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.

The IMDB dataset Used here was ethically collected, adheres to the University of Hertfordshire’s ethical guidelines, and does not involve personal data, exempting it from GDPR and ethics committee approval. As specified by the original authors, the dataset is cited appropriately in accordance with its terms of use.

### 3.1.2 EDA

### 3.1.3 Pre-processing

The IMDB dataset underwent extensive preprocessing to prepare it for the autocompletion task. The initial dataset of 50,000 reviews was subjected to a comprehensive cleaning pipeline to ensure data quality and consistency. The cleaning process addressed various textual irregularities, including HTML tags, inconsistent formatting, special characters, and varied punctuation styles. The preprocessing was implemented in three distinct phases: primary cleaning for fundamental text normalization, standardization of textual elements, and final formatting refinements.

The primary cleaning phase focused on HTML tag removal, foreign character handling, case normalization to lowercase, and preservation of basic punctuation. This was followed by standardization, where common abbreviations were converted (e.g., "U.S." to "USA"), number formats were standardized, contractions were expanded (e.g., "it's" to "it is"), and possessives were normalized. The final phase addressed punctuation and spacing inconsistencies, including the removal of multiple periods and proper formatting of numbered lists.

After cleaning, the dataset underwent length-based filtering, retaining reviews between the 10th and 90th percentiles (23-58 tokens), resulting in 488,822 training and 475,551 testing sequences. To optimize the dataset for training efficiency while maintaining representativeness, stratified sampling was employed, reducing the dataset to approximately 41% of its original size. This sampling maintained a perfectly balanced distribution across three length categories: short (23-34 tokens), medium (35-46 tokens), and long (47-58 tokens). The final pre-processed dataset consisted of 244,410 training sequences and 237,774 testing sequences, with each length category representing exactly one-third of the data. For the autocompletion task, each review was split into input-target pairs, with the first 100 characters serving as input and the remaining text as the target sequence.

## 3.2 BookCorpus

### 3.2.1 Overview

The BookCorpus dataset, originally introduced by (Zhu et al., 2015), presents an interesting case of data curation challenges in large-scale text datasets. While initially reported to contain 11,038 books, subsequent analysis revealed a more complex structure containing 7,185 unique books in plain text format. The dataset, accessed through HuggingFace, comprises downloaded files of 1.18 GB which generate into 4.85 GB of data, requiring a total disk space of 6.03 GB. For this research project, a subset of approximately 1,000 books (roughly 14% of the unique texts) has been selected to accommodate computational constraints while maintaining sufficient training data for the from-scratch language model.

The original dataset, as described in (Zhu et al., 2015)’s work, implemented a quality control measure by including only books exceeding 20,000 words, thereby ensuring content richness and filtering out potentially lower-quality shorter works. The texts span 16 distinct genres, with significant representation in Romance, Fantasy, and Science fiction categories. This genre diversity, combined with the substantial word count requirement, contributes to the dataset's suitability for training a general-purpose language model.

The dataset's plain text format and absence of personal information make it particularly appropriate for research applications. While using a subset of the full corpus represents a practical compromise, the selected portion maintains sufficient linguistic diversity and complexity to serve as foundational training data for the custom language model before fine-tuning on the domain-specific IMDB dataset.

The BookCorpus was ethically collected from freely available online books, adheres to the University of Hertfordshire's ethical guidelines, and contains no personal data, exempting it from GDPR and ethics committee approval.

### 3.2.2 EDA

### 3.2.3 Pre-processing

# 4. Methodology

## 4.1 IMDB dataset pre-processing

The preprocessing methodology for the IMDB dataset implemented a comprehensive pipeline combining text normalization, sequence preparation, and optimized sampling techniques. The implementation focused on both data quality and computational efficiency, utilizing batch processing and GPU acceleration where applicable.

The preprocessing pipeline was implemented through a hierarchical function structure, beginning with atomic cleaning operations and progressing to more complex transformations. The primary cleaning function, clean\_text(), executed a series of regular expression-based transformations: HTML tag removal via regex patterns, ASCII encoding/decoding for foreign character handling, and standardized text normalization. Contraction handling was implemented through a dictionary-based mapping system, expanding forms like "ain't" to "is not" while preserving semantic meaning. The pipeline maintained basic punctuation while standardizing spacing and case, utilizing regex substitutions for consistent formatting.

The sequence preparation phase implemented a sliding window approach through the create\_sequence\_pairs() function. The key parameters were configured with input length fixed at 100 characters, minimum target length at 20 characters, and a stride of 50 characters for window sliding. The function identified sentence boundaries using regex pattern matching for periods followed by spaces, optimizing split points to maintain semantic coherence. The process\_dataset() function handled batch processing of reviews, implementing progress tracking every 1000 reviews to monitor execution.

The stratified sampling implementation through create\_stratified\_sample\_fast() incorporated several optimization techniques. These included batch processing with size of 1000 sequences, numpy-based operations for length calculations, vectorized bin assignment using percentile-based boundaries, and memory-efficient index manipulation instead of full sequence copying. Length-based filtering retained sequences between the 10th and 90th percentiles, spanning 23 to 58 tokens, with stratification implementing three balanced categories. The sequences were categorized as short (23-34 tokens), medium (35-46 tokens), and long (47-58 tokens), ensuring equal representation in the final dataset.

The implementation utilized the DistilGPT-2 tokenizer with specific configurations, aligning pad\_token with eos\_token and setting maximum lengths of 512 tokens for input and 128 for target sequences. Training was configured with a batch size of 4 and gradient accumulation steps of 4. Data persistence was implemented through pickle serialization, with the final data structure maintaining tokenizer parameters and sequence lengths in a dictionary format containing the processed sequences, tokenizer name, and maximum length parameters for both input and target sequences.

The optimized implementation achieved processing times of approximately 6-7 minutes per dataset on T4 GPU infrastructure, though memory constraints necessitated batch processing. The final preprocessing pipeline reduced the dataset to 41% of its original size while maintaining balanced length distributions: 244,410 training sequences and 237,774 testing sequences, each with exact thirds distribution across length categories.

Validation was implemented at multiple stages throughout the pipeline, including input/target relationship verification during sequence pair creation, shape validation during batch processing, padding consistency checks in sampled sequences, and distribution verification across length categories. This methodology achieved a balance between preprocessing thoroughness and computational efficiency, though future implementations could benefit from more extensive memory usage tracking and compression techniques.

## 4.2 DistilGPT2 fine-tuning

## 4.3 Bookcorpus cleaning

## 4.4 Fine-tuning my model on the IMDB set

# 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 25 November 2024).

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A. and Fidler, S. (2015) Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books, [online] Available at: http://arxiv.org/abs/1506.06724.

# 9. Appendices