PROJECT REPORT

"CUSTOMER REVIEWS SENTIMENTAL ANALYSIS"

SUBMITTED TO:

JAYENDRA BARUA

SUBMITTED BY:

Kashika Jindal (102317154)

Kushali Gupta(102317148)

ABSTRACT

This project explores sentiment analysis using two machine learning algorithms: Naive Bayes and k-Nearest Neighbors (KNN). We processed a dataset of 1,000 customer reviews by tokenizing, removing stopwords, and lemmatizing the text. Both models were trained to classify reviews into Positive, Neutral, or Negative categories.

After evaluating the models on accuracy, KNN outperformed Naive Bayes, demonstrating the effectiveness of distance based methods in text classification. Despite Naive Bayes' simplicity, it served as a competitive baseline. This project highlights the application of machine learning for sentiment analysis and compares two classification techniques.

TABLE OF CONTENTS

INTRODUCTION

Sentiment analysis is a technique used to understand the emotions or opinions behind a piece of text. It's commonly applied to customer reviews, social media posts, and other types of feedback to automatically categorize the sentiment as positive, neutral, or negative. In this project, we focused on classifying customer reviews into these three categories using two well-known machine learning algorithms: Naive Bayes and k-Nearest Neighbors (KNN).

The project began with preprocessing the data, where we cleaned the text by removing unnecessary words, breaking down sentences into smaller pieces (tokens), and simplifying the words to their base forms (lemmatization). After preparing the data, we used both algorithms to classify the sentiment of the reviews.

Naive Bayes is a simple and efficient algorithm that works well with text data, while KNN uses the similarity between data points to make predictions. By applying both methods, we were able to compare their performances and see which one provides more accurate results.

The goal of this project is to demonstrate how machine learning can be used to analyse large sets of text data automatically, making it easier to understand customer opinions and feedback. We also aimed to explore how different algorithms work and identify their strengths and weaknesses in real-world applications.

By the end of the project, we will have a better understanding of which algorithm works best for sentiment analysis tasks and how it can be used to gain insights from customer reviews.

PROBLEM STATEMENT

In today’s digital world, businesses receive a large number of customer reviews through online platforms. These reviews hold valuable opinions about products, services, and customer experiences. However, manually reading and analyzing them is time-consuming, tiring, and not always practical.

To solve this problem, this project focuses on creating an automated system to classify customer reviews into three categories — Positive, Neutral, and Negative. The idea is to use machine learning techniques to understand the meaning behind the text and predict the sentiment of each review accurately.

For this, two popular algorithms, Naive Bayes and k Nearest Neighbors (KNN), are applied and compared based on their accuracy and performance. The goal is to find out which method works better in classifying reviews and can help businesses make faster, data-driven decisions from customer feedback.

OBJECTIVE

The main objective of this project is to build a machine learning-based system that can automatically classify customer reviews into Positive, Neutral, or Negative categories.

Another objective is to compare the performance of two popular algorithms — Naive Bayes and k-Nearest Neighbors (KNN) — in terms of accuracy and efficiency for sentiment analysis.

Lastly, the project aims to present these results visually through a simple user interface and heatmaps, making it easier for businesses to quickly understand customer opinions and improve their services.

METHODOLOGY

The project follows a structured approach to perform sentiment analysis on customer reviews:

1. **Data Collection:** A dataset containing customer reviews and their ratings is collected. The reviews are used as input text, and ratings help label sentiments as Positive, Neutral, or Negative.

2. **Data Preprocessing:** The text data is cleaned by converting it to lowercase, removing stopwords, punctuation, and applying lemmatization to convert words to their base forms. Tokenization is done to break sentences into individual words.

3. **Feature Extraction:** A vocabulary is created from the processed tokens, and each review is transformed into a numeric vector using a Bag-of-Words approach based on word counts.

4. **Model Training:** The dataset is split into 80% training and 20% testing sets. Two machine learning models —

Naive Bayes and k-Nearest Neighbors (KNN) — are trained using the training data.

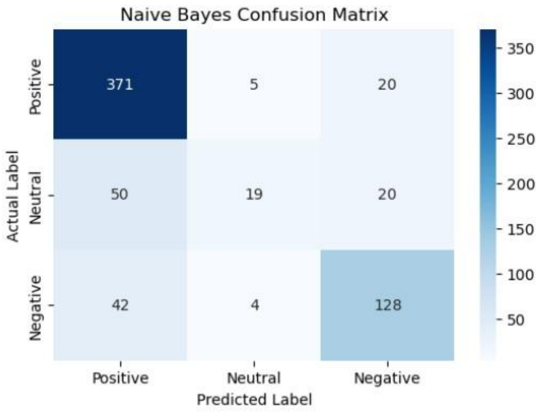
5. **Prediction and Evaluation:** Both models predict the sentiments for the test data. The performance is evaluated using accuracy scores, confusion matrices, and heatmaps to compare the models and understand their prediction patterns.

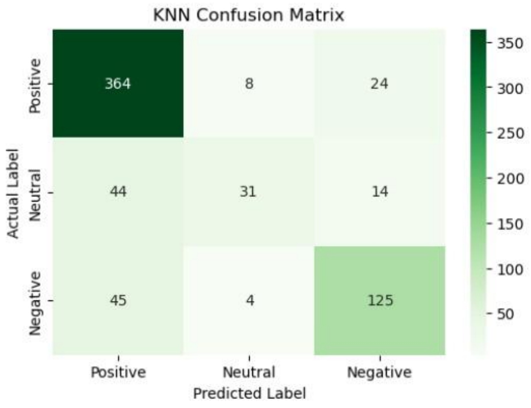
RESULT

After training and testing both models on the customer reviews dataset, the following results were observed:

- The Naive Bayes model achieved an accuracy of approximately 78.6% on the test data.
- The k-Nearest Neighbors (KNN) model achieved a slightly higher accuracy of approximately 78.91%, making it the better-performing model for this dataset.

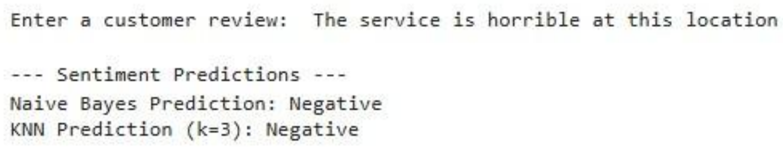
To better understand the prediction patterns, confusion matrices and heatmaps were generated for both models.





These visualizations showed that while both models performed well in identifying Positive and Negative reviews, they occasionally struggled with accurately predicting Neutral sentiments.

Additionally, the system was successfully able to take custom user input in the Jupyter Notebook and predict its sentiment using both models in real time.



```
Enter a customer review:  The service is horrible at this location

--- Sentiment Predictions ---
Naive Bayes Prediction: Negative
KNN Prediction (k=3): Negative
```

CONCLUSION

This project successfully applied Naive Bayes and k-Nearest Neighbors (KNN) algorithms for sentiment analysis of customer reviews. KNN slightly outperformed Naive Bayes in accuracy. Both models showed good performance in classifying reviews as Positive, Neutral, or Negative, with KNN providing more consistent results.

The confusion matrix and heatmaps helped visualize the models' performance. Future work could focus on improving accuracy by experimenting with different models and optimizing hyperparameters.

REFERENCES

1. https://www.geeksforgeeks.org/k-nearest-neighbours/

2. https://www.geeksforgeeks.org/naive-bayes-classifiers/

3. https://www.geeksforgeeks.org/natural-languageprocessing-overview/

4. https://www.geeksforgeeks.org/what-is-heatmap-datavisualization-and-how-to-use-it/