# Comprehensive Yelp Review Sentiment Analysis Report

## Executive Summary

This report presents a complete end-to-end Natural Language Processing (NLP) pipeline for multi-class sentiment analysis using the Yelp Review Full dataset. The project demonstrates the application of Transformer-based deep learning models to classify restaurant reviews into five sentiment categories. Despite computational constraints, the fine-tuned RoBERTa-based model achieved an overall accuracy of 62.2% and a weighted F1-score of 0.623, highlighting the effectiveness of modern NLP techniques.

## 1. Introduction

Sentiment analysis is a fundamental task in NLP that focuses on identifying opinions and emotions expressed in textual data. Online platforms such as Yelp generate large volumes of customer reviews, making automated sentiment analysis crucial for understanding user satisfaction and business performance.

This project addresses a challenging five-class sentiment classification problem. Unlike binary sentiment analysis, fine-grained sentiment classification requires capturing subtle linguistic differences between adjacent sentiment levels, such as neutral versus moderately positive reviews.

## 2. Dataset Description

The Yelp Review Full dataset is a widely used benchmark dataset containing approximately 650,000 training reviews and 50,000 test reviews. Each review is labeled with a star rating from 1 (very negative) to 5 (very positive). The dataset reflects real-world challenges, including informal language, emojis, and varying review lengths.

## 3. Text Preprocessing

A carefully designed preprocessing pipeline was applied to reduce noise while preserving semantic information. The steps included lowercasing, removal of HTML tags and URLs, emoji normalization, punctuation normalization, and whitespace cleanup.

Aggressive preprocessing techniques such as stop-word removal and stemming were intentionally avoided, as Transformer models are capable of learning contextual representations directly from raw text.

## 4. Model Selection and Training Strategy

The RoBERTa-base Transformer model was selected due to its proven performance in sentiment analysis tasks. RoBERTa improves upon BERT by leveraging optimized pretraining strategies and larger datasets.

Due to CPU-only computational constraints, the model was fine-tuned on a representative subset of the training data. This subsampling approach significantly reduced training time while still allowing the model to learn meaningful sentiment patterns.

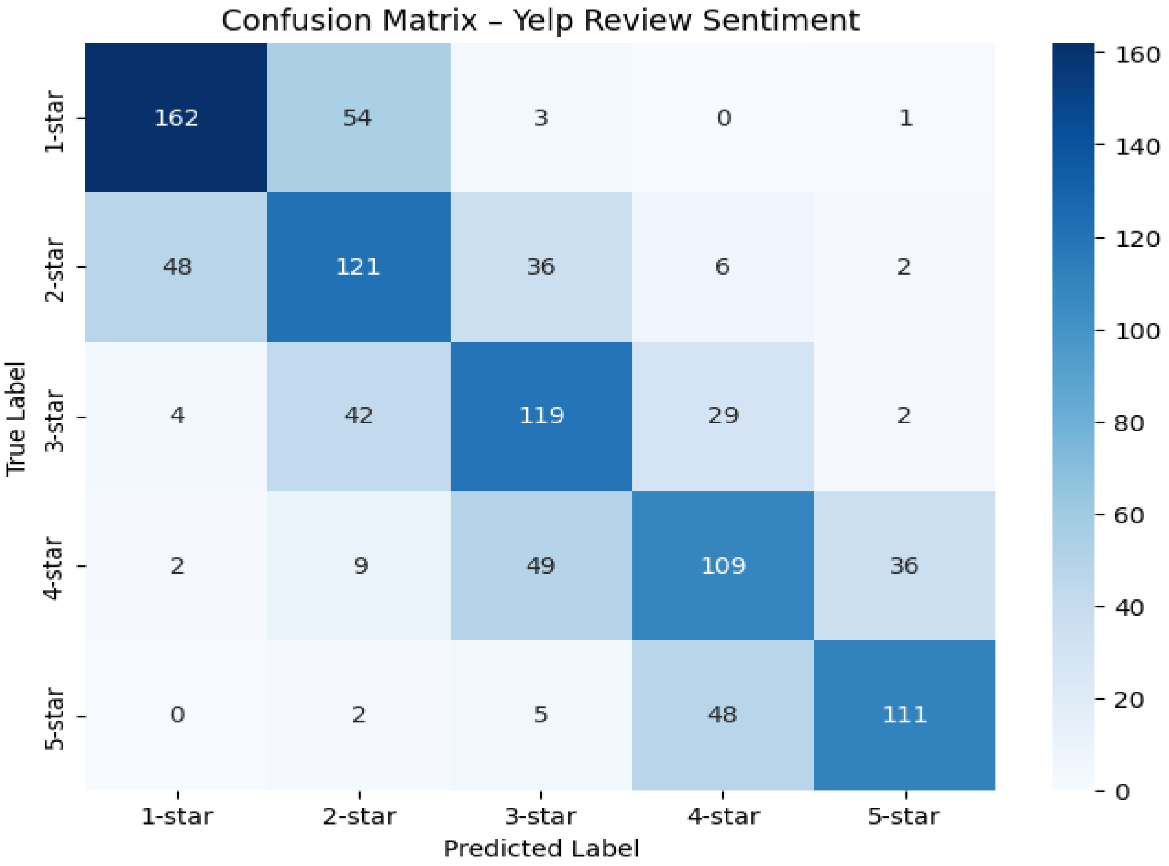
## 5. Evaluation Results and Analysis

Model evaluation was conducted on a randomly selected subset of 1,000 test reviews. Performance was measured using precision, recall, F1-score, accuracy, and a confusion matrix. The model achieved an overall accuracy of 62.2% and a weighted F1-score of 0.623.

| Sentiment Class | Star Rating | F1-score |
|---|---|---|
| Very Negative | 1-star | 0.74 |
| Negative | 2-star | 0.55 |
| Neutral | 3-star | 0.58 |
| Positive | 4-star | 0.55 |
| Very Positive | 5-star | 0.70 |

### Confusion Matrix Analysis

The confusion matrix illustrates that the model performs best on extreme sentiment classes, particularly 1-star and 5-star reviews. Most misclassifications occur between adjacent sentiment levels, such as 3-star and 4-star reviews, which reflects the subjective and overlapping nature of moderate sentiment expressions.



## 6. Conclusion and Future Work

This project successfully demonstrates a complete NLP workflow for multi-class sentiment analysis using modern Transformer-based models. Despite hardware limitations, the fine-tuned RoBERTa model achieved meaningful performance and provided valuable insights into sentiment classification on real-world review data.

Future work may include training on the full dataset using GPU resources, applying data augmentation techniques, and exploring ensemble models to further improve performance on intermediate sentiment classes.