International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 01 | Jan - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

# Review on Sentimental Analysis and Aspect Analysis with Codemix using LLM, BERT, and Naive Bayes

Dr. Keerthi H M

*Associate Professor,*
*B.E., M.Tech, Ph.D,*
*Dept. of Computer Science and Engg.*
*Malnad College of Engineering Hassan, India*

Mr. Sudarshan K P
*Dept of Computer Science and Engg.*
*Malnad College of Engineering*
*Hassan, India*

Mr. Siddarth S Mallik
*Dept of Computer Science and Engg.*
*Malnad College of Engineering*
*Hassan, India*

Mr. Sumeeth S M
*Dept of Computer Science and Engg.*
*Malnad College of Engineering*
*Hassan, India*

Mr. Sumit Chandra
*Dept of Computer Science and Engg.*
*Malnad College of Engineering*
*Hassan, India*

## Abstract

In the era of social media, where multilingual conversations are prevalent, analyzing code-mixed text poses unique challenges. This project presents a comparative analysis of sentiment analysis and aspect-based sentiment analysis on code-mixed data using advanced techniques like Large Language Models (LLM), BERT, and Naive Bayes.

Sentiment analysis categorizes text into positive, negative, or neutral sentiments, while aspect-based analysis identifies opinions on specific topics, such as "price" or "quality" in reviews. By focusing on code-mixed text, this study compares the effectiveness of each method in understanding sentiments and specific opinions, paving the way for improved applications in multilingual settings.

## I. Introduction

Code-mixed text, where multiple languages blend in a single conversation, is common in social media. This project aims to identify sentiments and specific opinions (aspects) in such multilingual texts using comparative techniques.

The models analyzed include LLMs, BERT, and Naive Bayes. Sentiment analysis identifies general sentiments, while aspect-based sentiment analysis focuses on specific features. This study's focus on code-mixed data highlights the complexities of multilingual contexts and evaluates the models' performances under these conditions

## II. Problem Statement

With the increasing prevalence of code-mixed languages, traditional sentiment analysis and aspect-based models struggle to process mixed-language data accurately. Existing models, including Naive Bayes, BERT, and LLMs, have varying levels of success in detecting overall sentiments and aspect-specific opinions. This project seeks to determine which model performs best in multilingual, mixed-language scenarios.

## III. Objective

The primary goal is to perform sentiment and aspect analysis on code-mixed text using LLMs, BERT, and Naive Bayes. This involves:

1. Handling complexities of code-switching and transliteration.

2. Ensuring accurate sentiment detection and aspect identification.

3. Comparing model performance based on accuracy, efficiency, and practical applications like social media monitoring and customer feedback analysis.

## IV. Literature Survey

### A. Challenges in Handling Code-Mixed Data

- J. Singh, M. Kapoor (2022) highlighted preprocessing challenges such as language identification and tokenization, and used custom tokenization and language detection tools.

### B. Use of Multilingual Embeddings

- Patel, H. S. Bhatia (2023) proposed leveraging cross-lingual embeddings and transformers for improved sentiment analysis in code-mixed data.

### C. RNNs and Attention Mechanisms

- S. Ray, L. Agarwal (2020) investigated RNNs with word embeddings for sentiment analysis in informal social media contexts.
- V. Sharma et al. (2021) focused on aspect extraction using attention mechanisms in transformers for multilingual settings.

### D. Naive Bayes for Efficiency

- R. Verma et al. (2019) demonstrated the application of Naive Bayes with TF-IDF and Word2Vec embeddings for efficient sentiment classification.

### E. Contribution to Current Research

- P. Joshi, R. Kumar (2021) fine-tuned multilingual BERT for code-mixed sentiment analysis using custom datasets and transfer learning.
- Gupta, S. Sharma (2020) explored aspect-based sentiment analysis using BERT-based models with attention mechanisms for deeper insights.

## V. System Workflow

The workflow for the system involves the following steps:

**1. Raw Data Collection:**
Data is collected from various sources, including social media platforms, where code-mixed text is prevalent.

**2. Preprocessing:**
The data undergoes normalization, stop word removal, stemming, and tokenization to prepare it for analysis.

**3. Feature Extraction:**
Features are extracted using embeddings such as TF-IDF, Word2Vec, and BERT-based embeddings.

**4. Model Training and Testing:**
Three models (LLM, BERT, and Naive Bayes) are trained using labeled datasets. Each model's performance is tested on unseen data to evaluate its capability in handling code-mixed text.

**5. Prediction and Analysis:**
The trained models predict sentiment (positive, negative, neutral) and aspects in the code-mixed text.

**6. Evaluation:**

Metrics such as accuracy, precision, recall, and F1-score are used to evaluate the models' performance.

## VI. System Architecture and Implementation

The system architecture is modular and consists of the following components:

• **Data Input:** A repository of code-mixed text datasets for training and testing.

• **Preprocessing Module:** Handles normalization, language detection, and tokenization.

• **Feature Extraction Module:** Extracts features from the preprocessed text using embeddings like TF-IDF and BERT.

• **Model Training Module:** Trains the models (LLM, BERT, and Naive Bayes) on labeled data.

• **Prediction Module:** Performs sentiment and aspect-based predictions.

• **Evaluation Module:** Evaluates the models' performance using standard metrics.

## VII. Model Training

### 1. Naive Bayes (Implemented using Scikit-learn)

• Converts input text into numerical feature vectors using methods like TF-IDF or Bag of Words.

• Trains the Naive Bayes classifier to predict sentiment and aspect categories based on labeled data.

• Advantage: Lightweight and efficient for baseline comparisons but less accurate for complex patterns.

### 2. BERT (Fine-Tuned using PyTorch)

• Prepares the dataset with tokenization using BERT's tokenizer.

• Adds a classification layer on top of the pre-trained BERT model for sentiment analysis.

• Fine-tunes the entire model using backpropagation and evaluates using metrics like F1-score and accuracy.

### 3. LLM (Developed using OpenAI's GPT Model)

• Fine-tunes GPT on labeled code-mixed datasets for sentiment and aspect analysis.

• Leverages prompt engineering and few-shot learning for task-specific performance.

• Provides superior context understanding for multilingual and complex text analysis.

## VIII. Results and Discussion

**Performance Metrics:**

• Naive Bayes: Accuracy: 78%, Precision: 74%, Recall: 70%.

• BERT: Accuracy: 88%, Precision: 85%, Recall: 83%.

• LLM: Accuracy: 92%, Precision: 90%, Recall: 88%.

**Challenges:**

• Handling multilingual text with transliteration.

• Managing slang and informal expressions.

**Comparative Analysis:**

• LLMs outperform BERT and Naive Bayes in terms of accuracy and scalability.

• Naive Bayes is computationally efficient but less accurate for complex patterns.

## IX. Conclusion

This study demonstrates the importance of using advanced models like LLMs and BERT for sentiment and aspect-based analysis of code-mixed text. While LLMs achieve the highest accuracy, BERT also performs well with slightly lower computational demands.

Naive Bayes, despite its simplicity, is effective for smaller datasets. Future work can involve expanding datasets, improving model architectures, and integrating real-time applications for social media monitoring and customer feedback analysis.

## X. References

[1] Kaustubh Yadav (2020) - A Comprehensive Survey on Aspect Based Sentiment Analysis

[2] B. Selvakumar and B. Lakshmanan(2022) - Sentimental analysison user's reviews using BERT

[3] Mickel Hoang, Oskar Alija Bihorac and Jacobo Rouces(2021) -Aspect-Based Sentiment Analysis Using BERT

[4] K. Rakshitha, Ramalingam H M, M. Pavithra, Advi H D andMaithri Hegde(2021) -Sentimental analysis of Indian regional languageson social media

[5] Aditya R. Pillai and Biri Arun(2021) - Sentimental analysis andoffensive text identification using deep learning

[6] Ankit Kumar Mishra, Sunil Saumya and Abhinav Kumar(2021)- Sentiment Analysis of Dravidian-CodeMix Language

[7] Hellwig, N. C., Fehle, J., Wolff, C. (2024) - Exploring largelanguage models for the generation of synthetic training samples foraspect-based sentiment analysis in low-resource settings

[8] Mamta,Asif Ekbal(2022) - Quality achhi hai (is good),satisfied!Towardsspecial aspect based sentiment analysis in code-mixed language.

[9]Kanwal Ahmed, Muhammad Imran Nadeem, Zhiyun Zheng aMuhammad Assam, Yazeed Yasin Ghadi, Heba G. Mohamed(2023) - Breakingdown linguistic complexities: A structured approach to aspectbasedsentiment analysis.

[10] Abdulrahman Radaideh,Fikri Dweiri,Mohammad Obaidat(2020)- A Novel Approach to Predict the Real Time Sentimental Analysis by Naive Bayes RNN Algorithm during the COVID Pandemic in UAE