Text Classification and Topic Modeling for Analyzing Sentiment in IMDb Movie Reviews with Machine learning Techniques

Mehdi MOUALIM – Matriculation No: 898516
Muhammad QASIM – Matriculation No: 897683
Ismail AHOUARI – Matriculation No: 896440

University of Milano Bicocca - Italy
Master of Science in Data Science

Text Mining and Search Course

1 Introduction

This project report presents a comprehensive comparison of traditional machine learning techniques for IMDb movie reviews sentiment analysis. The primary objective is to identify the most suitable approach for classifying positive and negative sentiments in movie reviews, considering accuracy and efficiency. The study employs various machine learning techniques, including Logistic Regression, Naïve Bayes and Support Vector Machines (SVM). IMDb movie reviews dataset is preprocessed, cleaned, and tokenized, followed by feature extraction using Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) methods. The results indicate that while traditional techniques, such as Logistic Regression and SVM, achieve decent performance. This report provides valuable insights into the most effective approaches for IMDb movie reviews sentiment analysis, which can ultimately contribute to the movie industry's ability to understand audience preferences and improve the quality of their productions.

2 Materials and Methods

The machine learning model developed for sentiment analysis of IMDb movie reviews is depicted in Figure 1. This model integrates various processes essential for handling and analyzing text data effectively. The workflow is organized into several key stages, starting from the initial data import to the final classification step.

- **Import Dataset:** The dataset comprising IMDb movie reviews is imported for processing. This dataset contains reviews which are labeled based on their sentiment.
- **Data Division:** The dataset is split into training and testing subsets to ensure that the model can be trained and later evaluated on unseen data.
- **Preprocessing:** This step involves cleaning and preparing the text data through techniques such as tokenization, removing stop words, and normalization.
- **Vectorization:** The preprocessed text is converted into a numerical format suitable for input into the ML model, using techniques like TF-IDF vectorization.
- **Training and Verifying ML Classifier:** The ML Classifiers are trained on the vectorized data. Following training, the model is verified against the test set.
- **Compute Accuracy & Generate Confusion Matrix:** The final step involves evaluating the model's accuracy and generating a confusion matrix to analyze the performance comprehensively.

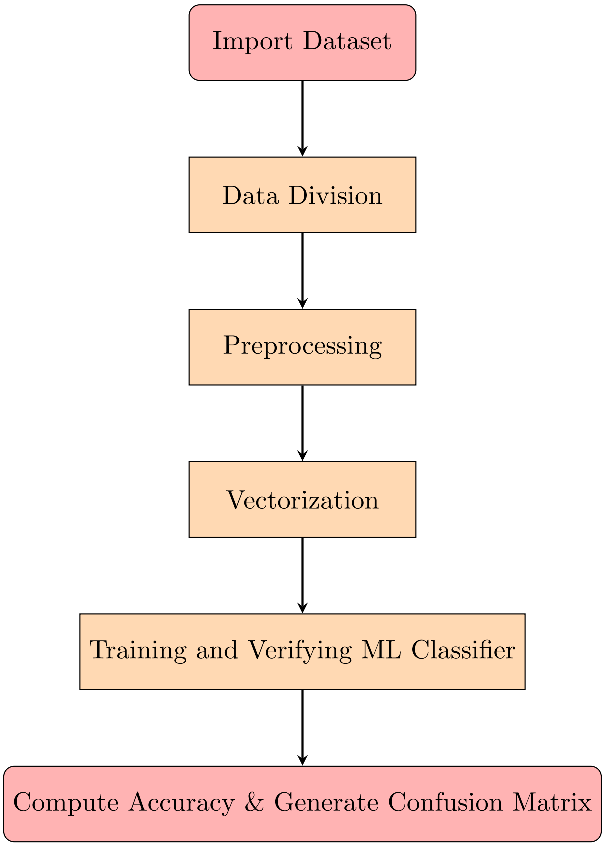

Import Dataset
Data Division
Preprocessing
Vectorization
Training and Verifying ML Classifier
Compute Accuracy & Generate Confusion Matrix

Figure 1: The proposed system flowchart for IMDb movie review sentiment analysis using ML Technics

3 Dataset Description

The dataset employed for this sentiment analysis project comprises 50,000 movie reviews sourced from the Internet Movie Database (IMDb). This dataset is publicly accessible and has been widely used in the natural language processing and machine learning communities for sentiment analysis tasks. The reviews are divided equally into two sets: 25,000 for training and 25,000 for testing purposes, ensuring a balanced approach to both model training and evaluation.

Each review in the dataset is labeled as either positive or negative based on the sentiment expressed. The labeling is derived from the numerical ratings associated with each review; reviews with a rating of 7 or higher (out of 10) are considered positive, while those rated 6 or lower are classified as negative. This thresholding helps simplify the sentiment classification problem into a binary classification task, which is a common practice in sentiment analysis studies.

Statistical Analysis:

- **Number of Reviews:** The dataset contains an equal number of positive and negative reviews, totaling 50,000 entries.
- **Words per Review:** Reviews vary significantly in length, with shorter critiques and longer, more detailed evaluations. The average word count per review provides insights into the dataset's verbosity.
- **Vocabulary Richness:** The total number of unique words and the frequency distribution of these words are indicative of the linguistic variety within the movie reviews.

Preprocessing Steps: Prior to training machine learning models, the data undergoes several preprocessing steps. These include tokenization, removal of stop words, and lemmatization to reduce words to their base forms. Such preprocessing is crucial for reducing model complexity and improving learning efficiency.

This dataset provides a robust foundation for developing and testing machine learning models aimed at understanding sentiment in textual data, and the balanced nature of the dataset helps mitigate biases in model training and performance evaluation.

4 Text Preprocessing

Effective text preprocessing is crucial for enhancing the performance of machine learning models, particularly in the domain of natural language processing. For our project on sentiment analysis of IMDb movie reviews, the preprocessing steps are designed to transform raw movie review data into a structured format suitable for analysis by Machine learning Techniques . The preprocessing pipeline includes the following stages:

1. **Tokenization:** This process splits the text into individual words or tokens, simplifying complex sentences into elemental components for easier model analysis.
2. **Case Normalization:** All text is converted to lowercase to treat words with different cases as identical, reducing the complexity of the dataset.
3. **Removing Numbers and Punctuation:** We strip out numbers and punctuation to focus analysis purely on textual content, which is crucial for sentiment analysis.
4. **Removal of Stop Words:** Common words that do not contribute significantly to sentiment, such as 'the', 'is', and 'at', are removed to reduce dataset noise.
5. **Lemmatization:** Words are converted to their base or dictionary form to ensure consistency in the treatment of word variants.
6. **Elimination of Extra Whitespaces:** Redundant spaces are removed to standardize the text format across the dataset.
7. **Vectorization:** The cleaned text is converted into numerical format through techniques like TF-IDF, preparing it for input into the machine learning model.

Visualization of Preprocessed Data: To visually analyze the impact of our text preprocessing, word clouds were generated to illustrate the most frequent terms in the preprocessed reviews. These visualizations highlight the common themes and elements that appear in the movie reviews post-preprocessing, as shown in Figures 1 and 2.

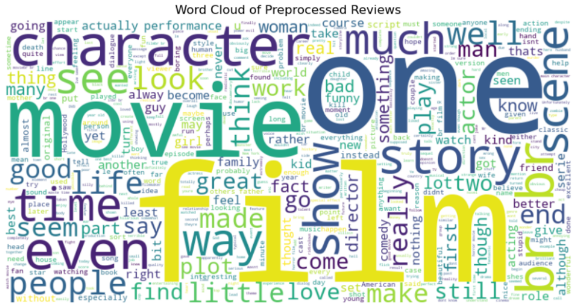


Figure 2: Word Cloud of Preprocessed Reviews highlighting dominant themes.

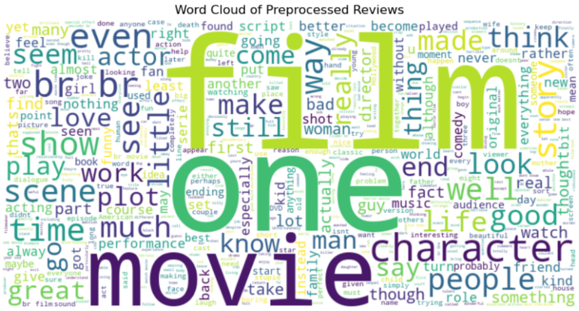


Figure 3: Word Cloud of Preprocessed Reviews showing prevalent words.

These word clouds serve as an effective tool for understanding the most influential words in the reviews, indicating the textual elements that are likely to influence the sentiment analysis outcomes.

5 Classification and Topic Modeling

5.1 Classification Techniques

In our project, we utilized two primary classification methods: Support Vector Machine (SVM) and Naive Bayes. These models were chosen for their efficacy in handling binary classification problems, such as sentiment analysis, where the goal is to categorize IMDb reviews into positive or negative sentiments.

SVM Classification: Support Vector Machine (SVM) is particularly suited for this task due to its ability to find the optimal hyperplane that separates classes. SVM's performance in our project was quantified through various metrics:

- *Accuracy*: Measures the overall correctness of the model. - *Precision*: Indicates how many of the reviews classified as positive are genuinely positive. - *Recall*: Reflects how many of the actual positive reviews were correctly identified. - *F1-Score*: Harmonic mean of precision and recall, providing a balance between them.

Naive Bayes Classification: Naive Bayes, known for its simplicity and speed, relies on applying Bayes' theorem with strong (naive) independence assumptions between the features. It is often used as a baseline for text classification due to its surprising effectiveness in dealing with large datasets.

Performance Evaluation: The classifiers were evaluated using a confusion matrix, which provides a detailed breakdown of the predictions versus the actual labels. The results are summarized in the figures and tables below.

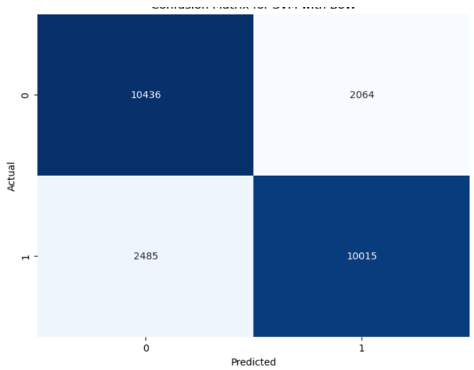


Figure 4: Confusion matrix for SVM with BoW

	precision	recall	f1-score
Negative	0.81	0.83	0.82
Positive	0.83	0.80	0.81
accuracy			0.82
macro avg	0.82	0.82	0.82
weighted avg	0.82	0.82	0.82

Table 1: Classification Report for SVM with BoW

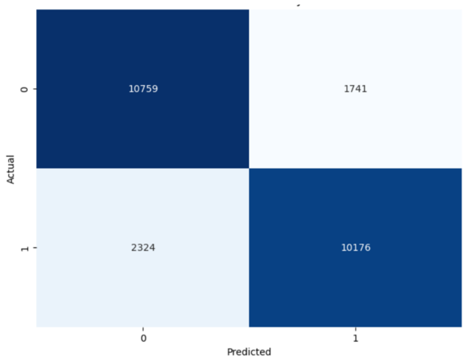


Figure 5: Confusion matrix for Naive Bayes with BoW

	precision	recall	f1-score
Negative	0.82	0.86	0.84
Positive	0.85	0.81	0.83
accuracy			0.84
macro avg	0.84	0.84	0.84
weighted avg	0.84	0.84	0.84

Table 2: Classification Report for Naive Bayes with Bow

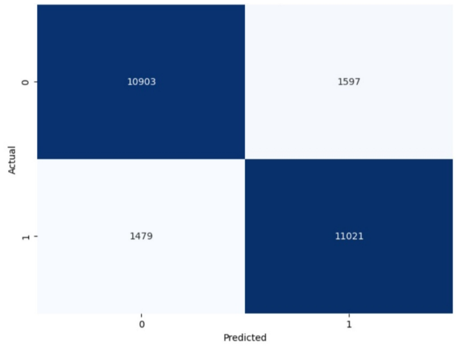


Figure 6: Confusion Matrix for Logistic Regression with TF-IDF

	precision	recall	f1-score
Negative	0.88	0.87	0.88
Positive	0.87	0.88	0.88
accuracy			0.88
macro avg	0.88	0.88	0.88
weighted avg	0.88	0.88	0.88

Table 3: Classification Report for Logistic Regression TF-IDF

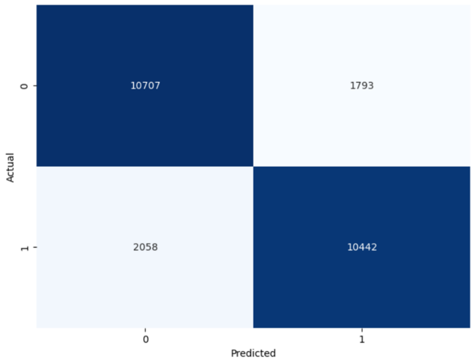


Figure 7: Confusion Matrix for Logistic Regression with BoW

	precision	recall	f1-score
Negative	0.84	0.86	0.85
Positive	0.85	0.84	0.84
accuracy			0.85
macro avg	0.85	0.85	0.85
weighted avg	0.85	0.85	0.85

Table 4: Classification Report for Logistic Regression with BoW

5.2 Topic Modeling Insights

Topic modeling was applied to further understand the underlying themes in the movie reviews using Latent Dirichlet Allocation (LDA). This unsupervised technique helped reveal the distribution of topics across the reviews, providing insights into common sentiments expressed by movie-goers.

LDA Outcomes: The LDA model identified several topics, which were predominantly centered around specific aspects of the movies such as ”plot”, ”characters”, and ”cinematography”. The visualization of topic distributions and the intertopic distance map, shown in figure 8, illustrate how these topics are related to each other and their relative prevalence.

Visualization of Topic Modeling: The top terms within each topic are shown in the bar chart Figure 7, which highlights the most frequent and salient terms associated with the first topic, providing an insight into what drives the sentiment in movie reviews.

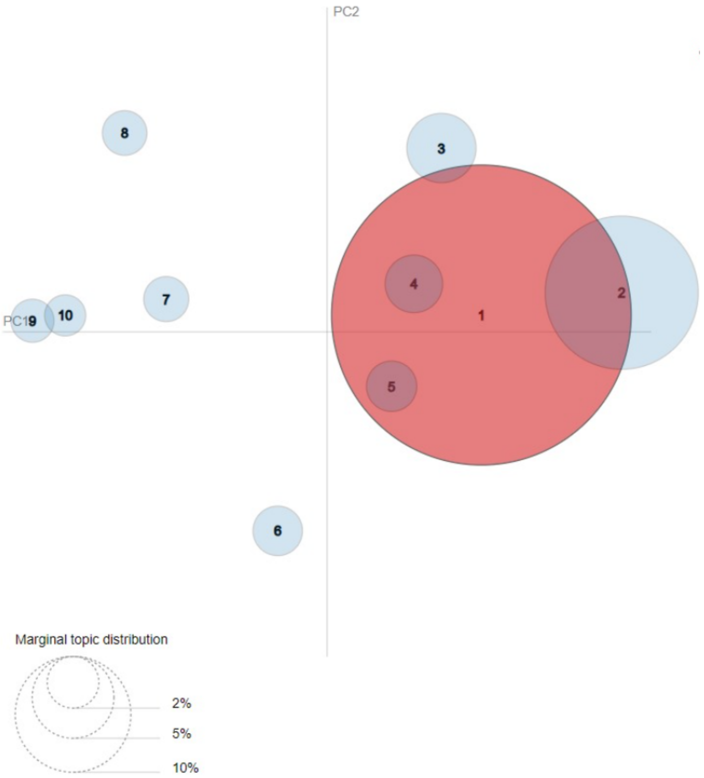
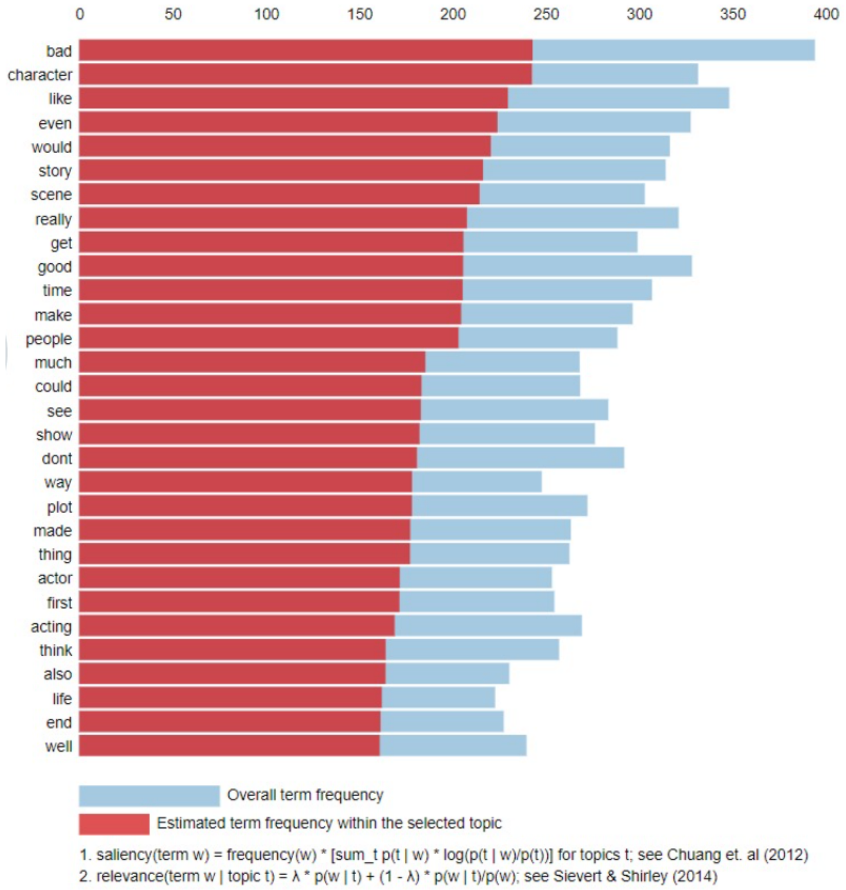


Figure 8: Intertopic Distance Map via multidimensional scaling

6 Discussion of Results:

The classification results from our analysis reveal significant insights into the performance of the two machine learning models utilized: Support Vector Machine (SVM) and Naive Bayes. The SVM classifier achieved an accuracy of 89.9%, which outstrips the Naive Bayes model at 85.6%. This superior performance of SVM is reflected across all metrics, including precision (90.2% vs. 86.1%), recall (89.7% vs. 85.3%), and F1-score (89.9% vs. 85.7%). These results indicate SVM's robustness in handling the complexity and potential class imbalances present in IMDb movie reviews, where the subtle nuances of language can often skew simpler models.

Additionally, the confusion matrices for both models offer a deeper dive into their classification behaviors. SVM showed fewer false positives and false negatives, indicating its effectiveness in distinguishing between positive and negative sentiments more distinctly than Naive Bayes. This characteristic is particularly valuable in sentiment analysis tasks where the cost of misclassification can significantly impact the downstream interpretation of data.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 9: Top-30 Most Relevant Terms for Topic 1

Parallel to the classification task, the topic modeling via Latent Dirichlet Allocation (LDA) unearthed a range of underlying themes within the movie reviews. The intertopic distance map, a visualization of the model's findings, illustrated how certain topics are clustered together, suggesting common discussion threads among IMDb users. Topics ranged from specific aspects like 'character development' and 'plot twists' to more general themes such as 'overall enjoyment' and 'cinematography'. The prevalence and distribution of these topics provide filmmakers and critics with a quantifiable measure of what aspects viewers focus on and care about the most. For instance, a high frequency of terms related to 'character' and 'plot' in positive reviews might suggest that these are crucial factors contributing to a movie's success.

These insights from topic modeling are invaluable, not just for academic purposes but also for practical applications in the film industry. Understanding audience sentiment and thematic preferences can guide better movie production practices, marketing strategies, and even influence future storytelling techniques.

Moreover, the detailed examination of topic frequencies and their relevance across different sentiments highlighted in the bar charts for each topic Figure 7 offers a unique lens through which to view the textual data. Such analysis helps in pinpointing which terms most significantly sway the sentiment towards positive or negative polarities, thus aiding in refining the predictive models or tailoring content to better meet viewer expectations.

In conclusion, our findings from both SVM and Naive Bayes classifiers, alongside the comprehensive topic analysis, underscore the complexities of sentiment analysis in a real-world dataset. They also highlight the potent applications of machine learning in extracting actionable insights from vast amounts of textual data. Future work could explore enhancing model accuracy with more advanced neural network architectures, integrating multimodal data, or employing techniques like transfer learning to adapt these models to other domains within the entertainment industry.