched Hindi-English sentiment analysis. These results suggest that the bert-base-uncased model generalizes better across datasets (Figures 3 and 4). However, this is not necessarily the case. Considering the dataset with "neutral" sentiment scores (Figure 5c), it is shown that google-muril with Bert performed at 92% accuracy while bert-base-uncased with Bert performed at 91%. Given a 1% variance, the models performed equally well for the neutral sentiment. Consequently, given a dataset with almost entirely neutral values, the returned accuracy would be quite similar. This indicates that the overall accuracy of a model does not always represent the accuracy of each individual sentiment. Furthermore, the choice of classifier did influence the corresponding model for the positive and negative sentiments but had little influence on the neutral sentiment.

The discrepancy in the accuracy of the two models could be attributed to the dataset used. Since the dataset consisted of Romanized Hindi-English tweets, the English-optimized model, as BERT-base-uncased, may have performed better. It is therefore important to recognize both the language of the text that is being worked with, as well as the actual lexicon that is being used.

Despite high-accuracy results, there are several limitations of the study which could impact the application of the findings. The dataset, although containing 3100 training rows as well as 776 testing and validation rows, still could be increased in size to reflect an even more accurate model comparison. Furthermore, the dataset is generally skewed towards neutral sentiment, revealing a limitation on the potential emotional complexity of online statements. Finally, the dataset consisted of tweets, rather than general consumer reviews or statements taken across multiple social media platforms, so the accuracy of the models could vary depending on the location from which the statements were taken.

5. Conclusion

The results from the study indicate that the selection of classifiers, as well as sentiment bias within the dataset, influence the accuracy of the model. The key findings from the study were that the BERT-base-uncased model with the Bert Sequence Classifier performed the highest with an accuracy of 89%, and the google-muril model with the Bert Sequence Classifier performed the lowest with an accuracy of 78%. This data demonstrates that one must be careful when deciding on a model and classifier to use during code-switched sentiment analysis, as well as which

Vol. 2025 (12) 971 – 977
ISSN 2688-3651 [online]

hyperparameters are being used, as they influence predictive performance and significantly affect model accuracies. In a dataset with a larger number of positive or negative sentiments, one must be careful about the classifier chosen due to variances of up to 22% with google-muril. Furthermore, as previously mentioned, the overall accuracy of models may not be representative of individual sentiments.

Although the results were promising, most machine learning models are still not yet equipped to handle low-resource languages such as Hindi at accuracies of 95% or higher. More varied datasets and model architectures should be developed to improve accuracy. Regardless, considering the relative newness of online Hindi-English code-switched data, it is important to move forward in advances regarding the field of Natural Language Processing.
Future research should aim to generate larger data sets taken from various sources and to develop new models based on these findings that would perform at even greater accuracies and have general applicability across platforms. Furthermore, these findings point toward which models and classification methods could be implemented into social media platforms and the mass market.

References

Aum, S., & Choe, S. (2021). srBERT: automatic article classification model for systematic review using BERT. *Systematic reviews*, 10(1), 285.

Bhat, I. A., Bhat, R. A., Shrivastava, M., & Sharma, D. M. (2018). Universal Dependency parsing for Hindi-English code-switching. *arXiv preprint arXiv:1804.05868*.

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). 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).

Dey, A., & Fung, P. (2012). A Hindi-English code-switching corpus. In *Proceedings of LREC 2014*. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2014/pdf/922_Paper.pdf

Gumperz, J. J. (1977). The sociolinguistic significance of conversational code-switching. *Studies in the Linguistic Sciences, 8*(2), 1–34. https://doi.org/10.1177/003368827700800201

Khanuja, S., Bansal, S., Mehta, A., Singh, A., Bhatia, P., Goyal, P., & Bhattacharyya, P. (2021). MuRIL: Multilingual representations for Indian languages. *arXiv*. https://arxiv.org/abs/2103.10730

Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of natural language processing, 2(2010), 627-666.

Singh, K., Sen, I., & Kumaraguru, P. (2018). A Twitter corpus for Hindi-English code mixed POS tagging. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media* (pp. 12–17). Association for Computational Linguistics. https://aclanthology.org/W18-3503.pdf

Merriam-Webster, Inc. (1984). Merriam-Webster's Dictionary of Synonyms: A Dictionary of Discriminated Synonyms with Antonyms and Analogous and Contrasted Words. Merriam-Webster.

Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network. Ieee Access, 10, 21517-21525.

Yadav, A., Garg, T., Klemen, M., Ulcar, M., Agarwal, B., & Sikonja, M. R. (2024). Code-mixed sentiment and hate-speech prediction. arXiv preprint arXiv:2405.12929.

Vol. 2025 (12) 971 – 977
ISSN 2688-3651 [online]

Yarullin, R., & Serdyukov, P. (2020, October). Bert for sequence-to-sequence multi-label text classification. In International Conference on Analysis of Images, Social Networks and Texts (pp. 187-198). Cham: Springer International Publishing.