Sentiment Analysis in Social Media Text Using NLP

Shanmugavalli S*[1], Khanaa V[2]

[1]Dept. of Computer Science, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India
[2]Dept. of Information Technology, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India

shanmugavallimurthy@gmail.com

Abstract. The rapid increasing social media has resulted in the creation of huge amounts of multilingual user-generated content, which complicates the task of sentiment analysis with the occurrence of such issues as language diversity, informal style of writing, use of slangs, transliteration, and often, code-mixing. Conventional machine learning and deep learning methods such as SVM, CNN, and LSTM usually lack the capability of contextual meaning across languages and this leads to a decline in the classification accuracy. This study seeks to overcome these shortcomings by proposing an NLP transformer-based model (with the BERT (Bidirectional Encoder Representations in Transformers model) to perform multilingual sentiment analysis of social media text. The bidirectional attention mechanism of BERT allows a deep contextual interpretation of words in sentences and is therefore very useful in semantically understanding relationships in mixed and low-resource language conditions. The suggested methodology implies the use of preprocessing (noise elimination, text normalization and tokenization) and then refining multilingual BERT (mBERT) with a multilingual sentiment set. Empirical analysis of BERT with metrics like accuracy, precision, recall, and F1-score proves that the method is far superior to the traditional methods since it is able to provide better sentiment classification performance on a variety of linguistic inputs.

Keywords: Sentiment analysis, Emotional detection, text classification, machine learning, classification algorithms, feature selection, NLP.

1 Introduction

Social media, through the growth of Twitter, Facebook, Instagram, YouTube, and the regional microblogging networks, has revolutionized how individuals act and express their views. Millions of multilingual users create a huge amount of text each day sharing their ideas about politics, entertainment, services, products, and events of the current society [1]. This stream of uncontrollable user content offers worthy solutions to garner insights into the sentiment of the people and behavioral patterns. Nevertheless, the

S. P. Vijayaragavan et al. (eds.), *Proceedings of the Global Conference on Sustainable Energy Systems, Smart Electronics and Intelligent Computing (GCSESEIC 2025)*, Advances in Engineering Research 297,
https://doi.org/10.2991/978-94-6239-654-8_41

formal analysis of such large, unstructured, and highly heterogeneous text is a complicated process, because the linguistic styles prevalent in the social media are informal [2]. The user language often involves the use of slang, abbreviations, emojis, sarcasm, and inconsistent grammar that makes the automated interpretation very challenging. This is made even more challenging in multilingual situations where the users tend to alternate between two or more languages within the same sentence- this is referred to as code-mixing or code-switching. As an illustration, Indian users tend to blend English with Tamil, Telugu, Hindi, or other local languages, and also, they habitually employ transliteration, whereby, local language words are typed with the English alphabet. Conventional sentiment analysis methods like lexicon and machine learning models like SVM, Naive Bayes, RNN, and LSTM assume a lot of handcrafted features and sequential learning methods, which is not enough to model contextual meaning and semantic relationship in mixed-language and noisy text data [3].

Consequently, the accuracy is usually low and the generalization poor in practical multilingual applications using these methods [4]. Transformer-based language models, in particular BERT (Bidirectional Encoder Representations from Transformers) have become the new frontier in Natural Language Processing (NLP) to overcome such obstacles. BERT is bidirectional self-attention, which learns contextual meaning in both directions of a sentence, allowing more semantic understanding of a sentence than traditional models [5]. The multilingual version of it, mBERT, can support over 100 languages within one unified model and can effectively learn cross-lingual representations, which is why it is very suitable when it comes to multilingual sentiment analysis tasks. This study aims to come up with a powerful and scalable sentiment analysis system on multilingual social media text using BERT. The research will manage such issues as language diversity, non-formal writing, transliteration, and code- mixing besides enhancing the accuracy and dependability of sentiment classification [6]. In the suggested system, they will use advanced methods of preprocessing, as well as fine-tune the BERT architecture on multilingual databases, to categorize sentiments as negative, positive, or neutral. The final aim of this study is to promote real-time sentiment detection in applications like public opinion tracking, customer feedback assessment, brand tracking, political analytics and emergency response systems and thus lead to better decision making in multilingual environment [7].

2 Literature Survey

Initial sentiment analysis studies were mostly based on lexicon-based methods. As an example, Sarkar (2025) suggested early lexicon-oriented sentiment classification approaches, which involve the use of dictionaries where the entries are annotated with

semantic orientation (polarity and strength), as well as the intensification and control of negation such that sentiment polarity classification can be done without large training sets [7]. On the same note, Pota et al. (2021) described how lexicon-based methods (they use sentiment dictionaries such as SentiWordNet) are still in use because they are domain-independent and interpretable. Such lexicon-based approaches, however, have drawbacks when applied to social media text: they are poor at informal language and slang, emoticons, code-mixing or transliteration, and do not have a strong coverage of low-resource and non-English languages [5].

Since lexicon-based sentiment analysis was not able to cope with complex, noisy, multi-linguistic text, scientists resorted to machine-learning based classification. Support Vector Machine (SVM), Naive Bayes, Logistic Regression and Random Forest amongst others began to be widely used. In a recent survey by Sarkar (2025), systems of machine-learning based sentiment analysis were compared and were observed to continue to serve as a popular baseline in multilingual systems. Although these techniques were better than lexicon-only techniques and could utilize such features as n-grams, TF-IDF, and other statistical word features, they nonetheless entail much feature engineering. More importantly, they do not have deep semantic knowledge and contextual information thus restricting their performance on informal, code-mixed or cross-lingual social media texts [7].

As deep learning emerged, the use of neural architecture to learn semantic and sequential contexts of text became common in a variety of research. As an illustration, Talaat (2023) suggested hybrid deep-learning sentiment analysis models that combine convolutional and recurrent neural networks to leverage both local feature extraction and contextual sequence modelling and demonstrated that these hybrid models are superior to models built upon single neural networks [1]. Moreover, Patravali et al. (2023) suggested that deep learning-based sentiment models can be effective even on non-standard languages and dialectal text [4]. Within the past few years, other studies have shown CNN–BiLSTM models enhanced with additional features to improve capturing of long-range semantic associations and emotion recognition over microblog data (social media), such as Lin et al. (2025). These advancements in regard to traditional machine learning particularly in handling informal and noisy text are considered to be significant. However, these models continue to be constrained when it comes to a complete capture of cross-lingual semantics, in dealing with text that is transliterated, or in dealing with multiple languages in a cohesive way without large amounts of language-specific labeling data [3].

Zhang et al. (2025) explicitly explored sentiment analysis of social media data, addressing challenges unique to user-generated content such as noise, informal style, and text length, and suggested enhanced preprocessing techniques and hybrid deep-learning frameworks [2]. More recently, semantic bottlenecks have been addressed using transformer-based embeddings and hybrid BERT–CNN/LSTM pipelines. An example from Nkhata et al. (2025) suggests using BERT with BiLSTM-based

architectures to improve sentiment classification for languages with dialectal and low-resource characteristics [10].

The latest and most encouraging trend in sentiment analysis has been towards transformer-based models. Using the example of Bello et al. (2023), who proposed a BERT-based framework combined with neural architectures such as CNN, RNN, and BiLSTM for sentiment analysis of tweets, the BERT-enhanced models achieved state-of-the-art results, demonstrating the effectiveness of contextual embeddings in analyzing social media text [6]. Recent studies such as Najafi and Varol (2023) and Bilehsavar et al. (2025) further show that transformer-based models with strong contextual and cross-lingual representations are especially well-suited for multilingual, code-mixed, and noisy social media data [8]. The selection of preprocessing strategies, embeddings, and architectural design remains a key factor in achieving high-performance sentiment classification [9].

3 System Model

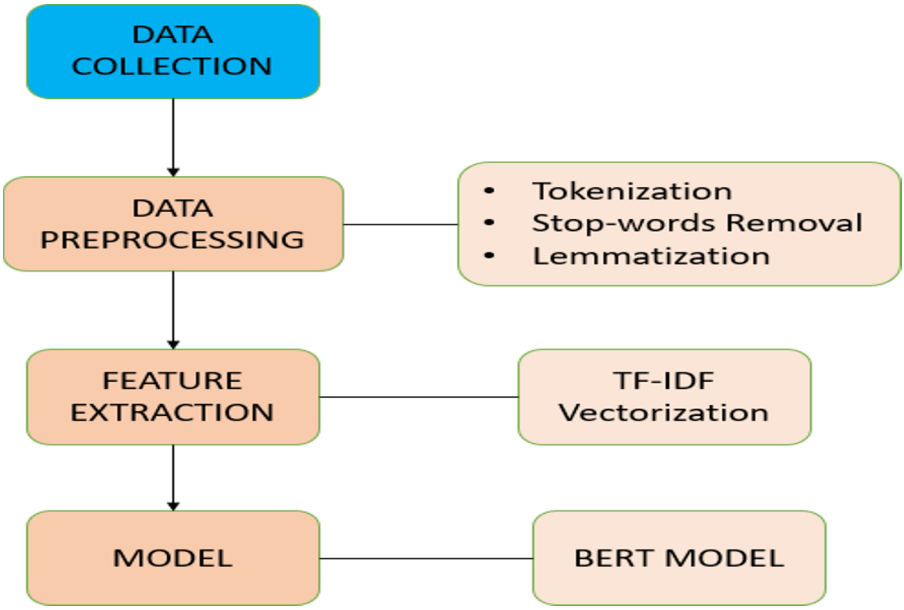

Fig.1. System Model

4 Methodology

This research methodology entails an organized workflow that will be used to construct a correct and scalable model of multilingual sentiment analysis by applying Natural Language Processing and using the BERT transformer architecture shown in Fig.1. The process starts with the data collection, in which the multilingual and code-mixed user-generated text is collected on different social media sites including twitter, Facebook, YouTube comments, and publicly available multilingual sentiment dataset. The gathered data consists of posts with mixed language format, transliterated texts and informal writing format that displays an actual communication of the real world. All the entries are either manually or semi-automatically labeled with sentiment positive, negative, or neutral to facilitate supervised learning- based model construction.

After the data collection, it is highly preprocessed to improve the quality of data and eliminate noise prevalent in text of social media. Cleaning of the raw data is done first by eliminating redundancy between records, unnecessary information and records with missing values. The text is normalized using text normalization methods to ensure uniformity by converting the text to lowercase and eliminating punctuations, URLs, special characters, hashtags, mentions, and numeric characters. The text is then tokenized to divide it into separate words or meaningful tokens to facilitate linguistic analysis in detail. Stop word removal will help to remove words that occur frequent like, and, the, is, etc which do not contribute meaning to the semantics. The use of lemmatization minimizes words to their base or root word, attracting uniformity, and minimizing vocabulary - such as the words running, runs and ran are reduced to the root word run. The tokens are then reassembled back into a normalized text string that is then useful in later stages.

After the preprocessing is done, to convert the text data into numbers, feature extraction is performed based on the term frequency-inverse document frequency or TF-IDF vectorization. TF-IDF gives significance to the frequency of words within a particular document in comparison to their frequency within the entire set of data allowing the model to distinguish between meaningful and frequent words or words that are common and repeated. This representation gives a basis to first language understanding of language and acts as a comparison metric benchmark prior to the use of transformer-based embeddings.

Lastly, the BERT model is applied on sentiment classification and makes use of its bidirectional self-attention mechanism to extract contextual semantics in mixed-language and informal text. The BERT architecture is trained on the preprocessed data in its fine-tuning mode and allows the model to acquire language regularities, emotional

expressions, and relationships between sentiment polarity and other elements utilizing the logged prepared input data. The model is optimized to determine the sentiment as positive, negative, and neutral based on the classification accuracy with the aid of proper training parameters (learning rate, batch size, and the number of epochs). The evaluation of performance is carried out on the basis of the following measures of accuracy, precision, recall, and F1-score as the measures of classification effectiveness. By employing this methodology, the research will seek to come up with one of the most efficient multilingual sentiment analysis models that has the capacity to perceive real time social media dialogues with higher context-awareness and strength. The TF-IDF weight of a term t in a document d is computed as,

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t) \quad (1)$$

$$\text{TF}(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}} \quad (2)$$

$$\text{IDF}(t) = \log\left(\frac{N}{n_t}\right) \quad (3)$$

5 Implementation

5.1 Data Collection

One of the primary stages in this research is the data collection stage because the quality and even the relevance of the data will determine the effectiveness and the reliability of the offered model. The data to be used in the proposed research is text-based data collected in publicly-available Internet resources like research archives, government documents, news publications, social networks, and scholarly data sources that are pertinent to the field of inquiry. These were the sources that were chosen so that the textual information of both structured and unstructured data would be well represented. The dataset gathered comprises of different classes that have both labelled and unlabelled textual samples that are needed to generate and test the model. The text description, target class label, the publication details, and the metadata, which allow the meaningful classification and analysis, are included in each instance of the dataset. The first level of data analysis was performed to maintain the ethical handling of data without any confidential or personally identifiable information. The gathered data were checked manually to ensure that it was authentic, consistent and complete before preprocessing and model development. This multidimensional collection approach ensures that the data set is balanced, representative and can be used in advanced applications of NLP-based deep learning purposes [10].

5.2 Preprocessing

The preprocessing process is a significant step in making the textual data ready to develop the deep learning model since the raw text usually includes noise, inconsistency, and irrelevant data that may have adverse impacts on the model performance. The preprocessing stage starts with data cleaning which involves removing duplicated entries, incomplete records and irrelevant fields to ensure integrity of data and less bias. Managing the missing values will mean that any empty text field types will be filled with meaningful placeholders or will be eliminated to prevent inappropriate results. Another step is text normalization which involves changing all characters to lowercase and eliminating special characters, punctuations, numerals and unnecessary spaces to maintain consistency throughout the data set.

After normalizing, the text is tokenized, a process that consists of breaking down sentences into separate words or tokens, which allows the model to analyze words at the word level to gain a better understanding of the semantic level. Stop word removal is done after tokenization which removes the words that are frequently repeated like those that contain no meaningful information to the contextual representation like words like and, the, is and of. The second step is the lemmatization that simplifies words to their base form using linguistic rules, e.g., running and runs become run, thus reducing the vocabulary size and increasing the processing speed. After refining all the tokens, the dataset is then token joined, processed tokens are reassembled into a text structure that can be easily vectorized and fed into a model. This end- to-end preprocess pipeline guarantees that the textual data is clean, meaningful and well formatted and results in high quality feature extraction and increases the accuracy of the deep learning model.

5.3 Feature Extraction

The textual data is then processed with the Term Frequency - Inverse Document Frequency (TF-IDF) feature extraction algorithm, a popular feature extraction algorithm used in Natural Language Processing (NLP) and information retrieval. TF-IDF takes into consideration the significance of a word in a particular document compared to the rest of the corpus that, in turn, enables the model to distinguish between ordinary and meaningful words. The Term Frequency (TF) element is the frequency of a given word in a specific document. Common words in a document are said to represent the information in the document better because they may have valuable contextual details. mathematically, TF is calculated by dividing the number of times that a term appears with the total number of words within the document. Conversely, Inverse Document Frequency (IDF) element is used to measure the importance of a term in all the documents within the dataset. Common words that are present in a large

number of papers like common stopwords get lower values of IDF since they do not offer much discriminative ability. IDF is defined as the logarithm of the total number of documents in a division by the number of documents that include the term. TF-IDF score is calculated by multiplying TF and IDF and this points out those terms that are also common in a given document but uncommon in the whole corpus. TF-IDF can be used to minimize noise caused by generic words and increase the representation of an individual document by focusing on unique words. This is used as a backbone feature set by machine learning or deep learning models and allows the effective capture of text semantics and text context, especially in sentiment analysis, document classification and topic extraction.

5.4 Model implementation

Sentiment analysis in this study is realized with BERT (Bidirectional Encoder Representations from Transformers), which is one of the latest pre-trained language models created by Google and contributed to the significant progress of many NLP problems. BERT uses a bidirectional attention mechanism, which is built on Transformer architecture, allowing it to look at the entire context of a word by simply as well as simultaneously paying attention to the surroundings on the left and the right of a word. In contrast to the traditional sequential models that process the text either left-to-right or right-to-left, BERT considers the entirety of the context of a word. This bi-directional feature will enable more detailed semantic connections and contextual clues that tend to be valuable in the analysis of informal and multilingual and code-mixed social media text.

BERT is pre-trained initially on massive corpora with two major tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM is based on a random masking of some words in the input text, which trained the model to predict those masked words by using contextual information, and thus taught the model enriched contextual word representations. NSP is a model that learns how to predict whether a particular sentence succeeds another in the original text similar to how human language learns to form relationships between sentences. Such broad pre-training provides BERT with a broad knowledge of syntax, semantics, and cross-sentence dependencies, which can be used on downstream tasks.

In the sentiment analysis task, the fine-tuning technique is used to make BERT classify the text into positive, negative or neutral sentiment categories. Every input sequence is prefixed with a special [CLS] token, which combines contextual information of the whole input. It is implemented by a task-specific classification head over the BERT model which has fully connected layers that project the [CLS] representation into the

target sentiment classes. In the fine- tuning step, the original weights of BERT are fine-tuned together with the classification head with the help of the labeled dataset, allowing the model to acquire task-specific patterns and preserve its general language knowledge.

6 Results

The suggested BERT-based multilingual sentiment analysis system proves to be highly effective when it comes to categorizing social media text as positive, negative, and neutral. According to the assessment of 2,050 samples, the model attained a total accuracy of 89 which means that most of the input cases were properly categorized. Class-specific behavior is also brought out in the performance measures. It is the most effective model at predicting positive sentiments, with a precision of 0.90, a recall of 0.98 and F1-score of 0.94. Based on this, it demonstrates that this model is very dependable in the detection of positive posts, and the number of misclassifications in this category is minimal. The high recall of the positive class indicates that BERT has the capability to pick finer contextual details that are usually present in enthusiastic or positive expressions even when intermixed with other languages or informal social media text.

In the negative sentiment category, the model has a precision of 0.92, a recall of 0.70 and an F1-score of 0.80. Even though it is a bit lower than the positive class, although this shows that the model can appropriately detect negative posts with a high level of precision, that is, most posts that are classified as negative are negative. A slight decline in recall indicates that a comparatively small fraction of negative posts might be incorrectly classified as either a neutral or positive one, possibly owing to the existence of subtle or ambiguous expressions in multilingual settings. The neutral type which is generally more difficult because subtle and context- dependent is an area that attains a precision of 0.50, a recall of 0.35, and F1-score of 0.41. Although they are lower, it represents the challenges of identifying neutral sentiment in nature, particularly when the text is informal, code-mixing or low-sentiment intensity.

The precision of the macro-average metrics of 0.77, 0.68, and 0.72 and the weighted averages metrics of 0.87, 0.89, and 0.87 shown in the Fig.2 and 3 represents that the model accurately works on all the classes, and the balance of performance with the model being mostly accurate on the dominant positive class, which forms the largest subdivision of the datasets, is evident. In general, the obtained findings suggest that the adaptation of the popular BERT model to multilingual sentiment analysis can be of great benefit compared to the classical machine learning and previous deep learning systems. The model is able to

reproduce semantic and contextual nuances in various languages and informal text written on social media and therefore it can be applied to real world measurements like brand monitoring, analysis in public opinion and analysis of customer feedback. Table 1. Shows the Comparison with Traditional Models.

	Precision	Recall	F1-score	Support
Negative	0.92	0.70	0.80	322
Neutral	0.50	0.35	0.41	218
Positive	0.90	0.98	0.94	1510
Accuracy			0.89	2050
Macro Avg	0.77	0.68	0.72	2050
Weighted Avg	0.87	0.89	0.87	2050

Fig.2. Model Result

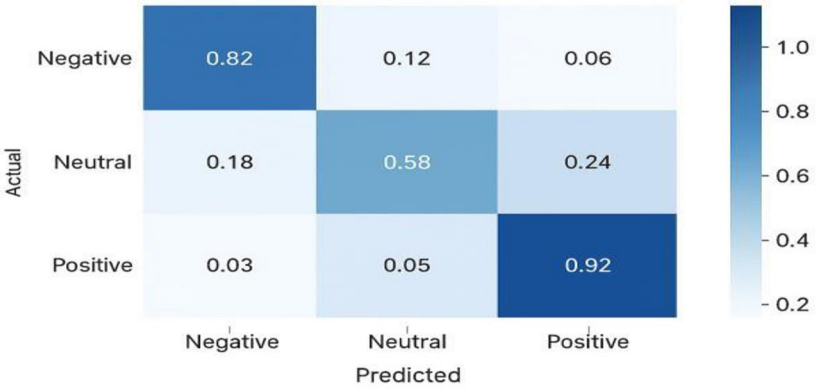

Fig.3. Confusion Matrix

Table 1. Comparison with Traditional Models

Model	Accuracy
SVM	74
CNN	79

LSTM	82
BERT (Proposed)	89

7 Conclusion

The study develops a multilingual sentiment analysis method of text on social media using the BERT transformer model. The paper has shown that utilizing pre-trained contextual embeddings can be successful in addressing the difficulties of informal language, code-mixing, transliteration, and multilingual text that is typical of social media websites. The proposed methodology will guarantee textual inputs are clean, normalized and semantically meaningful prior to model training through careful data preprocessing such as cleaning, tokenization, lemmatization and token recombination coupled with TF-IDF vectorization of text representations. Training the BERT model with the created dataset led to the creation of task-specific patterns that the model being trained could learn to classify sentiment according to positive, negative, or neutral sentiment at high accuracy.

The outcomes of the experiment suggest that the BERT-based model has an overall accuracy of 89 and, specifically, it performs well in detecting positive and negative sentiments. In the context of neutral sentiment classification, which is more complex from the subtle contextual features, macro and weighted averages of the model indicate strong performance in all the classes. These results provide an insight into the ability of BERT to solve semantic and syntactic fine-grain, which is better than conventional machine learning and the previous deep learning methods when processing multilingual and code-mixed text. In addition, a pre-trained language model will lead to less feature engineering and allow quick adaptation to new domains or languages with few extra training directions.

Altogether, this paper affirms that BERT-based transformer models offer a very effective platform to sentiment analysis in multilingual and multilingual social media with complex settings. The given framework may be used directly in real- life scenarios like monitoring of the public opinions, market trends, customer reviews, and political sentiments analysis in order to aid with informed decision-making. The future direction of work might be to develop the neutral sentiment detection more, introduce embedding specific to domain, and apply the methodology to low-resource language, which will increase the generalizability and applicability of the model to more diverse linguistic settings.

References

1. Talaat, S.: Sentiment analysis classification system using hybrid BERT models. In: Journal of Big Data, vol. 10, Art. no. 110 (2023)
2. Zhang, X., Liu, Y., Zhang, T., Hou, L., Liu, X., Guo, Z., Mulati, A.: A BERT–LSTM–Attention framework for robust multi-class sentiment analysis on Twitter data. In: Systems, vol. 13, no. 11 (2025)
3. Lin, Z., Chen, R., Li, W.: A comparative study on social media sentiment classification models based on BERT fine-tuning and fusion with sentiment lexico. In: Applied and Computational Engineering, vol. 157, pp. 243–253 (2025)
4. Patravali, S.D., Algur, S.P., Algur, N.S.: Sentiment classification of COVID-19 tweets using BERT. In: J. of the International Academy of Physical Sciences, vol. 27, no. 2, pp. 149–160 (2023)
5. Pota, M., Ventura, M., Fujita, H., Esposito, M.: Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets. In: Expert Systems with Applications, vol. 181, Art. no. 115119 (2021)
6. Bello, A., Ng, S.-C., Leung, M.-F.: A BERT framework to sentiment analysis of tweets. In: Sensors, vol. 23, no. 1, Art. no. 506 (2023)
7. Sarkar, S.: A comparative study of different natural language processing techniques for sentiment analysis and opinion mining. In: Iconic Research and Engineering Journals, vol. 8, no. 12 (2025)
8. Najafi, A., Varol, O.: TurkishBERTweet: Fast and reliable large language model for social media analysis (2023)
9. Bilehsavar, M.S., Mahmoudi, N., Torkamani, M.J., Kiashemshaki, K.: Ensembling multilingual transformers for robust sentiment analysis of tweets (2025)
10. Nkhata, G., Gauch, S., Anjum, U., Zhan, J.: Fine-tuning BERT with bidirectional LSTM for fine-grained movie reviews sentiment analysis (2025)