# Sentiment Analysis on Movie Reviews: Investigating the Impact of Pre-Training Improvements with DistilBERT

**Paul Dupire and Floriane Zanella.**
ENSAE Paris
paul.dupire@ensae.fr, floriane.zanella@ensae.fr
Github repository
(The README file of the repository gives precisions on how the work was divided between the authors).

## Abstract

This work investigates the performance of a DistilBERT model for a binary sentiment classification task on movie reviews from the ACL IMDb dataset. A baseline model is fine-tuned on the original reviews, then two improvements are tested: 1) using LIME, we identify a strong reliance on high-valence sentiment words in the model, which motivates an experimental variant where such words are systematically replaced with mitigated alternatives, and 2) we customize a tokenization function to manage the penalty caused by long reviews being automatically truncated with DistilBERT's initial tokenizer. We finally test a combination of both approaches. Classification performance is evaluated through accuracy, confusion matrices, ROC-AUC metrics, and observation of examples. The results indicate that mitigating overt sentiment words improves the classification of positive movie reviews, while custom tokenization improves overall metrics, in particular by better classifying long reviews. The mixed model gives the best recall but leads to less precision. These performances significantly surpass those of Maas et al. (2011), which was our latent goal.

## 1 Introduction

Online comments have become a rich source of public opinion, making automatic sentiment analysis essential for understanding users attitudes at scale. Sentiment Analysis enables the efficient extraction of emotions and opinions from text, supporting informed decision-making across various fields, and has seen a significant rise in academic interest over the past decade.

Maas et al. (2011) proposed a model for this task applied to online movie reviews, introducing an IMDb dataset - now a benchmark in the field- that we leverage to fine-tune a DistilBERT model for binary classification of sentiment.

Transformer-based models have indeed achieved substantial success in SA in recent years. However, the extent to which these models depend on highly explicit lexical markers remains an open question. The initial analysis reveals a strong reliance on overt sentiment words, motivating an experiment in which these words are neutralized to assess the model's ability to infer sentiment from broader contextual cues. Then, we first examine whether replacing explicit sentiment indicators (e.g., "excellent" becoming "good") affects classification accuracy and bias between positive and negative reviews, potentially improving the handling of nuanced expressions like sarcasm-quite common in reviews. On the other hand, DistilBERT models appear limited in their ability to classify long texts with mixed overall sentiment, raising the question of whether a more tailored encoding of reviews by a custom tokenization could improve the performance.

Our primary research question is whether these enhancements to the baseline model -applied separately or combined- lead to more robust classification of movie reviews, ensuring that they outperform the results of the original authors.

We achieve this latent goal, and find that both lexical mitigation and custom tokenization allow better predictions of sentiment polarity with the DistilBERT framework.

## 2 Literature review

Sentiment analysis has long served as a testbed for natural language understanding, particularly due to the interplay between lexical semantics, context, and pragmatic meaning.

Maas et al. (2011) introduced a sentiment-aware vector learning model, which combined unsupervised word representation learning with supervised sentiment annotations. Their work revealed that semantic similarity alone is insufficient for accurate sentiment inference and that models benefit from learning sentiment-specific representations. They also released the ACL IMDb dataset, which has been widely used in subsequent research [4].

Early work in the field emphasized machine learning approaches based on vector-space models and bag-of-words representations, which treat text as unordered collections of word counts [7]. While effective for simple classification tasks, these methods lack the ability to capture word meaning -and therefore sentiment-beyond surface co-occurrence statistics.

The introduction of distributed word representations, such as Word2Vec [6] and GloVe [8], provided more informative features by embedding words into continuous vector spaces. However, these embeddings are static and do not account for polysemy or contextual usage. To address this limitation, contextual embedding models were introduced, most notably ELMo [9], which produces word representations as a function of the entire sentence.

In parallel, interpretability has emerged as a critical concern for deep NLP systems. Ribeiro et al. (2016) proposed LIME, a model-agnostic technique for interpreting black-box predictions by learning local surrogate models around individual instances. LIME enables the identification of features that most influence a model's decision, which is particularly valuable in understanding the reliance on specific lexical items in sentiment classification tasks [10].

The development of BERT [2] marked a shift in NLP methodology by employing bidirectional transformers and masked language modeling pretraining. BERT's architecture facilitates deep contextual understanding, allowing it to outperform earlier methods on a range of tasks, including sentiment classification. The emergence of model compression techniques, particularly knowledge distillation, has addressed the computational constraints of deploying large language models in real-world scenarios. Sanh et al. (2020) demonstrated that DistilBERT, a distilled version of BERT, retains 97% of the language understanding capabilities while being 40% smaller and 60% faster at inference time [11]. Their approach leverages a triple loss combining language modeling, distillation, and cosine-distance losses during pre-training rather than applying task-specific distillation. This general-purpose pre-training distillation creates models suitable for edge computing and mobile applications without significant performance degradation. This advancement enables efficient deployment of transformer-based models in resource-constrained environments, making sophisticated NLP techniques more accessible for practical applications [5] or work such as ours.

Recent research has also expanded beyond traditional binary or multi-class sentiment classification to include finer-grained sentiment tasks, such as emotion detection and sentiment intensity prediction. Models like RoBERTa [3], an optimized variant of BERT, and XLNet [13], which incorporates autoregressive modeling, have been applied to these tasks, offering further improvements in performance. The integration of external knowledge, such as sentiment lexicons and commonsense reasoning, is also gaining attention. For example, models that incorporate commonsense knowledge, like COMET [1], are being explored for enhancing sentiment understanding in more complex, context-sensitive scenarios, such as sarcasm or irony.

## 3 Data

The dataset employed in this study is the Large Movie Review Dataset released by Maas et al. (2011), commonly referred to as the ACL IMDb dataset. It consists of 50,000 movie reviews collected from the Internet Movie Database (IMDb), and their sentiment label (0 for negative and 1 for positive), based on the rating the reviewers have associated to their comment. Positive examples are defined as those with original IMDb ratings of 7 or above, and negative reviews are those with ratings of 4 or below. Neutral reviews (ratings between 5 and 6) are excluded to ensure clear sentiment polarity. The authors split the whole dataset into a train set and a test set of same size (25,000), both evenly divided into positive and negative examples. We further split the train dataset into full train set (22,500 reviews) and validation set (2,500 reviews) ensuring the same polarity equilibrium, to help us through model fine-tuning. The reviews vary in length and writing style, containing both colloquial and formal expressions.

Statistical examination of the dataset revealed an average review length of approximately 230 words (standard deviation: 170 words), 7.6% of them exceeding 500 words. The vocabulary distribution followed a small number of frequent function words and a long tail of infrequent terms. Sentiment-laden words, though sparse, were disproportionately impactful, as preliminary analyses using LIME would later confirm. The distribution of ratings is almost identical in the test and train set, and symmetrical between the polarized fields (1 to 4 and 7 to 10). However, there are many more 1 and 10 ratings than other values.

## 4 Methodology

### 4.1 Baseline model and classification

Our baseline classifier is based on the distilbert-base-uncased model from the Hugging Face Transformers library. All reviews are preprocessed using the tokenizer associated with this model, which applies WordPiece segmentation and lowercases all input text, consistent with the pretraining configuration of DistilBERT (which justifies no additional preprocessing, such as stopword removal, lemmatization, or syntactic normalization). Each review is truncated to a maximum of 512 tokens to fit within the model's input constraints (shorter reviews being zero-padded as necessary) and converted into a sequence of token IDs and attention masks, where the attention mask distinguishes real tokens from padding.

Fine-tuning is performed using the Trainer API from the Hugging Face transformers library, using the following configuration: 2 training epochs, a batch size of 16 for both training and evaluation, a learning rate of 5e-5.[1], and AdamW as optimizer (which is integrated to the Trainer process). We tracked the training loss and the accuracy on the validation set every 500 steps. These baseline parameters are kept identical throughout all of our three variant models, and evaluation metrics remained also consistent to evaluate their performance on the test set : we used overall accuracy and class-wise metrics to assess performance balance across sentiment classes (precision, recall, f1 score), as well as ROC-AUC.

### 4.2 A LIME analysis of the baseline model leading to pre-training lexical mitigation

To investigate the model's internal decision-making process, a qualitative interpretability analysis is conducted using LIME (Local Interpretable Model-agnostic Explanations) on the baseline classifier. It provides local approximations of the model's decision boundary by perturbing input text and observing the effect on predictions, and then fits a sparse linear surrogate model to approximate the influence of each token on the output probability. The analysis is performed several times on misclassified examples selected from the test set. Half of them are the instances where the classifier produced incorrect predictions with the highest confidence, as measured by the softmax probability assigned to the predicted (but incorrect) class. The remaining half are randomly sampled from the set of misclassified reviews to provide comparison cases. The most influential tokens (for which LIME generated a token-level importance score indicating a high contribution to the prediction) are examined in their context to determine whether specific lexical features—particularly sentiment-laden

[1] We experimented several configurations of learning rate and size of batches on the baseline model to assess how it changed the performance on the validation set the training time, leading to this choice.

words —are disproportionately responsible for the classification output or if the model fails to resolve contextual cues such as negation, sarcasm, or compositional nuance.

A second instance of the DistilBERT classifier is then fine-tuned on a variant of the dataset where explicitly charged sentiment terms are mitigated through lexical substitution. In a curated list that we derived from multiple LIME outputs observations, each selected term is manually paired with a neutral or mitigated counterpart chosen to maintain syntactic plausibility and topic relevance. For example, "masterpiece" is replaced with "film," and "horrible" is replaced with "bad." These substitutions are applied uniformly across all the datasets (train, eval and test), using exact string replacement, and do not affect the grammatical structure of sentences beyond the replaced term to preserve the overall semantics of the reviews. The aim is to evaluate whether the classifier can maintain performance when deprived of direct lexical sentiment cues.

### 4.3 Truncation issue during tokenization

In parallel, the qualitative examination of some long reviews in our dataset indicated that high lengths seemed to make the presence of contradictory or nuanced sentiments more common (positive at the beginning then negative at the end, for example, associated with a negative rating) which DistilBERT struggles to classify correctly due to the token limit per review imposed by the model. This constraint requires truncation of overly long reviews, keeping only the first 512 tokens (including special tokens). Following the results of Sun et al. on this issue[12], we implement a tokenization function that truncates overly long reviews by keeping only the first 128 tokens and the last 382 tokens, and manually performs the padding and insertion of special tokens. We re-evaluate the performance of the DistilBERT model after this new pretraining step, which we first apply to the baseline model alone ('Custom tokenization model'), then combine with lexical mitigation ('Mixed model'). In particular, we focus on how the proportion of misclassified reviews evolves across the three models depending on whether the reviews were truncated (because they were too long) or not.

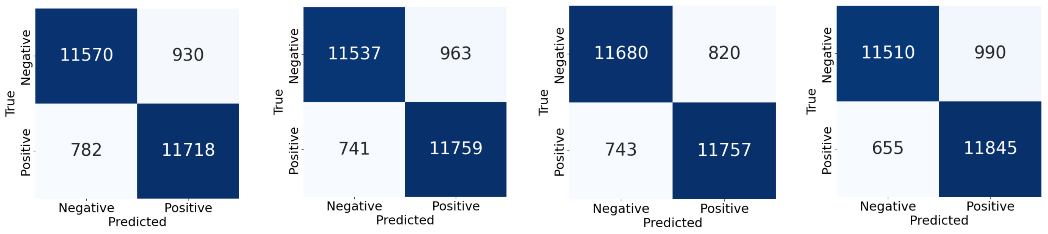
## 5 Results

All four DistilBERT models exhibit consistent improvements over the original benchmark from Maas et al. (2011). The standard evaluation metrics and confusion matrices are presented below for each variant.

Table 1: Classification metrics for the different DistilBERT models on the IMDb test set, with comparison to Maas et al.'s best result on the same test set.

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|
| Maas et al. | 0.8889 | - | - | - |
| Baseline | 0.9315 | 0.9265 | 0.9374 | 0.9319 |
| Mitigated | 0.9318 | 0.9243 | 0.9407 | 0.9324 |
| Custom tokenization | 0.9375 | 0.9348 | 0.9406 | 0.9377 |
| Mixed : mitigated + custom tokenization | 0.9342 | 0.9229 | 0.9476 | 0.9351 |

The baseline fine-tuned DistilBERT already achieves a substantial leap, with an accuracy of 93.15%, well-balanced precision and recall, and a ROC AUC of 0.98, reflecting a strong discriminative ability, with high sensitivity and specificity. The lexical mitigation strategy leads to a noticeable improvement in recall, rising to 94.07%, indicating fewer false negatives — i.e., more effective identification of positive reviews that might otherwise be misclassified, perhaps due to the mitigation of strong negative expressions in positive reviews that would be misinterpreted. The trade-off is a slight dip in precision.



(a) Baseline (b) Mitigated (c) Custom tokenization (d) Mixed

Figure 1: Confusion matrices for the four DistilBERT models. The mixed model is the less balanced between false positives and false negatives. Custom tokenization and Mitigated approaches are more balanced, still showing a good handling of the false positives.

The Custom tokenization reduces the number of false positives significantly, suggesting that this model is less prone to being misled by nuanced reviews beginning with positive statements thanks to the custom truncation. Finally, the Mixed model reaches the highest recall (94.76%) and F1 score (93.51%), while precision is the lowest of the four DistilBERT models, making it the less balanced model regarding the type of errors.

Applied on the baseline model results, the LIME analysis identifies highly polarized sentiment words such as "masterpiece", "wonderful" and "excellent" as the dominant contributors of some false positive examples. It shows that the classifier heavily depends on individual high-sentiment terms - sometimes used in contextual explanation or in sarcastic ways - and may not robustly account for compositional or contextual nuances.

Finally, we evidence a significant impact of the Custom tokenization function on the proportion of misclassified long reviews. Indeed, the Baseline model leads to misclassify 10.1% of long reviews (of more than 510 tokens), and only 6.4% of standard length reviews. These proportions are balanced when applying the Custom tokenization function, since the first proportion decreases to 7.1%, while the second one appears a bit lower too (6.1%).

However, when plotting the distribution of initial IMDb ratings for correctly and incorrectly classified reviews (see Fig.2 below), we notice that misclassified reviews are more likely associated with the most neutral ratings of the dataset (4 and 7). This emphasizes the difficulty of our sentiment analysis models to capture a correct overall polarized impression when mixed feelings are expressed.

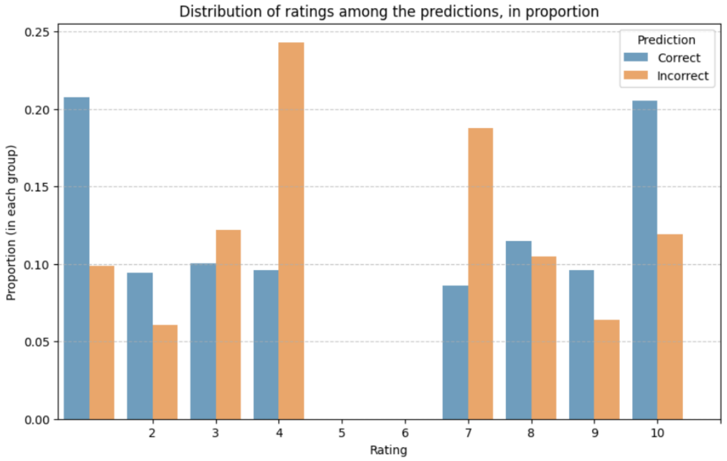


Figure 2: Distribution of true ratings according to the status of prediction dy the Custom Tokenization model, in proportion. Lecture : almost 25% of the incorrectly classified reviews are associated with a 4 rating.

## 6 Discussion and conclusion

This study sets out to evaluate and enhance the performance of a DistilBERT model on the task of binary sentiment classification of movie reviews using the ACL IMDb dataset. We examine two complementary avenues for improvement —lexical mitigation and custom tokenization—and assess their impact individually and in combination on prediction quality.

The baseline DistilBERT model performs strongly with an accuracy of 93.15% and a ROC-AUC score of 0.98. The mitigated model, which replaces sentiment-laden words with neutral alternatives, slightly reduces overall accuracy but improves performance on negative reviews by lowering false positives. The custom tokenization model, which addresses truncation issues, shows an improvement in recall and precision, demonstrating the value of preserving more contextual information in long reviews. Finally, the mixed model, combining both lexical mitigation and custom tokenization, provides the best recall but does not surpass the custom tokenization model in overall accuracy.

This work highlights the importance of addressing both lexical biases and tokenization limitations when pretraining models to enhance the performance of sentiment classification models. We show that relatively simple yet targeted pretraining interventions—such as vocabulary-level adjustments and input representation refinements—can improve the robustness and generalization of transformer-based sentiment classifiers.

However, our approach has notable methodological limitations that warrant consideration. The lexical mitigation strategy relies on manually curated word substitutions, which introduces subjectivity and potentially undermines generalizability across different domains. This manual intervention may inadvertently encode our own biases regarding what constitutes neutral or mitigated language. Similarly, our custom tokenization approach, while effective for this specific dataset, represents a heuristic solution rather than addressing the fundamental architectural limitations of transformer models in processing long texts. The fixed allocation of tokens (first 128 and last 382) assumes a consistent distribution of sentiment information across reviews—an assumption that likely varies across different writing styles and review structures. Furthermore, our binary classification framework oversimplifies the inherently continuous nature of sentiment, particularly evident in the disproportionate misclassification of reviews with moderate ratings (4 and 7). These limitations suggest that more sophisticated approaches are needed—perhaps integrating continuous sentiment scales or developing transformer architectures specifically designed for long-text sentiment analysis rather than relying on preprocessing adaptations, in order to handle the nuanced and sometimes misleading nature of online reviews better.

## References

[1] Antoine Bosselut et al. *COMET: Commonsense Transformers for Automatic Knowledge Graph Construction*. 2019. arXiv: 1906.05317 [cs.CL]. URL: https://arxiv.org/abs/1906.05317.

[2] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

[3] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL]. URL: https://arxiv.org/abs/1907.11692.

[4] Andrew L. Maas et al. "Learning Word Vectors for Sentiment Analysis". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Dekang Lin, Yuji Matsumoto, and Rada Mihalcea. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 142–150. URL: https://aclanthology.org/P11-1015/.

[5] Yanying Mao, Qun Liu, and Yu Zhang. "Sentiment analysis methods, applications, and challenges: A systematic literature review". In: *Journal of King Saud University - Computer and Information Sciences* 36.4 (2024), p. 102048. ISSN: 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2024.102048. URL: https://www.sciencedirect.com/science/article/pii/S131915782400137X.

[6] Tomas Mikolov et al. *Distributed Representations of Words and Phrases and their Compositionality*. 2013. arXiv: 1310.4546 [cs.CL]. URL: https://arxiv.org/abs/1310.4546.

[7] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. “Thumbs up? Sentiment Classification using Machine Learning Techniques”. In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Association for Computational Linguistics, July 2002, pp. 79–86. DOI: 10.3115/1118693.1118704. URL: https://aclanthology.org/W02-1011/.

[8] Jeffrey Pennington, Richard Socher, and Christopher Manning. “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162/.

[9] Matthew E. Peters et al. *Deep contextualized word representations*. 2018. arXiv: 1802.05365 [cs.CL]. URL: https://arxiv.org/abs/1802.05365.

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. 2016. arXiv: 1602.04938 [cs.LG]. URL: https://arxiv.org/abs/1602.04938.

[11] Victor Sanh et al. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. 2020. arXiv: 1910.01108 [cs.CL]. URL: https://arxiv.org/abs/1910.01108.

[12] Chi Sun et al. *How to Fine-Tune BERT for Text Classification?* 2020. arXiv: 1905.05583 [cs.CL]. URL: https://arxiv.org/abs/1905.05583.

[13] Zhilin Yang et al. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. 2020. arXiv: 1906.08237 [cs.CL]. URL: https://arxiv.org/abs/1906.08237.

# A Appendix

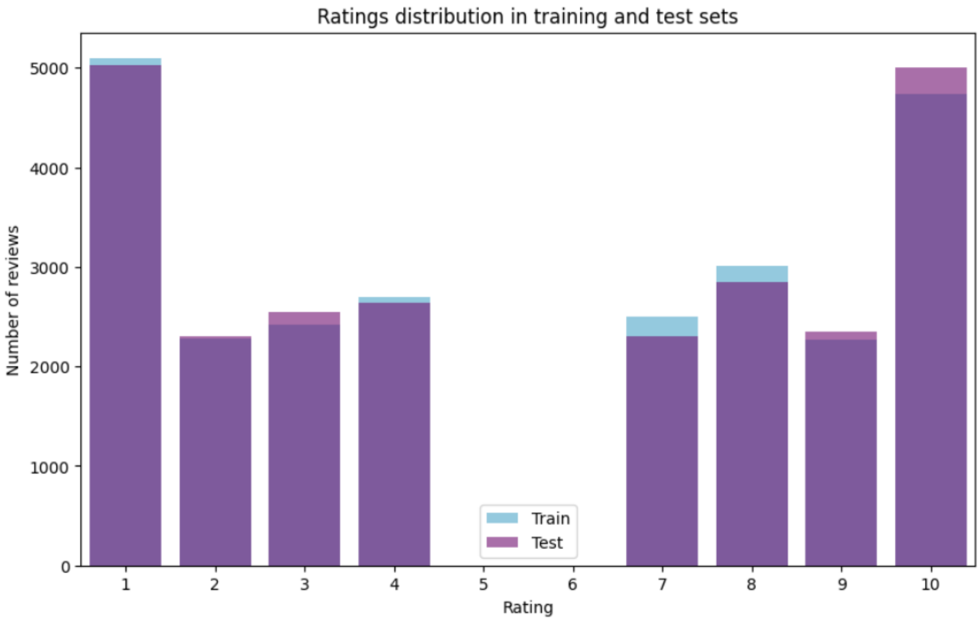


Figure 3: Distribution of ratings given by IMDb users along their reviews in train and test datasets. Extreme notes are the most represented (1 and 10), which means the dataset contains highly polarized reviews. The two datasets are balanced regarding this distribution.

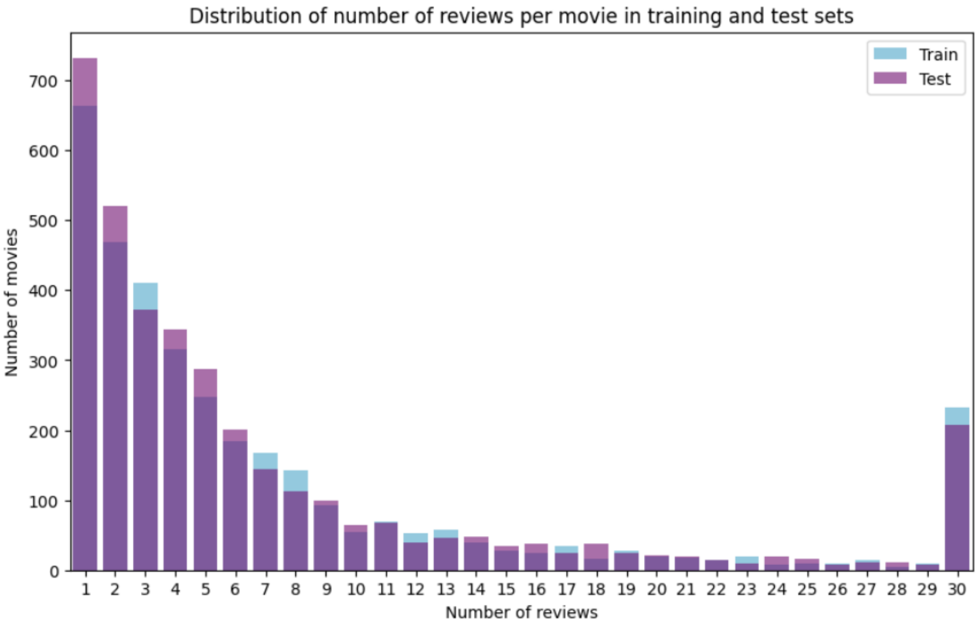


Figure 4: Distribution of the number of reviews per movie. The creators of the dataset ensured no more than 30 reviews per movie were present in the data.

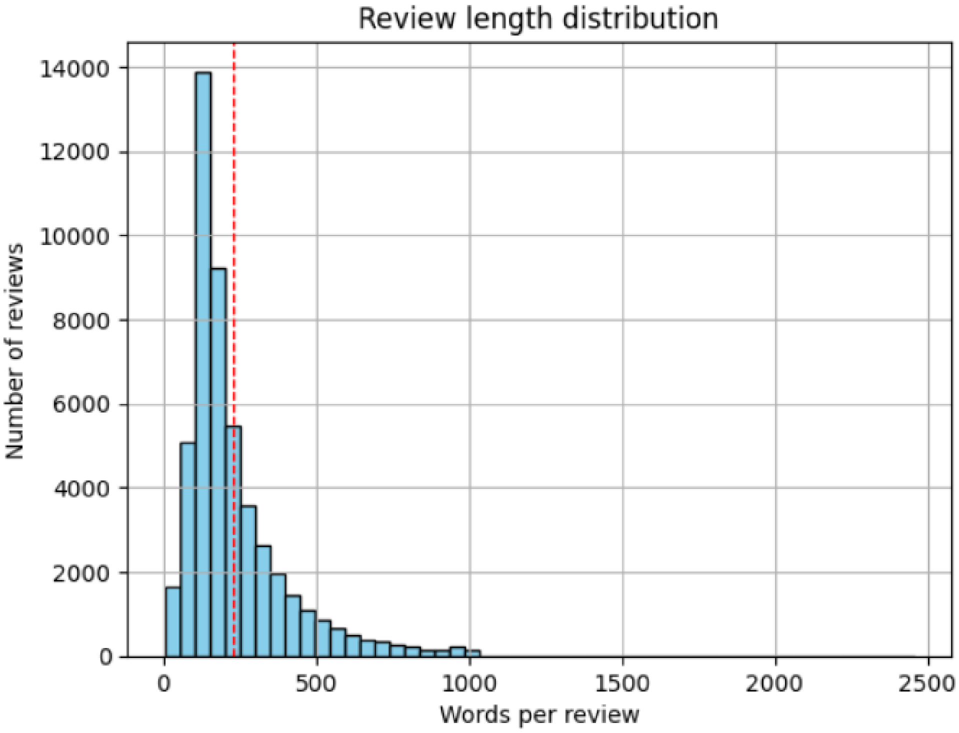


Figure 5: Distribution of the number of words per review. The red line indicates the mean.

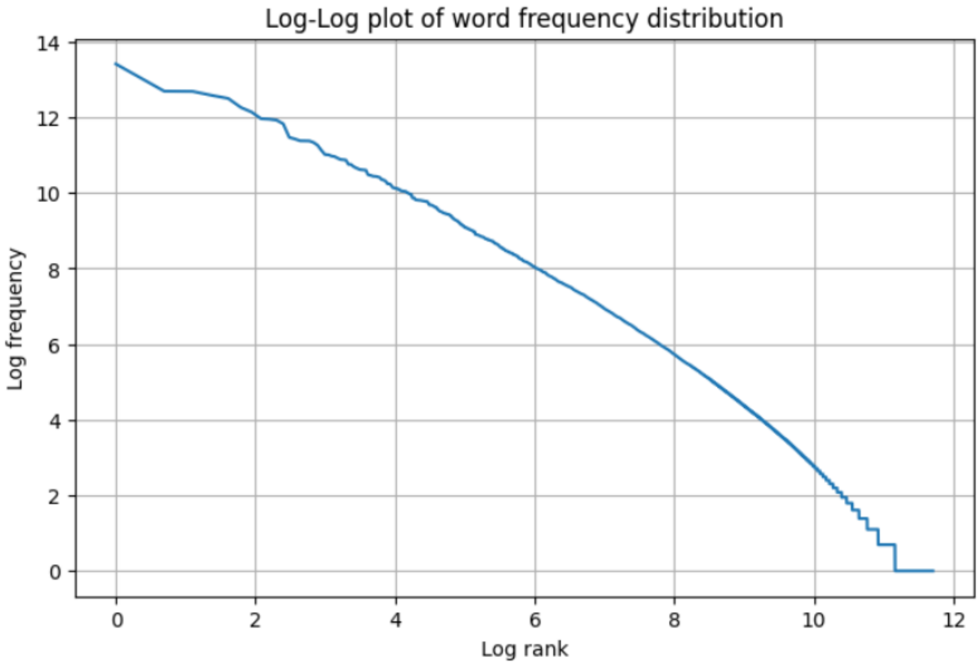


Figure 6: Log-Log Plot of Word Rank vs. Frequency. A small number of high-frequency words—predominantly function words such as "the, and, is" dominate the corpus, while the majority of words occur infrequently.

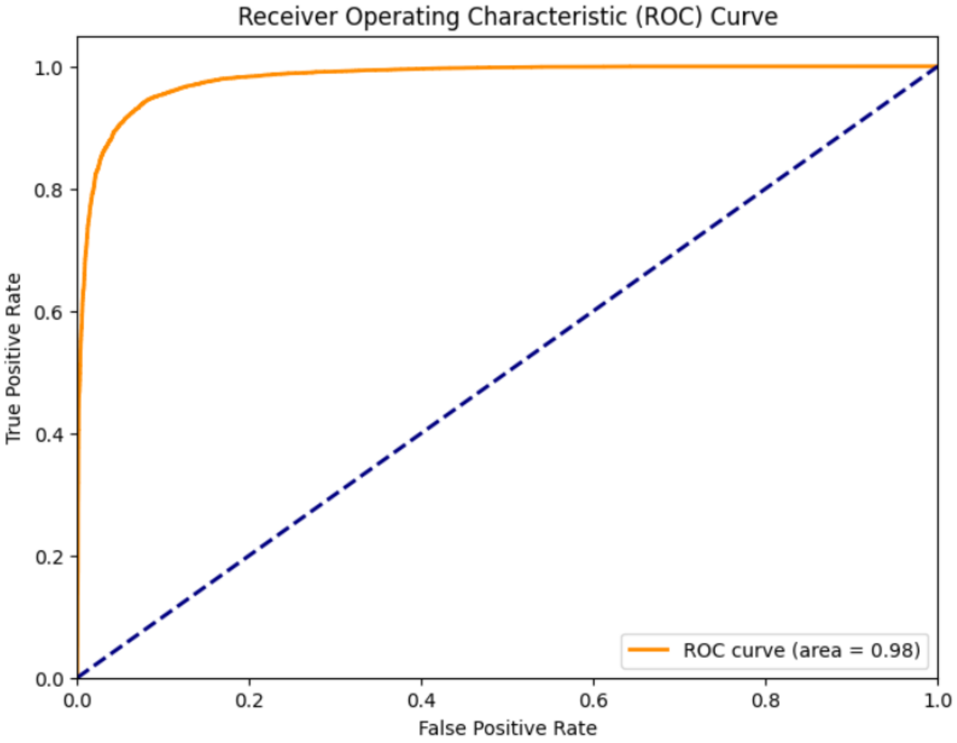


Figure 7: ROC curve for the baseline model.