Mini Project ML for NLP: Sentiment Analysis on Movie Reviews

Sofia Vaca Angulo
Master Data and Economics for Public Policy

Abstract

This current paper delves into how the sentiment analysis models have developed via a comparison between traditional machine learning models and a current state-of-the-art large language model (LLM). Utilizing the IMDB movie review dataset introduced by Maas et al. (2011), the subsequent benchmark models were utilized:TF-IDF and Delta-IDF subsequent to SVM or Logistic Regression, LSA and LDA in combination used along with SVM, and Word2Vec-based classifiers. These were later compared with LLaMA 3, a LLM, employed in a zero-shot setting via the Groq API. Results show that while baseline models are resilient and computationally efficient, the LLM is more accurate than them without using feature engineering or training. Analysis also highlights the application of prompt engineering and addresses the trade-off between interpretability, scalability, and practical realization. This comparison sheds light on the abilities of LLMs for real-world sentiment classification problems.

1 Introduction

Ever since the arrival of the internet, web user-generated content has attained huge significance. Individuals nowadays share their views and experiences openly on web forums generating a large amount of text data. Sentiment analysis, in this context, has become an important method for recognizing as well as analyzing public opinions, emotions, and attitudes regarding diverse subjects, products, and services (Birjali et al., 2021). This method can provide great benefit to businesses, policymakers, and researchers by allowing them to track public opinion, aid in decision-making, and more clearly comprehend social attitudes and patterns.

Essentially, sentiment analysis comes down to determining if some text has a positive, negative, or neutral sentiment. Throughout the years, a variety of computational methods have been developed to approach this task, from traditional statistical models to more recent deep learning and language modeling approaches. More specifically, recent advances in Artificial Intelligence and Large Language Models (LLMs) have expanded the possibilities of this type on analysis.

This paper seeks to compare and investigate the performance of both traditional and state-of-the-art models in sentiment classification. Firstly, it reproduces and compares conventional methods reported in Maas et al. (2011), such as TF-IDF, Delta-IDF, Latent Semantic Analysis (LSA), and Latent Dirichlet Allocation (LDA), with each being combined with Support Vector Machine (SVM) or Logistic Regression classifiers. Secondly, the research is complemented by adding other more updated methods, such as zero-shot LLM-based classification and Word2Vec-based embeddings. This experimental sequence enables a valid comparison of previous and current approaches in sentiment analysis.

The experiments proposed are carried out on Maas et al.'s (2011) IMDB Large Movie Review Dataset, a popular benchmark for binary sentiment classification. The report follows the following

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

organization: Section 2 overviews the state of the art for sentiment analysis, both traditional machine learning approaches and recent methods on word embeddings and large language models (LLMs). Section 3 discusses the models selected for implementation, why they were selected, and their theoretical underpinnings. Section 4 explains the dataset used for experimentation. Section 5 explains the execution pipeline, from preprocessing to model evaluation. Section 6 explains the experimental outcomes and compares the performance of every approach. Section 7 concludes the report by providing an overview of the key findings and explaining the limitations of the study

2 State of the Art in Sentiment Analysis

Sentiment analysis is a widely studied task in natural language processing (NLP) which has evolved significantly in the last decades. To approach this task of categorizing the emotional tone of a piece of text, researchers have proposed a wide range of computational methods. As Mao et al. (2024) explain, sentiment analysis methods can be broadly categorized into the following groups:

1. **Lexicon-based approaches**: rely on pre-computed sentiment dictionaries or corpus-based frequency signals.
2. **Traditional machine learning classifiers**: rely on hand-designed features (e.g., bag-of-words or TF-IDF) in combination with classifiers like Support Vector Machines (SVM), Naive Bayes (NB), Logistic Regression (LR), K-nearest Neighbors (KNN), and Decision Trees (DT).
3. **Hybrid approaches**: which combine elements of rule-based and statistical methods.
4. **Deep learning-based classifiers**: including models based on Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Gated Recurrent Units (GRUs), and Long Short-Term Memory networks (LSTMs).
5. **Other approaches**: such as transformer-based models, aspect-based sentiment analysis (ABSA), and transfer learning.

A decade ago, traditional machine learning classifiers and hybrid approaches were at the center of sentiment analysis activities. A prominent example is the work of Maas et al. (2011), who proposed a supervised method to train word embeddings on document-level sentiment labels. The embeddings were also combined with traditional techniques like TF-IDF, Delta-IDF, Latent Dirichlet Allocation (LDA), and Support Vector Machines (SVMs). Their experiments showed that it is possible to incorporate sentiment information in vector representations when supervision is available during training, resulting in embeddings that capture semantic similarity in addition to sentiment polarity. Their findings also highlighted the strength of hybrid models, where combinations of supervised embeddings with linear classifiers like SVMs offered better accuracy. The paper also released a nice and popular benchmark dataset (IMDB).

Sentiment analysis over the past few years has been transformed by advances in deep learning and pre-trained embeddings. Hand-engineered features were substituted with dense vector representations with models such as Word2Vec, GloVe, and fastText that are more capable of capturing semantic relations. These embeddings were then usually fed through deep models such as CNNs or LSTMs for classification.

Most recently, Large Language Models (LLMs) like LLaMA, BERT, RoBERTa, GPT-3.5 and GPT-4, have achieved new benchmarks by enabling zero-shot or few-shot sentiment classification through prompt-based interfaces. These models enable sentiment analysis without task-specific training, relying on contextual intelligence derived from massive corpora.

3 Proposal of model

The core proposal of this project is to evaluate the performance of a Large Language Model (LLM) for sentiment classification, using LLaMA 3 (8B) accessed through the Groq API. The model is applied in a zero-shot classification setting, meaning it predicts the sentiment of movie reviews based entirely on prompt-based instructions, without any additional training on labeled data.

To analyze the capabilities of this modern approach, its performance is compared against a series of benchmark models introduced by Maas et al. (2011), including TF-IDF with SVM, Delta-IDF with

Logistic Regression, and Word2Vec-based classifiers. All models are evaluated on the same dataset (IMDB).

This experiment is motivated by the growing body of research showing that LLMs can perform remarkably well on tasks with minimal supervision, especially when given carefully designed prompts. Unlike traditional models, which require extensive preprocessing, manual feature engineering and explicit training, LLMs can be a more flexible alternative.

In addition, the experiment allows us to discuss broader trade-offs between traditional and modern approaches including interpretability, computational cost, inference time, and data requirements. Given the increasing relevance of LLMs across domains, this comparison offers timely insight into their potential as viable tools for sentiment analysis in practical applications.

4 Data

As mentioned, the data used in this project is the IMDB Large Moview Review Dataset put together by Maas et al. (2011). The dataset was created for binary sentiment classification and was relevant since it contained substantially more data than previous benchmarking datasets. It includes 25,000 labeled movie reviews for training and 25,000 for testing, with a balanced split between positive and negative sentiments in both subsets. The reviews are highly polarized to allow for well-defined sentiment differences also it has a limit number of at most 30 reviews per movie. Along with the labeled data, the dataset also includes a large unlabeled set of reviews for unsupervised or semi-supervised learning experiments. As it can be seen in the following figures, the databases were balanced and the distribution of review lengths in words is very similar among both test and train data.

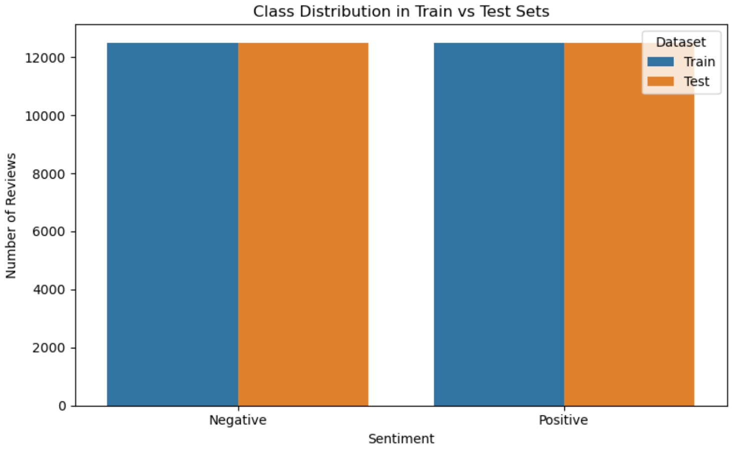

Figure 1: Distribution of sentiment labels.

5 Implementation of the model

5.1 Implementation: Benchmark models

For all the benchmark model, I applied the preprocessing technique recommended by Maas et al. (2011), with the goal of retaining as much emotional and linguistic richness of the original text. The aim was to retain expressive properties in a manner such that the models can learn from the way human writers write. This is why they suggest not to remove stopwords like "not," "is," and "was" because these are most likely to express sentiment, especially when they include negation, just like emoticons, punctuation and exclamation marks. Also, they suggest not to perform stemming and lemmatization. Therefore, the only preprocessing operation performed was HTML tag substitution. As the IMDB dataset was already pre-split between training and test sets, splitting and sampling was not necessary here.

The following benchmark models were employed:

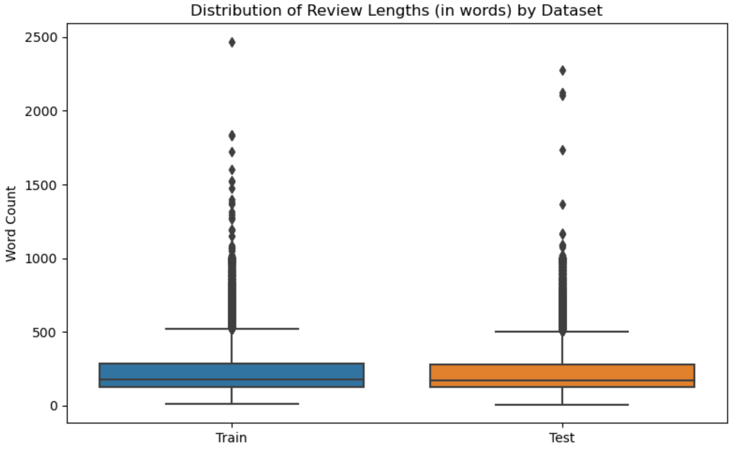

Figure 2: Distribution of review lengths in words

- **Bag of Words + SVM Classifier with TF-IDF**: TF-IDF vectors were computed to express documents in terms of term importance, and a linear SVM was employed for classification.
- **BoW + Logistic Regression using Delta-IDF**: Delta-IDF weights were applied to the BoW matrix to emphasize more discriminative words across classes; the weighted vectors were then predicted using logistic regression due to computational constraints with SVM.
- **Latent Semantic Analysis (LSA) + SVM**: LSA was applied as a dimensionality reduction technique on the TF-IDF matrix using SVD, projecting documents into a lower-dimensional semantic space, then classified with an SVM.
- **Latent Dirichlet Allocation (LDA) + SVM**: LDA models documents as mixtures of latent topics and maps each review to a vector of topic proportions, which are then used as input to an SVM classifier.
- **Word2Vec + SVM**: Word2Vec is trained on the training data, and each review is represented by the mean of its word vectors; these semantic vectors are classified using an SVM, although the averaging weakens sentiment-specific signals like emphasis or word order.

5.2 Implementation: LLM for sentiment analysis

The configuration of the large language model (LLM) was different from that of the other models. While the others were both trained and tested on the various datasets, the LLM was utilized for inference alone in the test set. This is because large language models, by nature, cannot be trained further since they perform zero-shot classification solely based on prompt instructions.

I ran the model with the LLaMA 3 (8B) architecture via the Groq API. Due to daily usage constraints and the prohibitively computational cost of querying an LLM, it was not feasible to run predictions on all test data. Due to this, analysis was run on 400 randomly sampled test reviews to facilitate a comparison within API limitations.

In this setup, every review was passed through the LLM with a sentiment classification prompt, and the model produced a predicted label (positive or negative). The output was stored and matched against the true labels in the data.

6 Analysis of results

Table 1 presents the performance of the different models tried. The LLM model (LLaMA 3) performs better than the conventional methods with 93% accuracy. The top-performing benchmark models have an accuracy of 87%, achieved by TF-IDF + SVM and Delta-IDF + Logistic Regression, which aligns with the competitive baselines established by Maas et al. (2011).

As expected, LDA + SVM and Word2Vec + SVM did worse, with accuracies of 73% and 78%, respectively. This is consistent with the nature of these methods. LDA, as a topic modeling technique, is very good at revealing hidden thematic structure in text but these are not necessarily sentiment polarity aligned. Therefore, when the topic proportions are used as classification features, they may overlook the emotional tone of the review. Word2Vec, however, summarizes reviews through averaging word embeddings, which excels at capturing global semantic relations but tends to ignore sentiment indications (especially word order or stress-based ones). This estimation is likely the cause of its lower accuracy compared to TF-IDF-based models.

Table 1: Performance comparison of sentiment analysis models on the IMDB dataset.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
TF-IDF + SVM	0.87	0.87	0.87	0.87
Delta-IDF + Logistic Regression	0.87	0.87	0.87	0.87
LSA + SVM	0.85	0.85	0.85	0.85
LDA + SVM	0.73	0.73	0.73	0.72
Word2Vec + SVM	0.78	0.78	0.78	0.78
LLM (LLaMA 3, zero-shot)	0.93	0.93	0.92	0.93

The competitiveness of the LLM's performance can also be seen from the confusion matrix given below. Its high precision further shows the ability of modern language models to well read sentiment, even without fine-tuning. One limitation of this test is, however, that fewer observations were used to test the LLM than were used to test the other models. As mentioned, with usage quota per day and cost of computing API calls, only a subset of the test set (400 reviews) was examined. This introduces a limitation on the comparability of results, as the LLM's performance was evaluated under different conditions.

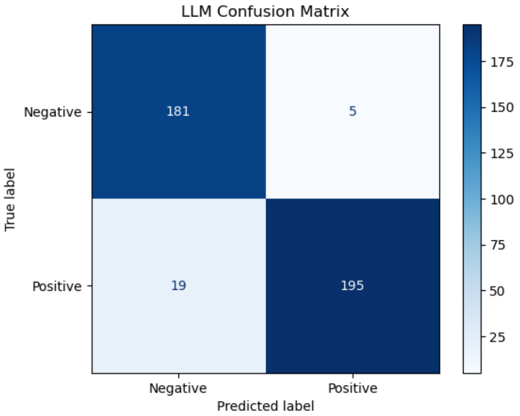

Figure 3: LLM Confusion Matrix

In addition, when comparing Figure 4 and Figure 5, it can be seen that although both models perform similarly overall, the Delta-IDF + Logistic Regression model makes fewer incorrect predictions in both classes. This suggests that Delta-IDF weighting can, in fact, produce a more sentiment-sensitive representation than standard TF-IDF, which aligns closely to the proposal of Maas et al. (2011). Finally, Logistic Regression—albeit computationally lighter—performs similarly to SVM, which validates its applicability when using sentiment-aware features.

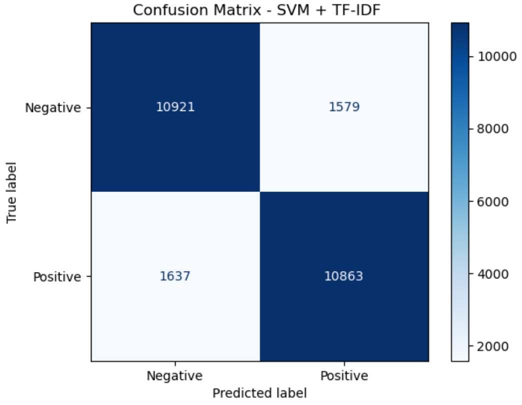

Figure 4: SVM + TF-IDF Confusion Matrix

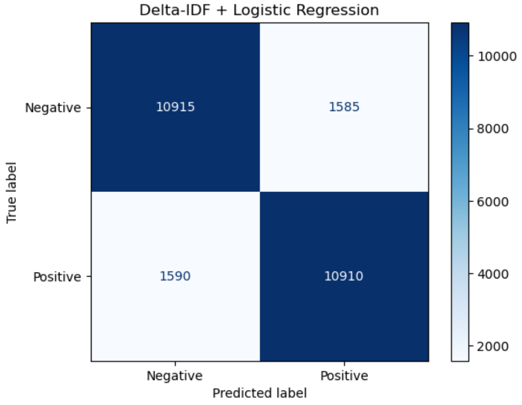

Figure 5: Delta-IDF + Logistic Regression Confusion Matrix

7 Conclusion

Sentiment analysis has had an evolution of methods in the last decades. It has transformed from simple lexicon-based to more traditional machine learning methods to more deep models such as LLMs and embeddings. All these methods have their own strengths and trade-offs.

This project showed hybrid methods such as TF-IDF or Delta-IDF combined with SVMs can be fast, computationally efficient, and interpretable but can have a lower accuracy. In contrast, LLMs like LLaMA 3 outperformed these models in terms of accuracy, without having the need to do any manual feature engineering and with a straightforward implementation. However, LLMs are slower to execute because of API calls, require more computational power, and can cause interpretability challenges as their internal workings are not easily explained.

Another important takeaway from this project is the importance of prompt engineering. The way a prompt instruction is written can significantly affect how an LLM responds. In one version of this experiment, a first non so specific prompt made the model return unexpected outputs like -1, 2, and 7 instead of just binary sentiment labels. This shows that LLMs are sensitive to prompt formulation, and how careful design is essential to ensure a high performance.

It is also important to point out one of the main limitations of the experiment: the LLM-based approach was only feasible for a significantly smaller dataset (500 reviews) compared to the classical models, because there were API quota and computational power constraints. This variation limits comparison across results, and as such, it is impossible to definitively state about one being categorically better than the other. The results must be interpreted as exploratory and not definitive.

Overall, LLMs open new and promising avenues for sentiment analysis, especially when combined with good prompts. Nevertheless, the traditional models are still there, particularly when there are few resources or when interpretability and transparency are essential.

Furthermore, as future task, it would be interesting applying these methods to understand sentiment behind political opinions or debates.

References

[1] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., & Potts, C. (2011) Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150.

[2] Mao, Y., Wang, L., Zhang, H., Yu, J., & Zhang, X. (2023) A comprehensive survey on sentiment analysis: State of the art and emerging trends. *Information Fusion* **90**, pp. 230–270.
`https://doi.org/10.1016/j.inffus.2022.09.001`

[3] Birjali, M., Beni-Hssane, A., & Erritali, M. (2021) A comprehensive survey on sentiment analysis: Approaches, challenges, and trends. *Knowledge-Based Systems* **226**, 107134.
`https://doi.org/10.1016/j.knosys.2021.107134`