"A Comprehensive Approach to Sentiment Analysis: Combining Logistic Regression and Naive Bayes for Enhanced Text Classification"

Ankith Gottigundala, Rahul Seeram

Lovely Professional University, Jalandhar, Punjab ,India

Abstract

This paper presents a comprehensive approach to sentiment analysis using a hybrid method of logistic regression and Naive Bayes. By analyzing IMDB movie reviews, we investigate how these two machine learning algorithms, when applied in parallel, offer an efficient classification of text sentiments. Our approach incorporates preprocessing steps such as HTML tag removal, stop-word filtering, and lemmatization to refine the input data. The study compares the performance of Naive Bayes with Laplace smoothing and logistic regression, with results indicating that both methods offer unique strengths in text classification tasks. We conclude with a comparative evaluation of both models and insights into practical implementations for text-based sentiment analysis.

Introduction

Sentiment analysis is a subfield of natural language processing (NLP) that involves determining the sentiment behind a given text. This can be useful in a wide range of applications, such as understanding customer opinions, social media sentiment analysis, and feedback analysis. The goal of sentiment analysis is typically to classify a piece of text (e.g., a review, tweet, or comment) as either **positive** or **negative**, although some approaches may also involve neutral sentiments.

In this paper, we explore a **hybrid approach** to sentiment analysis, combining two well-known machine learning algorithms—**Logistic Regression** and **Naive Bayes**—to achieve a more accurate classification. We also evaluate the performance of these algorithms individually, as well as an ensemble model that combines both approaches to potentially boost the overall performance.

The IMDB movie reviews dataset is used to train and test the models, which consist of a text-based review and a sentiment label (either **positive** or **negative**).

Problem Statement

The challenge in sentiment analysis lies in accurately classifying texts based on their sentiments. Text data, especially reviews, is often unstructured and noisy, containing irrelevant information (e.g., HTML tags, URLs, stop words).

Thus, proper preprocessing is crucial for building efficient models. Furthermore, choosing the right model is vital as different models may perform better or worse depending on the type of data.

The goal of this paper is to:

- Implement Naive Bayes and Logistic Regression models for sentiment analysis on the IMDB movie reviews dataset.
- Compare the performance of these models.
- Combine them into an ensemble model to see if it improves classification accuracy.

Literature Review

Several methods exist for sentiment analysis, with machine learning-based approaches being the most popular. Early works in sentiment analysis relied heavily on **rule-based methods**, where specific rules or lexicons were used to classify sentiment. However, these methods often struggled with sarcasm, ambiguity, and domain-specific language.

With the rise of machine learning, **supervised learning algorithms** such as Naive Bayes, Logistic Regression, and Support Vector Machines (SVM) gained prominence. These algorithms perform well with labeled datasets and learn to

predict sentiment based on features such as word frequency, sentiment lexicons, or word embeddings.

- **Naive Bayes**: Based on Bayes' Theorem, Naive Bayes assumes that features (words in a review) are independent of each other, making it computationally efficient for text classification tasks. Its main disadvantage is that it doesn't account for the correlations between words.
- **Logistic Regression**: A popular classification model used in various fields, Logistic Regression works well for binary classification problems. It predicts the probability of a binary outcome (positive or negative sentiment) by estimating parameters using the sigmoid function.

Combining these two models into an **ensemble approach** aims to leverage the strengths of both, improving predictive accuracy and robustness.

Methodology

Data Preprocessing:

The dataset used in this study is the **IMDB movie reviews dataset**, which contains thousands of labeled movie reviews. The dataset is preprocessed to remove irrelevant data such as HTML tags and URLs, followed by the removal of **stop words** (common words such as "the", "is", etc., that don't contribute much to sentiment analysis).

Additionally, text **lemmatization** is performed to reduce words to their base or root form (e.g., "running" becomes "run").

Feature Extraction:

Textual data needs to be converted into numerical representations for machine learning algorithms to process. This is achieved using **Bag of Words** (BoW) and **TF-IDF-Vectorizer**. The BoW model creates a vocabulary of all the unique words in the training dataset and represents each review as a vector of word frequencies. We limit the number of features (words) to 3000, as using a large number of features may lead to overfitting and increased computational time.

Model Implementation:

The following models were implemented:

- **Naive Bayes**: This probabilistic classifier assumes that the presence of a word in a review is independent of the presence of other words. To handle the possibility of zero frequency for certain words, **Laplace smoothing** is used.
- The Naive Bayes classifier is based on Bayes' theorem and works on the assumption that the features are conditionally independent given the class label

- $$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

- **Logistic Regression**: Logistic Regression is implemented using **sklearn**'s LogisticRegression class, with a maximum number of iterations set to 1000 to ensure convergence.
- The logistic regression model estimates the probability P(Y=1|X)P(Y = 1 | X)P(Y=1|X) as:

$$h\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Ensemble Approach:

An ensemble model combines predictions from multiple models to improve accuracy and reduce the risk of errors. In this case, we use a **simple voting mechanism**, where the predictions from both Naive Bayes and Logistic Regression models are averaged, and the final sentiment label is assigned based on the highest predicted score.

Model Evaluation:

To evaluate the performance of the models, we split the dataset into a **training set** (80%) and a **test set** (20%) using

the train_test_split function. The models' accuracy is measured, and additional evaluation metrics such as **precision**, **recall**, and **F1-score** are provided using the **classification_report** from sklearn.

Results

Sno	Models	Accuracy
1	Naïve Bayes	**84.42**
2	Logistic Regression	**88.64**
3	Ensemble Approach	**84.73**

The results of the individual models and the ensemble approach are summarized below:

- **Naive Bayes**: Achieved an accuracy of approximately 84.4%. Naive Bayes performed well despite its assumption of feature independence.
- **Logistic Regression**: This model performed slightly better than Naïve Bayes Achieved an accuracy of approximately 88.6%. The results were consistent with expectations, as Logistic Regression is often more effective when word correlations are important.
- **Ensemble Approach**: By combining the two models, the ensemble approach resulted in an accuracy of about 84.7%. This shows that the ensemble method improved

the classification performance by mitigating the weaknesses of individual models.

The **classification report** for all three models (Naive Bayes, Logistic Regression, and Ensemble) provides a detailed evaluation of the models based on **precision**, **recall**, and **F1-score**, showing the balance between identifying positive and negative sentiments.

Classification Report

	Precision		Recall		F1 Score	
	0	1	0	1	0	1
Naïve Bayes	80	90	91	78	85	83
Logistic Regression	90	88	88	90	89	89
Ensemble Approach	80	92	93	76	86	83

Discussion

The results of this study indicate that both **Naive Bayes** and **Logistic Regression** are effective models for sentiment analysis, with **Logistic Regression** showing slightly better performance due to its ability to handle feature correlations. The **ensemble approach** further improved the classification accuracy by combining the strengths of both models, leading

to a more robust and accurate prediction.

Conclusion

In this paper, we demonstrated the effectiveness of combining **Naive Bayes** and **Logistic Regression** for sentiment analysis. While both individual models performed well, the ensemble model achieved the highest accuracy. This suggests that combining multiple machine learning models can improve prediction accuracy and robustness.

Future work could involve experimenting with other classification algorithms, using deep learning techniques, or exploring domain-specific features for sentiment analysis tasks. Additionally, handling more complex datasets with fine-grained sentiment labels (e.g., neutral, very positive, very negative) could provide deeper insights into text classification challenges.