Project Report: Text Analysis Using Naive Bayes Models

Introduction

This project aims to classify customer reviews as either positive or negative using natural language processing (NLP) techniques. We employed Naive Bayes models to analyze textual reviews from a dataset of user feedback, covering all stages from data cleaning and exploration to model building and performance evaluation.

Dataset Used

The dataset comprises customer reviews with ratings from 1 to 5. To simplify the analysis, we transformed the ratings into binary classifications: ratings of 4 and 5 were considered positive, and ratings of 1 and 2 were considered negative. Neutral ratings (3) were excluded.

Data Exploration

We conducted an initial exploration of the data to understand the distribution of ratings and the nature of the reviews. It was observed that ratings with a score of 5 were the most frequent, indicating a general positive bias in the dataset. This distribution was illustrated using a bar chart, which highlighted that most customers expressed high satisfaction with their experiences by assigning the highest possible rating.

Analytical Steps

1. Text Data Cleaning:

- Standardizing Text: All text was converted to lowercase to ensure uniformity.

- Removing Unnecessary Elements: Special characters, links, and numbers were removed using regular expressions.

- Stopword Removal and Lemmatization: The spacy library was utilized to perform lemmatization and remove stopwords, simplifying the text to its essential terms.

2. Data Analysis:

- Review Length Analysis: A new column representing the length of each review was added, and the distribution of review lengths was analyzed to understand the variation in text length.

- Rating Distribution: A bar chart illustrated the distribution of reviews across rating categories, allowing us to visualize the overall sentiment in the dataset.

- Word Cloud Analysis: Frequent words for each rating category were visualized, helping us understand the common terms associated with positive and negative sentiments.

3. Transforming Text into Numeric Features:

- We applied TfidfVectorizer to transform text into numerical features based on the frequency and importance of words, which enhances the model's ability to distinguish between positive and negative reviews.

4. Model Building and Evaluation:

- Data Splitting: We used train_test_split to split the data into 80% training and 20% testing sets.

- Training Models: Two models were trained: Multinomial Naive Bayes and Bernoulli Naive Bayes.

- Evaluating Model Performance Using Confusion Matrix and ROC Curve: The confusion matrix and ROC curve were used to evaluate each model's accuracy and to calculate the Area Under the Curve (AUC) for further insight into model performance.

Results

1. Confusion Matrix Analysis

- Multinomial Naive Bayes:

  - True Positives: 16,092 cases.

  - True Negatives: 560 cases.

  - False Positives: 2,617 cases.

  - False Negatives: 69 cases.

- Bernoulli Naive Bayes:

  - True Positives: 14,881 cases.

  - True Negatives: 1,860 cases.

  - False Positives: 1,317 cases.

  - False Negatives: 1,280 cases.

2. Performance Metrics and ROC Curve

Model	Accuracy	Precision	Recall	AUC
Multinomial Naive Bayes	86.11%	86.01%	99.57%	0.92

| Bernoulli Naive Bayes | 86.57% | 91.87% | 92.08% | 0.90|

- Multinomial Naive Bayes: This model has a high recall rate (99.57%), which makes it effective at identifying most positive cases but results in a higher number of false positives.

- Bernoulli Naive Bayes: This model showed higher precision, reducing the number of false positives, and achieved a slightly higher overall accuracy, though with a lower recall.

Conclusion

This project provided a comprehensive analysis of customer reviews using Naive Bayes models to classify sentiments as positive or negative. Through initial exploration, we gained insight into the distribution of ratings, noting that positive reviews (rated 5) were the most frequent, indicating high customer satisfaction overall.

The models presented varying results, with Multinomial Naive Bayes excelling in recall (capturing the majority of positive cases) and Bernoulli Naive Bayes achieving higher precision and a slightly higher overall accuracy. For applications prioritizing reduced false positives, Bernoulli Naive Bayes is recommended, while Multinomial Naive Bayes would be better suited for applications focusing on capturing all positive cases.

This project demonstrates how NLP techniques can be effectively applied to analyze text data and gain insights into customer sentiment, thereby supporting data-driven decision-making based on textual analysis.