**Quantifying Sentiment: A Comparative Evaluation of Language Model Performance on the Rotten Tomatoes Corpus**

**Winter 2025 CS 545 Final Project Report  
Alec Rogers, Luis Becerra, Mini Sengupta**

Contents

[Abstract 2](#_Toc191997121)

[Team members and roles 2](#_Toc191997122)

[1.0 Introduction & Research Motivation 3](#_Toc191997123)

[2.0 Literature review 3](#_Toc191997124)

[3.0 Methodology 3](#_Toc191997125)

[3.1 Dataset Evaluation 3](#_Toc191997126)

[3.2 Embedding (Alec) 3](#_Toc191997127)

[3.3 Flan (Luis) 3](#_Toc191997128)

[3.4 RoBERTa (Mini) 3](#_Toc191997129)

[3.5 Llama (Mini) 3](#_Toc191997130)

[4.0 Results and Discussion 3](#_Toc191997131)

[5.0 Conclusions and Future Work 4](#_Toc191997132)

[References 5](#_Toc191997133)

# Abstract

Sentiment analysis, the task of determining the emotional tone behind text, is a crucial application of Natural Language Processing (NLP). With the rapid advancement of Large Language Models (LLMs) and the closed nature of their design, there are many researchers attempting to develop assessments and frameworks for assessment (e.g. bias [1], x-metrics [2]) to characterize their performance. This project conducted a comparative analysis of LLMs (RoBERTa, Llama, Flan and a Custom) on the Rotten Tomatoes movie review dataset, a widely recognized resource for sentiment classification with a diverse range of expressions. By comparing their accuracy, F1-score, precision, or recall, we hope to gain insights into current LLM-driven sentiment analysis.

# Team members and roles

* Alec Rogers: Programming lead. Embedding baseline and a custom sentiment analysis engine owner.
* Luis Becerra: Flan algorithm owner and custom prompt engineering.
* Mini Sengupta: Bert and Llama algorithm owner.

# 1.0 Introduction & Research Motivation

To be filled in

# 2.0 Literature review

To be filled in

# 3.0 Methodology

To be filled in

## 3.1 Dataset Evaluation

To be filled in

## 3.2 Embedding (Alec)

To be filled in

## 3.3 Flan (Luis)

To be filled in

## 3.4 RoBERTa (Mini)

**To be filled**

## 3.5 Llama (Mini)

To be filled in

# 4.0 Results and Discussion

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | average score | total wins | win rate |
| MiniMax | 4.92 | 1354 | 58.49% |
| Monte Carlo | 4.53 | 1832 | 79.14% |
| Entropy Max | 4.54 | 2061 | 89.03% |
| Reinforcement Learning | 4.47 | 1699 | 73.39% |

**Table 1**. Average Score (average number of guesses to solve a game), total wins, and win rate for the 4 algorithms focused in this study. The total number of games played was 2315 and all algorithms used “SLATE” as the starting guess.

To be filled

# 5.0 Conclusions and Future Work

To be filled.

# References

|  |  |
| --- | --- |
| [1] | J. Echterhoff, Y. Liu, A. Alessa, J. McAuley and Z. He, "Cognitive Bias in Decision-Making with LLMs," 2024. [Online]. Available: https://arxiv.org/abs/2403.00811v3. |
| [2] | M. Kahng, I. Tenney, M. Pushkarna, M. X. Liu, J. Wexler, E. Reif, K. Kallarackal, M. Chang, M. Terry and L. Dixon, "LLM Comparator: Visual Analytics for Side-by-Side Evaluation of Large Language Models," 2024. [Online]. Available: https://arxiv.org/html/2402.10524v1. |
| [3] | J. Alammar and M. Grootendorst, Hands-On Large Language Models: Language Understanding and Generation, O'Reilly Media, 2024. |