Draft FPR 😊

# Table of Contents

Contents

[Table of Contents 2](#_Toc183443849)

[1. Introduction 3](#_Toc183443850)

[2. Background and Literature Review 4](#_Toc183443851)

[2.1 Background and overview 4](#_Toc183443852)

[2.2 Literature review 4](#_Toc183443853)

[3. Data 5](#_Toc183443854)

[3.1 IMDB Dataset 5](#_Toc183443855)

[3.1.1 Overview 5](#_Toc183443856)

[3.1.2 EDA 6](#_Toc183443857)

[3.1.3 Pre-processing 6](#_Toc183443858)

[3.2 BookCorpus 6](#_Toc183443859)

[3.2.1 Overview 6](#_Toc183443860)

[3.2.2 EDA 6](#_Toc183443861)

[3.2.3 Pre-processing 6](#_Toc183443862)

[4. Methodology 7](#_Toc183443863)

[5. Results 8](#_Toc183443864)

[6. Analysis and discussion 9](#_Toc183443865)

[7. Conclusions 10](#_Toc183443866)

[8. References 11](#_Toc183443867)

[9. Appendices 12](#_Toc183443868)

# 1. Introduction

I will be writing the introduction later (Hopefully)

# 2. Background and Literature Review

## 2.1 Background and overview

## 2.2 Literature review

# 3. Data

## 3.1 IMDB Dataset

### 3.1.1 Overview

The IMDB Large Movie Review Dataset is a comprehensive collection of film reviews, having 50,000 highly polar reviews equally distributed between training and testing sets, with an additional 50,000 unlabelled reviews available for extended research purposes. Originally developed for binary sentiment classification research by (Maas et al., 2011) this dataset was chosen for the film review autocompletion task due to its extensive coverage, quality, and structured composition.

The dataset's construction followed a detailed process outlined in Maas et al.'s research. The authors gathered reviews from IMDB while implementing specific constraints to ensure data quality and balance. They imposed a limit of 30 reviews per movie to prevent any single film from dominating the dataset, resulting in a diverse collection spanning various genres, periods, and review styles as the demographic for each film genre are often specific. This careful curation process contributed to the dataset's robustness for natural language processing tasks other than sentiment analysis.

As for technical specifications, the dataset had considerable scale with downloaded files totalling 84.13 MB and a generated dataset size of 133.23 MB, combining to be 217.35 MB of total size. The data is stored in plain text format, with each instance containing two primary fields: the 'text' field storing the review content as a string, and the 'label' field indicating sentiment classification (0 for negative, 1 for positive). This straightforward structure facilitates efficient data handling and preprocessing for my task.

While the dataset was initially designed for sentiment analysis, its selection for my film review autocompletion task can be justified by a number of compelling factors. The substantial volume of reviews written by a diverse group of people, as opposed to a specific group such as professional film critics, provides rich linguistic patterns and domain-specific vocabulary essential for generating contextually appropriate completions. The consistent quality and natural language patterns present in the reviews offer valuable training material for language models focused on generating coherent and contextually relevant text continuations. Furthermore, the dataset's balanced nature and diverse movie coverage ensure that the trained models can generate completions across various film genres and review styles.

The primary limitation of this dataset lies in the its original intended purpose differing from my current autocompletion task. However, this constraint is effectively mitigated through appropriate pre-processing techniques and does not significantly impact the dataset's effectiveness for my research objectives. The dataset's accessibility through the Huggingface repository ensures reliable and standardized access to the research materials.

### 3.1.2 EDA

### 3.1.3 Pre-processing

## 3.2 BookCorpus

### 3.2.1 Overview

### 3.2.2 EDA

### 3.2.3 Pre-processing

# 4. Methodology

# 5. Results

# 6. Analysis and discussion

# 7. Conclusions

# 8. References

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y. and Potts, C. (2011) *Learning Word Vectors for Sentiment Analysis*, [online] Available at: https://aclanthology.org/P11-1015 (Accessed November 25, 2024).

# 9. Appendices