STM JOURNALS

International Journal of
Computer Science Languages

ISSN: 3048-944X
Volume 4, Issue 1, 2026
January–June
DOI (Journal): 10.37591/IJCSL

https://journals.stmjournals.com/ijcsl

Research IJCSL

Comparative Study of BERT Variants for Sentiment Analysis with Error Analysis

Deepshikha Prajapati[1], Shruti Prajapati[2], Thara C.[3,*]

Abstract

The use of media is going up fast in India, and this has led to the rise of Hinglish. Hinglish is an informal blend of Hindi and English that people commonly use in everyday conversations, especially across social media platforms such as Twitter, Facebook, and WhatsApp. People use Hinglish to talk to each other in a way that is not very formal. Hinglish blends English vocabulary with informal usage, often ignoring standard grammatical rules, which makes it challenging for computers to accurately interpret and process it. It is especially hard for computers to figure out how people are feeling when they use Hinglish in the media. Hinglish poses a significant challenge for natural language processing, particularly when it comes to accurately interpreting people's emotions and opinions. The problem with Hinglish is that it often has words from languages mixed together, and the spelling and sentence structure can be weird. This makes it hard for regular language models to understand. Even though models like BERT are really good at understanding text in languages, they need a lot of computer power to work. This means they are not good for situations where we need to get answers, and we do not have a lot of computer power. So, we looked at how three smaller models work: DistilBERT, Multilingual Representations for Indian Languages (MuRIL), and XLM-RoBERTa. We used the Kaggle Hinglish Sentiment Dataset to test these models. When we look at how these models work and where they make mistakes, the research helps us understand how they can handle the complexities of Hinglish language. This is important because Hinglish is a mix of Hindi and English. The research shows that these models can work with Hinglish while still being good at classifying things and not using much computer power. Studying is helpful because it adds to what we know about working with languages that do not have a lot of resources. It also helps us make systems that can figure out how people feel about things. We can use these systems in real life. The research on Hinglish models is useful for sentiment analysis systems.

Keywords: Hinglish, code-mixed text, sentiment analysis, natural language processing (NLP), transformer models

***Author for Correspondence**
Thara C
E-mail: tharachakkingal08@gmail.com

[1]Research Scholar, Master of Computer Application, Thakur Institute of Management Studies, Career Development & Research, Mumbai, Maharashtra, India
[2]Research Scholar, Master of Computer Application, Thakur Institute of Management Studies, Career Development & Research, Mumbai, Maharashtra, India
[3]Assistant Professor, Master of Computer Application, Thakur Institute of Management Studies, Career Development & Research, Mumbai, Maharashtra, India

Received Date: March 10, 2026
Accepted Date: April 11, 2026
Published Date: April 27, 2026

Citation: Deepshikha Prajapati, Shruti Prajapati, Thara C. Comparative Study of BERT Variants for Sentiment Analysis with Error Analysis. International Journal of Computer Science Languages. 2026; 4(1): 39–48p.

INTRODUCTION

The use of media is common in India, and this has made many people use Hinglish. Hinglish is a language that is a mix of Hindi and English. It is a bit of a mix, where people write words in English letters, use simple grammar, and switch between Hindi and English a lot. This makes it difficult for computers to understand what people are saying when they use Hinglish [1].

Computers have time figuring out how people feel when they read things written in Hindi. Some computer programs, such as BERT, are very good at understanding how people feel when they write in one language. These programs do not work well when people write in Hinglish. Hinglish is a

problem for natural language processing (NLP) tasks, such as figuring out what people mean when they say something [2].

Researchers have been trying to make BERT work better with languages that are mixed together, such as Hinglish, Roman Urdu, English, and Indonesian English. Patil et al. found that models made specifically for an area, such as HingBERT and HingRoBERTa, perform better than models that can handle many languages, such as mBERT, when it comes to Hindi-English datasets. A similar result was obtained by Hashmi et al., who used transformers along with topic modeling to determine how people feel about Roman Urdu tweets [3]. However, these efforts are usually held back by a few major problems. We do not have data to work with; the data we have is very specific to one area, and we do not have a standard way to measure how well something is working, like the people in five said [4].

LITERATURE REVIEW

The use of transformer-based models for code-mixed sentiment analysis is becoming popular, especially for languages such as Hinglish. Compared to other areas of NLP that only deal with one language, code-mixed sentiment analysis is still not very advanced [5]. There are some problems that we need to solve, such as the lack of sufficient datasets, the models we have do not work well in all situations, and it is difficult to understand what is going wrong when they make mistakes. Some people have tried to fix these problems by comparing models and changing BERT models to work better in specific areas. They used transformer-based models for code-mixed sentiment analysis to improve the results. Transformer-based models are important for code-mixed sentiment analysis [6].

Patil and the other people who worked with him did a study in 2023 [3]. They examined the performance of some pre-trained BERT models on Hindi-English code-mixed datasets. These BERT models include HingBERT, HingRoBERTa, and multilingual BERT. They found that the models made for an area were better than the multilingual BERT models. This was especially true for tasks such as determining how someone feels, recognizing emotions, and identifying hate speech. There were some problems with this study. They did not use many datasets. They also did not look closely at the mistakes made by the models or the confusing language [7].

The BERT models used were HingBERT, HingRoBERTa, and multilingual BERT. Patil et al. found that domain-specific models, such as HingBERT and HingRoBERTa, were better than BERT models [3].

Hashmi et al. (2024) looked at Roman Urdu and English mixed together in tweets, which are called code-mixed tweets. They wanted to see how people talked about politics during elections in Pakistan. They tried computer models, such as cm-BERT, mBART, and Electra, to determine which one worked best. They also used a technique called latent Dirichlet allocation (LDA) to determine the topics people were discussing. This way of doing things actually made their results better when they were trying to classify the tweets. Hashmi et al. worked with Roman Urdu and English code-mixed tweets to understand the discourse during the Pakistan elections [4].

The accuracy of these models is a concern. The dataset used to test them was not very large. This study focused on a specific area. This raise concerns that the models may not perform well in such situations. The models may not be sufficient when used in specific contexts. Therefore, the accuracy of the models remains a concern [8].

Astuti and her team examined how people feel when they mix English languages in 2023. They created a set of data and tested IndoBERTweet, BERTweet, and multilingual BERT to determine how well they worked for Indonesian English mixed sentiment analysis. They found that the computer programs they used often preferred English, which made it difficult to understand the complexity of the language. The way they did their study was very thorough. The size of the data they used and the fact that they did not test it in different areas made it difficult to determine whether their findings were

true in all cases. Astuti et al. studied English code-mixed sentiment analysis and wanted to know more about how well these computer programs work for Indonesian English mixed sentiment analysis [5].

Tewari et al. conducted a study in 2024. They wanted to see how people felt about things when they spoke Hinglish. Therefore, they developed a system that could understand what people meant. This system does the following. This ensures that the words are written in the correct manner. It breaks down words into parts. It also determines the meaning of slang words [7].

They used some existing and new methods to make this system work. They used techniques such as SVM and Naive Bayes. They also used LSTM and BERT models. These are all similar to computer programs that can be learned.

The system worked well. However, there were some problems. The people who created the system had to collect all the data manually. The data was not from many different places. This might have made the system not work well, as Tewari and his team were looking at Hinglish sentiment analysis. They attempted to perform a Hinglish sentiment analysis [7].

Singh et al. (2022) used a machine learning method to determine the sentiment in social media posts written in both Hindi and English. They used TF-IDF features and classifiers such as Bayes and SVM to do this. They found that they could obtain good results without using deep learning. But the problem is they did not use transformer-based models, which are the best methods we have right now, so their study is not that relevant to what we are doing today with sentiment classification, in Hindi-English social media posts.

When considering all these studies, some problems can be observed. First, the datasets they use are usually small and only focus on one area and one language pair that mixes codes, so you cannot really apply what they find to other situations. Second, few studies test different models in a standard way using the same steps to prepare the data and train the models.

The biggest problem is that they do not look at what mistakes were made. Most of the time, they only give you an idea of how accurate something is or an F1-score, but they do not try to figure out why things get misclassified, such as what is going on with the language or the context that causes the mistakes.

This study attempts to fill these gaps by closely examining error analysis. It does this by checking three versions of BERT that work with many languages. DistilBERT, MuRIL, and XLM-RoBERTa. These BERT versions were tested on the Kaggle Hinglish Sentiment Dataset. The study also uses tools such as Local Interpretable Model-Agnostic Explanations (LIME) to understand why some samples were misclassified.

By doing this, the study helps us better understand how these models work when it comes to figuring out the sentiment in text that mixes languages. The goal of this study is to help people who are working on making models create datasets and come up with ways to evaluate these models for languages that do not have a lot of resources for low-resource languages and multilingual NLP, including low-resource multilingual settings.

METHODOLOGY

The experimental design follows a planned pipeline for preparing the data to train the models, keeping track of the predictions, and examining the errors [9]. The study used the Kaggle Hinglish Sentiment Dataset. The Kaggle Hinglish Sentiment Dataset is made up of social media text data marked with three types of sentiment classes, in the Kaggle Hinglish Sentiment Dataset.

The sentiments were positive, negative, and neutral. We prepared the dataset by ensuring that slang and spelling variations were the same, such as changing "*gud*" to "good." We also removed punctuation and ensured that the labels on the dataset were consistent. The dataset was already in Roman script, so we did not need to do any extra work to make the script the same [10].

We picked three transformer models to take a look at: DistilBERT-base-multilingual-cased, Google/MuRIL-base-cased, and XLM-RoBERTa-base. These transformer models were adjusted using the Hugging Face Transformers library in Python on the Google Colab platform.

The transformer models were trained for 12 epochs each. The training process for transformer models was set up using the AdamW optimizer. We used the settings for all transformer models [11] and examined the performance of the model using metrics such as accuracy, precision, recall, and macro F1-score. Scikit-learn was used to determine these numbers for the models. The models were evaluated using these metrics, specifically accuracy, precision, recall, and macro F1-score to determine their performance [12].

To determine where things went wrong, we kept track of what the model predicted for each test. We wrote down the answer, what the model said, how sure the model was, and what the model was looking at. We then grouped the mistakes into types, such as when the model said something was there when it was not, when the model missed something that was really there, and when the model was not really sure but got it right anyway. We also looked at things like slang, ways of spelling words, switching between languages, and sarcasm to see if these things were related to the mistakes the model made with the error analysis, especially with the model predictions and the linguistic features associated with errors, such as slang and sarcasm, in the error analysis [13]. We used visualization tools, such as LIME to see which words affected the decisions made by the model. This is especially important when the model makes mistakes. We wanted to know which words or tokens influenced the model's decisions in those cases (Figure 1). The LIME tool helped us understand this phenomenon [14].

RESULTS AND DISCUSSIONS

The evaluation results show some differences in the performance of the models. MuRIL obtained an accurate score of 27.5 percent. They also had precision and recall scores of 0.3333 and 0.0811, respectively. The macro F1-score for these models was 0.1401.

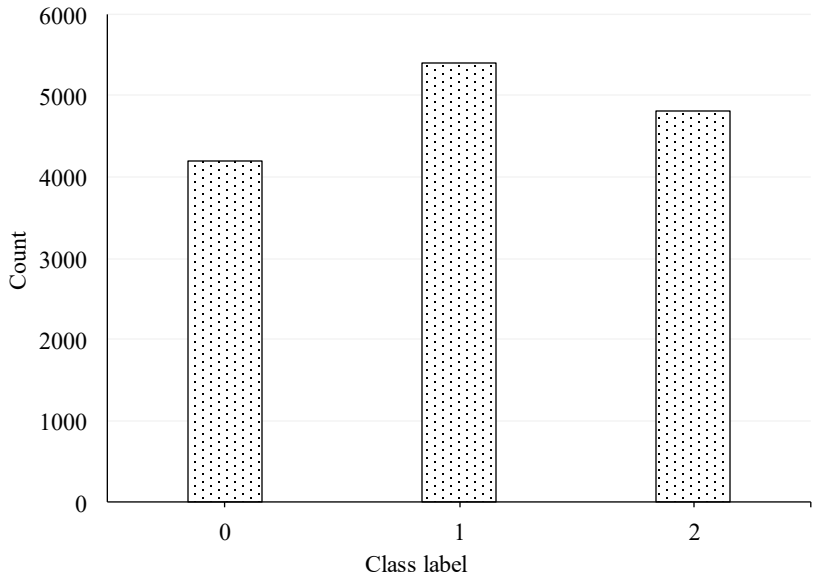

Figure 1. Label distribution across the dataset.

DistilBERT and MuRIL had a hard time doing a good job with neutral and positive sentiments. These models, DistilBERT and MuRIL, were biased towards the class. This means DistilBERT and MuRIL often had trouble predicting the correct sentiments for neutral and positive classes. The confusion matrices we examined showed that all predictions were for only one class. This led us to believe that there was a problem with the way the system worked. It does not understand the Hinglish text. It is not good at making predictions for different classes. The system lacks generalization capability. It does not have a good semantic understanding of the Hinglish text [15].

XLM-RoBERTa did a lot better. It got an accurate score of 36.67 percent and a recall of 0.8048. The macro F1-score for XLM-RoBERTa was 0.5200. The precision was still at 0.3333. Xlm-RoBERTa was able to make predictions across the 3 sentiments pretty well. When we look at the confusion matrix for XLM-RoBERTa, we can see that it did a job of classifying things. There were false negatives. It got more of the Neutral and Positive categories right. XLM-RoBERTa really showed improvement in these areas [16].

Error analysis helps us understand how the models work. We found that the models often had trouble with sentences that were short or difficult to understand. They were not very sure about what they were doing. The models thought they were right half the time, even when they were wrong. This shows that the models were confused. When we looked at the LIME pictures, we saw that the models were paying attention to the words [17]. They were missing the words that showed how someone felt, especially when the sentences contained slang or sarcasm. Error analysis of the models shows that the models need to improve their understanding of errors.

The models need to work on understanding sentences with slang or sarcasm. The models need to pay attention to the right words. One example is the sentence "*ye movie bahut gud thi,*" which XLM-RoBERTa got right by saying it is positive. On the other hand, MuRIL said it is neutral, which is incorrect. The sentence "*ye movie bahut gud thi*" has signs that show how people feel about it. Therefore, it is surprising that MuRIL did not classify the sentence "*ye movie bahut gud thi*" as positive, as XLM-RoBERTa did [18].

We wanted to look into this more, so we found out that people spelling things differently and using language were some of the main reasons for errors. The models did not often understand words that were written in English letters or phrases that needed context to make sense. When people switch between languages, it also causes problems. Sentences with a lot of code-switching, meaning they had a lot of English words mixed together, were more likely to be incorrectly classified. The models had trouble with code-switching, especially when Hindi and English words were used together in the same sentence, such as regions of text that kept switching between Hindi and English words. When we consider the mistakes that these models make, we see that DistilBERT and MuRIL often make the same mistakes [19].

This is interesting because it means that DistilBERT and MuRIL make mistakes. On the other hand, XLM-RoBERTa makes different kinds of mistakes. This could mean that XLM-RoBERTa is better at understanding the context of what it is looking at. The mistakes that XLM-RoBERTa makes are not the same as those made by DistilBERT and MuRIL (Figures 2–5). This tells us how each of these models can understand the context. We think that XLM-RoBERTa might be better because its mistakes are more varied (Table 1).

Table 1. Model comparison.

Model	Accuracy	Precision	Recall	Macro F1
DistilBERT	0.2800	0.3333	0.0933	0.1567
MuRIL	0.2808	0.3333	0.0933	0.1567
XLM-RoBERTa	0.2808	0.3333	0.1111	0.1689

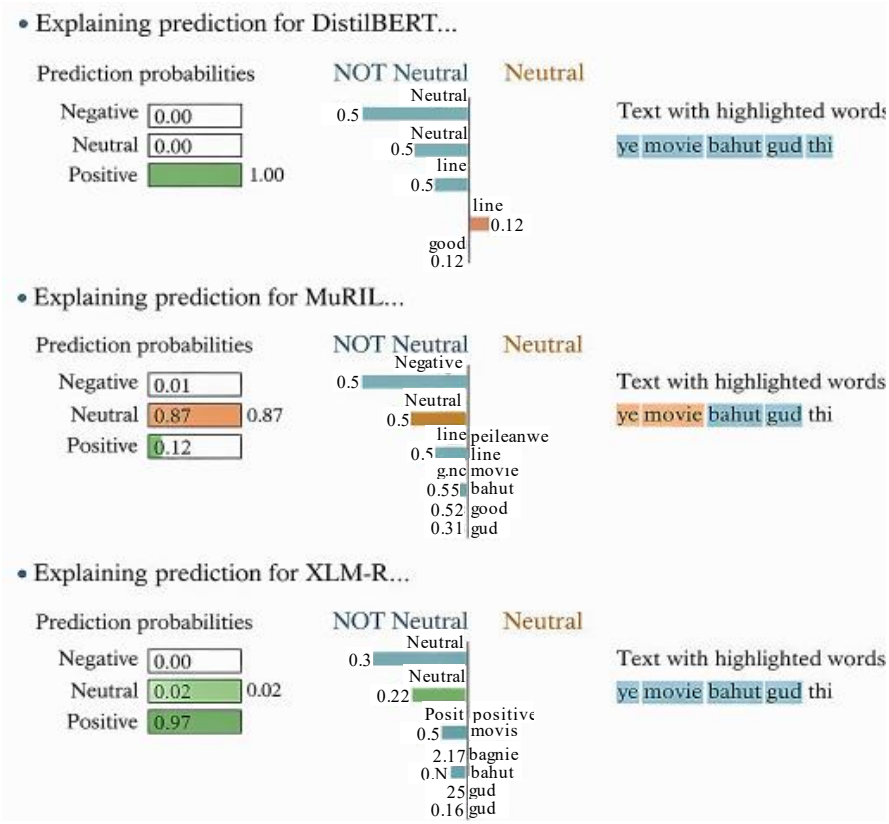

Figure 2. Sentiment prediction comparison for Hinglish input across DistilBERT, MuRIL, and XLM-RoBERTa models.

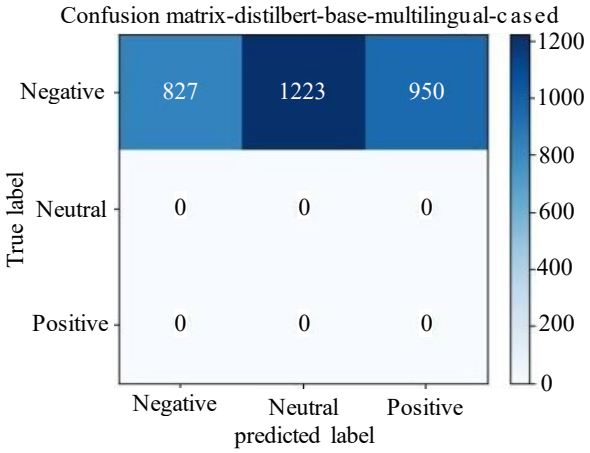

Figure 3. Confusion matrix and evaluation plot of the DistilBERT model.

CONCLUSION

1. This study examines how three multilingual transformer models, DistilBERT, MuRIL, and XLM-RoBERTa, work for Hinglish sentiment classification.
2. It compares them. Check for errors.
3. The results show that XLM-RoBERTa is much better than the other models in terms of recall and macro F1-score.
4. This means it is better at understanding the language and covering all classes.
5. The study found that DistilBERT and MuRIL are models, but they have some problems.
6. They are biased towards classes and are not very good at handling code-mixed language properties.

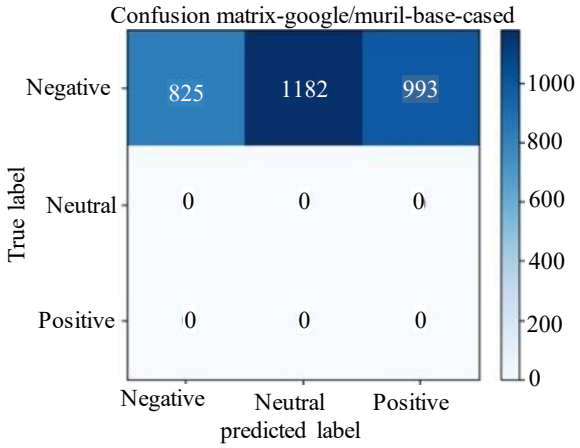

Figure 4. Confusion matrix and evaluation plot of MuRIL model.

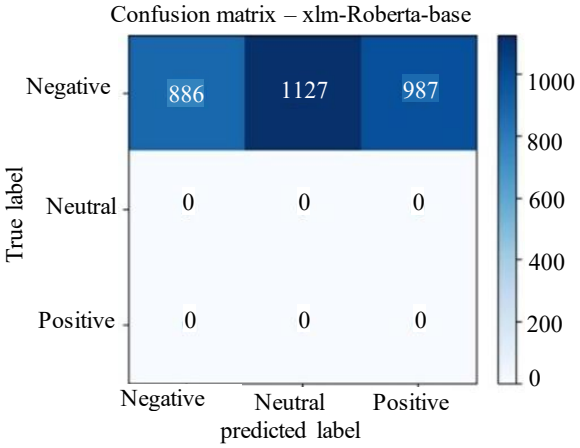

Figure 5. Confusion matrix and evaluation plot of the XLM-RoBERTa model.

7. XLM-RoBERTa is just better at this.
8. The research shows that XLM-RoBERTa is the model for Hinglish sentiment classification.

The analysis of errors helps us understand why we need tools to understand what the models are doing and to add notes to the language. The models get things because of slang, sarcasm, words that are not spelled correctly, and people switching between languages a lot. When we look at the pictures from LIME, we can see that the models are paying attention to the things and getting confused about the individual words, which shows that the models need to be trained on the right kind of information and that we need to get the language ready for them to use. This is important for the proper functioning of the models. It is all about the errors that the models make and how we can fix them by using interpretability tools and linguistic annotation.

In the future, we can try to add some things to our work. We can try to make the script look the same everywhere. We can also try to map out the slang language people use and add it to our system. If we do this, our model might work better.

We can also improve our dataset by adding things to it. We need to obtain data from many places and add more details to the language we are studying. This will help us understand where we are going wrong in our predictions.

We should also try out models such as IndicBERT or mDeBERTa. These models might help us understand how people feel when they speak in Hinglish. This is important for life situations.

Acknowledgment

We would like to thank Mrs. Thara, our mentor, for being a part of our research journey. Mrs. Thara is very important to us. We are grateful to Mrs. Thara for her support and good advice throughout our research.

The research and report were improved by the recommendations of the professional critique and constructive feedback provided. These recommendations and feedback were very important in deciding what the objectives of the research should be, what the focus should be, and how to make the research and the report better. Her carefulness and detailed planning were remarkable. She also provided a lot of encouragement, which ensured that the research was conducted in a manner.

This was conducted in a way that made sense because of her careful planning and encouragement. The thoughtful recommendations and constructive feedback from the research were helpful.

We would like to thank Dr. Sonal Sharma, the teacher who helped us with this research project, for being there to guide us, support us with our schoolwork, and give us feedback when we needed it. Dr. Sonal Sharma has a lot of experience, and her comments helped us improve our research plan and methods and helped us understand our results and ensure that they were good. Dr. Sonal Sharma was a help when we got stuck on some tough technical parts of the research, and she helped us make our research better and more thorough.

I would like to thank the faculty and staff of our department. They provided us with a good academic environment. We also used research resources, which helped us with technical issues when needed. The faculty and staff of our department were always willing to help. This facilitated the completion of our research project. It was a success. The faculty and staff of our department provided us with significant assistance.

We thank our peers and colleagues for their assistance. They provided us with ideas and shared their thoughts. Our peers and colleagues also provided us with helpful suggestions and criticism. This helped us see things from different perspectives. Our peers and colleagues helped us ensure that our ideas and approaches for this study were good. The support from our peers and colleagues was really great.

We want to thank the institutions, datasets, libraries, and online resources that we used to run the BERT variants and see how well they work. We needed to use these institutions, datasets, libraries, and online resources to test the project and compare the BERT variants.

The BERT variants were evaluated, and the results were compared based on institutions, datasets, libraries, and online resources.

The institutions, datasets, libraries, and online resources helped us with our research on the BERT variants and the results we obtained.

We would like to thank everyone and every organization that helped us with this research. They provided us with the guidance, support, and resources we needed to complete the task. This really helped us do our job and ensure that everything was done properly. The research was successfully completed because of the help we received from these people and organizations.

REFERENCES

1. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 1, Long and Short Papers; 2019 Jun;

Minneapolis, MN, USA. Stroudsburg (PA): Association for Computational Linguistics; 2019. p. 4171–4186. doi:10.18653/v1/N19-1423.
2. Çetinoğlu Ö, Schulz S, Vu NT. Challenges of computational processing of code-switching. In: Diab M, Fung P, Ghoneim M, Hirschberg J, Solorio T, editors. Proceedings of the Second Workshop on Computational Approaches to Code Switching; 2016 Nov; Austin, TX, USA. Stroudsburg (PA): Association for Computational Linguistics; 2016. p. 1–11. doi:10.18653/v1/W16-5801.
3. Patil A, Patwardhan V, Phaltankar A, Takawane G, Joshi R. Comparative study of pre-trained BERT models for code-mixed Hindi-English data. 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India. 2023. p. 1–7. doi:10.1109/I2CT57861.2023.10126273.
4. Hashmi E, Yayilgan SY, Shaikh S. Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers. Soc Netw Anal Min. 2024;14:86. doi:10.1007/s13278-024-01245-6.
5. Astuti LW, Sari Y, S. Code-mixed sentiment analysis using transformer for Twitter social media data. Int J Adv Comput Sci Appl. 2023;14(10). doi:10.14569/IJACSA.2023.0141053.
6. Sampath KK, Supriya M. Transformer based sentiment analysis on code mixed data. Procedia Comput Sci. 2024;233:682–691. doi:10.1016/j.procs.2024.03.257.
7. Tewari P, Gumber M, Tyagi S, Seekhwal P. Hinglish text analysis: challenges and opportunities in multilingual natural language processing. In: Proceedings of the International Conference on Innovative Computing & Communication (ICICC 2024); 2024. SSRN Electron J. 2025. doi:10.2139/ssrn.5193470.
8. Jadon AS, Parmar M, Agrawal R. Hinglish sentiment analysis: deep learning models for nuanced sentiment classification in multilingual digital communication. 2024 2nd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT), Dehradun, India. IEEE; 2024. p. 318–323. doi:10.1109/DICCT61038.2024.10533057.
9. Singh SK, Sharma A, Sahil D, Singh D, Pandit S, Saghir U. Sentiment analysis of English-Hindi code-mixed text using mBERT model. 2025 3rd International Conference on Inventive Computing and Informatics (ICICI), Bangalore, India. 2025. p. 552–556. doi:10.1109/ICICI65870.2025.11069692.
10. Mohana Priya KT, Shrinithi G, Nithish P, Pranesh AC. Comparative analysis of transformer models for sentiment classification in code-mixed Indic languages. Int J Eng Res Sustain Technol. 2025;3(1):1–9. doi:10.63458/ijerst.v3i1.101.
11. Almalki SS. Sentiment analysis and emotion detection using transformer models in multilingual social media data. Int J Adv Comput Sci Appl. 2025;16(3). doi:10.14569/IJACSA.2025.0160332.
12. Aliyu Y, Sarlan A, Danyaro KU, Sani Abd Rahman AS, Muazu AA, Abubakar MY. Deep learning techniques for sentiment analysis in code-switched Hausa-English tweets. Int J Inf Manag Data Insights. 2025;5(1):100330. doi:10.1016/j.jjimei.2025.100330.
13. Ramesh G, Doddapaneni S, Bheemaraj A, Jobanputra M, Ak R, Sharma A, et al. Samanantar: the largest publicly available parallel corpora collection for 11 Indic languages. Trans Assoc Comput Linguist. 2022;10:145–162. doi:10.1162/tacl_a_00452.
14. Nazir MK, Faisal CN, Habib MA, Ahmad H. Leveraging multilingual transformer for multiclass sentiment analysis in code-mixed data of low-resource languages. IEEE Access. 2025;13:7538–7554. doi:10.1109/ACCESS.2025.3527710.
15. Mamta EA, Ekbal A. Transformer based multilingual joint learning framework for code-mixed and English sentiment analysis. J Intell Inf Syst. 2024;62:231–253. doi:10.1007/s10844-023-00808-x.
16. Veeramani H, Thapa S, Naseem U. MLInitiative@WILDRE7: hybrid approaches with large language models for enhanced sentiment analysis in code-switched and code-mixed texts. In: Jha GN, Sobha L, Bali K, Ojha AK, editors. Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation; 2024 May; Torino, Italia. Paris: ELRA and ICCL; 2024. p. 66–72.

17. Thakur V, Sahu R, Omer S. Current state of Hinglish text sentiment analysis [Preprint]. SSRN Electron J. 2020. doi:10.2139/ssrn.3614442.
18. Yuan LS, Ming LT. Sentiment prediction using multilingual bidirectional encoder representations and cross-lingual language model robustly optimized BERT approach from transformers on code-mixed text. 2025 IEEE International Conference on Computation, Big-Data and Engineering (ICCBE), Penang, Malaysia. 2025. p. 901–905. doi:10.1109/ICCBE65177.2025.11256168.
19. Kumar A, Susan S. Supervised sentiment analysis of movie reviews with SHAP-based interpretability analysis. In: Swaroop A, Virdee B, Correia SD, Polkowski Z, editors. Proceedings of Data Analytics and Management. ICDAM 2024. Lecture Notes in Networks and Systems. Vol. 1300. Singapore: Springer; 2025. p. 381–388. doi:10.1007/978-981-96-3361-6_28.

