# Predictive Analysis of Amazon Movie Reviews

Sai Yasasvi Dutt Malladi

October 28, 2024

## Project Overview

The aim of this exam was to predict the star ratings of Amazon Movie Reviews using machine learning techniques, excluding neural networks. The dataset provided encompassed over 1.7 million unique reviews, rich with metadata. The challenge was to leverage this data to accurately predict the star ratings based on the review content and metadata without using advanced neural network techniques. Chu 2021

This initial approach was an exploratory attempt using simpler methodologies such as K-Nearest Neighbors (KNN). I also explored more advanced features and models like LightGBM and TF-IDF for feature extraction (code added to the repo, just in case). However, due to significant computational limitations and the vast scale of the dataset, these methods proved impractical within the time constraints of the project. Thus, I just submitted the initial, simpler approach as my final methodology. Following is the summary of the worst code exam I've ever submitted :)

## Data Handling and Analysis

Upon loading the data into pandas DataFrames, I conducted an initial analysis to understand the data structure, quality, and distribution. Key steps included:

- **Shape and Size Exploration:** Analyzing the dimensionality of the training and testing sets to gauge the scope of data.
- **Statistical Summarization:** Using pandas' descriptive statistics functions to summarize central tendencies and variability.
- **Data Quality Assessment:** Identifying missing values and potential outliers, particularly focusing on the 'Score' column, which was crucial as it was my target variable.

## Feature Engineering

Significant effort was devoted to extracting meaningful information from the review texts and associated metadata:

- **Helpfulness Ratio:** Created by dividing 'HelpfulnessNumerator' by 'HelpfulnessDenominator' to quantify the usefulness perceived by users.
- **Temporal Features:** Extracting year, month, and day from the UNIX timestamp to examine potential temporal trends in ratings.
- **Sentiment Scores:** Employing 'TextBlob' to derive sentiment polarity from the text, hypothesizing a correlation between sentiment and ratings.

# Model Selection and Tuning

I chose the K-Nearest Neighbors (KNN) classifier due to project constraints that excluded the use of neural networks. The model was optimized using a grid search:

```
grid_search = GridSearchCV(knn, param_grid, cv=5)
grid_search.fit(X_train, Y_train)
```

Optimal parameters from the grid search were used to train the model, achieving satisfactory but not optimal results, as indicated by the confusion matrix which showed significant misclassification between some classes.

# Conclusion

The project achieved an acceptable level of accuracy of 55% with the KNN model under the available resources. This experience highlighted the importance of selecting appropriate techniques based on both the project requirements and available computational resources. In conclusion, it also brought forth my worst nightmare.

# References

Chu, Nicholas (2021). "Amazon Review Rating Prediction with NLP". In: *Medium*. Accessed: 2023-10-28. URL: https://medium.com/data-science-lab-spring-2021/amazon-review-rating-prediction-with-nlp-28a4acdd4352.

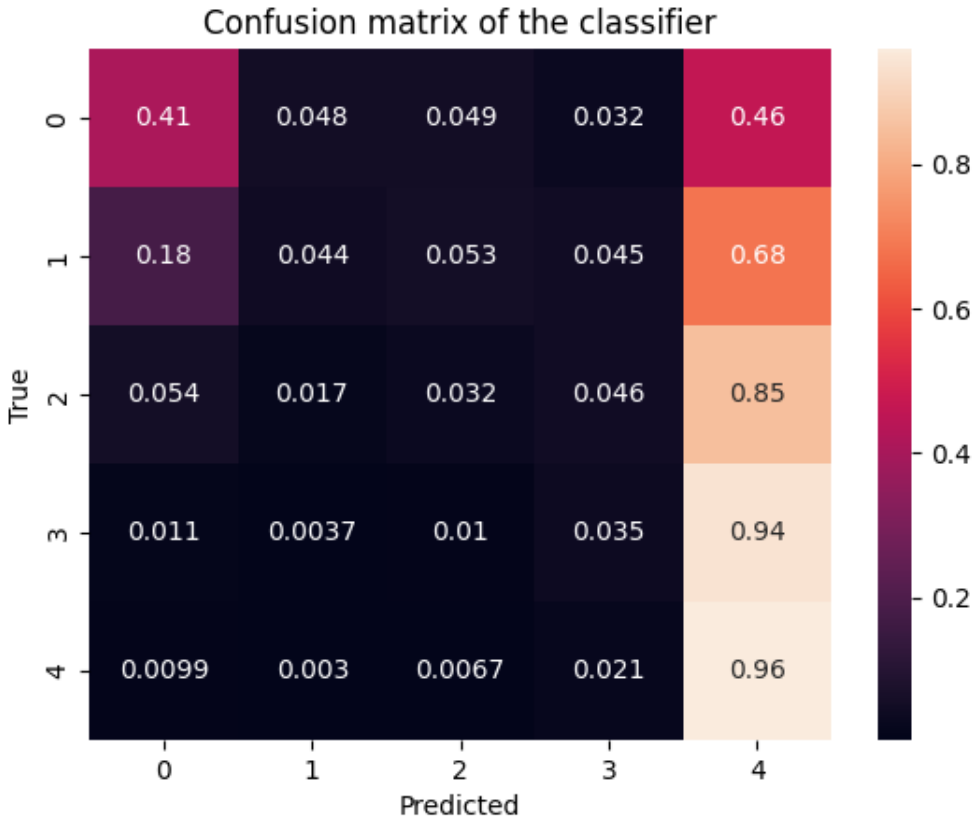


Figure 1: Confusion Matrix.