Sentiment Analysis on Yelp Restaurant Reviews Using Transformer-Based Models

MSc Computer Science and Data Science
Natural Language Processing
(MCSCIN5A0625)

Submitted by:
Jayasri DHANAPAL
Divya JAYAPRAKASH
Reshma KARTHIKEYAN NAIR

Abstract

Sentiment analysis is a key application of Natural Language Processing (NLP) that aims to identify opinions and emotions expressed in textual data. In this project, we address the problem of **multi-class sentiment classification** on the Yelp Restaurant Review dataset, where reviews are labeled on a 1-star to 5-star scale. A transformer-based language model, **RoBERTa-base**, is fine-tuned using the **entire Yelp Review Full dataset consisting of 650,000 training samples**. To ensure feasibility under limited computational resources (Google Colab), multiple optimization strategies such as mixed-precision training, gradient accumulation, gradient checkpointing, and parallelized preprocessing were applied. The proposed model is evaluated against a strong baseline model, *AdhamEhab/fine-tuned-bert-yelp*. Experimental results demonstrate that the fine-tuned RoBERTa model significantly outperforms the baseline in terms of accuracy and weighted F1-score, validating the effectiveness of transformer-based approaches for large-scale sentiment analysis.

1. Introduction

Online reviews play a critical role in shaping consumer decisions, particularly in the restaurant industry. Manual analysis of large-scale review data is infeasible, making automated sentiment analysis an essential tool for extracting actionable insights. Traditional machine learning approaches often fail to capture contextual meaning, sarcasm, and long-range dependencies present in natural language.
Transformer-based architectures, such as BERT and RoBERTa, have achieved state-of-the-art performance on a wide range of NLP tasks by leveraging self-attention mechanisms and large-scale pretraining. This project explores the application of **RoBERTa-base** for sentiment classification and evaluates its effectiveness against a well-established BERT-based baseline model.

2. Problem Definition

- **Task Type**: Multi-class text classification
- **Input**: Yelp restaurant review text
- **Output**: One of five sentiment classes (1–5 stars)

The objective is to correctly predict the sentiment rating associated with each review and to outperform an existing baseline model.

3. Dataset Description

The **Yelp Review Full Dataset** is used in this project.
Dataset Characteristics:

- **Source**: Hugging Face Datasets
- **Training samples**: 650,000
- **Test samples**: 50,000
- **Labels**:
    - 1 star (Very Negative)
    - 2 stars (Negative)
    - 3 stars (Neutral)
    - 4 stars (Positive)
    - 5 stars (Very Positive)

The dataset is balanced across classes in the test set, ensuring fair evaluation.

4. Data Preprocessing

Text preprocessing was performed to standardize and clean the input data:

- Conversion to lowercase
- Removal of URLs
- Normalization of whitespace
- Trimming of leading and trailing spaces

To efficiently process the large dataset, preprocessing was parallelized using multiple CPU processes. This significantly reduced preprocessing time while maintaining data integrity.

5. Tokenization

The **RoBERTa tokenizer** was employed to convert text into token IDs.
Tokenization Strategy:

- Maximum sequence length: 128 tokens
- Truncation applied for longer reviews

This sequence length was chosen based on empirical observation that most Yelp reviews fall within this limit, while also reducing the quadratic computational cost of self-attention.

6. Model Architecture

6.1 RoBERTa-base

- Transformer-based encoder model
- Approximately 125 million parameters
- Pre-trained on large-scale English corpora
- Classification head added for 5 sentiment classes

6.2 Optimization Techniques

To enable training on the **full dataset** under Colab constraints, the following optimizations were applied:

- **Mixed-precision (FP16) training**
- **Gradient accumulation** to simulate larger batch sizes
- **Gradient checkpointing** to reduce GPU memory usage
- **Parallel data preprocessing and tokenization**
- Reduced logging and checkpoint overhead

These techniques allowed large-scale training without sacrificing model performance.

7. Training Configuration

Parameter	Value
Model	RoBERTa-base
Learning Rate	2e-5
Batch Size (per device)	32
Gradient Accumulation Steps	2
Effective Batch Size	64
Epochs	2
Weight Decay	0.01
Precision	FP16
Device	Tesla T4 (Google Colab)

8. Training Time Analysis

Training was performed on Google Colab with a Tesla T4 GPU. Due to Colab session limits, the training process was interrupted before completing all planned steps. The best-performing checkpoint was automatically selected based on weighted F1-score.
Training Runtime Summary:

- **Total training time**: 13,649 seconds (3 hours 47 minutes)
- **Planned steps**: 20,000
- **Stopped at step**: 14,000
- **Best checkpoint**: Step 12,000
- **Precision**: FP16

Early termination was caused by platform time constraints rather than model convergence issues.

9. Evaluation Metrics

The following metrics were used to evaluate model performance:

- Accuracy

- Precision (weighted)
- Recall (weighted)
- F1-score (weighted)
- Confusion matrix

Weighted metrics were chosen to account for class distributions.

10. Experimental Results

10.1 RoBERTa Model Performance

- **Accuracy**: 0.6701
- **Weighted F1-score**: 0.6692

Class-wise Results:

Class	Precision	Recall	F1-score
1 star	0.78	0.79	0.78
2 stars	0.61	0.62	0.62
3 stars	0.63	0.58	0.60
4 stars	0.59	0.58	0.59
5 stars	0.74	0.77	0.75

The model performs best on strongly polarized sentiments (1-star and 5-star), while neutral reviews (3-star) remain more challenging due to inherent ambiguity.

10.2 Baseline Model Performance (BERT)

- **Accuracy**: 0.6123
- **Weighted F1-score**: 0.6153

10.3 Model Comparison

Model	Weighted F1-score
Baseline BERT	0.615
RoBERTa (Proposed)	**0.669**

The RoBERTa model achieves a **~5.4% absolute improvement** in weighted F1-score over the baseline.

11. Discussion

The experimental results confirm that RoBERTa's robust pretraining and optimized fine-tuning strategy significantly improve sentiment classification performance. The use of the full dataset strengthens the reliability and generalizability of the model. Runtime optimizations ensured feasibility without compromising scientific rigor. Early stopping due to Colab limits does not invalidate results, as the best checkpoint was selected based on evaluation metrics.

12. Conclusion

This project demonstrates the successful application of transformer-based models for large-scale sentiment analysis. By fine-tuning RoBERTa on the full Yelp Review dataset and employing efficient training strategies, superior performance over a strong baseline was achieved. The results highlight the effectiveness of modern NLP models in real-world opinion mining tasks.

13. Future Work

- Training with RoBERTa-large or DeBERTa
- Addressing neutral sentiment ambiguity using ordinal regression
- Hyperparameter tuning for further performance gains
- Deployment as a real-time sentiment analysis service