Sentiment analysis on movie reviews

GUIBERT Julien
julien.guibert@ensae.fr

Abstract

In this work, I focus on improving sentiment analysis for movie reviews by addressing the challenge posed by reviews expressing mixed feelings. Traditional classification algorithms often struggle with such ambiguous cases, as illustrated by review #757: “I loved most of the script but did not like the ending. [...] Seeing the completed movie I have to say I am amazed. [...] I have two main problems with the movie.” Assigning a clear positive or negative label to these reviews is not only difficult but arguably meaningless.

To address this, I propose an approach that detects and removes mixed-sentiment reviews prior to evaluating classification performance. By excluding these inherently ambiguous cases, we obtain a more meaningful measure of the model’s true discriminative ability. I present an analysis of how classification accuracy changes when mixed reviews are excluded, emphasizing the importance of handling sentiment ambiguity in performance evaluation. The code is publicly available at `https://github.com/guibertj/EnsaeNLP2025.git` fd

1 Introduction

We base our experiments on the IMDB movie reviews dataset, which contains 50,000 labeled reviews evenly split between positive and negative sentiments. Reviews were preprocessed by removing noise and tokenized into individual words.

A central issue uncovered during this phase was the presence of *mixed-sentiment reviews*, where both positive and negative expressions coexist. These reviews blur the decision boundary and compromise model evaluation. For instance if we take the two following reviews we can see this phenomenon :

Example of Mixed Review

> “Seeing the completed movie I have to say I am amazed. Reading the script I had not imagined the acting (and Gorman’s directing) would be as powerful and moving as it is here. The two leads did an amazing job. They were very believable and this movie is well worth watching just for the performances of these two amazing stars. [...] I have two main problems with the movie. First off, and I’m sure I mentioned this when I read the script, the wife should have died. If she were dead the ending would be much more believable. As it is, we have this guy who has been agonizing over a woman who left him ’years ago.’ If she left him and he is still in love after so long he’d have to be seriously delusional. But he doesn’t come across that way in the script. (There is also the problem of why he did not expend all this energy going after the wife he still loves.)”

Example of Non-Mixed Review

> “I went and saw this movie last night after being coaxed to by a few friends of mine. I’ll admit that I was reluctant to see it because from what I knew of Ashton Kutcher he was only able to do comedy. I was wrong.

> Kutcher played the character of Jake Fischer very well, and Kevin Costner played Ben Randall with such professionalism. The sign of a good movie is that it can toy with our emotions. This one did exactly that.
>
> The entire theater (which was sold out) was overcome by laughter during the first half of the movie, and were moved to tears during the second half. While exiting the theater I not only saw many women in tears, but many full grown men as well, trying desperately not to let anyone see them crying.
>
> This movie was great, and I suggest that you go see it before you judge.”

To detect whether a review is mixed (meaning it contains both positive and negative opinions), we break the review into individual sentences and analyze the sentiment of each one. First, each review is split into sentences. If a review has only one sentence, we skip it since there’s not enough context to judge a mix of sentiments.

Then, each sentence is cleaned (removing noise like punctuation or special characters), and a sentiment classifier is used to predict whether the sentence is positive or negative. If the classifier is highly confident (e.g., more than 70% confident it’s positive, or less than 30% confident, meaning likely negative), we count that sentence accordingly. Sentences that fall in between are considered neutral and ignored for this purpose. Once all the sentences are analyzed, we look at how many are clearly positive and clearly negative. If a review has at least two strongly positive and two strongly negative sentences, and each type makes up at least 30% of the clearly polarized sentences, we mark it as mixed.

In short: a review is considered mixed if it contains a significant and balanced amount of both strongly positive and strongly negative statements. By applying this methodology, we identified that **14.39%** of the reviews are mixed. An example of this can be seen in the next graph for one review.

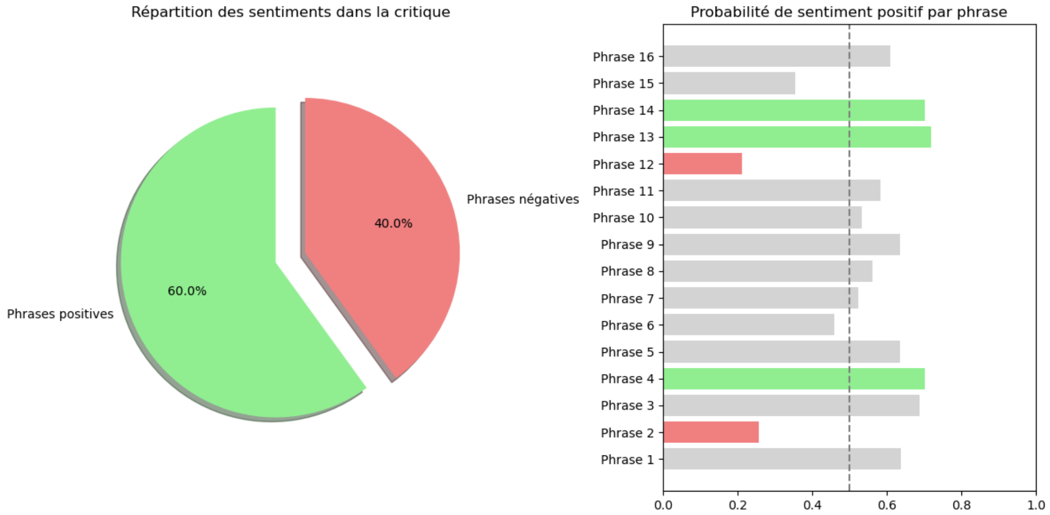

Figure 1: Sentence sentiment analysis in one review

> “Even if you’re a fan of Jean Rollin’s idiosyncratic body of work, you will be caught off guard by this exceptional foray into science fiction territory. For once, there’s not a single diaphanously gowned vampire girl in sight! True to tradition, the budget proved way too tight to realize the director’s vision entirely. Yet this is largely compensated by his obvious love of genre cinema, dedication to his craft and sheer ingenuity. [...]
>
> The final shot offers a particularly heartbreaking variation on that of Chaplin’s *MODERN TIMES* as Elisabeth, approaching complete meltdown by now, and a wounded Robert stumble along the railroad bridge, clumsily clasping each other’s outstretched hands.”

2 Model

The model used in our analysis is a standard supervised sentiment classifier. It consists of the following pipeline:

- Preprocessing: text cleaning, tokenization, and vectorization using TF-IDF.
- Classifier: logistic regression trained on the labeled dataset.

To identify mixed-sentiment reviews, we employed a function based on polarity lexicons. Each review was scored based on the co-occurrence of positive and negative words as it is explained in the introduction. Reviews exceeding a threshold in both categories were flagged as *neutral*.

3 Results

We evaluated the model in two settings:

- Using the full test set (including mixed-sentiment reviews).
- Using a filtered version excluding mixed reviews.

The results of the experiments are presented in the table below: the complete dataset contains 25,000 reviews, while the dataset without mixed reviews consists of 21,402 reviews.

Metric	**Complete Dataset**	**Dataset without Mixed Reviews**
Accuracy	88.3481%	90.7111%
F1 Score	88.3480%	90,7104%

Table 1: Model Evaluation Metrics for the Complete Dataset and the Dataset without Mixed Reviews

Removing mixed reviews has a notable impact on the overall performance: the accuracy increases **by 2.36%** once these ambiguous cases are excluded. This suggests that mixed reviews are significantly harder for the sentiment analysis model to classify correctly.

Further experiments conducted in our notebook show that when evaluating both groups separately, the model achieves an accuracy of 74.29% on mixed reviews, compared to 90.71% on clearly positive or negative reviews—a gap of 16.42%. This highlights the challenge posed by nuanced sentiments and the importance of considering them when assessing model performance. Indeed this supports the hypothesis that model performance is underestimated when ambiguous reviews are included because even humans can't distinguish them—it is just an impossible task.

4 Conclusion

Our findings highlight the importance of treating ambiguous reviews differently during both training and evaluation. When left unaddressed, these reviews can skew performance metrics and lead to misleading model comparisons. Filtering out uncertain cases is not just about making the task easier; it helps separate the tricky parts so that models can be judged more fairly in my opinion.

This perspective also encourages the development of more sophisticated models that go beyond binary sentiment classification. By estimating uncertainty or working along a continuous sentiment scale, future models could deliver richer, more informative predictions. Ultimately, ambiguity is an inherent part of language, and acknowledging it leads to more reliable evaluations and more expressive sentiment analysis tools.