Multilingual Sentiment Analysis for E-Commerce Platform

"A Deep Learning Approach for Analyzing Multilingual Customer Sentiments in E-commerce"

[1]CH. Satyananda Reddy, [2]B. Manisha Ratnam

[1]Professor, [2]Student

Department of IT & CA, AU College of Engineering

Department of Information Technology and Computer Applications Andhra University,

Visakhapatnam, India

Abstract— In the era of global e-commerce, understanding customer sentiment across diverse languages is vital for enhancing user experience and business intelligence. This project, titled "Multilingual Sentiment Analysis in E-commerce Platform", focuses on predicting customer sentiment—positive, negative, or neutral—based on product reviews submitted in multiple languages. The core objective is to bridge the language gap in online feedback interpretation using advanced machine learning and natural language processing techniques. To achieve this, a hybrid approach leveraging both deep learning and traditional models is implemented—specifically, BERT (Bidirectional Encoder Representations from Transformers) for robust text embeddings and contextual understanding, and Random Forest for efficient classification.

Index Terms— Multilingual Sentiment Analysis, E-commerce, BERT, Random Forest, Natural Language Processing, Product Review Classification, Customer Feedback.

I. INTRODUCTION

In the digital age, e-commerce platforms have revolutionized how products and services are bought and sold, enabling businesses to reach customers around the globe. With this expansion comes an influx of user-generated content, particularly product reviews, which provide valuable insights into customer satisfaction, preferences, and concerns. However, a significant challenge arises when these reviews are written in multiple languages, reflecting the diversity of the user base. Traditional sentiment analysis systems are often language-specific, limiting their ability to understand and process reviews written in different linguistic contexts. This gap highlights the need for a robust, scalable solution that can analyze multilingual feedback accurately.

The project titled "Multilingual Sentiment Analysis in E-commerce Platform" aims to address this challenge by developing an intelligent system capable of predicting the sentiment—positive, negative, or neutral—of product reviews written in various languages. This system is a hybrid approach combining the Natural Language Processing (NLP) and machine learning techniques. Specifically, the model leverages BERT (Bidirectional Encoder Representations from Transformers), a deep learning model known for its contextual understanding of text, and the Random Forest algorithm, known for its robustness and interpretability in classification tasks.

II. ABBREVIATIONS AND ACRONYMS

BERT – Bidirectional Encoder Representation from Transformer,

SVM – Support Vector Mechanism

NLP – Natural language Processing

III. LITERATURE REVIEW:

In the rapidly growing e-commerce sector, analyzing customer feedback has become crucial for improving services and products. Traditional sentiment analysis techniques—such as *Naive Bayes, Support Vector Machines (SVM)*, and *Decision Trees*—have primarily focused on English-language reviews and lacked the ability to handle the complexity of multilingual inputs [1]. With customers increasingly providing feedback in various native languages, these models have proven insufficient for real-world, global applications.

Language-specifics such as grammar and idiomatic expressions, often lead to inaccurate translations, which in turn affects sentiment interpretation [2], [3]. This challenge has encouraged researchers to explore models that can process multiple languages natively, without depending solely on translation tools.

The introduction of *Bidirectional Encoder Representations from Transformers (BERT)*, especially its multilingual version (mBERT), revolutionized text processing. mBERT supports over 100 languages and can capture context across linguistic boundaries, making it highly effective for multilingual sentiment analysis [4], [5].

However, transformer models like mBERT are still sensitive to overfitting and class imbalance. To mitigate this, ensemble learning methods—especially Random Forest—have been employed to enhance prediction stability and reduce model variance [6]. Combining BERT-based embeddings with ensemble classifiers has shown improved performance in sentiment classification, especially in noisy or high-dimensional datasets [7].

While major e-commerce platforms such as Amazon and Alibaba have adopted sentiment analysis tools, many of these systems are still focused on English or rely heavily on manual translation. This limits automation and scalability [8]. To address these limitations, this project integrates mBERT with Random Forest in a hybrid architecture that enables efficient and scalable sentiment analysis across multiple languages.

IV. METHODOLOGY:

The approach adopted in this project is carefully designed to develop an effective multilingual sentiment analysis system by combining deep learning and machine learning methods. The process is broken down into distinct phases that work together to achieve accurate sentiment prediction. It starts with gathering and cleaning multilingual review data, followed by translating non-English content when required. The next phase involves converting text into numerical features using TF-IDF. These features are then used to train a Random Forest classifier. Additionally, a separate pipeline uses a fine-tuned BERT model to better understand the contextual meaning of the reviews, ensuring a deeper and more reliable analysis of sentiments across languages.

1. Data Preparation:

A multilingual dataset containing product reviews in various languages (such as Hindi, Tamil, Spanish, Telugu, etc.) was used. Each review was labelled with a sentiment category: *Positive*, *Negative*, or *Neutral*. The dataset was cleaned to remove duplicates, null values, and noise such as HTML tags or special characters.

2. Language translation:

To maintain consistency during analysis, reviews written in non-English languages were translated into English. This was done using the *MarianMT Model* from Hugging Face's *transformers* library. A custom dictionary was also used to map frequently occurring multilingual phrases into their English equivalents for consistent labelling.

3. Text Preprocessing:

Text preprocessing helps clean and prepare the review data for analysis. In this project, once the reviews were translated into English, the text was converted to lowercase to maintain consistency. Words that do not add much meaning, known as stop words such as "is", "the", "and" etc are removed. The cleaned text was then split into individual words through tokenization, making it easier for the model to understand and analyze. Tokenization was done using methods from *NLTK* and Hugging Face's *BertTokenizer*. This step ensured that only relevant and meaningful information was passed to the sentiment classification models.

4. Feature Extraction using TF-IDF:

The translated reviews were converted into numerical features using *Term Frequency–Inverse Document Frequency (TF-IDF)* vectorization. This approach highlights important terms that are unique to individual reviews, thereby improving model performance in sentiment classification.

$$IDF(t) \; = \; log\left(\frac{1+N}{1+df(t)}\right) + 1$$

Where:

t = term (word)

N = total number of documents

$df(t)$ = number of documents where the term t appears

5. Sentiment Classification using Random Forest:

The TF-IDF values are used to train a *Random Forest classifiers*. This algorithm constructs different decision trees and analysis the result of each and every output. The majority value will become the output of the model. This method improves the system's robustness by handling overfitting and ensuring better generalization to unseen data.

6. Fine-Tuning of Multilingual BERT

In parallel, a multilingual BERT (mBERT) model was fine-tuned on the same sentiment dataset. mBERT, pre-trained on over 100 languages. It was fine-tuned using Hugging Face's Trainer API to perform sequence classification, allowing the model to understand and classify sentiment-rich expressions in context.

7. Hybrid Model Integration and Evaluation:

To maximize classification performance, predictions from both the Random Forest and mBERT models were combined. The ensemble strategy leveraged the efficiency of Random Forest along with the deep contextual understanding of BERT. The model's performance was evaluated using precision, recall, F1-score and accuracy.

8. Web Interface:

Created a web application for this project in which the user can give feedback and get result whether it is a positive, negative or neural review. For this if the user is new to the web application, they should register to the portal and login using credentials.

V. RESULT AND ANALYSIS

The proposed multilingual sentiment analysis system was evaluated using a diverse dataset containing product reviews written in multiple languages such as Hindi, Tamil, Telugu etc. Each review was labelled with its corresponding sentiment: Positive, Negative, or Neutral. The performance of the model was assessed using standard classification metrics including accuracy, precision, recall, and F1-score.
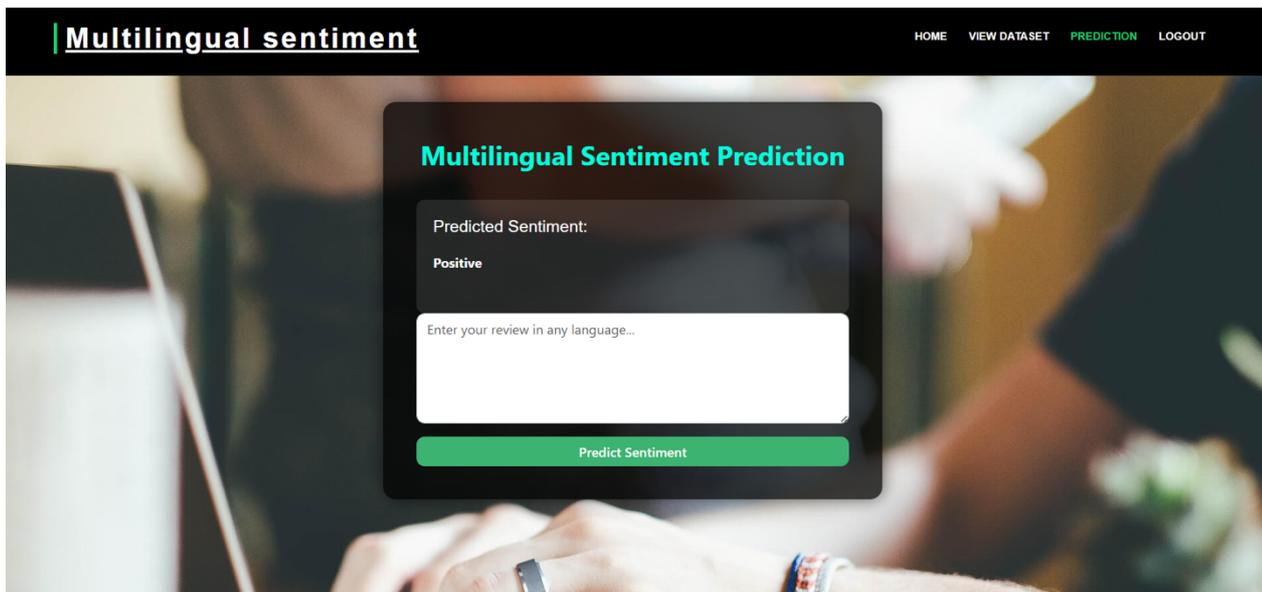The user needs to login to the portal using login_id and password which are created during registration. After login to the portal the user can analyze the review whether it is positive, negative or neutral.
Example review prediction:
  • This is the prediction where you can give review text and get the analysis.
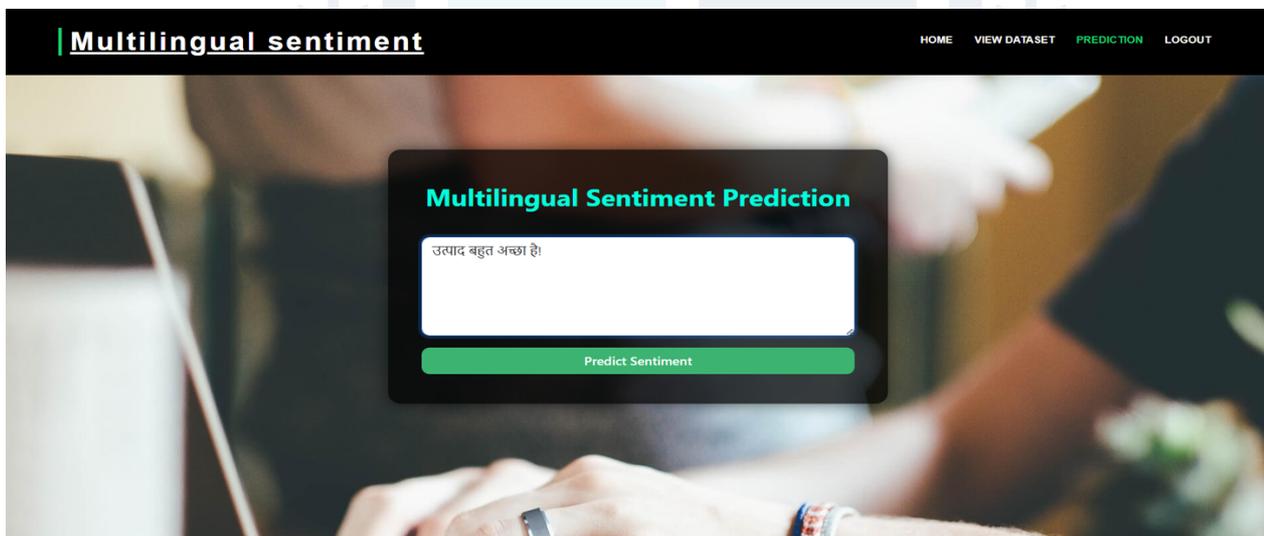
  • This is the output for the above review

Performance Result:

Table 1 summarizes the precision, recall and F1-score of multilingual sentiment analysis for e-commerce platform

Table 1: Review performance

Feature	Precision	Recall	F1-score
Random Forest	1.00	1.00	1.00
mBERT	88.9	88.3	88.6

VI. CONCLUSION:

The multilingual sentiment analysis system in this project showcases the effective combination of deep learning and machine learning to address challenges in global e-commerce. Using Multilingual BERT for contextual understanding and Random Forest for classification, it accurately identifies sentiment in product reviews across various languages. This removes the need for language-specific models or translation, enhancing both performance and scalability. The system enables inclusive, automated feedback analysis, helping e-commerce platforms gain insights from diverse customer opinions. It supports better decision-making, boosts user satisfaction, and improves product offerings, while also paving the way for smarter multilingual systems

VII. REFERENCES

[1] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[2] S. Akhtar, D. Gupta, and A. Ekbal, "A multi-task learning framework for multilingual and multi-attribute sentiment analysis," *Knowledge-Based Systems*, vol. 187, 2020.

[3] J. Balahur and M. Turchi, "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis," *Computer Speech & Language*, vol. 28, no. 1, pp. 56–75, 2014.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of NAACL-HLT*, 2019, pp. 4171–4186.

[5] P. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?," in *Proc. ACL Workshop on Representation Learning for NLP*, 2019.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[7] Y. Zhang, M. Rao, and J. Jiang, "Combining BERT with ensemble learning for text classification," *IEEE Access*, vol. 8, pp. 103949–103960, 2020.

[8] S. Sun, "Sentiment analysis for e-commerce product reviews in multilingual environments," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, pp. 487–494, 2020