A Comparative Analysis of SVM and DistilBERT with Polarity-aware Features

Ismael DEMBELE
ENSAE Paris
ismael.dembele@ensae.fr

Abstract

Sentiment analysis is a fundamental task in natural language processing, with applications in customer feedback, content moderation, and market analysis. In this work, we explore the effectiveness of combining classical machine learning techniques with modern transformer-based models to improve sentiment classification on the IMDB dataset. We also examine the contribution of an external lexical polarity score, integrated as a feature alongside traditional or contextualized representations. Our study contrasts a TF-IDF + SVM approach with a fine-tuned DistilBERT model, and evaluates how the lexical polarity score impacts each. We find that while polarity features enhance classical models, they offer no benefit to DistilBERT and may even reduce its performance. Our findings highlight the importance of feature-model compatibility in sentiment classification.

1 Introduction

Understanding sentiment expressed in text is a core problem in natural language processing (NLP), and plays a crucial role in opinion mining, product reviews, and social media analysis. Given its relevance and availability of labeled data, the IMDB movie reviews dataset has become a standard benchmark for evaluating sentiment classification systems.

Historically, approaches to sentiment classification have relied on bag-of-words representations and linear classifiers such as Naive Bayes and Support Vector Machines (SVM). These models are simple and efficient, but they lack contextual understanding and struggle with subtle linguistic phenomena such as irony or negation. More recently, pre-trained language models like BERT have achieved state-of-the-art performance by capturing deep contextual representations of text through transformer architectures.

In parallel, other approaches have attempted to inject external knowledge, such as lexicons of sentiment-bearing words, into learning pipelines. One such source is the polarity lexicon provided by the IMDB dataset creators, which assigns a sentiment score to each word based on supervised learning.

This project investigates two primary questions:

1. How do classical models like SVM compare to transformer-based models like DistilBERT on binary sentiment classification?
2. Can an external lexical polarity score improve model performance in either case?

To answer these, we replicate traditional pipelines using TF-IDF and SVM, with and without a polarity-aware feature (‘er_score‘), and we fine-tune a DistilBERT model on the same task, evaluating the effect of the same polarity signal. We perform extensive comparisons and error analysis to identify the strengths and limitations of each approach.

Preprint. Under review.

2 Related Work

Sentiment classification has been extensively studied in the NLP literature. Early work by Pang et al. [5] applied machine learning classifiers such as Naive Bayes, Maximum Entropy, and SVMs using unigram and bigram features. These methods demonstrated the viability of supervised learning for sentiment, but were limited by their inability to capture context or word order.

Subsequent efforts, notably Maas et al. [3], introduced the IMDB dataset and proposed learning word embeddings that encode sentiment dimensions. Their method combined unsupervised learning on unlabeled data with supervised polarity information, highlighting the value of integrating semantic and affective knowledge.

More recently, transformer-based architectures like BERT [1] have revolutionized NLP by providing deep contextual representations through self-attention mechanisms. Fine-tuning pre-trained models on specific tasks, such as sentiment classification, has yielded significant improvements over classical methods. DistilBERT [6], a compressed variant of BERT, maintains much of the performance with lower computational cost.

A growing body of work has also explored the fusion of pre-trained language models with external lexical knowledge, including polarity lexicons [2, 4]. These approaches attempt to reinforce the affective dimension of word usage, but the effectiveness of such integration remains mixed and task-dependent.

Our work revisits this question in the context of binary sentiment classification. We compare a traditional SVM-based pipeline, with and without a polarity-aware score, to a fine-tuned DistilBERT model. We also analyze whether the lexical polarity signal can complement or interfere with the performance of transformer-based models.

3 Dataset

We use the IMDB Large Movie Review Dataset introduced by Maas et al. [3]. It contains 50,000 labeled movie reviews: 25,000 for training and 25,000 for testing, balanced across positive and negative sentiment. Each review is stored in a text file, with its rating encoded in the filename (e.g., `123_10.txt`). Ratings range from 1 to 10.

Following previous work, we remove neutral reviews (ratings 5 and 6) and retain only those rated 1–4 (negative) and 7–10 (positive), ensuring a clear polarity boundary.

3.1 Text length distribution

The reviews vary widely in length, from a few dozen to over 2,000 words. This is important for modeling choices such as input truncation for BERT.

As shown in Figure 1, most reviews cluster between 100 and 400 words. This distribution justifies limiting transformer input length to 512 tokens, while preserving most content.

3.2 Sentiment and rating structure

Figure 2 shows how sentiment labels correspond to rating values. Ratings 1 and 10 dominate, reflecting user tendency toward extremes.

This skew suggests that many reviews are strongly opinionated, which can aid learning but may reduce generalizability to subtler expressions.

3.3 Polarity Lexicon Feature (`er_score`)

The dataset includes two useful resources:

- `imdb.vocab`: a list of all words used in the corpus, sorted by frequency;
- `imdbEr.txt`: a file containing a polarity score for each word, in the same order.

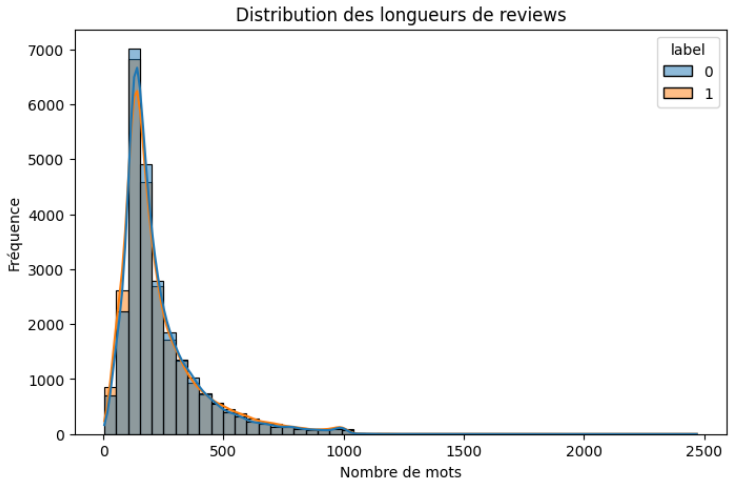

Figure 1: Distribution of review lengths (in number of words). Most reviews fall under 500 words.

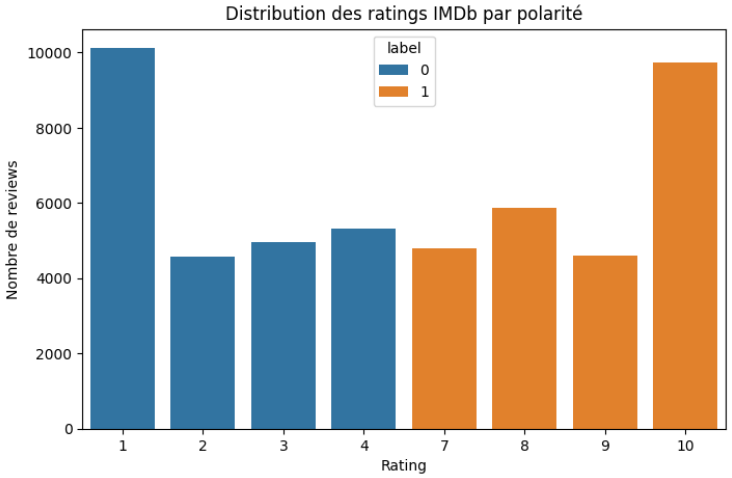

Figure 2: Distribution of IMDb ratings by sentiment class. Extreme values are overrepresented.

Using these, we compute a scalar polarity score for each review as the average of known word polarities:

$$\text{er\_score}(x) = \frac{1}{|x|} \sum_{w_i \in x} s(w_i)$$

where $s(w_i)$ is the score from the lexicon if $w_i \in$ `imdb.vocab`.

This score is later used as a standalone feature in some models (SVM, BERT variant), aiming to inject sentiment priors explicitly.

3.4 Frequent terms by sentiment

Figure 3 shows the most frequent words in positive and negative reviews. Words like *excellent*, *funny*, or *enjoyable* dominate positive reviews, while *boring*, *waste*, and *worst* are prominent in negative ones.

Figure 3: Word clouds for positive (left) and negative (right) reviews. Larger words appear more frequently.

These terms reflect strong emotional polarity and validate the potential usefulness of word-level sentiment features.

4 Methodology

We evaluate two complementary families of models for binary sentiment classification: a classical pipeline using sparse lexical features (TF-IDF) and a linear Support Vector Machine (SVM), and a contextualized transformer-based pipeline based on fine-tuning DistilBERT. We also assess the influence of a scalar sentiment feature, `er_score`, across both paradigms.

4.1 TF-IDF + SVM

Our baseline approach follows a traditional pipeline:

1. Text preprocessing (lowercasing, stopword removal, punctuation stripping);
2. TF-IDF vectorization over a vocabulary of 20,000 terms;
3. Classification via a linear SVM.

The SVM attempts to find a hyperplane separating the two sentiment classes:

$$\hat{y} = \text{sign}(w^\top x + b)$$

where $x \in R^d$ is the TF-IDF vector, w is the weight vector learned by the model, and b is the bias.

We consider two variants:

- **TF-IDF only**: standard textual features;
- **TF-IDF + er_score**: a scalar polarity score appended to x as an additional dimension.

This enrichment aims to complement sparse lexical features with a sentiment prior aggregated from the lexicon.

4.2 DistilBERT Fine-tuning

While bag-of-words models treat tokens as independent and unordered, transformer-based models leverage self-attention to capture contextual dependencies. We fine-tune DistilBERT [6], a compact variant of BERT [1], on our binary classification task.

Each review is tokenized using the pre-trained DistilBERT tokenizer and truncated to a maximum of 512 tokens. The resulting input is passed through the encoder, and we use the final [CLS] embedding $h_{[\text{CLS}]} \in R^{768}$ for classification via a softmax layer.

We train the model using the HuggingFace `Trainer` API with the following settings:

- Dataset: 12,000 positive and 12,000 negative reviews (balanced);
- Batch size: 16, learning rate: 2×10^{-5};
- Number of epochs: 2;
- Evaluation: accuracy, F1-score, precision, recall on a 20% held-out test set.

4.3 Incorporating `er_score` into BERT

To investigate whether the external polarity score can enhance contextual models, we design a variant where the scalar `er_score` is concatenated to the [CLS] representation before classification:

$$\tilde{h} = [h_{[\text{CLS}]}; \text{er\_score}] \quad \rightarrow \quad \text{MLP} \quad \rightarrow \quad \hat{y}$$

This modified representation is passed through the classification head.

Despite this design, we observe a degradation in performance (detailed in Section 5), suggesting that DistilBERT may already encode sentiment internally, and that naive scalar fusion can act as noise.

4.4 Evaluation Strategy

For both families of models, we compare:

- Global classification metrics: accuracy, F1, precision, recall;
- The impact of adding `er_score`;
- Error types and model disagreements;
- Structure of learned representations (via t-SNE on embeddings).

This multi-angle evaluation allows us to assess both quantitative performance and qualitative behavior.

5 Results

We now present the performance of our models on the IMDB sentiment classification task. All evaluations are conducted on a held-out test set representing 20% of the labeled dataset (4,800 examples). We analyze the contribution of lexical polarity, compare classical and contextual models, and study their learned representations.

5.1 SVM with TF-IDF and Polarity Score

Our baseline SVM trained on TF-IDF features achieves an accuracy of 89.3%. When augmented with the scalar `er_score`, performance improves modestly to 89.6%, suggesting that lexical polarity helps classical models capture affective signals not fully represented in TF-IDF vectors.

Interestingly, precision and recall both benefit from the additional feature, resulting in a balanced improvement in F1-score.

5.2 Fine-tuned DistilBERT

Fine-tuning DistilBERT for two epochs on 24,000 labeled reviews yields our best results, with an accuracy of 90.7% and an F1-score of 0.9066. This confirms the strength of transformer-based contextual representations for sentiment classification, particularly on long or complex reviews.

Compared to SVM, the gain may appear small numerically, but error analysis (see Section 6) reveals that BERT consistently succeeds on reviews where context, negation, or subtlety dominate.

5.3 Attempt to Inject Polarity into BERT `er_score`

We also experimented with enriching BERT by appending the scalar `er_score` to the [CLS] embedding. Contrary to expectations, this variant performs worse (88.9% accuracy), with a slight drop across all metrics.

This suggests that DistilBERT already encodes sentiment effectively from context, and that simple scalar concatenation may introduce noise rather than signal. A more structured integration (e.g., gated fusion or adapters) might be required.

5.4 Comparative Metrics

Table 1 reports the main evaluation metrics—accuracy, F1-score, precision, and recall—for all four model configurations. It provides a numerical baseline to assess the impact of model choice and the effect of the polarity score `er_score`.

Model	Accuracy	F1	Precision	Recall
TF-IDF + SVM	89.3%	0.893	0.890	0.896
TF-IDF + SVM + er_score	89.6%	0.896	0.892	0.900
DistilBERT (2 epochs)	**90.7%**	**0.9066**	0.9041	0.9091
DistilBERT + er_score	88.9%	0.889	0.884	0.895

Table 1: Performance comparison of models on the IMDB sentiment classification task.

Several trends emerge from this table. First, we observe that adding the polarity score improves the SVM pipeline across all metrics. The F1-score increases from 0.893 to 0.896, and recall improves slightly as well. This confirms that simple sentiment priors can benefit linear models, especially when based on frequency-driven representations like TF-IDF.

In contrast, integrating `er_score` into DistilBERT leads to a performance degradation: accuracy drops from 90.7% to 88.9%, and F1-score from 0.9066 to 0.889. This suggests that deep contextual models may already internalize sentiment cues, and that appending scalar values without dedicated integration mechanisms can be counterproductive.

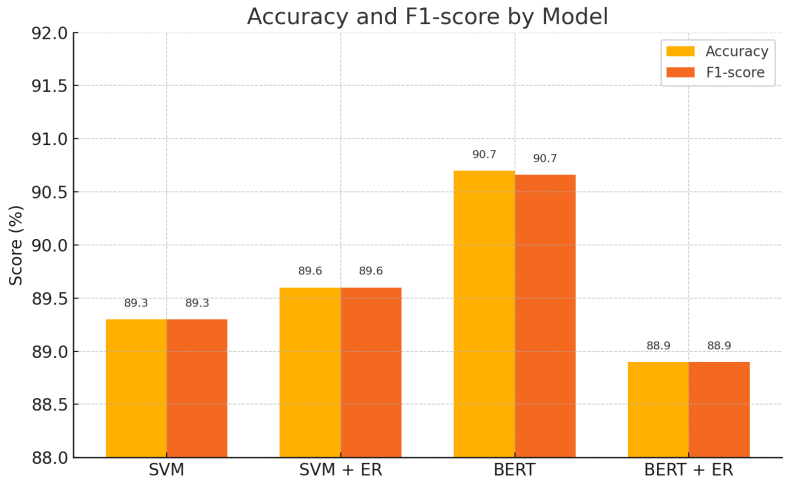

Figure 4: Accuracy and F1-score comparison across models.

Figure 4 provides a visual summary of accuracy and F1-score for each model. It highlights that BERT outperforms SVM overall, but also that the polarity feature benefits classical models while impairing contextual ones. Notably, the gap between SVM + ER and BERT is reduced in terms of F1, showing that lexical sentiment information remains a powerful signal when well-matched to the model type.

5.5 Embedding Visualization

To better understand how DistilBERT structures sentiment information internally, we extract the [CLS] token embeddings from the final layer of the model for all test examples. We then apply t-distributed stochastic neighbor embedding (t-SNE) to project these 768-dimensional vectors into 2D space.

As shown in Figure 5, the embeddings naturally separate into two main clusters corresponding to positive and negative reviews. This demonstrates that the [CLS] representations learned by the model encode a strong sentiment signal, even without explicit supervision on representation shape.

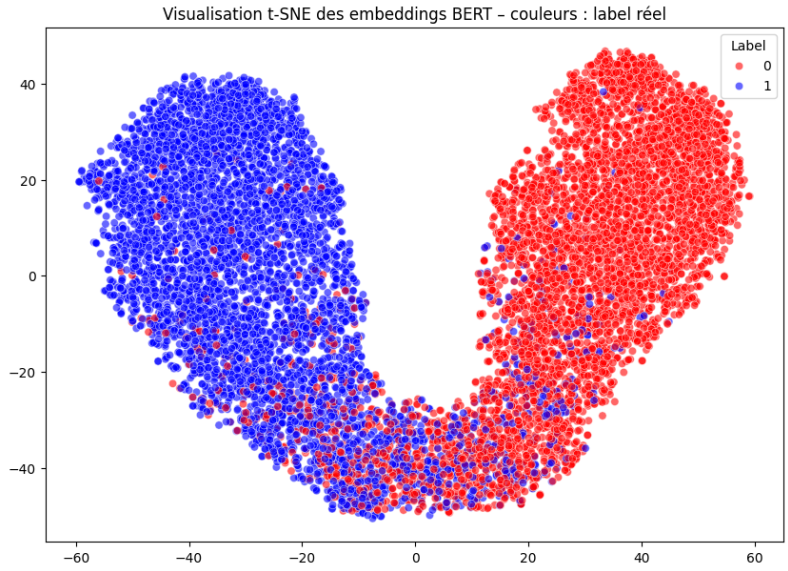

Figure 5: t-SNE projection of DistilBERT [CLS] embeddings on the test set. Color indicates sentiment label.

Interestingly, we observe a few outliers and boundary cases where the separation is less clean. These correspond to ambiguous or mixed reviews, where sentiment is expressed more subtly or contradictorily. This aligns with some of the errors made by BERT, as we explore in Section 6.

This visualization provides further evidence that DistilBERT internally organizes sentiment information in a geometrically meaningful way, supporting its high classification accuracy. It also suggests that future work could benefit from explicitly modeling this latent geometry—e.g., via clustering or embedding regularization—to further enhance interpretability and robustness.

6 Discussion

While the overall performance of our models is high, a closer inspection reveals meaningful differences in behavior and limitations. We structure this discussion around three axes: error patterns, model behavior with polarity injection, and lessons for future integration of lexical knowledge.

6.1 Error Patterns and Confusion Analysis

To better understand DistilBERT's misclassifications, we analyze its confusion matrix (Figure 6). The model correctly classifies over 93% of positive and negative reviews, but shows a slight asymmetry: false positives are slightly more frequent than false negatives.

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.324700	0.256663	0.903542	0.905026	0.882753	0.928451
2	0.156100	0.254788	0.910000	0.909244	0.907718	0.910774

Figure 6: Confusion matrix of DistilBERT on the test set.

This asymmetry suggests that the model is more prone to interpreting neutral or subtle content as positive. This could stem from the review distribution in the training set, where very negative reviews are often explicit, while positive reviews may rely on emotional nuance.

6.2 Case Study: When BERT Gets It Wrong

Examining individual examples sheds light on common pitfalls. Consider the following review, labeled positive but predicted negative:

"This film is the perfect satire of how power can corrupt... brilliant and disturbing."
True label: Positive **BERT Prediction:** Negative **er_score:** 0.091

Here, the word "disturbing" may have misled the classifier, despite the clear overall praise. This illustrates how contextual ambiguity or polysemy can interfere with BERT's sentiment perception.

Inversely, short, emotionally charged reviews are sometimes correctly classified by SVM but missed by BERT:

"Worst movie ever. Total waste of time."
True label: Negative **BERT Prediction:** Positive **er_score:** 0.33

This suggests that BERT's reliance on token context may reduce sensitivity to highly polarized keywords in short texts.

6.3 The Challenge of Feature Fusion

Our attempt to improve DistilBERT by injecting the `er_score` into the [CLS] embedding led to decreased performance. This highlights an important lesson: scalar-level sentiment information, even if useful in isolation, may not integrate naturally into pretrained embedding spaces without careful design.

While SVM benefits directly from the polarity signal—likely because it operates in a flat, interpretable feature space—BERT relies on hierarchical, token-level attention. Simply appending a scalar to a deep embedding vector may break the model's inductive biases.

6.4 Lessons and Opportunities

These observations suggest several directions for future improvement:

- **Better feature fusion:** Instead of concatenating the polarity score, we could explore fusion via gating mechanisms or auxiliary attention.
- **Multiview modeling:** Combine contextual embeddings and symbolic features in a two-branch architecture, where BERT and a sentiment feature encoder are trained jointly.
- **Error-aware training:** Use examples where SVM and BERT disagree to fine-tune ensemble strategies or confidence calibration.
- **Prompt-based approaches:** Reformulate sentiment detection as a textual inference task using instruction-tuned models.

Overall, this project underscores that the value of a feature depends as much on how it is injected as on its intrinsic quality. Classical models and LLMs see the world differently—and bridging their representations requires more than concatenation.

7 Conclusion

In this work, we conducted a comparative study of sentiment classification methods using the IMDB movie review dataset. We contrasted a classical pipeline based on TF-IDF features and a linear SVM with a transformer-based model, DistilBERT. In both setups, we evaluated the impact of an external polarity signal (`er_score`) derived from a sentiment lexicon.

Our results show that DistilBERT significantly outperforms the SVM baseline, achieving 90.7% accuracy and an F1-score of 0.9066. However, we also find that the usefulness of `er_score` depends on the model: it benefits SVM but impairs BERT when naively appended. This highlights the importance of alignment between model architecture and external features.

Our qualitative and quantitative analyses reveal strengths and limitations on both sides. BERT handles context and nuance effectively, while SVM remains more sensitive to explicit polarity terms. We also observe clear sentiment structure in the embedding space of BERT.

Future work could explore more structured fusion mechanisms—such as gated attention, auxiliary classifiers, or multi-view learning—to integrate symbolic sentiment features into deep models. More broadly, our findings emphasize that feature engineering remains relevant, but must be tailored to the architecture in use.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

[3] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, 2011.

[4] Saif M Mohammad and Peter D Turney. Nrc emotion lexicon. *National Research Council, Canada*, 2013.

[5] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 10:79–86, 2002.

[6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.