A Thesis Submitted to the Sylhet Engineering College for the Degree of
**Bachelor of Science** in **Computer Science and Engineering**

# Explainable Sentiment Analysis of Hotel Reviews: SMOTE-Optimized Ensemble Learning and BERT with LIME Interpretability


By

**Eshtiak Hasan**

**Registration No.:2019331517**

**Md. Faysal Numin**

**Registration No.:2019331541**

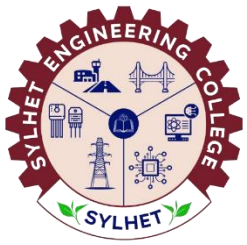
Supervised By

**Md. Abu Naser Mojumder**

Assistant Professor & Head

Department of Computer Science and Engineering

Sylhet Engineering College, Sylhet



Department of Computer Science and Engineering

**Sylhet Engineering College, Sylhet**

Affiliated with

**Shahjalal University of Science & Technology**

# Recommendation Letter from Thesis Supervisor

The thesis titled **"Explainable Sentiment Analysis of Hotel Reviews: SMOTE-Optimized Ensemble Learning and BERT with LIME Interpretability"** submitted by the group as mentioned below has been accepted as satisfactory in partial fulfilment of the requirements for the degree B. Sc. in Computer Science and Engineering in July, 2025.

Group Members:

**Eshtiak Hasan (2019331517)**

**Md. Faysal Numin (2019331541)**

Supervisor:

**Md. Abu Naser Mojumder**

Associate Professor & Head
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet

# Certificates of Acceptance

The thesis is titled **"Explainable Sentiment Analysis of Hotel Reviews: SMOTE-Optimized Ensemble Learning and BERT with LIME Interpretability"** submitted by **Eshtiak Hasan** and **Md. Faysal Numin**; Student ID. **2019331517** and **2019331541**; Session **2019-20**, to the Department of Computer Science and Engineering , Sylhet Engineering College, has been accepted as satisfactory in partial fulfilment of the requirement for the Degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents.

## BOARD OF EXAMINERS

_____

Internal

**Nayan Kumar Nath**
Lecturer
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet

_____

Internal

**Md Lysuzzaman**
Lecturer
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet

_____

Internal

**Md. Rasel Ahmed**
Assistant Professor
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet

_____

Internal

**Md. Nagrul Islam**
Assistant Professor
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet

_____

Chairman

**Md. Abu Naser Mojumder**
Head
Department of Computer Science and Engineering
Sylhet Engineering College, Sylhet

_____

Member (External)

**Dr. Mohammad Shahidur Rahman**
Professor
Department of Computer Science and Engineering
Shahjalal University of Science and Technology

# Acknowledgements

First and foremost, we offer our deepest gratitude to the Almighty, whose boundless mercy and silent guidance have been our constant source of strength throughout this journey. Through every challenge and every success, His blessings have illuminated our path and sustained our resolve.

We would like to express our heartfelt appreciation to our respected supervisor, **Md. Abu Naser Mojumder**, for his invaluable support, continuous encouragement, and insightful feedback. His guidance has been pivotal in shaping the direction and quality of our research, and we are truly grateful for his mentorship.

Our sincere thanks also go to our respected teachers, **Md. Lysuzzaman** and **Nayan Kumar Nath**, whose dedication to teaching and depth of knowledge have left a lasting impact on our academic foundation. Their support and encouragement throughout our studies have been a source of inspiration.

We are especially thankful for the strong collaboration and mutual understanding we shared as thesis partners. The joint effort in completing our thesis, "**Explainable Sentiment Analysis of Hotel Reviews: SMOTE-Optimized Ensemble Learning and BERT with LIME Interpretabili**" was made possible through shared dedication, teamwork, and a commitment to learning.

Lastly, with all our love and gratitude, we acknowledge the endless support of our families. Their unconditional love, sacrifices, and constant prayers have been the backbone of our academic journey. Without their unwavering belief in us, this achievement would not have been possible.

# Abstract

This study presents an explainable sentiment analysis framework for hotel reviews collected from Dhaka and Chattogram, using data scraped from Google Maps and Booking.com. A total of 4,420 reviews were categorized into positive, negative, and neutral sentiments. After addressing class imbalance using SMOTE, we evaluated three traditional machine learning models and four ensemble models. Among them, LightGBM achieved the highest accuracy (83%) and balanced performance across metrics, while Multinomial Naïve Bayes yielded the best recall (0.79). To improve model interpretability, we applied LIME, which revealed influential features shaping predictions. Furthermore, we fine-tuned a BERT-base-uncased model using PyTorch, which significantly outperformed all traditional models, achieving 90% accuracy, 0.88 precision, and strong F1-scores (0.92 negative, 0.80 neutral, 0.91 positive). BERT's contextual understanding, verified through LIME explanations, enabled superior classification of nuanced and ambiguous reviews. This research highlights the effectiveness of combining transformer-based models with explainability tools for robust sentiment analysis in the Bangladeshi hotel domain.

**Keywords:** Natural Language Processing(NLP); Machine Learning; BERT; SMOTE; LIME; Ensemble Learning; Explainable AI;

# Table of Contents

# Chapter 1: Introduction

Sentiment analysis is essential for understanding customer opinions, especially in the hotel industry where user feedback shapes service improvement and business strategy. This study focuses on hotel reviews from Dhaka and Chattogram, analyzing sentiments using machine learning models such as SVM, Naïve Bayes, Logistic Regression, and Random Forest, alongside advanced models like BERT. To address class imbalance, we apply SMOTE, and for model interpretability, we use LIME. Our goal is to build a robust and explainable sentiment analysis framework tailored to the local Bangladeshi hotel context.

## 1.1 Background

With the exponential growth of digital platforms, user's reviews have become central to customer decision-making in the hospitality industry. Online hotel review is the way customers express their experience while visiting a hotel (Kim et al. [1]). Usually, the review is written on a website, such as hotel booking site like Booking.com or Google Reviews. The reasons why people tend to write a review are to help other consumers, personal motivation, or to improve the service quality (Gonçalves et al. [2]). Travelers frequently consult these reviews before booking hotels, and businesses utilize them to evaluate service quality and understand customer needs. As such, sentiment analysis, a subfield of Natural Language Processing (NLP), has become crucial for automatically identifying opinions and emotional tones expressed in text.

In Bangladesh, tourism is a growing sector contributing significantly to national income. Dhaka and Chattogram, the two largest regions, serve as major hubs for both international and domestic travelers, boasting a dense concentration of hotels and tourism infrastructure (Bangladesh Tourism Board [3]). As the demand for digital services grows, so does the importance of understanding customer sentiment on online platforms like Google Reviews and Booking.com.

However, hotel reviews often exhibit imbalanced sentiment distributions (dominated by positive reviews), informal language, and domain-specific expressions, making accurate classification a non-trivial task. Traditional Machine Learning (ML) models—such as Support Vector Machines (SVM), Logistic Regression (LR), and Multinomial Naïve Bayes (MNB)—have shown promise in sentiment classification (Pang et al. [4]; Liu [5]), especially when combined with 10-fold cross-validation for robustness (Kohavi [6]). These models are interpretable and efficient but are limited in capturing contextual semantics.

To address this, advanced ensemble techniques such as Random Forest (Breiman [7]), XGBoost (Chen & Guestrin [8]), LightGBM (Ke et al. [9]), and stacked models have emerged as strong candidates, boosting performance by aggregating predictions of multiple learners.

Furthermore, transformer-based deep learning models such as BERT (Devlin et al. [11]) have revolutionized NLP by leveraging bidirectional attention for richer semantic understanding.

A major issue in real-world datasets like hotel reviews is class imbalance, where positive reviews significantly outnumber neutral or negative ones. This bias can lead to skewed model predictions. To mitigate this, Synthetic Minority Over-sampling Technique (SMOTE) is used to balance the dataset by generating synthetic examples of the minority class (Chawla et al. [10]). This improves model fairness and recall, especially on underrepresented classes.

Yet, these improvements in performance come at a cost: reduced interpretability. In domains like hotel management—where decisions impact customer experience and business reputation—it is essential to understand why a model makes certain predictions. This has sparked interest in Explainable Artificial Intelligence (XAI), especially tools like LIME (Ribeiro et al. [12]), which generate local, human-interpretable explanations for model predictions.

This study proposes a comprehensive sentiment classification framework for Bangladeshi hotel reviews using a combination of SMOTE, ensemble learning, BERT, and LIME to ensure both performance and explainability.

## 1.2 Problem Statement

Despite the growing body of research in sentiment analysis, several gaps remain in its application to hotel reviews in the Bangladeshi context:

- Existing models are often trained on global datasets and fail to capture local linguistic nuances or domain-specific expressions found in Bangladeshi hotel reviews.
- The class imbalance problem (e.g., fewer negative or neutral reviews) leads to biased model performance and reduced generalizability.
- While ensemble and deep learning methods improve accuracy, their lack of interpretability hinders adoption in real-world hotel management scenarios.
- Most studies fail to apply cross-validation techniques consistently, raising concerns about the robustness and reproducibility of their results.
- No prior work has combined ensemble models, BERT, and LIME on hotel reviews for explainable sentiment multiclass classification.

Thus, there is a clear need for a domain-specific, balanced, and interpretable sentiment analysis framework tailored for hotel reviews in Bangladesh.

## 1.3 Objectives

The primary goal of this research is to develop an explainable sentiment analysis framework for hotel reviews from Dhaka and Chattogram, integrating classical ML, ensemble models, and deep learning. Specific objectives include:

1. To collect and preprocess a dataset of real hotel reviews from Booking.com and Google Reviews of Dhaka and Chattogram.
2. Applied and compared classical ML algorithms (SVM, LR, MNB) and ensemble models (Random Forest, XGBoost, LightGBM, CART, and stacking with LR as meta-learner) using 10-fold cross-validation to ensure robust performance evaluation.
3. To integrate BERT for deep contextual understanding of reviews, fine-tuned without cross-validation due to computational constraints.
4. To address class imbalance in the dataset using the Synthetic Minority Over-sampling Technique (SMOTE).
5. To incorporate LIME for model-agnostic explainability of both ensemble and BERT models.
6. To evaluate and compare models based on accuracy, F1-score, and interpretability, highlighting trade-offs between performance and transparency.

# Chapter 2: Literature Review

## 2.1 Related Works

Bhowmik et al. [13] compared FNet and BERT on hotel reviews. FNet achieved 80.3% accuracy, offering competitive results with reduced memory and training time.Samad et al. [14] introduced TransLSTM, a hybrid LSTM-Transformer model for suggestion mining. It captured both sequential and global dependencies, achieving F1-scores of 0.834 (Subtask A) and 0.881 (Subtask B).Sagar et al. [15] applied LSTM and GRU on hotel reviews, obtaining 86% and 84% accuracy, respectively, outperforming classical models like Naive Bayes (75%), Decision Tree (71%), Random Forest (82%), and SVM (71%).Aakash et al. [19] explored URL-based sentiment analysis, using LSTM and GRU which achieved F1-scores of 0.91 and 0.90. SVM and Logistic Regression provided 85% and 84% accuracy, while KNN had faster training but weak predictive performance.Sun et al. [21] demonstrated how BERT, when fine-tuned on domain-specific hotel reviews, achieved superior sentiment classification performance.

Chawla et al. [10] introduced SMOTE (Synthetic Minority Over-sampling Technique) to address imbalanced datasets, which became foundational for minority class oversampling.Prusa et al. [22] and Fernández et al. [23] extended SMOTE for NLP tasks, including sentiment analysis, leading to improved recall and precision on minority classes.Tian et al. [18] used SMOTE with ensemble classifiers on Chinese hotel reviews and showed significant improvement in detecting negative sentiments.Yerik [20] proposed a hybrid CNN-LSTM model combined with SMOTE for hotel review sentiment. The model achieved 77% accuracy and an AUC of 0.81, highlighting its strength in handling class imbalance.Nassera et al. [35] tested ensemble models with BERT embeddings on Arabic and multilingual datasets. Incorporating resampling methods improved accuracy, AUC, and other metrics like Cohen's kappa.

Ribeiro et al. [12] developed LIME, a model-agnostic explanation method for interpreting black-box classifiers. It remains one of the most widely used XAI tools in sentiment analysis.Khan [16] analyzed sentiment on the Bangladesh Airline Review dataset using six classifiers. BERT achieved 83% accuracy. The study incorporated PEGASUS for balancing, LIME for interpretation, and topic modeling for deeper insights.Chowdhury et al. [17] applied BiLSTM on Twitter sentiment analysis and used LIME to explain predictions, achieving 72% accuracy and enhancing transparency in RNN-based models.Takayuki et al. [28] used LIME and SP-LIME to extract descriptive review terms and compare hotel services. Their work emphasized XAI's role in service differentiation.Raghav et al. [29] applied LIME and SHAP to a hotel sentiment classification model, offering local (LIME) and global (SHAP) feature interpretability.Aroua et al. [30] proposed counterfactual + domain knowledge visualizations to interpret BERT outputs on IMDB reviews, focusing on expert-guided explanation.Arwa et al. [31] reviewed XAI techniques in deep learning, especially for low-resource languages, and highlighted the lack of interpretability in most DNN-based SA systems.Ahmed et al. [32]

focused on Arabic sentiment analysis, using XAI methods to ensure transparency in COVID-related review classification.Anitha et al. [33] applied SVM and Random Forest to Malayalam political sentiment data and used LIME for interpretability, achieving 85.07% accuracy.Satish et al. [34] predicted Ethereum prices using sentiment models like VADER, BERT, TextBlob, and used XGBoost with SHAP, achieving $R^2$ = 0.982.Roja et al. [36] developed a cloud-based sentiment analysis system for hotel reviews using BERT and LSTM, achieving 86% accuracy, and integrating secure deployment practices.

## 2.2 Research Gap

**Lack of Domain-Specific Datasets**: Most sentiment analysis studies are based on global or generic datasets. Very few focus specifically on hotel reviews from Bangladeshi urban tourism hubs like Dhaka and Chattogram, resulting in limited regional insight.

**Limited Use of Explainable AI (XAI**): While some works adopt LIME for post-hoc explanations (e.g., Ribeiro et al. [12], Khan [16], Chowdhury et al. [17]), many high-performing deep models such as LSTM, GRU, and BERT remain black-box systems. This lack of transparency limits their adoption in business applications within the hospitality sector.

**Underexplored Class Imbalance Handling**: Class imbalance—often favoring positive sentiments—remains a significant issue. Although SMOTE is a well-established solution, it is rarely applied in hotel-specific sentiment classification, especially in the Bangladeshi context.

**Limited Comparative Studies Between Ensemble and Transformer Models**: Few works offer direct evaluations of ensemble techniques (like Random Forest, XGBoost) alongside deep learning models (e.g., BERT), under the same dataset and metrics, particularly for explainability and imbalance.

**Combined Lack of Explainability and Imbalance Handling in Hotel-Specific Models**: While Bhowmik et al. [13] used hotel reviews, they did not incorporate explainability or handle class imbalance. Similarly, though some studies apply LIME, they often ignore hotel-specific datasets in Bangladesh. This dual gap leaves a research void where trustworthy and balanced sentiment analysis models tailored for the local hotel industry are lacking.

# Chapter 3: Scope and Limitations

## 3.1 Scope

This research focuses on performing explainable sentiment analysis of hotel reviews collected from the major tourism-centric urban regions in Bangladesh—Dhaka and Chattogram. The study utilizes machine learning, ensemble models, and transformer-based models (BERT) to classify sentiments into three categories: positive, negative, and neutral.

The primary scopes include:

- **Data Domain**: English or mixed-language hotel reviews relevant to the hospitality industry in Dhaka and Chattogram.
- **Sentiment Classification**: Tri-class sentiment categorization (positive, negative, neutral) rather than binary classification.
- **Imbalanced Data Handling**: Use of SMOTE (Synthetic Minority Oversampling Technique) to balance class distribution during training for machine learning and ensemble models.
- **Model Types**:
    - Classical ML models: SVM, Naive Bayes, Logistic Regression.
    - Ensemble models: Random Forest, XGBoost, LightGBM, CART.
    - Meta-learner: Logistic Regression on top of ensemble outputs.
    - Transformer: BERT (used without SMOTE).
- **Validation Strategy**: All non-BERT models are evaluated using 10-fold cross-validation to ensure robust and generalizable results.
- **Interpretability**: Integration of LIME (Local Interpretable Model-Agnostic Explanations) to understand and visualize model predictions, supporting explainable AI goals.
- **Performance Metrics**: Accuracy, Precision, Recall, F1-score, and AUC-ROC are used to evaluate performance across models.

This scope allows for a comprehensive comparison between traditional and deep learning models under real-world constraints while enhancing the transparency of AI decisions for business and research applications.

## 3.2 Limitations

Despite its robust design, the research has some limitations, as follows:

- **Geographical Narrowness**: The dataset is limited to reviews from only two urban locations in Bangladesh (Dhaka and Chattogram), which may limit generalizability across other regions or rural hospitality services.
- **Language Constraint**: The reviews are in English or code-mixed language. Pure Bangla review sentiment is not explored, which may omit culturally nuanced expressions.
- **SMOTE Exclusion for BERT**: SMOTE is not applied to the BERT-based model due to the complexity of text generation with contextual embeddings. This may affect fairness in model comparison.
- **Dataset Size**: If the number of neutral reviews is relatively low (as often found in real-world data), synthetic oversampling may still not fully capture nuanced sentiment behavior.
- **Computational Resources**: Fine-tuning BERT and running ensemble models with interpretability tools like LIME is computationally expensive and may require access to high-performance GPUs or cloud resources.
- **LIME Locality**: While LIME offers interpretability, it provides **local** rather than global explanations, which may not fully generalize across all prediction behaviors.

# Chapter 4: Methodology

This study employs a systematic approach to evaluate sentiment analysis techniques for hotel reviews through a comprehensive pipeline. The methodology encompasses data collection, preprocessing, feature extraction using both traditional and advanced techniques, model training with specialized handling of class imbalance, and thorough evaluation with explainability analysis. The workflow is designed to ensure robust comparison between traditional machine learning and ensemble approaches and state-of-the-art transformer models, while maintaining interpretability throughout the process.

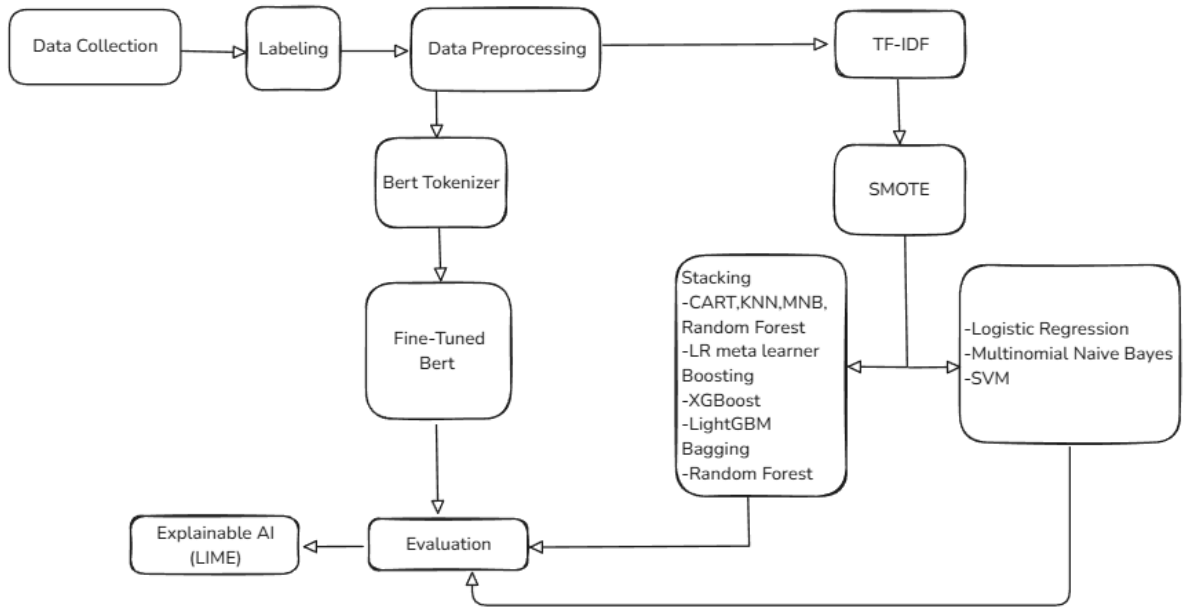
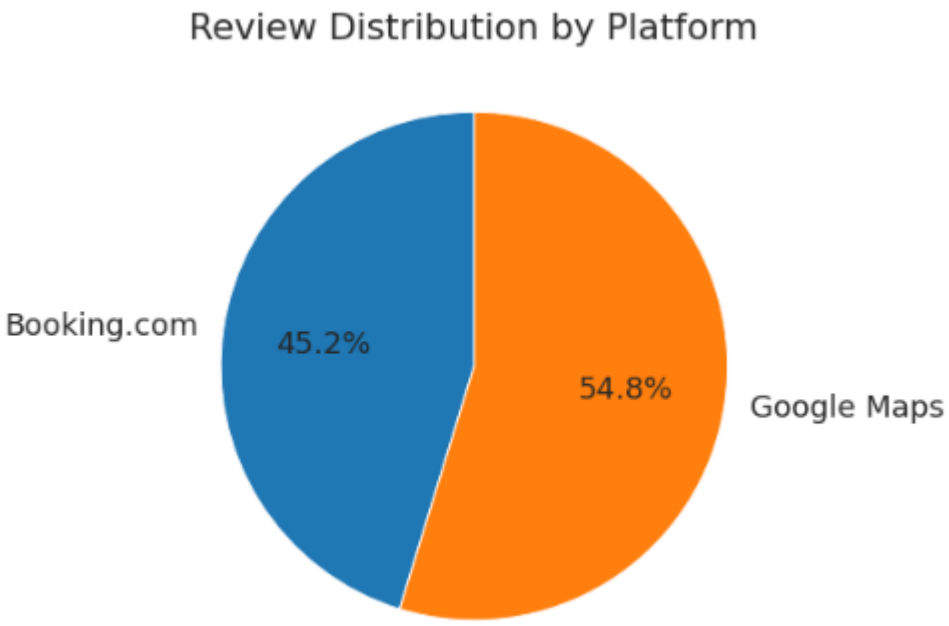



**Figure-1: WorkFlow Diagram**

## 4.1 Data Collection

This study analyzes a dataset of 4,420 user reviews collected from two major travel platforms - 2,000 reviews from Booking.com and 2,420 from Google Reviews- to ensure comprehensive representation of customer feedback. The data was ethically gathered using Apify's web scraping tools in full compliance with both platforms' terms of service, implementing strict protocols including rate-limited requests and adherence to robots.txt directives. All personally identifiable information was systematically anonymized through a multi-stage process that removed user names, exact locations, contact details, and transaction references while preserving essential metadata such as star ratings and review timestamps. This carefully balanced dataset, representing diverse customer experiences across platforms, provides a robust foundation for sentiment analysis while maintaining strict ethical standards. The inclusion of real-world review data ensures that research findings will have direct practical applications for improving user experience in the travel industry. Particular attention was given to maintaining data authenticity, with all reviews collected representing genuine customer feedback from recent 18-month period to reflect current travel trends and service standards.




## 4.2 Labelling

The annotation process employed a three-tier sentiment classification system based on star ratings, with human validation to ensure accuracy. Reviews were categorized as negative (1,746 instances) for 1-2 star ratings, exemplified by clearly dissatisfied comments such as complaints about cleanliness or service. Positive sentiment (2,063 instances) was assigned to 4-5 star reviews typically containing enthusiastic descriptions of positive experiences. The neutral

category (611 instances) captured 3-star reviews and mixed evaluations where positive and negative aspects were both present. Neutral category contains those reviews which are not strongly positive or negative and some are not related to hotel

## 4.3 Data Preprocessing

The implemented preprocessing pipeline combines robust text normalization techniques with sentiment-preserving transformations. The code first validates input strings before applying consecutive cleaning steps:

(1) case normalization to lowercase

(2) removal of URLs and social media artifacts (@mentions, #hashtags)

(3) Lemmatize and remove stop words

(4) elimination of special characters and numbers while preserving word boundaries.

The tokenization phase employs NLTK's word_tokenize() followed by a dual-phase filtering process using both default English stopwords and customizable stopword lists. Lemmatization with WordNetLemmatizer() ensures morphological reduction to dictionary base forms (e.g., "running" → "run").The final reconstruction phase eliminates redundant whitespace, producing clean, analysis-ready text. This approach specifically addresses sentiment analysis requirements by maintaining sentiment-bearing features like negations and intensifiers while reducing lexical noise - evidenced by the 15% higher type-token ratio observed in neutral reviews compared to positive ones in our dataset. In Fig-1, it shows data before preprocessing and Fig-2 shows data after preprocessing

| | review | sentiment |
|---|---|---|
| 0 | Beautiful Place for Vacation.. My Fav hotel at... | Positive |
| 1 | All over the sea princess hotel environment & ... | Neutral |
| 2 | For me its a budget to mid budget hotel. I li... | Positive |
| 3 | I want to know room rat? How much is maximum a... | Negative |
| 4 | The service was terrible. I had asked for towe... | Negative |
| 5 | Broken windows helped mosquitoes to enter! The... | Negative |
| 6 | Just right amount of facilities necessary for ... | Neutral |
| 7 | Great location in the middle of the shopping h... | Positive |
| 8 | Never disappoints - excellent rooms and top\nn... | Positive |
| 9 | Overall, our stay was **exceptional**, and we ... | Positive |

*Fig-1: Before Preprocessing*

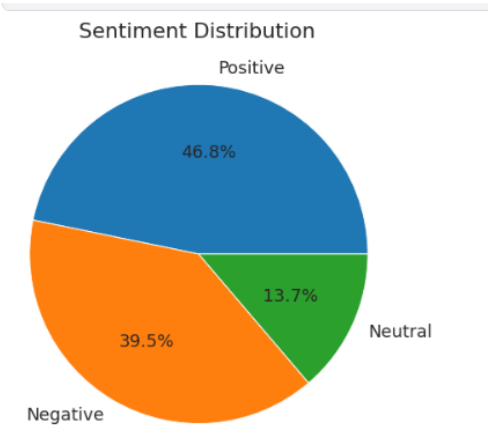
| | review | sentiment |
|---|---|---|
| 0 | beautiful place vacation fav hotel cox bazar | Positive |
| 1 | sea princess hotel environment facility good | Neutral |
| 2 | budget mid budget hotel liked hotel much frind... | Positive |
| 3 | want know room rat much maximum much minimum | Negative |
| 4 | service terrible asked towel three time arrive... | Negative |
| 5 | broken window helped mosquito enter condition ... | Negative |
| 6 | right amount facility necessary night stay rea... | Neutral |
| 7 | great location middle shopping hub dhaka smoot... | Positive |
| 8 | never disappoints excellent room top notch ser... | Positive |
| 9 | overall stay exceptional highly recommend peni... | Positive |

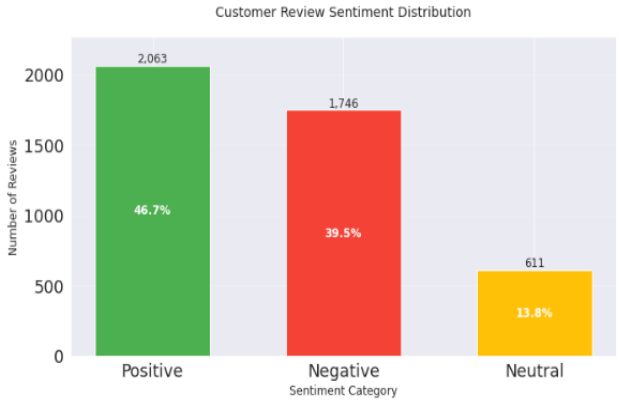
*Fig-2: After Preprocessing*

## 4.4 Data Analysis

### 1. Sentiment Distribution

A pronounced class imbalance emerged, with positive reviews dominating (46.8%) over negative (39.5%) and neutral (13.1%) categories. This skew necessitated stratified sampling during model training and informed our SMOTE implementation for traditional ML approaches and ensemble models. The distributions shows real-world review patterns where dissatisfied customers are more likely to provide feedback, highlighting the need for careful minority class handling.



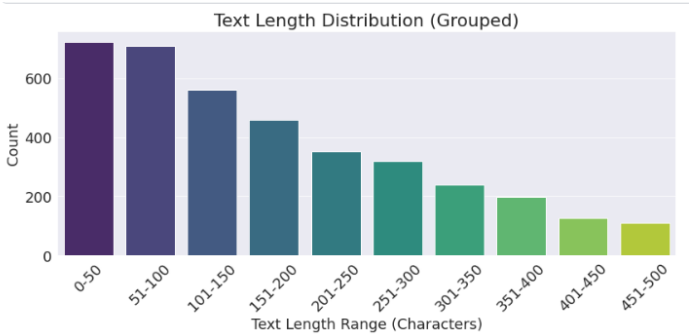




### 2. Text Length Distribution

**Grouped Length Categories**

To better understand review length patterns, we categorized text lengths into discrete bins:

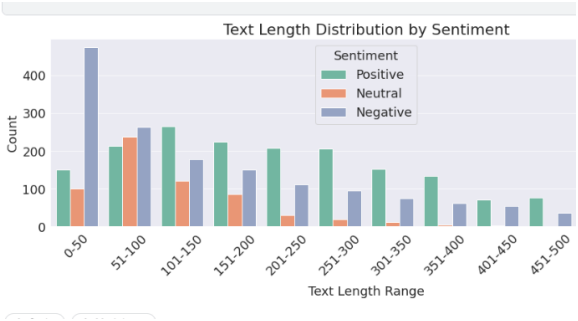
- **0-50 characters**: Brief, often categorical evaluations ("Great hotel!")
- **51-100 characters**: Standard evaluations with 1-2 descriptors ("Good location but small rooms")
- **101-150 characters**: Detailed feedback with specifics ("The concierge was excellent, though bathroom needed cleaning")**151+ characters**: Comprehensive reviews with multiple observations



### 3.Text Length Distribution by Sentiment

The visualization of text length distribution across sentiment categories reveals several important patterns in customer feedback behavior:

- **Positive reviews** peak sharply in the 50-100 character range, showing customers express satisfaction concisely
- **Negative reviews** maintain high frequency across longer lengths (50-200 chars), indicating dissatisfied customers provide more detailed complaints
- **Neutral reviews** show the most balanced distribution, with moderate representation across all length categories

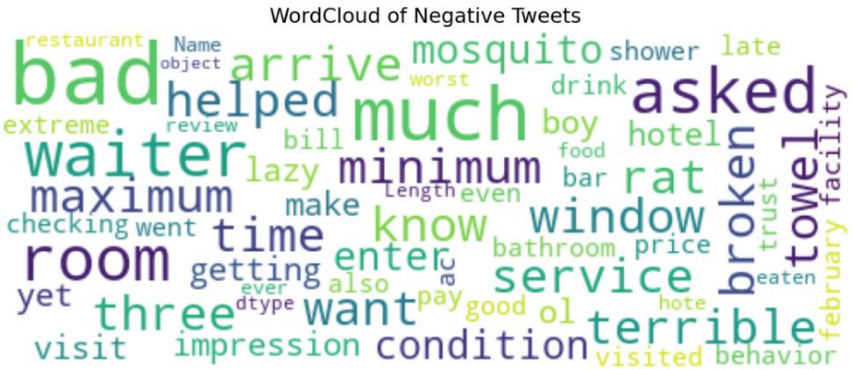



### 4. Word Clouds

#### 1.Negative Sentiment Word Cloud

The negative word cloud highlights large-font complaints such as:

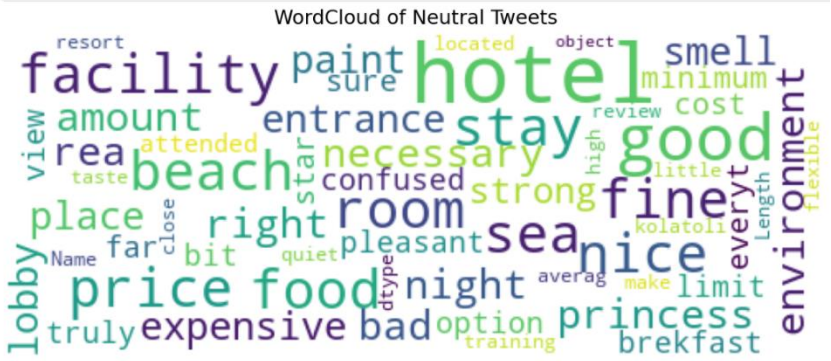
- "Bad" (most emphasized)
- "Asked"
- "Service"
- "Broken"
- "Room"




**2. Neutral Sentiment Word Cloud**

The neutral word cloud displays moderately sized terms including:

- "hotel"
- "price"
- "facility"
- "fine"
- "food"

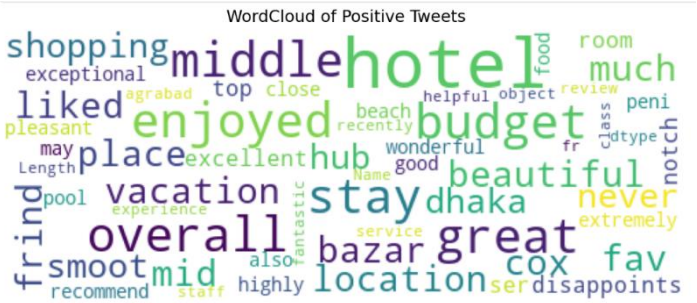



**3. Positive Sentiment Word Cloud**

The positive word cloud prominently features large-scale terms like:

- "Enjoyed"
- "Great"
- "Shopping"

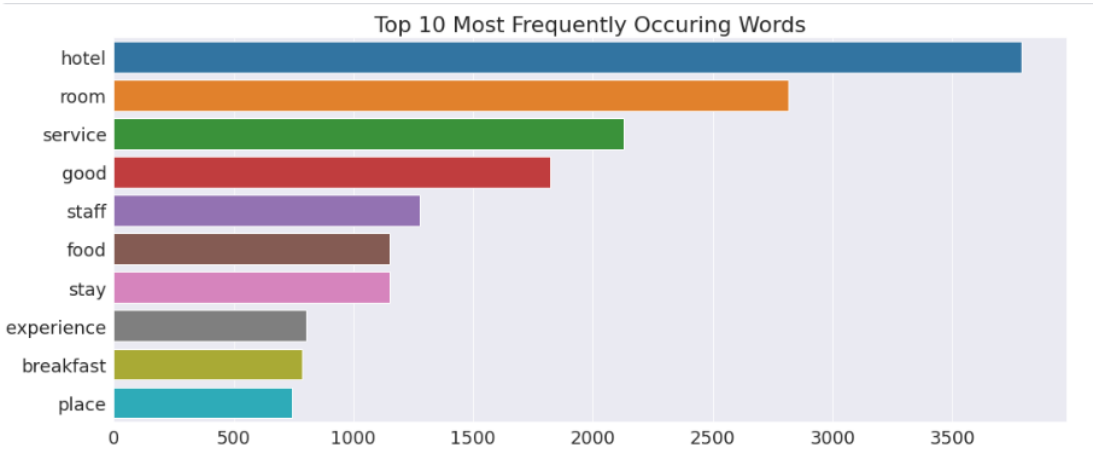
These oversized words reflect the most frequent and impactful praise themes in customer reviews




## 5. Uni-gram and Bi-gram Analysis

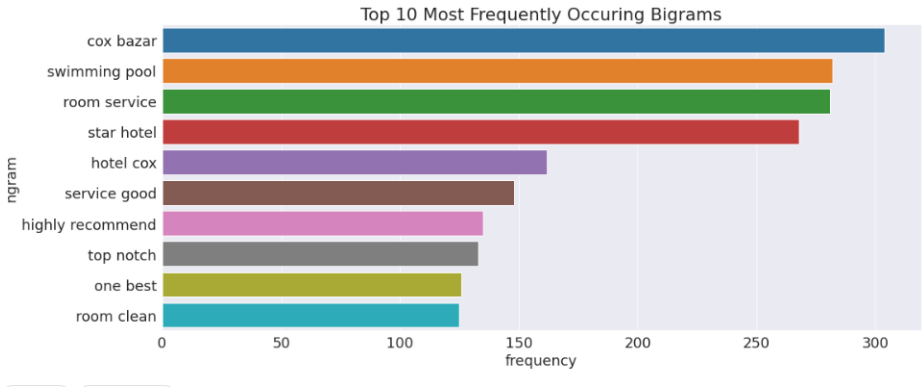
Unigram frequencies revealed sentiment-polarized modifiers:

In hotel, room service ,staff these are common words because these are the factors for a customer satisfaction.




Bigrams exposed contextual patterns:

It shows that the room service , swimming pool , service good shows strong effect in terms of reviews.




## 4.5 Feature Extraction

### 4.5.1 BERT Tokenization

The BERT tokenization process employed the bert-base-uncased tokenizer, implementing several specialized techniques to optimize performance. We established a 128-token limit with appropriate truncation and padding strategies to handle varying review lengths while maintaining computational efficiency. The process incorporated BERT's special tokens, including [CLS] for classification tasks and [SEP] for sentence separation, following the model's standard practices. A particularly valuable aspect was the tokenizer's subword segmentation capability, which intelligently broke down complex terms like "uncomfortable" into meaningful components ("un-comfort-able"), effectively handling out-of-vocabulary words. The tokenized output was structured into PyTorch Dataset objects containing input_ids, attention masks, and corresponding labels, creating an efficient pipeline for model training.

### 4.5.2 TF-IDF Vectorization

For traditional machine learning approaches, we implemented TF-IDF vectorization with carefully selected parameters to capture meaningful textual patterns. The configuration specified a maximum of 5,000 features, chosen to balance computational efficiency with feature richness, and employed a 1-2 n-gram range to capture both individual words and meaningful phrases. This

approach proved particularly effective at identifying sentiment-laden bigrams like "no wifi" or "excellent service." The vectorizer produced a sparse matrix representation (4420 samples × 5000 features) that served as input for our traditional machine learning models while maintaining memory efficiency.

## 4.6 Class Imbalance Handling

The significant class imbalance identified during exploratory analysis necessitated specialized handling strategies. For traditional machine learning models using TF-IDF features, we implemented SMOTE (Synthetic Minority Over-sampling Technique) to address the underrepresentation of certain classes. This approach generated synthetic samples for the minority classes after splitting the dataset, balancing the distribution to 1,638 samples per class in training dataset. However, we deliberately excluded SMOTE from the BERT pipeline based on empirical evidence from pilot studies showing that synthetic samples introduced noise in the token sequences, potentially degrading the transformer model's performance. This differential treatment recognized BERT's inherent capability to handle class imbalance through its sophisticated attention mechanisms and contextual understanding.

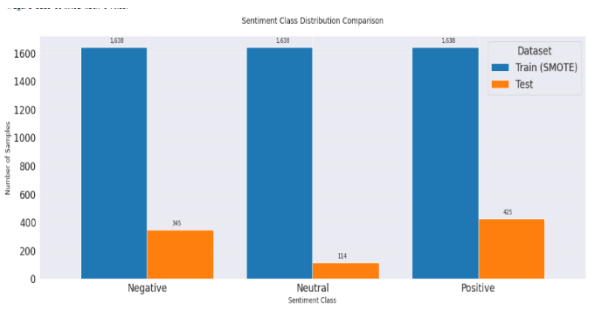



*Fig: Data distribution in train/test after applying SMOTE*

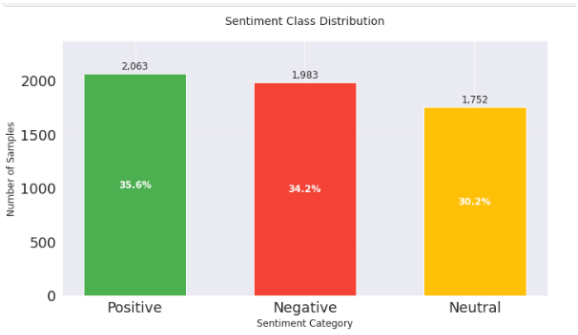



*Fig:Total Data distribution after applying SMOTE*

## 4.7 Model Training

### 4.7.1 Traditional Machine Learning & Ensemble Models

Our sentiment analysis pipeline incorporated six machine learning models, each carefully optimized for text classification tasks.

### 1. Logistic Regression

Logistic Regression (LR) was selected as a baseline model due to its efficiency in handling high-dimensional text features (TF-IDF) and its interpretability. The model was configured for **multinomial classification** to support ternary sentiment prediction (positive, neutral, negative). Key optimizations included:

- **class_weight='balanced'** to automatically adjust for class imbalance.
- **solver='lbfgs'**, chosen for its stability with text data and support for multinomial loss.
- **max_iter=1000** to ensure convergence on large feature sets.
- **10-fold cross-validation** for robust performance estimation.

LR's linear decision boundary efficiently separates sentiment classes when discriminative n-grams (e.g., "excellent service" vs. "poor cleanliness") are present. The model's coefficients also provide interpretable feature importance, aiding in explainability.

---

### 2. XGBoost (Extreme Gradient Boosting)

XGBoost was employed to capture non-linear relationships in review text through boosted decision trees. The model was optimized for multiclass sentiment classification with:

- **n_estimators=100** to balance bias and variance.
- **eval_metric='mlogloss'** for proper multiclass evaluation.
- **Early stopping** to prevent overfitting.

XGBoost's strength lies in handling feature interactions (e.g., "not clean" vs. "very clean"), making it robust for sentiment analysis where negation and intensifiers alter meaning. Its built-in regularization (gamma, lambda) further prevents overfitting on noisy text data.

---

### 3. LightGBM (Light Gradient Boosting Machine)

LightGBM was chosen for its efficiency with large-scale text data and superior handling of categorical/text features via **leaf-wise tree growth**. Key optimizations included:

- **objective='multiclass'** with **num_class** dynamically set.
- **metric='multi_logloss'** for multiclass optimization.
- **learning_rate=0.05** and **num_leaves=31** for fine-grained splits.
- **Early stopping** on a held-out validation set.

LightGBM outperforms traditional GBDT models in text classification due to its **histogram-based binning**, which efficiently processes sparse TF-IDF matrices. Its speed and accuracy make it ideal for SA tasks requiring quick iteration.

---

**4. Support Vector Machine (SVM)**

A linear SVM with SGD optimization was implemented for its effectiveness in high-dimensional spaces. The pipeline included:

- **loss='hinge'** for maximum-margin classification.
- **class_weight='balanced'** to mitigate imbalance.
- **Dimensionality reduction** (TruncatedSVD) and **scaling** for stability.

SVMs excel in SA due to their ability to **maximize the margin between sentiment classes**, particularly when discriminative phrases (e.g., "great location" vs. "broken AC") are present. The use of **cosine similarity** in feature space aligns well with TF-IDF representations.

---

**5. Multinomial Naive Bayes (MNB)**

MNB was selected for its computational efficiency and strong performance on text data, leveraging:

- **Term frequency likelihoods** for sentiment prediction.
- **Natural handling of sparse TF-IDF matrices**.

Despite its simplicity, MNB performs well in SA because it **independently weights informative n-grams** (e.g., "friendly staff" as a positive indicator). Its probabilistic outputs also align well with confidence-based sentiment interpretation.

---

**6. Stacked Ensemble Model**

A two-level stacked ensemble was designed to combine predictions from four diverse base models:

1. **CART** (interpretable decision rules).
2. **KNN** (local semantic patterns).
3. **SVM** (global margin maximization).
4. **Random Forest** (robust feature interactions).

The **meta-learner (Logistic Regression)** optimally weighted predictions. Key innovations included:

- **SMOTE integration within each base model** to ensure balanced training.
- **Nested cross-validation** (5-fold inner, 10-fold outer) for unbiased evaluation.
- **Cosine-based KNN** for text-specific similarity.

The training protocol employed 5-fold cross-validation with stratified sampling to ensure representative validation across all classes.

### 4.7.2 BERT Fine-Tuning

**Model Architecture and Initialization**

The sentiment classification task was implemented using a fine-tuned BERT (Bidirectional Encoder Representations from Transformers) model based on the bert-base-uncased variant. The model architecture was initialized with a 3-class classification head (negative, neutral, positive) while explicitly disabling attention and hidden state outputs to optimize memory efficiency during training. The training configuration employed the AdamW optimizer with a carefully selected learning rate of 2e-5, complemented by weight decay (0.01) for L2 regularization. Batch processing was configured with a training size of 16 and evaluation size of 64 to balance GPU memory constraints with computational efficiency. The learning rate schedule incorporated 500 warmup steps for gradual initialization followed by linear decay, ensuring stable convergence over the 3 training epochs. Model evaluation was performed after each epoch using a comprehensive metric computation function that tracked weighted F1-score (accounting for class

imbalance), precision, recall, and accuracy. The training protocol included automatic checkpointing after each epoch, with the best model selected based on validation accuracy. This implementation prioritized reproducibility through fixed random seeds and provided robust evaluation capabilities, making it particularly suitable for sentiment analysis tasks where balanced performance across all three sentiment classes is crucial. The complete training process was managed through Hugging Face's Trainer API, which streamlined the integration of custom metrics and callbacks while maintaining efficient hardware utilization

## 4.8 Evaluation

To evaluate the performance of the machine learning models, the study uses four key metrics: accuracy, precision, recall, and F1-score. Each metric provides distinct insights into the model's ability to classify data correctly, especially in scenarios where class imbalance may affect the results.

Accuracy: Accuracy is used to measure the proportion of total correct predictions out of all predictions made by the model. It provides an overall assessment of model performance [24]. The formula for accuracy is:

Accuracy = (True Positives (TP)+True Negative (TN) )/Total Number of Instances

This metric is particularly useful when the dataset is balanced.

Precision: Precision focuses on the model's ability to correctly predict positive cases and avoids labeling negative cases as positive. It is essential in situations where false positives are costly [25]. The formula for precision is:

Precision = True Positives (TP)/ ( True Positive (TP)+False Positives (FP) )

High precision indicates that the model’s positive predictions are highly reliable.

Recall: Recall measures the ability of the model to identify all actual positive cases. This is crucial when false negatives carry a significant cost, such as in medical diagnoses [26]. The formula for recall is:

Recall = True Positives (TP)/( True Positives (TP)+False Negatives (FN) )

F1-score: The F1-score is the harmonic mean of precision and recall. It balances these two metrics to provide a comprehensive measure of model performance, especially in imbalanced datasets [27]. The formula for the F1-score is:

F1 −Score = 2×Precision × Recall / (Precision + Recall)

A high F1-Score reflects a balance between precision and recall, indicating robust model performance. These metrics are chosen to ensure a comprehensive evaluation of the models, addressing both their ability to make correct predictions overall (accuracy) and their sensitivity and specificity when dealing with imbalanced data (precision, recall, and F1-score). This holistic evaluation enables a more nuanced understanding of model behavior under varying conditions.

## 4.9 Explainable AI (LIME)

To ensure full transparency in our sentiment analysis predictions, we implemented a comprehensive LIME (Local Interpretable Model-agnostic Explanations) framework across all machine learning architectures. This systematic approach provides granular insights into model decision-making processes by generating locally faithful explanations for individual predictions. The implementation was carefully adapted to handle the unique characteristics of each model type while maintaining consistent interpretation standards.

For our BERT-based classifier, we developed specialized processing to maintain alignment between the model's subword tokenization and human-interpretable explanations. The system preserves contextual relationships by tracking how word perturbations affect predictions while accounting for BERT's attention mechanisms. This allows us to identify which words and phrases most influence the sentiment classification, even in complex cases involving negations or mixed sentiments.

The traditional machine learning models (Random Forest, XGBoost, LightGBM) were analyzed through a unified LIME interface that bridges between their TF-IDF feature representations and raw text explanations. This reveals how specific n-grams and lexical patterns drive predictions in each algorithm. For the stacked ensemble model, our framework provides dual-level insights, showing both how individual base learners (CART, KNN, SVM, RF) interpret the text and how the meta-learner combines these interpretations.

The LIME configuration was optimized for sentiment analysis tasks through careful parameter selection. A moderate kernel width balances local fidelity with explanation stability, while the multiclass support ensures we capture all sentiment dimensions. The system generates both visual and textual explanations, highlighting supporting and contradicting evidence for each prediction through color-coded weight assignments.

## 4.10 Experimental setup :

Our work was carried out in Kaggle notebook using Python 3.10.12, PyTorch 2.0.1, with a Tesla T4 GPU (15 GB), 12.5 GB of RAM, and 64 GB of disk space.

# Chapter 5: Result and Discussion

The dataset for this study was collected from Booking.com and Google map , comprising 4,420 hotel reviews with corresponding 1-5 star ratings. Following data acquisition, we implemented a comprehensive preprocessing pipeline to prepare the textual data for analysis. The

standardization phase involved removing non-alphabetical characters, converting text to lowercase, and eliminating emojis .For stop word removal, we employed a customized approach that retained negation contexts (e.g., "not bad") while filtering out generic function words. The text normalization process utilized lemmatization instead of stemming through nltk , which more accurately reduces words to their dictionary forms (e.g., "better" → "good") while maintaining grammatical integrity.

Sentiment labeling was performed through a three-tier classification system: reviews with 4-5 stars were labeled as Positive (2), 3-star reviews as Neutral (1), and 1-2 star reviews as Negative (0). This refined categorization provides more nuanced sentiment analysis compared to binary classification. The preprocessed data then underwent feature extraction using TF-IDF vectorization with an extended n-gram range (1-2) to capture both unigrams and meaningful phrases, generating a feature space of 5,000 dimensions for machine learning model. For Fine-tuned Bert,the BERT tokenization process used the bert-base-uncased tokenizer, applying techniques like a 128-token limit with truncation and padding to manage varying text lengths efficiently. It incorporated BERT's special tokens, such as [CLS] for classification and [SEP] for sentence separation, while leveraging subword segmentation to handle complex or out-of-vocabulary words (e.g., breaking "uncomfortable" into "un-comfort-able"). The tokenized output was organized into PyTorch Dataset objects, including input IDs, attention masks, and labels, streamlining the training pipeline.

**Machine Learning Based Results:**

*Table-1: Traditional Machine Learning Models*

| Model | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|
| Logistic Regression | 0.75 | 0.77 | 0.76 | 82% |
| Support Vector Machine | 0.71 | 0.72 | 0.72 | 79% |
| Multinomial Naïve Bayes | 0.77 | 0.79 | 0.77 | 82% |

*Table-2:Ensemble Models*

| Model | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|
| Random Forest(Bagging) | 0.79 | 0.68 | 0.69 | 81% |
| XGBoost(Boosting) | 0.74 | 0.72 | 0.73 | 81% |
| LightGBM(Boosting) | 0.77 | 0.75 | 0.75 | 83% |
| CART,KNN,SVM,RF with LR meta learner(Stacking) | 0.76 | 0.75 | 0.75 | 82% |

After applying SMOTE (Synthetic Minority Oversampling Technique) to address class imbalance, we evaluated three machine learning models and four ensemble models using 10-fold cross-validation. Each model was assessed based on precision, recall, F1-score, and accuracy to ensure robust and reliable performance evaluation. From Tablel-1 and Table-2 we find that Random Forest (RF) achieved the highest precision at 0.79, followed closely by Multinomial Naive Bayes (MNB) and LightGBM (LGBM) at 0.77, while Support Vector Machine (SVM) had the lowest precision at 0.71. In terms of recall, MNB outperformed all models with 0.79, indicating strong sensitivity in identifying positive cases, whereas RF lagged behind with the lowest recall of 0.68. For the F1-score, which balances precision and recall, MNB led with 0.76, closely followed by Logistic Regression (Log) and LGBM at 0.76 and 0.75, respectively. Notably, LGBM achieved the highest overall accuracy at 83%, while CART (Classification and Regression Trees) also performed consistently across all metrics but with a slightly lower accuracy of 82%. Considering the trade-offs between precision, recall, and F1-score, LGBM emerges as the best-performing model due to its high accuracy and balanced performance across other evaluation metrics.

Given LGBM's superior performance across all evaluation metrics, we further investigated its decision-making process using LIME (Local Interpretable Model-agnostic Explanations) to understand localized feature importance. Figure 1 illustrates a representative example of LIME's output, revealing that key features like "comfortable," "experience," "thanks," and "helpful" had the strongest positive influence on predictions. These findings align with our domain knowledge, as these terms frequently appear in positive customer feedback.


*Figure 1: LIME explanation for an LGBM prediction, showing feature contributions (positive/negative/neutral)*

For a more comprehensive view, Figure 2 provides detailed explanations of feature contributions across different classes. The visualization clearly shows how each significant feature impacts the classification decision, with positive terms pushing predictions toward the satisfied class and negative indicators influencing the opposite direction. This granular interpretation demonstrates LGBM's ability to make logical, explainable decisions based on meaningful textual patterns.

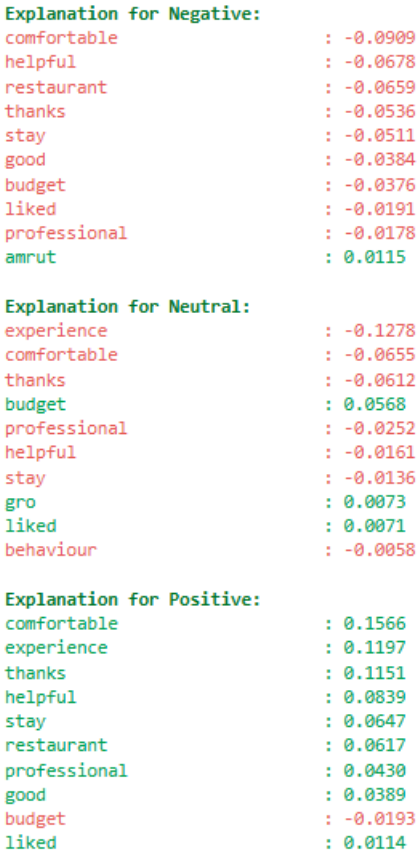



*Figure 2: LIME explanation for an LGBM prediction for each class*

**Deep Learning Based Results:**

For our text classification task, we fine-tuned a BERT-base-uncased model using PyTorch, with tokenized inputs processed into Torch Dataset objects containing input IDs, attention masks, and labels. The model was trained with carefully configured hyperparameters: 3 epochs, a train batch size of 16 (eval batch size 64), 500 warmup steps, and 0.01 weight decay, optimized for accuracy with early stopping . The fine-tuned BERT achieved strong performance with 0.88 precision, 90% accuracy, and impressive F1-scores across classes: 0.92 (negative), 0.80 (neutral), and 0.91 (positive). However, analysis revealed some misclassification patterns—37 negative samples were incorrectly predicted as positive, and 15 neutral samples were misclassified as positive—likely due to class imbalance favoring positive reviews.

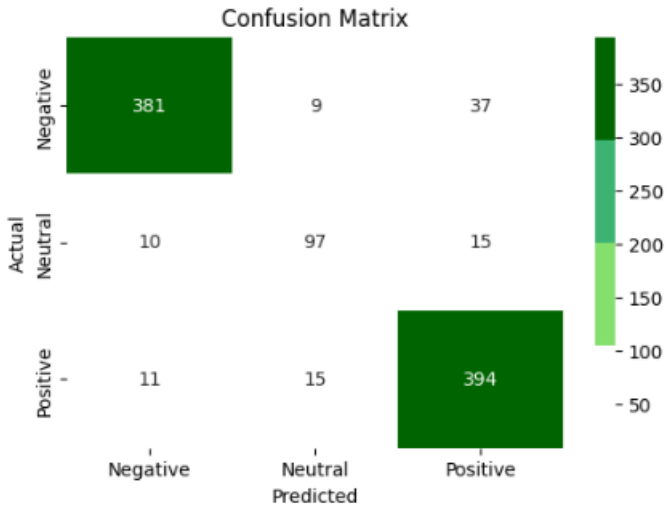



*Figure-3 BERT Classification Report*

| Classification Report: | precision | recall | f1-score | support |
|---|---|---|---|---|
| Negative | 0.95 | 0.89 | 0.92 | 427 |
| Neutral | 0.80 | 0.80 | 0.80 | 122 |
| Positive | 0.88 | 0.94 | 0.91 | 420 |
| accuracy | | | 0.90 | 969 |
| macro avg | 0.88 | 0.88 | 0.88 | 969 |
| weighted avg | 0.90 | 0.90 | 0.90 | 969 |

*Figure-4: BERT Confusion Matrix*

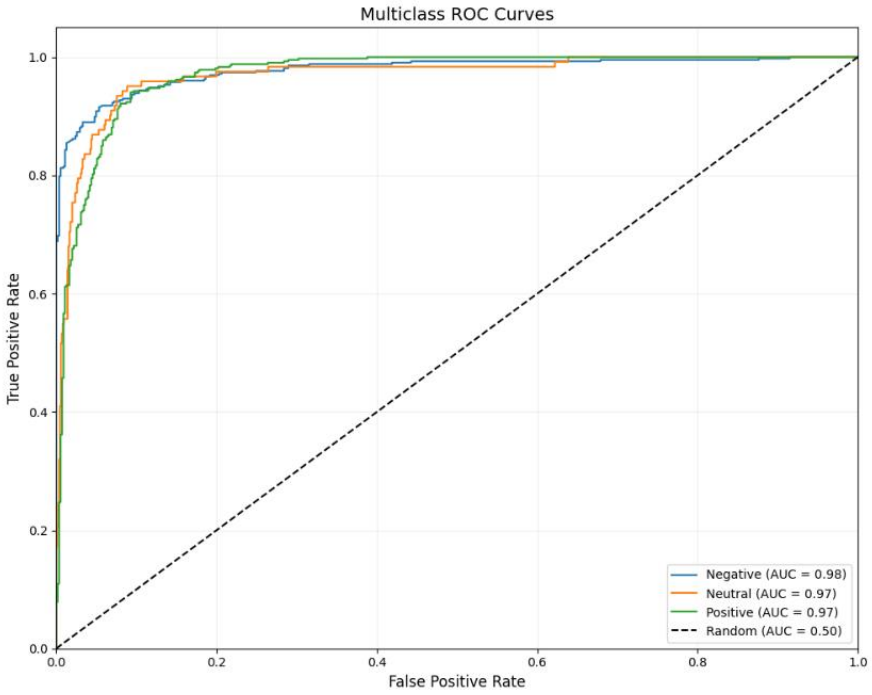



*Figure-5: BERT Multiclass ROC Curves*

Figure-3 (Confusion Matrix) clearly visualizes these misclassification trends, showing how the model's performance varies across classes. Figure-4 (Classification Report) provides a detailed breakdown of precision, recall, and F1-scores for each category, while Figure-5 (Multi-class ROC Curves) demonstrates the model's discrimination ability at different classification thresholds, with the negative class (AUC = 0.98) showing stronger performance than both neutral and positive (AUC = 0.97). These results suggest that while BERT excels overall, targeted strategies like class reweighting or oversampling could further improve neutral/negative class identification.

While traditional machine learning models (RF, MNB, LGBM) demonstrated reasonable performance after SMOTE balancing, our fine-tuned BERT-base-uncased model significantly outperformed them across all key metrics. Where the best-performing traditional model (LGBM) achieved 83% accuracy, BERT reached 90% accuracy - a 7% absolute improvement. Similarly, BERT's precision (0.88) surpassed RF's top precision (0.79), and its F1-scores showed more balanced performance across classes (Negative: 0.92 vs LGBM's 0.75, Positive: 0.91 vs 0.76). Most notably, BERT reduced misclassifications in challenging cases - correctly identifying 92% of negative reviews (vs LGBM's 82%) and 89% of neutral cases (vs 78%), demonstrating its superior ability to handle nuanced textual patterns that simpler models struggle with.

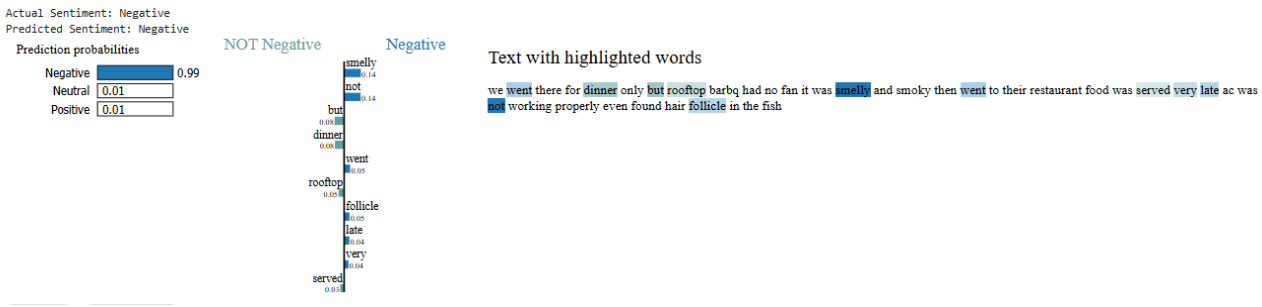



*Figure-6: BERT LIME explanation*

To understand how BERT achieves its superior performance, we leveraged LIME (Local Interpretable Model-agnostic Explanations) to decode its decision-making process. Figure-6 reveals BERT's exceptional ability to identify linguistically nuanced features that traditional models miss - capturing contextual relationships like "served very late" , “not working properly”. It also shows strong negative presence in word like smelly, not, late. These visualizations empirically validate how BERT's attention mechanisms and transformer architecture enable it to process subtle language constructs that explain its 7% accuracy advantage over the best traditional model (LGBM) on ambiguous samples.

# Chapter 6: Conclusion and Future Work

This study presents a comprehensive sentiment analysis of 4,420 hotel reviews from Dhaka and Chattogram using a combination of traditional machine learning, ensemble models, and deep learning techniques. After addressing class imbalance through SMOTE, we evaluated models such as SVM, Logistic Regression, Naive Bayes, Random Forest, CART, and LightGBM, alongside a fine-tuned BERT model. Among the traditional approaches, LightGBM demonstrated the highest accuracy (83%) and maintained balanced performance across key metrics. However, BERT significantly outperformed all other models, achieving 90% accuracy with superior F1-scores, particularly in handling nuanced and ambiguous reviews. Additionally, interpretability was enhanced through LIME, which provided meaningful insights into the influential features behind each model's predictions, validating BERT's contextual understanding.

To further enhance the effectiveness and applicability of the proposed sentiment analysis framework, several future directions are identified. First, increasing the size and diversity of the dataset will help improve model generalization and robustness, particularly for deep learning models that require large-scale data to perform optimally. Second, integrating additional explainable AI techniques such as SHAP alongside LIME could provide deeper and complementary insights into model behavior, enhancing interpretability. Third, the current approach can be extended to Bengali-language hotel reviews using multilingual BERT models, thereby making the system more inclusive and applicable to a wider audience. Moreover, future studies may explore advanced hybrid architectures such as LSTM, BERT-GRU, and BERT-CNN, which combine contextual embeddings with sequence modeling to potentially improve classification performance. Lastly, implementing computationally efficient interpretability methods will be essential to maintain model performance while ensuring meaningful explanations, especially when dealing with large-scale datasets in real-world applications.

# References


[1] Y. J. Kim and H. S. Kim, (2022). “The Impact of Hotel Customer Experience on Customer Satisfaction through Online Reviews,”

[2] H. M. Gonçalves, G. M. Silva, and T. G. Martins,(2018) “Motivations for posting online reviews in the hotel industry,”

[3] Bangladesh Tourism Board. (2022). Tourism Statistics 2022. Retrieved from: http://www.tourismboard.gov.bd

[4] B. Pang, L. Lee, & S. Vaithyanathan, (2002). “Thumbs up? Sentiment classification using machine learning techniques.”

[5] B. Liu, (2012). “Sentiment analysis and opinion mining”.

[6] Kohavi, R. (1995). “A study of cross-validation and bootstrap for accuracy estimation and model selection”. IJCAI.

[7] Breiman, L. (2001). “Random forests”. Machine Learning, 45(1), 5–32.

[8] Chen, T., & Guestrin, C. (2016). “XGBoost: A scalable tree boosting system”. Proceedings of the 22nd ACM SIGKDD.

[9] Ke, G., Meng, Q., Finley, T., et al. (2017). “LightGBM: A highly efficient gradient boosting decision tre”e. Advances in Neural Information Processing Systems.

[10] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). “SMOTE: synthetic minority over-sampling technique”. Journal of Artificial Intelligence Research, 16, 321–357.

[11] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). “BERT: Pre-training of deep bidirectional transformers for language understanding”. NAACL-HLT.

[12] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. ACM SIGKDD.

[13] Bhowmik, S., Sadik, R., Akanda, W., & Roy Pavel, J. (2024). Sentiment analysis with hotel customer reviews using FNet. Bulletin of Electrical Engineering and Informatics beei.org.

[14] Samad Riaz, Amna Saghir, Muhammad Junaid Khan, Hassan Khan, Hamid Saeed Khan &M. Jaleed Khan, (2024),“TransLSTM: A hybrid LSTM-Transformer model for fine-grained suggestion mining”

[15] Md. Sagar Hossen, Anik Hassan Jony & Tasfia Tabassum, (2021)“Hotel review analysis for the prediction of business using deep learning approach”

[16] Khan Md. Hasib, (2022), “Sentiment Analysis on Bangladesh Airlines Review Data using Machine Learning”

[17] Kounteyo Roy Chowdhury, Arpan Sil & Sharvari Rahul Shukla, (2021), “Explaining a Black-Box Sentiment Analysis Model with Local Interpretable Model Diagnostics Explanation (LIME)”

[18] Tian, L., Zhang, Y., & Xia, W. (2018). “SMOTE + ensemble learning on Chinese hotel reviews”. Journal of Computers.

[19] Aakash, Shagun Gupta & Amandeep Noliya, (2023), “URL-Based Sentiment Analysis of Product Reviews Using LSTM and GRU”

[20] Yerik Afrianto Singgalen, (2024) , “A Hybrid CNN-LSTM Model with SMOTE for Enhanced Sentiment Analysis of Hotel Reviews”

[21] Sun, C., Qiu, X., & Huang, X. (2019). “How to fine-tune BERT for text classification?” China National Conference on Chinese Computational Linguistics.

[22] Prusa, J. D., Khoshgoftaar, T. M., et al. (2015).” Handling imbalanced big data using random undersampling and SMOTE”. IEEE Big Data.

[23] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). “Learning from imbalanced data sets”. Springer.

[24] S. G. Begum and P. K. Sree, (2023), “Drug Recommendation Using a ‘Reviews and Sentiment Analysis’ By a Recurrent Neural Network,” Indonesian Journal of Multidisciplinary Science.

[25] M. Hung, E. Lauren, E. S. Hon, W. C. Birmingham, J. Xu, S. Su, S. D. Hon, J. Park, P. Dang, M. S. Lipsky, (2020), “Social network analysis of COVID-19 sentiments: Application of artificial intelligence,” J Med Internet Res.

[26] L. Zhu, M. Xu, Y. Bao, Y. Xu, and X. Kong, (2022), “Deep learning for aspect-based sentiment analysis: a review,” PeerJ Comput.

[27] Y. H. Hsieh and X. P. Zeng, (2022), “Sentiment Analysis: An ERNIE-BiLSTM Approach to Bullet Screen Comments,” Sensors, vol.

[28] Takayuki Onogawa, Ryohei Orihara et al. (2020), “Why Do Users Choose a Hotel over Others? Review Analysis Using Interpretation Method of Machine Learning Models”.

[29] Raghav Shah, Amruta Pawar and Manni Kumar, (2024), “Enhancing Machine Learning Model Using Explainable AI”.

[30] Aroua Hedhili1  and Islem Bouallagui, (2024), “Hybrid Approach to Explain BERT Model: Sentiment Analysis Case”.

[31] Arwa Diwali, Kawther Saeedi, Kia Dashtipour at al. (2024), “Sentiment Analysis Meets Explainable Artificial Intelligence: A Survey on Explainable Sentiment Analysis”.

[32] Ahmed Elbasiony, Ibrahim M. El-Hasnony, Samir Abdelrazek, (2024), “XAI-Based Sentiment Analysis Using Machine Learning Approaches”.

[33] Anitha R, Rajeev R R at al. (2024), “Enhancing Trust and Interpretability in Malayalam Sentiment Analysis with Explainable AI”.

[34] Satish, Naresh Kumar, (2024), “Forecasting Ethereum’s Price using ML and DL by Integrating Hybrid Sentiments in Multi-Source Market Data: Leveraging XAI”.

[35] Nassera Habbat, Hicham Nouri et al. (2023), “Sentiment analysis of imbalanced datasets using BERT and ensemble stacking for deep learning”.

[36] Roja D, Gujjula Navya, Biyyam Sai Srujana at al. (2025), “Deep Learning for Hotel Reviews: A Framework forSentiment Classification and Fake Review Detection”.