RoBERTa-Based Sentiment Analysis for Predicting Hotel Review Scores

Bardia Dehbasti
Dept. of Electrical and Computer Engineering
Toronto Metropolitan University
Toronto, Canada
bardia.dehbasti@torontomu.ca

***Abstract*—This project explores the application of transformer models, specifically the RoBERTa model, for sentiment analysis to predict review scores for hotel reviews. By leveraging transformer-based models, the study aims to enhance the accuracy of sentiment classification for binary prediction (positive and negative sentiments). For this project, the hotel review data is preprocessed to remove noise and standardize the text. RoBERTa is fine-tuned on this dataset to capture the nuanced sentiments expressed in the reviews with the aim to perform binary classification. Additionally, RoBERTa's performance is compared with traditional sentiment analysis models, including the NLTK Naive Bayes classifier, a Long Short-Term Memory (LSTM) neural network, and a simple Artificial Neural Network (ANN).**

Experimental results indicate that the LSTM model outperforms the other models in overall performance, followed by the ANN model. RoBERTa shows high precision but requires further improvement to surpass traditional models like Naive Bayes in terms of recall and overall performance.

The goal is to assess the effectiveness of transformer-based models in a binary classification task and to compare their performance against older models. The findings highlight the potential of transformer models in improving sentiment analysis, offering deeper insights into customer satisfaction and areas for improvement.

***Index Terms*—sentiment analysis, RoBERTa, transformer models, NLP, NLTK, Naive Bayes, LSTM, ANN**

I. Introduction

Many applications in natural language processing (NLP) require the ability to accurately classify sentiments expressed in text. Sentiment analysis techniques are commonly used to determine whether a piece of text conveys a positive or negative sentiment. This project focuses on predicting binary sentiment (positive or negative) for hotel reviews.

Sentiment analysis, also known as opinion mining, is a computational study of people's opinions, sentiments, emotions, and attitudes expressed in written language. It involves various tasks such as identifying the polarity of the text (positive or negative), extracting subjective information, and determining the strength or intensity of the sentiment [11]–[14]. This field of study is crucial for understanding public opinion, improving customer service, and enhancing decision-making processes.

Traditional sentiment analysis techniques, such as those based on lexicons or basic machine learning models, often fall short in capturing the complex and nuanced sentiments present in customer reviews. Models like the Naive Bayes classifier from the NLTK library [15] have been widely used for sentiment classification tasks. However, these models may struggle with context-dependent sentiments and idiomatic expressions commonly found in hotel reviews. Additionally, they are often limited to binary classification, which can still be challenging for certain applications.

In recent years, transformer-based models have revolutionized the field of NLP by achieving state-of-the-art performance across various tasks. Among these, BERT (Bidirectional Encoder Representations from Transformers) [17] and its extension RoBERTa (Robustly optimized BERT approach) [18] have shown remarkable capabilities in understanding context and capturing semantic nuances. These models are pre-trained on a large corpus and fine-tuned for specific tasks, making them highly effective for sentiment analysis. Their architecture allows for better handling of context-dependent sentiments and complex language patterns, making them suitable for binary classification tasks.

This project aims to leverage the advanced capabilities of RoBERTa for predicting binary sentiment from hotel reviews. The study involves preprocessing a large dataset of hotel reviews to remove noise and standardize the text. RoBERTa is then fine-tuned on this dataset to accurately classify sentiments as positive or negative. Additionally, the performance of RoBERTa is compared with traditional sentiment analysis models, including the Naive Bayes classifier from the NLTK library, a Long Short-Term Memory (LSTM) neural network, and a simple Artificial Neural Network (ANN).

The rest of the project report is organized as follows: Section II is a literature review on related work on sentiment analysis techniques using BERT-based transformers. Section III describes the dataset and the preprocessing steps required to make a suitable input for the transformer model. Section IV details the problem statement. Finally, Section V describes the RoBERTa model and its architecture.

II. Literature Review

In a study by Ferdoshi et al., VADER and RoBERTa models were compared for sentiment analysis of learner reviews from MOOCs, showing that RoBERTa outperforms VADER in accuracy, precision, and F1-score, although VADER captures finer sentiment nuances within the reviews [10].

Rahmania et al. evaluated VADER and RoBERTa for customer sentiment analysis of Amazon Go store reviews, revealing that RoBERTa surpasses VADER in overall accuracy and F1-score for positive sentiment classifications [9].

Kumar et al. investigated sentiment analysis of product reviews using VADER and RoBERTa, finding that RoBERTa achieved higher accuracy (91%) compared to VADER, which provided useful insights into sentiment intensity and valence [8].

Joshy and Sundar compared BERT, DistilBERT, and RoBERTa for sentiment analysis on movie reviews and tweets, concluding that BERT outperformed the other models in both datasets due to its robust architecture and comprehensive training [7].

Prasanthi and colleagues utilized BERT and RoBERTa for sentiment analysis on social media, demonstrating that RoBERTa's enhanced training techniques provided better performance metrics and emphasized its potential for handling large datasets efficiently [6].

Wang et al. improved sentiment classification accuracy on Chinese comments using RoBERTa with data augmentation techniques, achieving superior performance compared to traditional models like TextCNN, BiLSTM, and BiGRU [1].

Kumar et al. used the RoBERTa model for sentiment analysis of Twitter data related to the Russo-Ukrainian War, showing that RoBERTa provided more confident and accurate sentiment classifications compared to BERT [2].

Ferdoshi et al. analyzed Twitter sentiments using VADER and RoBERTa, finding that RoBERTa outperformed VADER in capturing contextual and semantic nuances, providing more reliable sentiment classification [3].

Tumuluru et al. proposed an ensemble model incorporating Transformer-XL, RoBERTa, and XGBoost for Twitter sentiment analysis, achieving higher accuracy and demonstrating the benefits of combining advanced NLP techniques and machine learning algorithms [4].

Abdal and colleagues introduced a robust method for Twitter sentiment analysis using RoBERTa, which surpassed other models like Decision Tree, SVM, and LSTM, achieving an accuracy of 96.78% [5].

III. Dataset Description

The dataset used in this project is a publicly available dataset from https://data.world/datafiniti/hotel-reviews. This dataset contains a comprehensive collection of hotel reviews, including various attributes such as review text, review title, review ratings, and more.

A. Loading the Data

The dataset is loaded into a Pandas DataFrame for easy manipulation and analysis.

B. Selecting the Necessary Columns

The columns `reviews.text` and `reviews.title` were selected to be the inputs to our model. These columns contain the actual review content and title, which will be merged to form the input to the sentiment analysis model.

C. Handling Missing Values

Any row with a null value in the `reviews.text` column is dropped from the dataset to ensure the quality of the input data.

D. Cleaning Reviews

The reviews are cleaned to remove nonsensical content and foreign language reviews:

1) Language Detection: The `langdetect` library is used to detect and keep only English reviews.

2) Removing Gibberish: Reviews that are nonsensical (e.g., containing only symbols) are removed.

3) Short Reviews: Reviews with fewer than four characters are removed.

E. Merging Columns

The `reviews.text` and `reviews.title` columns are merged into a single column. This combined text serves as the input for the sentiment analysis model.

F. Setting the Target Variable

The `reviews.rating` column is used as the target variable for the classification task, representing the review scores to be predicted, which are ratings ranging from 1 through 5. For this project, the ratings are converted to binary labels (positive: rating ≥ 3, negative: rating < 3).

G. Tokenization and Embeddings

For the NLTK Naive Bayes classifier, tokenization and feature extraction are crucial preprocessing steps. Tokenization involves splitting the text into individual words or tokens. The NLTK library provides various tokenizers, and a dictionary-based feature extraction method is used to represent each token.

RoBERTa uses subword tokenization (Byte-Pair Encoding) to handle the input text. The text is tokenized into subword units, which are then converted into embeddings by the model. This process allows RoBERTa to effectively manage large vocabularies and capture semantic nuances in the text.

H. Example Data

Figure 1 shows the first five rows of the dataset after preprocessing, including the combined review text and the target review rating.

The preprocessing steps ensure that the dataset is clean, standardized, and suitable for input to the sentiment analysis models. By focusing on the review text and titles, the models can better understand the context and sentiment expressed in each review, leading to more accurate sentiment classification.

IV. Model Description

The RoBERTa model, short for "Robustly optimized BERT approach," is an advanced transformer-based model designed for a wide range of natural language understanding tasks, including sentiment analysis. RoBERTa is an extension of the BERT (Bidirectional Encoder Representations from Transformers) model, which is known for its bidirectional training of Transformer models.

Combined Review Text	Review Rating
Pleasant 10 min walk along the sea front to the Water Bus. restaurants etc. Hotel was comfortable breakfast was good - quite a variety. Room aircon didn’t work very well. Take mosquito repelant!	1
It was a bit far from city center	0
Great hotel, staff where amazing and the location and views make this a great Base to visit Venice!	1
Perfect, best hotel over others in the area	1
Excellent Staff	1

Fig. 1. First five rows of the preprocessed dataset.

A. BERT Model Overview

BERT is a transformer-based model that relies on a stack of encoders to build contextualized word representations. Each encoder consists of multi-head self-attention mechanisms and feed-forward neural networks. BERT’s key innovation is its bidirectional approach, which allows it to capture context from both the left and right of each token in a sentence [17].

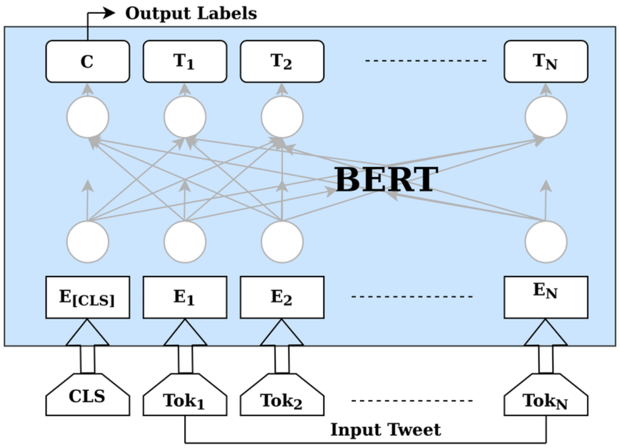

Fig. 2. Structure of the BERT Model for Sentiment Analysis [19]

Figure 2 shows the structure of the BERT model for sentiment analysis. The input embeddings consist of token embeddings, segment embeddings, and positional embeddings. These embeddings are fed into multiple layers of encoders, each containing self-attention and feed-forward layers. The output layer is used for classification tasks, such as sentiment analysis.

B. RoBERTa Model Overview

RoBERTa builds on BERT by incorporating several modifications and optimizations, including:

- Training with larger mini-batches and learning rates.
- Removing the next sentence prediction (NSP) objective.
- Training on a larger dataset with more steps.
- Dynamically changing the masking pattern during pre-training.

These improvements allow RoBERTa to achieve better performance across various NLP tasks, including sentiment analysis [18].

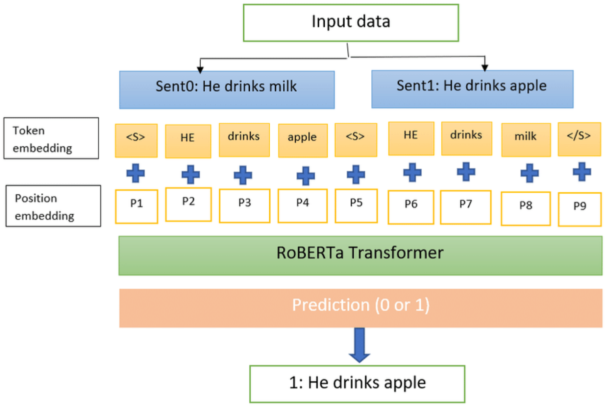

Fig. 3. Structure of the RoBERTa Model for Sentiment Analysis [20]

Figure 3 shows the structure of the RoBERTa model for sentiment analysis. The input embeddings are similar to BERT but optimized for better performance. The model consists of multiple layers of encoders with self-attention and feed-forward layers. The output layer is used for multi-class classification tasks, such as predicting review scores.

C. Fine-Tuning for Sentiment Analysis

To adapt RoBERTa for sentiment analysis of hotel reviews, the model undergoes a fine-tuning process. This involves the following steps:

1) **Preprocessing:** The hotel review text is cleaned by removing non-English reviews, gibberish, and very short reviews. The cleaned text is then tokenized using RoBERTa’s subword tokenization (Byte-Pair Encoding). The tokenized text is converted into embeddings that serve as the model’s input.
2) **Model Architecture:** RoBERTa’s architecture includes multiple layers of encoders, each with self-attention mechanisms that allow the model to capture contextual information. The final layer produces a representation for each token, which is aggregated to form a single representation for the entire review.
3) **Classification Layer:** A fully connected layer is added on top of RoBERTa to classify the sentiment of each review. The model is fine-tuned using labeled hotel review

data to optimize the classification layer for predicting binary sentiment (positive or negative).

4) **Training:** The model is trained using a cross-entropy loss function, with optimization performed using the Adam optimizer. The training process involves adjusting the model's weights to minimize the loss and improve classification accuracy.

The fine-tuning process allows RoBERTa to learn the specific nuances of hotel review sentiments, leading to improved performance in predicting review scores compared to traditional sentiment analysis models.

V. Results

In this section, we compare the performance of different sentiment analysis models applied to hotel reviews. The models compared include the Naive Bayes classifier from NLTK, a Long Short-Term Memory (LSTM) neural network, a simple Artificial Neural Network (ANN), and the RoBERTa model.

Table I summarizes the accuracy, precision, recall, and F1-score for each model.

TABLE I
Performance Comparison of Sentiment Analysis Models

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Naive Bayes (NLTK)	0.79	0.96	0.77	0.85
LSTM	0.88	0.93	0.92	0.93
ANN	0.86	0.90	0.93	0.92
RoBERTa	0.83	0.94	0.83	0.88

A. Discussion

The results indicate that different models have their own strengths and weaknesses:

- **Naive Bayes (NLTK):** This model achieved an accuracy of 79%, with a very high precision of 96% but lower recall at 77%, resulting in an F1-score of 85%. This suggests that while the Naive Bayes model is good at identifying positive sentiment (high precision), it misses more positive cases (lower recall).
- **LSTM:** The LSTM model outperforms the other models in accuracy (88%), precision (93%), recall (92%), and F1-score (93%). This indicates that LSTM is highly effective in both identifying positive sentiment and minimizing false positives, making it the best overall performer in this comparison.
- **ANN:** The ANN model also performs well with an accuracy of 86%, precision of 90%, recall of 93%, and an F1-score of 92%. This model shows a balanced performance, slightly lower than LSTM but still very effective.
- **RoBERTa:** The RoBERTa model achieved an accuracy of 83%, precision of 94%, recall of 83%, and an F1-score of 88%. While RoBERTa has a high precision, its recall is comparable to the Naive Bayes model, indicating it is very precise in its predictions but misses a significant number of positive cases.

These results highlight the advanced capabilities of LSTM in handling sequential data and capturing long-term dependencies, leading to superior performance in sentiment analysis tasks. While RoBERTa shows high precision, its recall needs improvement, suggesting it is highly reliable in predicting positive sentiments but may require more data or fine-tuning to capture a broader range of sentiments.

Although RoBERTa demonstrates strong potential, its current performance indicates that it still requires further improvement to be more useful compared to the traditional Naive Bayes classifier. Enhancements in training data, model tuning, or hybrid approaches might help in leveraging RoBERTa's advanced capabilities more effectively.

The superior performance of the LSTM model across all metrics suggests that it is particularly useful for applications requiring high accuracy and balanced precision and recall, such as customer feedback analysis and automated review moderation. RoBERTa's high precision makes it valuable in scenarios where minimizing false positives is crucial. However, further optimization might enhance its overall performance and recall.

The comparison underscores the importance of selecting the appropriate model based on the specific requirements of the sentiment analysis task, leveraging the strengths of each model to achieve the best results.

VI. Conclusion

This project explores the application of various sentiment analysis models to predict the sentiment of hotel reviews. We compared the performance of the Naive Bayes classifier from NLTK, a Long Short-Term Memory (LSTM) neural network, a simple Artificial Neural Network (ANN), and the RoBERTa model.

The results indicate that the LSTM model outperforms the other models in terms of accuracy, precision, recall, and F1-score. This suggests that LSTM is highly effective in handling sequential data and capturing long-term dependencies, making it the best overall performer in this comparison. The ANN model also performs well, demonstrating a balanced performance across all metrics.

The RoBERTa model, while showing high precision, has a lower recall compared to LSTM and ANN, indicating that it misses a significant number of positive cases. Although RoBERTa has strong potential due to its advanced transformer-based architecture, it requires further improvement to be more useful compared to the traditional Naive Bayes classifier. Enhancements in training data, model tuning, or hybrid approaches might help in leveraging RoBERTa's advanced capabilities more effectively.

The Naive Bayes classifier, although having high precision, falls short in recall and overall accuracy. This highlights the limitations of traditional models in capturing the complex and nuanced sentiments expressed in customer reviews.

In conclusion, the LSTM model stands out as the most effective for binary sentiment classification of hotel reviews. However, RoBERTa's high precision and potential for improvement suggest that with further optimization, it could become a valuable tool for sentiment analysis tasks. The comparison underscores the importance of selecting the appropriate model based on the specific requirements of the sentiment analysis task, leveraging the strengths of each model to achieve the best results.

Future work could focus on improving RoBERTa's recall and overall performance through enhanced training techniques and exploring hybrid models that combine the strengths of different approaches.

REFERENCES

[1] X. Wang, S. Xue, J. Liu, J. Zhang, J. Wang, J. Zhou, ”Sentiment Classification Based on RoBERTa and Data Augmentation,” Proceedings of CCIS2023.

[2] N. P. Kumar, et al., ”Sentiment Analysis of Russo-Ukrainian War using Twitter Text Corpus,” Department of Information Science and Engineering, R.V Institute of Technology and Management, Bangalore, India, 2023.

[3] J. Ferdoshi, et al., ”Unveiling Twitter Sentiments: Analyzing Emotions and Opinions through Sentiment Analysis on Twitter Dataset,” Department of Computer Science and Engineering, School of Data and Sciences, Brac University, Dhaka, Bangladesh, 2023.

[4] P. Tumuluru, et al., ”Advancing Twitter Sentiment Analysis: An Ensemble Approach with Transformer-XL, RoBERTa, and XGBoost,” Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India, 2023.

[5] M. N. Abdal, et al., ”A Transformer-Based Model for Twitter Sentiment Analysis using RoBERTa,” Computer Science and Engineering Discipline, Khulna University, Khulna, Bangladesh, 2023.

[6] K. N. Prasanthi, R. E. Madhavi, D. N. S. Sabarinadh, B. Sravani, ”A Novel Approach for Sentiment Analysis on Social Media using BERT and RoBERTa Transformer-Based Models,” Department of Computer Science and Engineering, Lakireddy Bali Reddy College of Engineering (Autonomous), Mylavaram, India, 2023.

[7] A. Joshy, S. Sundar, ”Analyzing the Performance of Sentiment Analysis using BERT, DistilBERT, and RoBERTa,” Centre for Artificial Intelligence, TKM College of Engineering, Kollam, Kerala, India, 2023.

[8] B. Kumar, A. D'Souza Jacintha, S. Sunil, V. S. Badiger, ”Sentiment Analysis for Products Review based on NLP using Lexicon-Based Approach and RoBERTa,” Department of Computer Science, Presidency College, Bengaluru, India, 2024.

[9] A. H. Azeemi, A. Waheed, ”Performance Analysis of VADER and RoBERTa Methods for Smart Retail Customer Sentiment on Amazon Go Store,” Information Technology University, Lahore, Pakistan, 2024.

[10] J. Ferdoshi, et al., ”Unveiling Sentiments: Analyzing Learners' Experience Using VADER and RoBERTa Models,” Department of Computer Science and Engineering, School of Data and Sciences, Brac University, Dhaka, Bangladesh, 2023.

[11] ”What Is Sentiment Analysis?” IBM, www.ibm.com.

[12] ”Sentiment analysis,” Wikipedia, en.wikipedia.org.

[13] ”8 Applications of Sentiment Analysis,” MonkeyLearn, www.monkeylearn.com.

[14] H.D. Sharma and P. Goyal, ”An Analysis of Sentiment: Methods, Applications, and Challenges,” Eng. Proc., 2023, 59(1), 68.

[15] S. Bird, E. Loper, and E. Klein, ”Natural Language Processing with Python,” O'Reilly Media, Inc., 2009.

[16] C.J. Hutto and E. Gilbert, ”VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text,” in Proc. 8th Int. AAAI Conf. Weblogs Soc. Media (ICWSM), Ann Arbor, MI, 2014, pp. 216-225.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. 2019 Conf. North American Chapter Assoc. Comput. Linguistics: Human Lang. Technol., Minneapolis, MN, USA, Jun. 2019, pp. 4171-4186.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, ”RoBERTa: A Robustly Optimized BERT Pretraining Approach,” in Proc. 2019 Conf. Empirical Methods Natural Lang. Process. (EMNLP), Hong Kong, China, Nov. 2019, pp. 2455-2465.

[19] ”BERT model architecture,” ResearchGate. [Online]. Available: https://www.researchgate.net/publication/348214408/figure/fig2/AS:976491001159681@
model-architecture.ppm

[20] ”RoBERTa model architecture,” ResearchGate. [Online]. Available: https://www.researchgate.net/profile/Heba-Aljarrah-2/publication/348754985/figure/fig1/AS:983989435265025@1611612771423/The-architecture-of-our-model.ppm