EVALUATING MACHINE LEARNING MODELS IN SENTIMENT ANALYSIS: A COMPARATIVE STUDY ON YELP REVIEWS

Saadeddine Yassine Mohamad Daaboul Zein Khamis

Abstract

Sentiment analysis is essential to comprehending client feedback and improving customer service and corporate strategy. The increasing number of online reviews, particularly on Yelp and other platforms, makes getting insights from user feedback more important than ever. This project aims to harness machine learning to accurately categorize Yelp reviews into positive or negative sentiments, addressing the need for businesses to effectively understand and respond to customer opinions. By utilizing a dataset of 10,000 Yelp reviews, with features including text, user ratings, and review counts, the project explores various machine learning models to determine the most effective approach for sentiment classification. It involves preprocessing the data, extracting features using TF-IDF vectorization, and employing various ML models such as Naive Bayes, Logistic Regression, SVM, Random Forest, and Decision Tree. These models are trained and evaluated using metrics including accuracy, precision, recall, and F1-score. Among these, the Logistic Regression and SVM models demonstrate notable performance, achieving an accuracy of 84% and a weighted F1-score of 83%.

Contents

1 INTRODUCTION

Given the massive volume influx of raw text data on media, and more than 260 million reviews on Yelp as of 2004 [1], automated sentiment analysis is ever more pronounced. In the realm of machine learning and natural language processing, sentiment analysis stands out as a pivotal tool for interpreting vast amounts of unstructured textual data. The challenge, however, lies in effectively processing and extracting meaningful insights from such data, which is often loaded with emotion, informal language, sarcasm, and context-specific nuances. This project specifically targets Yelp reviews, aiming to differentiate the underlying sentiments expressed by users. The objective is to accurately classify these sentiments into positive or negative categories, thereby providing valuable insights into customer satisfaction and areas of improvement for businesses. To achieve this, we employ various machine learning models, each with its unique strengths and approaches to handling textual data.

2 DATASET DESCRIPTION

In 2017, Yelp compiled and released a database of reviews published on their platform. The database contains 10,000 data points, and 10 feature columns. The features consist of: (1) *business_id* the unique id of the business to which the review is attributed to, (2) *date* the date on which the review was published, (3) *review_id* the unique id of the review published, (4) *stars* the number of stars given to the review by users, (5) *review* the review given by the user, (6) *type* the type of text entered – Review, (7) *user_id* the unique user id of the user publishing the review, (8) *cool* the number of cool votes that the review received, (9) *useful* the number of useful votes that the review received, and (10) *funny* the number of funny votes the review received. Therefore, the database will be used to predict whether a given review is (1) good or (0) bad.

3 METHODOLOGY

3.1 Overview

In this study¸we evaluated various models, including Naive Bayes, Logistic Regression, Support Vector Machine (SVM), Random Forest, and Decision Tree, to determine their effectiveness in sentiment analysis. The experimental setup was designed to systematically evaluate and compare the performance of various machine learning models in sentiment analysis of Yelp reviews. The process involved preprocessing the dataset of 10,000 Yelp reviews, transforming the textual data into a machine-readable format using TF-IDF vectorization, and then applying different classification models. Each model was assessed using a range of metrics to determine its effectiveness in accurately classifying reviews into positive or negative sentiments. The setup aimed to identify the most suitable model for sentiment analysis in the context of Yelp data. As such, our goal is

to predict the sentiment bases solely on the review itself and no other feature.

The methodologies chosen for this study were grounded in natural language processing and machine learning. The primary focus was on text preprocessing, which involved cleaning, normalizing, and vectorizing the Yelp reviews. The TF-IDF vectorization technique was employed to convert text data into numerical form. The algorithms selected for sentiment classification included Naive Bayes, Logistic Regression, SVM, Random Forest, and Decision Tree. Each of these models has distinct characteristics: Naive Bayes is known for its simplicity and speed, Logistic Regression for handling binary classification effectively, SVM for its robustness in high-dimensional spaces, Random Forest for its ensemble approach in improving accuracy, and Decision Tree for its interpretability and ease of use. The choice of these models allowed for a comprehensive comparison across different algorithmic approaches in sentiment analysis.

3.2 Data Preprocessing

The data preprocessing involved intricate techniques to refine the Yelp reviews for analysis. Firstly, the text data was cleaned by removing punctuation and converting all characters to lowercase to ensure uniformity. Then, stop-words in the English language, which are common words adding little semantic value, were removed from the text. Also, we apply lemmatization to the text to reduce words to their base or dictionary form. For example, "running" becomes "run," and "better" becomes "good." This process helps in normalizing the text data and can be particularly useful for reducing the feature space (number of unique words) and for dealing with different forms of the same word. The crucial step was the application of TF-IDF vectorization. This method converted the cleaned text into numerical vectors by calculating the frequency of each word in a review relative to its frequency across all the reviews, and assigning higher weights to words unique to a review. These steps were crucial in extracting meaningful features from the text, and enabled the models to analyze the sentiments more effectively.

3.3 Training and Evaluation Metrics

Training and Evaluation
The training and evaluation of the models followed a 70-30 train-test split of the data, ensuring a significant representation for both training and testing phases. This split allowed for robust model training while providing a substantial test set to evaluate model performance. The evaluation metrics used were accuracy, precision, recall, and F1-score. Detailed accuracy graphs and tables comparing these metrics across different models will be included. These visuals serve to clearly illustrate each model's performance, providing insights into their effectiveness in sentiment analysis. The comparative analysis is essential to understanding the performance of each model.

Evaluation Metrics
The models' performance was evaluated using several key metrics, including:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{Precision} = \frac{TP}{TP+FP}$$
$$\text{Recall} = \frac{TP}{TP+FN} \qquad \text{F1-score} = \frac{2(precision+recall)}{precision+recall}$$

Accuracy: Measures the proportion of correctly predicted sentiments out of all predictions. It's a primary indicator of overall model performance.

Precision: Indicates the proportion of positive identifications that were actually correct. Precision is crucial when the cost of falsely identifying a review as positive (when it is not) is high, which could mislead understanding customer satisfaction.

Recall (Sensitivity): Assesses the proportion of actual positives that were correctly identified. Recall is key in scenarios where missing out on true positive sentiments (i.e., failing to identify a genuinely positive review) could lead to overlooking valuable customer feedback.

F1-Score: A harmonic mean of precision and recall, providing a balance between these two metrics. It is useful in balancing precision and recall, particularly important in datasets where there might be an imbalance between positive and negative reviews.

These metrics collectively offer a comprehensive understanding of the models' strengths and weaknesses in sentiment classification.

4 PRESENTATION OF RESULTS AND DISCUSSION OF FINDINGS

The results indicate that Logistic Regression and SVM performed best with a test accuracy of 0.84. Their high precision and recall in both classes suggest effective handling of linear separability in text data. Naive Bayes, despite its simplicity, showed high precision in the negative class but lower recall, indicating a tendency to miss negative sentiments. Random Forest and Decision Tree, being non-linear models, had moderate success, with Random Forest outperforming Decision Tree, likely due to its ensemble approach reducing overfitting. The results reflect the varying capabilities of each model in managing the complexities of sentiment classification in text data. It is worthy to note that overall, Logistic Regression performs best although it has the same accuracy and similar metrics to the Support Vector Machine, however, the Logistic Regression model takes about $\frac{1}{10}SVM_{training\ time}$. So if we take into perspective the exponentially growing volume of data, Logistic Regression performs best.

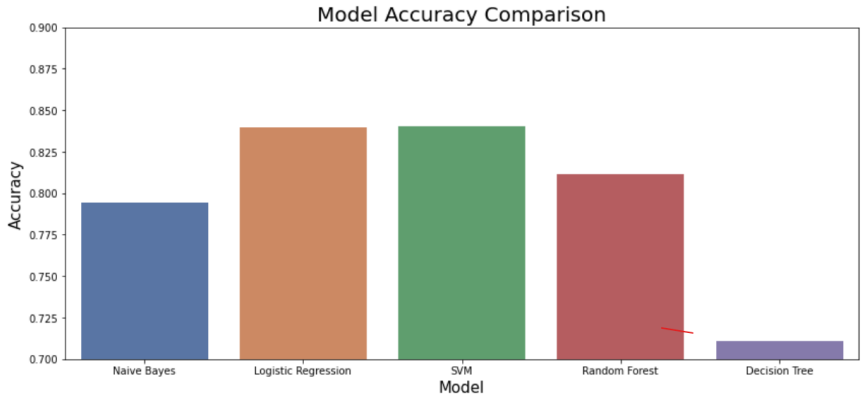

Figure 1. Comparing Models Based on Test Accuracy

	Model	Precision	Recall	F1- Score
Class 0	SVM	0.846276	0.582334	0.689922
	Logistic Regression	0.844937	0.582334	0.689477
	Random Forest	0.817360	0.492912	0.614966
	Decision Tree	0.525441	0.551799	0.538298
	Naïve Bayes	0.905405	0.365322	0.520591

Table 1. Evaluating Metrics Based on Class 0

	Model	Precision	Recall	F1- Score
Class 1	SVM	0.838328	0.953433	0.892183
	Logistic Regression	0.838260	0.952952	0.891934
	Random Forest	0.809971	0.951512	0.875055
	Decision Tree	0.778707	0.983197	0.869086
	Naïve Bayes	0.798233	0.780605	0.789320

Table 2. Evaluating Metrics Based on Class 1

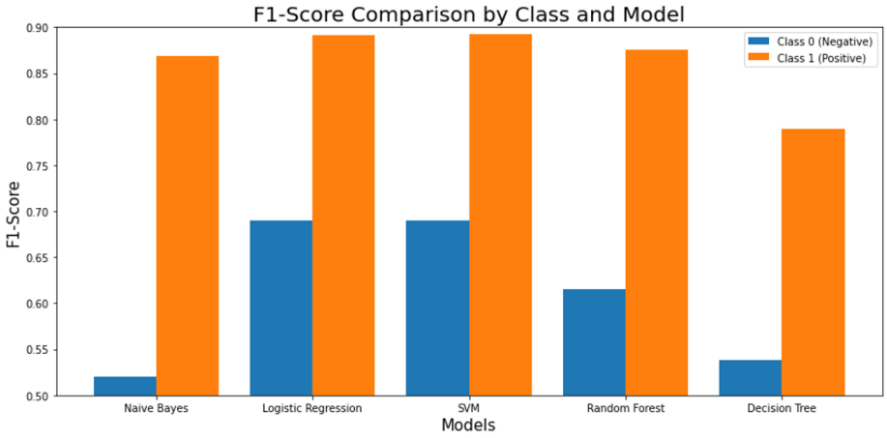

Figure 2. Comparing Models Based on F1-Score

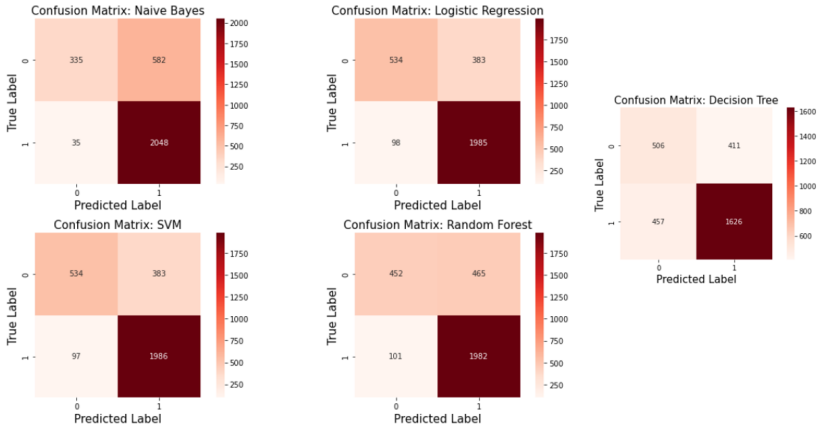

Figure 3. Comparing Confusion Matrices of the Different Models

5 CONCLUSION

The study successfully implemented and compared multiple machine learning models for Yelp sentiment analysis. Logistic Regression and SVM emerged as top performers with 84% accuracy, effectively categorizing sentiments. These models demonstrated a balanced handling of both positive and negative classes, as shown in their precision, recall, and F1-scores. This research underlines the critical role of sentiment analysis in interpreting customer feedback on platforms like Yelp. Accurately analyzing sentiments is vital for businesses to enhance customer satisfaction and inform strategic decisions. The findings emphasize the potential of machine learning in extracting actionable insights from vast textual data, an essential tool in today's data-driven world.

References

[1] https://vpnalert.com/resources/yelp-statistics/