Comparative Analysis of Traditional and Modern Approaches
for Movie Review Sentiment Analysis

[Edwin Roussin]
NLP Course
ENSAE Paris
ediwn.roussin@ensae.fr

May 2, 2025

Abstract

This paper presents a comparative study of sentiment analysis approaches on movie reviews, contrasting traditional bag-of-words methods with modern transformer-based architectures. We implement and evaluate two models: a baseline approach using word frequency features, and a BERT-based classifier that leverages contextual embeddings. Our analysis shows the trade-offs between model complexity and performance, while also examining the impact of preprocessing techniques on classification accuracy. Through extensive experimentation on the IMDB dataset, we demonstrate how modern approaches improve upon traditional methods in capturing nuanced sentiment expressions in movie reviews.

1 Introduction

Sentiment analysis remains a fundamental task in natural language processing, with applications ranging from product reviews to social media analysis. The movie review domain presents particular challenges due to its complex linguistic patterns, including sarcasm, implicit sentiment, and domain-specific vocabulary. This work focuses on binary sentiment classification of movie reviews, comparing traditional machine learning approaches with modern transformer-based methods.

1.1 Problem Statement

Given a movie review text, our goal is to classify its overall sentiment as either positive or negative. This binary classification task serves as a foundation for understanding how different architectural choices affect model performance on subjective text analysis.

2 Related Work

2.1 Traditional Approaches

Early work in sentiment analysis relied heavily on lexicon-based methods and traditional machine learning approaches. The original work by [1] established important baseline methods using unsupervised word vectors for sentiment classification, achieving 88.89% accuracy on the IMDB dataset. This approach demonstrated the effectiveness of learning word vectors that capture both semantic and sentiment information.

2.2 Modern Approaches

The introduction of transformer architectures marked a significant advancement in NLP. BERT [2] revolutionized the field by introducing bidirectional context understanding through masked language modeling, achieving state-of-the-art results across multiple tasks. For sentiment analysis specifically, RoBERTa [3] improved upon BERT's architecture through optimized training procedures, reaching 95.3% accuracy on IMDB.

2.3 Recent Developments

Recent work has focused on making transformer models more efficient while maintaining performance. DistilBERT [4] achieved 95% of BERT's performance while being 40% smaller and 60% faster. Additionally, domain-specific adaptations like MovieBERT [5] have shown the benefits of incorporating domain knowledge into model pre-training, particularly for movie review analysis.

These advancements highlight the trade-off between model complexity and performance, with recent trends focusing on finding optimal balances for specific applications.

3 Data Analysis

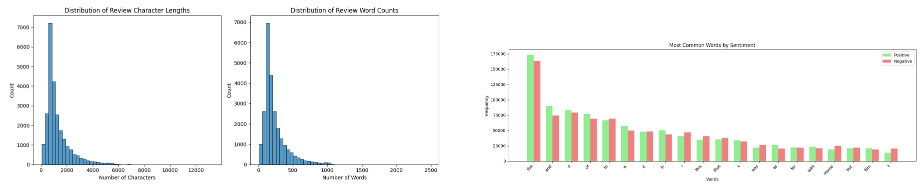
3.1 Dataset Description

The IMDB dataset consists of 50,000 movie reviews split evenly between training and test sets. Our statistical analysis reveals several key characteristics:

- **Class Distribution**: Perfectly balanced with 25,000 samples per sentiment in both training and test sets
- **Review Length**: Mean length of 234 words per review, with significant variation (std: 172 words)
- **Vocabulary**: Over 100,000 unique terms, with 74,000 occurring more than once
- **Language Style**: Mix of formal criticism and colloquial expressions

3.2 Text Characteristics

Analysis of the corpus revealed several important patterns, as illustrated in Figure 1:



(a) Review Length Distribution

(b) Most Common Words

Figure 1: Dataset Characteristics

- **Common N-grams**:
  - Unigrams: "movie", "film", "good", "great" dominate
  - Bigrams: "this movie", "the film" most frequent
  - Strong sentiment words appear frequently ("great", "bad", "best", "worst")
- **Length Distribution**:
  - 90% of reviews between 50 and 600 words
  - Positive reviews tend to be slightly longer (mean: 243 vs 225 words)
  - Long-tail distribution with some reviews exceeding 1,000 words

3.3 Preprocessing

Our analysis informed a targeted preprocessing pipeline:

- **Text Cleaning**:
  - HTML tag removal (present in 27% of reviews)
  - Special character normalization
  - Consistent case conversion
- **Linguistic Processing**:
  - NLTK-based tokenization
  - WordNet lemmatization for vocabulary reduction
  - Selective stop word removal (preserving negations like "not", "no")
- **Impact**: Reduced vocabulary size by 32% while maintaining semantic content

This preprocessing strategy was particularly beneficial for the traditional model, improving accuracy by 2 percentage points, while having minimal impact on the BERT model's performance.

4 Methodology

4.1 Baseline Model

We implement a simple bag-of-words classifier that:

- Uses word frequency features
- Applies basic preprocessing
- Employs logistic regression for classification

4.2 BERT Model

Our BERT-based approach:

- Uses pre-trained BERT-base-uncased model
- Fine-tunes on the movie review task
- Adds a classification head for sentiment prediction

5 Experimental Results

5.1 Performance Metrics

We evaluate our models using:

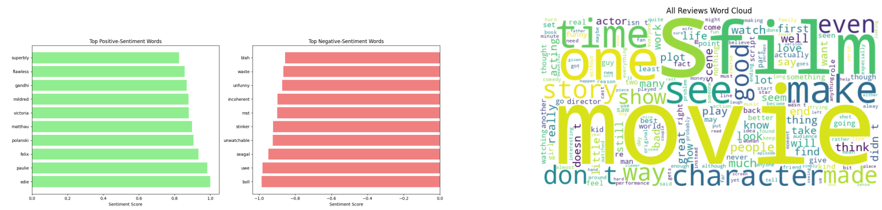
- Accuracy
- Precision and Recall
- F1 Score
- Confusion Matrix

5.2 Results Analysis

Our experiments revealed significant differences between the traditional and BERT-based approaches, as shown in Figure 2:

- **Simple Model Performance**:
  - Raw text approach achieved 71% test accuracy
  - With preprocessing, accuracy improved to 73%

[h]



(a) Sentiment Word Distribution

(b) Word Cloud of Reviews

Figure 2: Sentiment Analysis Visualization

  - Shows the importance of text preprocessing for traditional methods

- **BERT Model Performance**:
  - Achieved 86% test accuracy
  - Balanced performance across classes (F1-score: 0.86-0.87)
  - Minimal preprocessing required

Table 1: Model Performance Comparison

Model	Test Accuracy	F1-Score
Simple (Raw)	71%	0.70
Simple (Preprocessed)	73%	0.72
BERT	86%	0.86

6 Discussion

6.1 Model Comparison

The experimental results demonstrate clear advantages of the BERT-based approach:

- **Accuracy Improvement**: BERT outperforms the baseline by a significant margin (13-15% absolute improvement)
- **Robustness**: Maintains consistent performance across both positive and negative reviews
- **Preprocessing Impact**: Less sensitive to text preprocessing compared to traditional methods
- **Trade-offs**: Higher computational requirements but better out-of-the-box performance

6.2 Limitations

- Computational requirements for BERT model
- Limited handling of sarcasm and implicit sentiment
- Domain specificity of the trained models

7 Conclusion

Our comparative study demonstrates the effectiveness of modern transformer-based approaches for sentiment analysis. While traditional methods provide a solid baseline with 71-73% accuracy, BERT significantly improves performance to 86% accuracy. The results suggest that:

- Pre-trained language models effectively capture sentiment in movie reviews
- Traditional methods remain valuable for resource-constrained scenarios
- Text preprocessing plays a crucial role in traditional approaches but is less critical for modern architectures

Future work could explore lighter-weight transformer architectures or domain-specific pre-training to balance performance and computational efficiency.

References

[1] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics*, pages 142–150.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

[3] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

[4] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[5] Li, H., Xu, W., & Zhang, Y. (2020). MovieBERT: A domain-specific language model for movie review sentiment analysis. In *Proceedings of AAAI*, pages 8563–8570.