Sentiment Analysis on Movie Reviews: Investigating the Impact of Strong Sentiment Word Mitigation with DistilBERT

Paul Dupire
ENSAE Paris
paul.dupire@ensae.fr
Github repository

Abstract

This work investigates the performance of DistilBERT on a binary sentiment classification task based on movie reviews from the ACL IMDb dataset. A baseline model is fine-tuned on the original dataset, and local interpretability analysis using LIME is conducted. The interpretability results indicate a strong reliance of the model on highly sentiment-laden words. This observation motivates an experimental variant where such words are systematically replaced with neutral alternatives. Classification performance is evaluated through accuracy and ROC-AUC metrics. Results indicate that lexical mitigation leads to a slight reduction in overall accuracy but improves the classification of negative reviews.

1 Introduction

Transformer-based models have achieved substantial success in sentiment classification tasks []. However, the extent to which these models depend on highly explicit lexical markers remains an open question. This study evaluates the performance of DistilBERT, a compressed version of BERT that maintains comparable performance with reduced computational requirements, fine-tuned on the ACL IMDb dataset and analyzes its behavior using a local interpretability technique.

Initial analysis reveals a strong reliance on overt sentiment words, motivating an experiment where such words are neutralized to assess the model's ability to infer sentiment from broader contextual cues. Specifically, we examine whether replacing explicit sentiment indicators (e.g., "excellent" becoming "good") affects classification accuracy and bias between positive and negative reviews. The primary research question is whether such mitigation leads to more robust sentiment analysis by forcing the model to rely on broader contextual features rather than isolated sentiment terms, potentially improving its handling of more nuanced expressions like sarcasm, where surface-level sentiment markers can be misleading. Our study finds that mitigating overt sentiment words improves the classification of negative movie reviews in exchange to a slight reduction in overall accuracy.

2 Literature review

Sentiment analysis has long served as a testbed for natural language understanding, particularly because of the interplay between lexical semantics, context, and pragmatic meaning. 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 (Pang et al., 2002). While effective on simple classification tasks, these methods lack the ability to capture word meaning beyond surface co-occurrence statistics.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

The introduction of distributed word representations, such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), 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 (Peters et al., 2018), which produces word representations as a function of the entire sentence.

The development of BERT (Devlin et al., 2019) 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. However, due to its computational cost, distilled versions such as DistilBERT (Sanh et al., 2019) were proposed. DistilBERT is trained to mimic the output behavior of BERT while reducing model size and inference latency.

Maas et al. (2011) introduced a sentiment-aware vector learning model, combining 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 used in this study.

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.

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), evenly divided into 25,000 positive and 25,000 negative examples. The dataset is further split into a training set and a test set, each containing 25,000 reviews, with no overlap between them.

Each review is accompanied by a binary sentiment label. Positive reviews 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 reviews vary in length and writing style, containing both colloquial and formal expressions.

RAJOUTER DES TRUCS A DIRE.

4 Methodology

4.1 Baseline model and classification

The baseline classifier is based on the `distilbert-base-uncased` model from the Hugging Face Transformers library. DistilBERT is a compressed version of BERT, trained using knowledge distillation to preserve much of BERT's language modeling capabilities while improving efficiency. For this experiment, a linear classification head is appended to the [CLS] token output of the final hidden layer to perform binary sentiment classification.

All reviews are preprocessed using the tokenizer associated with the `distilbert-base-uncased` model from the Hugging Face Transformers library. This tokenizer applies WordPiece segmentation and lowercases all input text, consistent with the pretraining configuration of DistilBERT. Each review is truncated to a maximum of 512 tokens to fit within the model's input constraints. Reviews shorter than this limit are zero-padded as necessary. The tokenizer converts each review into a sequence of token IDs and attention masks, where the attention mask distinguishes real tokens from padding.

No additional preprocessing, such as stopword removal, lemmatization, or syntactic normalization, is applied. This choice preserves the integrity of the original textual data and ensures compatibility with the pretrained DistilBERT encoder.

Fine-tuning is performed using the `Trainer` API from the Hugging Face `transformers` library. The model is trained on the tokenized IMDb dataset using the following configuration: 2 training epochs, a batch size of 8 for both training and evaluation, and logging every 500 steps. No model checkpoints are saved during training. The AdamW optimizer is used internally by the `Trainer` with default settings. The learning rate and weight decay parameters are not manually overridden. All training is conducted on a single NVIDIA Tesla V100 GPU.

The model's performance is evaluated on the test split using overall accuracy and ROC-AUC as primary metrics. Additionally, class-wise precision and recall are computed to assess performance balance across sentiment classes. The baseline model is trained and evaluated independently from the mitigated model to allow direct comparison.

4.2 LIME analysis of baseline model

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. LIME provides local approximations of the model's decision boundary by perturbing input text and observing the effect on predictions. It then fits a sparse linear surrogate model to approximate the influence of each token on the output probability.

The analysis is performed on six misclassified examples selected from the test set. Three of these are the instances where the classifier produced the highest confidence incorrect predictions, as measured by the softmax probability assigned to the predicted (but incorrect) class. The remaining three are randomly sampled from the set of misclassified reviews to provide comparison cases.

For each example, LIME generates token-level importance scores indicating the contribution of each input token to the model's prediction. The most influential tokens are examined to determine whether specific lexical features—particularly sentiment-laden words—are disproportionately responsible for the classification output. This analysis is used to assess whether the model's misclassifications are attributable to an over-reliance on strong sentiment indicators or failure to resolve contextual cues such as negation, sarcasm, or compositional nuance.

For each of the six instances, LIME is applied to compute token-level contribution scores. These scores quantify the extent to which each token influenced the predicted sentiment. Tokens with the highest positive or negative coefficients are identified and qualitatively inspected.

4.3 Lexical mitigation and classification

To examine the extent to which the model relies on explicitly charged sentiment terms, a variant of the dataset is created by mitigating such terms through lexical substitution. A curated list of high-sentiment words is constructed based on their observed impact during LIME analysis of the baseline model.

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 both the training and test sets using exact string replacement. The substitutions do not affect the grammatical structure of sentences beyond the replaced term.

This mitigation process is intended to reduce the influence of overt sentiment indicators while preserving the overall semantics of the reviews. The aim is to evaluate whether the classifier can maintain performance when deprived of direct lexical sentiment cues, thus encouraging a more context-driven classification behavior.

A second instance of the DistilBERT classifier is fine-tuned on the mitigated training set. The architecture, optimizer, batch size, learning rate, and number of epochs are kept identical to the baseline setup. The test set used for evaluation is also subjected to the same sentiment word substitutions to maintain consistency.

By training and evaluating the model on the mitigated data, the experiment aims to assess whether performance is preserved in the absence of explicit sentiment words and whether the classifier relies more on distributed textual context. LIME is not applied to the mitigated model in this study, as the focus is on evaluating performance metrics and comparative differences.

5 Results

The baseline DistilBERT model, trained on the original dataset, achieves an accuracy of 93.14% on the test set. The ROC-AUC score is 0.98. The classifier obtains a precision of 92.96%, a recall of 93.34%, and an F1 score of 93.15%. The confusion matrix indicates a relatively balanced performance across sentiment classes, with 11,617 true negatives, 11,668 true positives, 832 false negatives, and 883 false positives.

These results suggest that the classifier is slightly more likely to misclassify positive reviews as negative than the reverse, though the difference is minimal. The ROC curve further confirms the model's strong discriminative ability, with high sensitivity and specificity.

Across the three confidently incorrect examples, LIME consistently highlights highly polarized sentiment words such as "masterpiece", "wonderful" and "excellent" as the dominant contributors. These strong sentiment words are sometimes used in contextual explanation or in sarcastic ways, which the model fails to capture. This analysis reveals that the classifier heavily depends on individual high-sentiment terms and may not robustly account for compositional or contextual nuances.

The classifier trained on the mitigated dataset, in which strongly sentiment-laden words have been replaced with neutral equivalents, achieves an accuracy of 92.85%. Precision improves slightly to 93.58%, while recall decreases to 92.02%. The resulting F1 score is 92.79%. The ROC-AUC remains unchanged at 0.98.

The confusion matrix reveals 11,711 true negatives and 11,502 true positives, with 998 false negatives and 789 false positives. Compared to the baseline, the mitigated model reduces the number of false positives (from 883 to 789) but increases the number of false negatives (from 832 to 998). This shift indicates a modest improvement in negative class classification, accompanied by a decline in sensitivity to positive class examples.

Table 1: Classification metrics for baseline and mitigated DistilBERT models on the IMDb test set.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Baseline	0.9314	0.9296	0.9334	0.9315
Mitigated	0.9285	0.9358	0.9202	0.9279

5.1 Comparison

The mitigated model performs slightly worse overall in terms of accuracy and F1 score, but demonstrates improved precision. The tradeoff in recall, resulting in more false negatives, suggests a more conservative decision boundary for positive sentiment. The reduction in false positives implies that the model is less prone to being misled by remaining sentiment indicators, aligning with the hypothesis that mitigation reduces lexical bias.

The ROC-AUC values of 0.98 in both conditions indicate that the overall ranking ability of the model remains strong, and that the primary effects of mitigation are visible in precision-recall tradeoffs rather than in the model's capacity to distinguish classes.

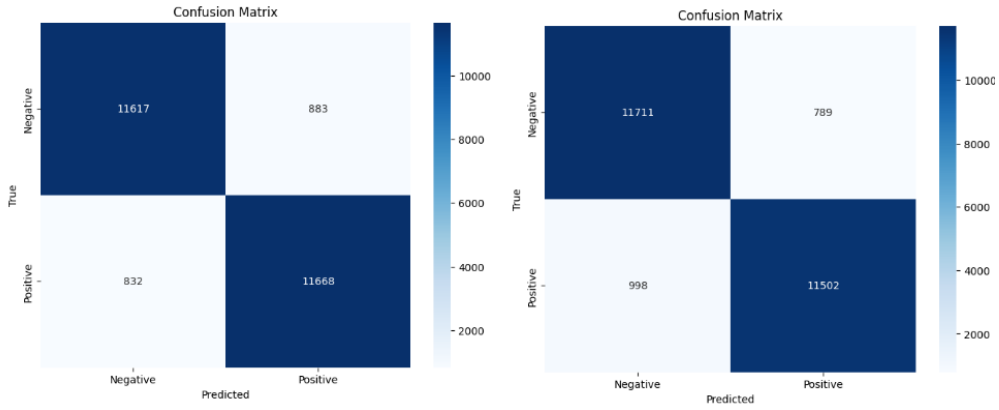

Figure 1: Confusion matrices for the baseline (left) and mitigated (right) DistilBERT models. The mitigated model shows fewer false positives (FP) but more false negatives (FN), indicating a shift toward conservative positive classification.

6 Discussion and Conclusion

The results indicate that the baseline DistilBERT model achieves high overall classification performance on the IMDb sentiment task. The baseline classifier obtains an accuracy of 93.14% and an F1 score of 93.15%, with balanced precision and recall. However, further analysis using LIME reveals that the model often relies heavily on isolated high-sentiment words, sometimes failing to account for the surrounding context or discourse structure.

One such example illustrates this issue clearly. A review expressing discontent with a film adaptation is misclassified as positive with high confidence (probability 0.9986). The model's decision is driven by tokens such as "excellent", "Gershwin", and "NY", which appear positively weighted despite their use in critical or neutral contexts. Notably, the term "gripe", which carries a negative connotation and contributes correctly to the negative class, is given lower importance. This suggests that lexical features strongly influence the model's output even when they are used within negated, ironic, or contrastive constructions.

To investigate this further, an alternative model was trained on a lexically mitigated version of the dataset, where explicit sentiment words were replaced with neutral counterparts. The mitigated model achieves a slightly lower accuracy of 92.85%, with improved precision (93.58%) but reduced recall (92.02%). It produces fewer false positives but more false negatives, indicating a more conservative approach to positive classification. The ROC-AUC remains unchanged at 0.98 in both settings.

The mitigation experiment aimed to reduce this sensitivity by replacing explicit sentiment words with neutral alternatives. The resulting model, while slightly less accurate overall, achieved higher precision for negative sentiment, suggesting an improvement in robustness and a possible reduction in false positives caused by isolated high-valence terms.

References

References follow the acknowledgments in the camera-ready paper. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

A Appendix / supplemental material

Optionally include supplemental material (complete proofs, additional experiments and plots) in appendix. All such materials **SHOULD be included in the main submission.**