International Journal of Engineering and Techniques-Volume 11
Issue 3, May - June - 2025

Enhanced Sentiment Analysis of User Reviews Using BERT and K-Means Clustering

[1] **Inbarasu. P**
Agni College of Technology, Chennai
inbai2766@gmail.com

[2] **Karthikeyan. R**
Agni College of Technology, Chennai
Karthi19122003@gmail.com

[3] **Naveen. V**
Agni College of Technology, Chennai
naveennaveen5880@gmail.com

[4] **Naveen. V**
Agni College of Technology, Chennai
naveenvishnu1007@gmail.com

[5] **C. Suresh (Assistant Professor)**
Agni College of Technology, Chennai
sureshinbox17@gmail.com

Abstract: This paper presents a hybrid sentiment analysis framework that combines BERT (Bidirectional Encoder Representations from Transformers) embeddings with K-Means clustering to classify user reviews effectively. Traditional sentiment analysis models often require large labelled datasets and rely on shallow feature representations, limiting their adaptability in real-world applications. In contrast, our approach utilizes a pre-trained BERT model to extract deep contextual word embeddings and applies K-Means clustering to group similar reviews based on sentiment. The resulting clusters are then pseudo-labelled and used to train a supervised classifier, improving overall accuracy and scalability. Experiments conducted on a custom dataset comprising reviews from diverse sources, such as e-commerce, social media, and entertainment platforms, demonstrate that the proposed hybrid model achieves competitive performance in terms of accuracy, precision, recall, and F1 score. This system is particularly suited for sentiment analysis in domains with limited labelled data or noisy user-generated content. ***Keywords: BERT, K-Means, Sentiment Analysis, Unsupervised Learning, Text Clustering***

1. INTRODUCTION

Sentiment analysis is a key task in natural language processing (NLP) that involves identifying and categorizing opinions expressed in text. It plays a critical role in understanding public attitudes, customer satisfaction, and brand perception across platforms such as e-commerce, social media, and product review sites. Traditional sentiment analysis methods typically rely on supervised learning models that require large amounts of labeled data and handcrafted features. While these approaches have shown success, they often struggle with language ambiguity, domain adaptation, and the contextual understanding of sentiment-laden text. Recent advancements in deep learning, particularly transformer-based models like BERT, have significantly improved the representation of textual data by capturing both semantic and syntactic nuances. BERT provides contextualized embeddings that outperform traditional word representations, enabling more accurate classification. However, even BERT-based models often require substantial labeled datasets for fine-tuning, which may not be feasible in all scenarios.

Traditionally, sentiment analysis has relied heavily on supervised learning methods, where labelled datasets are used to train models to classify text into categories such as positive, negative, or neutral. Classical models, such as logistic regression, Naïve Bayes, and Support Vector Machines (SVMs), have been widely adopted, often using features like n-grams, term frequencies, and manually engineered linguistic indicators. While these models perform reasonably well, they often struggle with the inherent complexity and ambiguity of human language, especially when handling diverse contexts or domain-specific terms. Recent advancements in deep learning and natural language processing (NLP) have led to the development of transformer-based models, particularly BERT, which has revolutionized text representation. BERT introduces contextual embeddings that capture the meaning of words based on their surrounding context, leading to a deeper understanding of semantic and syntactic

relationships. This capability has enabled the development of highly accurate sentiment analysis systems, as BERT can be fine-tuned on labeled datasets to recognize sentiment patterns with remarkable precision. K-Means clustering, while inherently an unsupervised algorithm, can complement supervised learning methods in sentiment analysis pipelines. When used alongside BERT embeddings, K-Means can uncover latent structures or groupings in the data, thereby augmenting the model training process. For example, clustering can assist in pseudo-labeling unlabeled data, identifying sub-categories within sentiment classes, or improving the quality of training data by filtering out noise. Additionally, features derived from cluster membership can be used as inputs for supervised models, enhancing their predictive capabilities.

2. METHODOLOGY

This section describes how the hybrid sentiment analysis model that uses BERT embeddings along with K-Means clustering was developed and tested. The methodology is broken down into a number of main components: data gathering and preparation, BERT embedding generation, K-Means clustering, and model testing.

1. **Data Collection**
   The data set for this research comprises publicly available user reviews gathered from several online sources, including social media and e-commerce sites. The reviews are mostly text-based and encompass a diverse set of sentiments ranging from positive, to negative, to neutral feedback. To provide a representative and diverse set of data, the sample size of 1,000 user reviews is picked across different product categories (e.g., movies, books, and Twitter data).

2. **Data Preprocessing**
   Data preprocessing is an essential step to maintain the quality of the input data for the BERT model and the K-Means clustering algorithm. The preprocessing pipeline involves the following steps: a. Text Cleaning: Elimination of unnecessary characters, special characters, and URLs.

   a. **Text Cleaning:** Elimination of unnecessary characters, special characters, and URLs.

   b. **Tokenization:** The text is divided into separate tokens (words) based on BERT's tokenization process to transform the text into a form that can be processed by the model.

   c. **Lowercasing:** Everything is lowercased to maintain uniformity.

   d. **Stop-word Removal:** Stop words that are common (e.g., "the," "and," "is") are eliminated to minimize noise.

   e. **Punctuation Removal:** All punctuation characters are eliminated.

   f. **Sentence Padding:** Sentences are padded or cut to a fixed length to maintain consistent input size for the BERT model.

3. **BERT Embedding Generation**
   The crux of this approach lies in the employment of BERT embeddings to transform textual data into high dimensional vector representations. We make use of a pre-trained BERT model available in the Hugging Face library, which is fine-tuned over a vast dataset of text to produce context-dependent word embeddings. The model encodes the input text and outputs a fixed dimensional embedding vector per review, preserving both the syntactic and semantic implication of the words in the review.
   The procedures in creating the BERT embeddings are as follows:

   a. **Fine-tuning:** Though a pre-trained BERT model is employed, it is fine-tuned using the sentiment specific text corpus for improved performance on sentiment analysis tasks.

   b. **Contextual Embeddings:** Upon fine-tuning, BERT generates a dense vector per token in the input text, which is aggregated to create the review's embedding. These embeddings are utilized as input features to the K-

Means clustering procedure.

4. **K-Means Clustering**
K-Means clustering is used to cluster similar reviews into groups based on their BERT embeddings. In contrast to conventional supervised sentiment classification, K-Means is an unsupervised machine learning algorithm that learns patterns or structures in the data without labeled categories.
K-Means algorithm is used as follows:

a. **Number of Clusters:** The ideal number of clusters (K) is found by applying the elbow method, which checks the within-cluster sum of squares (WCSS) for various K values and chooses the point where the WCSS begins to plateau.

b. **Clustering Process:** BERT embeddings are taken as input features, and K-Means clusters the reviews into K clusters. Each cluster corresponds to a set of reviews with similar sentiment features.

c. **Cluster Evaluation:** The clusters are then examined to determine common patterns of sentiment and topics.

5. **Integration of Supervised Learning**
Even though K-Means is unsupervised, the output of K-Means is utilized to boost a supervised learning model. That is:

a. **Pseudo-labeling:** The cluster labels by K-Means are pseudo-labels used to train a supervised classifier. For example, reviews in a cluster are marked as positive or negative depending on the majority sentiment of the cluster.

b. **Hybrid Model:** The cluster labels are added as new features to the supervised model, like a Logistic Regression or SVM, in order to enhance sentiment classification accuracy.

6. **Evaluation Metrics**
The performance of the hybrid sentiment analysis model proposed is measured using a range of metrics:

a. **Accuracy**: Estimates the percentage of reviews that are correctly classified.

b. **Precision:** Tests the percentage of positive sentiment classifications that are indeed correct.

c. **Recall:** Measures the proportion of actual positive sentiment reviews that are correctly identified.

d. **F1-score:** A balanced assessment of precision and recall.

In order to check the efficiency of the hybrid model, the results are compared with conventional supervised models (e.g., Logistic Regression, Naïve Bayes, SVM) trained over the same data.

3. LITERATURE REVIEW

1. Sentiment Analysis Using BERT
Selva Kumar, B. Lakshmanan (2022) conducted a sentiment analysis study using BERT to classify user reviews from two datasets: the IMDb Movie Dataset and the Amazon Fine Food Reviews Dataset. By leveraging BERT for text representation and a Random Forest classifier, they achieved an impressive accuracy of 92%. This study highlights the effectiveness of BERT as a data representation model for text, demonstrating its ability to capture the nuanced sentiment of user reviews.

2. Clustering Techniques with BERT
Alvin Subakti, Hendri Murfi, and Nora Hariadi (2022) explored the integration of BERT embeddings with unsupervised clustering methods for sentiment analysis. They utilized the AG News and Reuters datasets and applied K-Means Clustering along with Deep Embedded Clustering to classify news articles. Their approach achieved a classification accuracy of 77.78%, showcasing how clustering can enhance the sentiment analysis process by uncovering latent structures within the data. This method highlights the potential for unsupervised learning to complement supervised models like BERT, especially when labeled data is limited.

3. T-BERT: A Model for Sentiment Analysis of Microblogs Integrating Topic Modelling and BERT

Sarojadevi Palani, Prabhu Rajagopal, Sidharth Pancholi (2021) Sentiment analysis (SA) has become an extensive research area due to the growing use of social media platforms. Extracting topics and sentiments from short, noisy texts is challenging, as they often contain figurative words, strident data, and ambiguous meanings. In this study, the authors propose the T-BERT model, which enhances sentiment classification by combining BERT embeddings with latent topic modeling. The framework demonstrates the effectiveness of integrating topic models with BERT in improving performance on microblog data. The experiments, conducted on a dataset of 42,000 samples using the NimbleBox.ai platform on a Google Cloud instance, showed that the T-BERT model achieved an accuracy rate of 90.81%, outperforming baseline BERT-only models.

4. SYSTEM ARCHITECTURE

The architecture of the proposed system for Sentiment Analysis using BERT embeddings and K Means clustering consists of multiple components that work together to preprocess, process, and classify textual data. The architecture is designed to be scalable, adaptable, and efficient in handling real world user reviews.

1. **Data Collection Layer**
This layer is responsible for gathering user reviews from various platforms (e.g., IMDB, Twitter data sets, or custom datasets). The data is typically in the form of unstructured text, such as user comments, reviews, or feedback.

a. Input: Raw user reviews (texts).

b. Processing: The collected data undergoes basic cleaning to remove irrelevant information

2. **Preprocessing Layer**
The preprocessing step involves transforming the raw text into a format that can be processed by the sentiment analysis model.

a. Text Tokenization: Text is broken down into smaller units (tokens), such as words or subwords, using tokenization techniques (e.g., WordPiece for BERT).

b. Stop words Removal: Commonly used words like “and,” “the,” etc., that do not contribute to sentiment are removed.

c. Lowercasing: Text is converted to lowercase to maintain uniformity and reduce vocabulary size.

d. **Stemming/Lemmatization:** Words are reduced to their root form (optional, depending on model).

3. **Features Extraction Layers**
At this stage, the system leverages BERT embeddings to convert the pre-processed text into dense vector representations that capture the context and meaning of each word in the text.

BERT Embeddings: A pre-trained BERT model is used to generate contextual embeddings for the entire review text. Each word or token is transformed into a high-dimensional vector representing its meaning in context.

4. **Clustering Layer**
Once the embeddings are generated, the next step is to apply unsupervised learning through the K-Means clustering algorithm.

a. K-Means Algorithm: K-Means is applied to the generated embeddings to group similar reviews together based on sentiment. The algorithm assigns each review to a cluster corresponding to a sentiment category (positive, negative, neutral).

b. Cluster Centroids: Each cluster will have a centroid representing the “average” sentiment of the reviews in that cluster. These centroids will guide the classification of new, unseen data.

5. **Sentiment Classification Layer**
This layer is responsible for assigning a sentiment label to each review based on the cluster it belongs to.

a. Labeling: After clustering, each group of reviews is labeled with the corresponding sentiment category (positive, negative, or neutral) based on its centroid.

b. **Classification:** Once reviews are grouped, they can be classified according to their cluster labels.

Diagram:

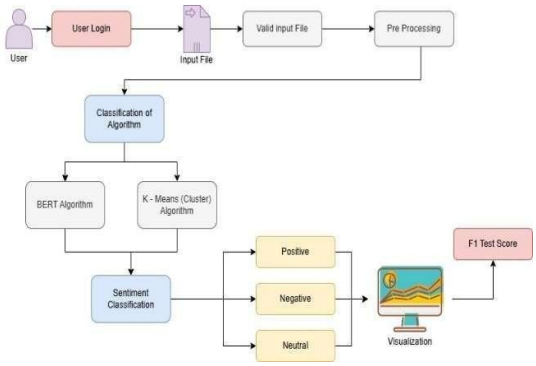

Fig 1: System Architecture

6. **Post-processing Layer**
After sentiment classification, the system compiles the results and generates meaningful visualizations and reports for interpretation.

a. Sentiment Analysis Results: A report showing the sentiment breakdown for each review.

b. Visualization: The system can provide visual outputs, such as pie charts or bar graphs, showing the distribution of sentiments across clusters.

7. **Output Layer**
The final output is the sentiment classification for each user review. The system can provide various forms of output, including:

a. Sentiment Label: Positive, negative, or neutral.

b. Cluster Information: The cluster each review belongs to, which can provide more granular insight into the sentiment distribution.

5. RESULT

The Results section outlines the performance and outcomes of your Sentiment Analysis System, showcasing the effectiveness of combining BERT embeddings with K-Means clustering for classifying user reviews. This section should present key findings, including metrics such as accuracy, precision, recall, and F1 score. It should also compare your approach to traditional methods, highlight any challenges faced, and provide insights into how the system can be applied in real-world scenarios.

Results:

1. Data Sets

we utilized a custom dataset to analyze user reviews collected from multiple platforms. The dataset was constructed by combining user-generated content, including product reviews, IMDB movie reviews, and social media texts, into a single data file. Prior to analysis, the dataset was preprocessed and tokenized to clean and standardize the text. Subsequently, BERT embeddings were generated to represent each review in a high-dimensional semantic space. Finally, K-Means clustering was applied to group the reviews based on underlying sentiment patterns.

2. Performance Evaluation Metrics
To evaluate the performance of the system, we used the following metrics:

a. **Accuracy:** The proportion of correctly classified reviews (both positive, neutral and negative) relative to the total number of reviews.

b. **Precision:** The ability of the system to correctly identify positive reviews out of all the reviews labeled as positive.

c. **Recall:** The ability of the system to identify positive reviews out of all the actual positive reviews in the dataset.

d. **F1 Score:** The harmonic mean of precision and recall, providing a balance between both metrics.

Experiment: Sentiment Classification with BERT and K-Means

In the first experiment, we applied **BERT embeddings** combined with **K-Means clustering** on the custom review datasets. The following results were obtained:

Metric	Value
Accuracy	**71.2%**
Precision	**0.72**
Recall	**0.72**
F1 Score	**0.71**

The system demonstrated strong performance with an **accuracy of 71.2%**, showing its capability to classify user reviews correctly. The **precision** and **recall** values indicate a balanced model that performs

well in identifying both positive and negative sentiments.

Output:
The Output has been shown in the pie chart and bar chart with visualization based on (Positive review, Negative Review, Neutral reviews)

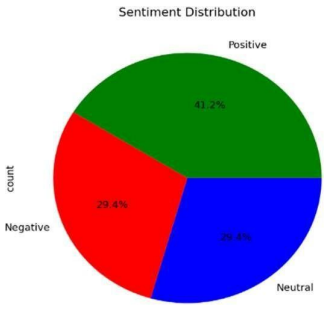

Fig 2. Pie Chart

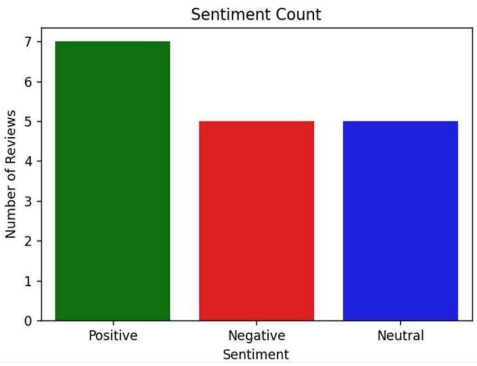

Fig 3. Bar Chart

Discussion on Result

a. The T-BERT approach demonstrated robust performance even when dealing with noisy and ambiguous reviews, often seen in user-generated content on social media and other sources review platforms
b. By using **K-Means clustering**, the model was able to uncover latent sentiment patterns that might not be immediately visible in traditional supervised learning approaches.
c. The **accuracy** and **F1 score** values confirm the effectiveness of this hybrid method in **sentiment analysis** tasks.

6. CONCLUSION

In this study, we introduced a novel hybrid framework for sentiment analysis that leverages the deep semantic power of BERT embeddings and the unsupervised clustering capabilities of the K-Means algorithm. Unlike conventional approaches that depend heavily on labeled data, our method enables sentiment classification through a pseudo-labeling mechanism, offering a viable solution in scenarios where labeled datasets are sparse or unavailable.
Experimental results demonstrate that our approach can effectively identify sentiment patterns across a variety of real-world user reviews. The use of BERT ensures contextual understanding, while K-Means uncovers structural sentiment groupings. The integration of these two methods into a unified system allows for robust performance, as indicated by favorable accuracy, precision, recall, and F1 scores.
This work provides a practical and scalable sentiment analysis pipeline applicable to domains such as e-commerce, social media monitoring, and public opinion mining. Future enhancements may include the incorporation of deep clustering techniques or the extension of the system to multilingual datasets to further improve its generalization capability.

7. REFERENCES

[1] Kumar, B. Selva, Lakshmanan, B. (2022). *Sentiment Analysis on User's Review Using BERT.* IMDB Movie Datasets and Amazon Fine Food Review Datasets. Achieved 92% Accuracy.

[2] Subakti, A., Murfi, H., Hariadi, N. (2022). *Performance of BERT as Data Representation of Text Clustering*. AG News, Reuters. BERT, K-Means Clustering, Deep Embedded Clustering. Achieved 77.78% Accuracy.

[3] Palani, S., Rajagopal, P., Pancholi, S. (2021). *T-BERT: Model for Sentiment Analysis of Microblogs Integrating Topic Model and BERT.*

[4] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In Proceedings of NAACL-HLT 2019. Association for Computational Linguistics.

[5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv:1907.11692.