Text Analysis Report

Objective

Developing a text classification model to classify Yelp reviews as either positive or negative is the aim of this task. This model aids in automating sentiment analysis and extracting valuable information from user reviews.

Dataset Overview

The Yelp reviews dataset, extracted from the Yelp Dataset Challenge 2015 data, is used as a text classification benchmark in a paper by Xiang Zhang. The dataset consists of 560,000 training samples and 38,000 testing samples. The files train.csv and test.csv contain training samples as comma-sparated values, with columns corresponding to class index and review text. Review texts are escaped using double quotes, and internal double quotes are escaped with double quotes. New lines are escaped with a backslash followed by an "n" character.

The dataset consisted with two key columns:

- An Integer column to index raws.

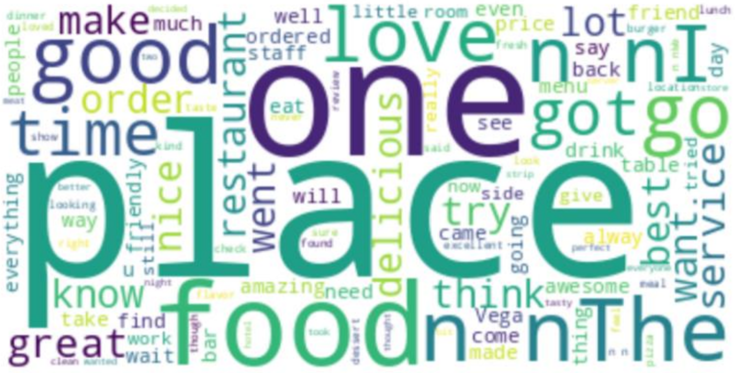
- a string column contain the reviews.

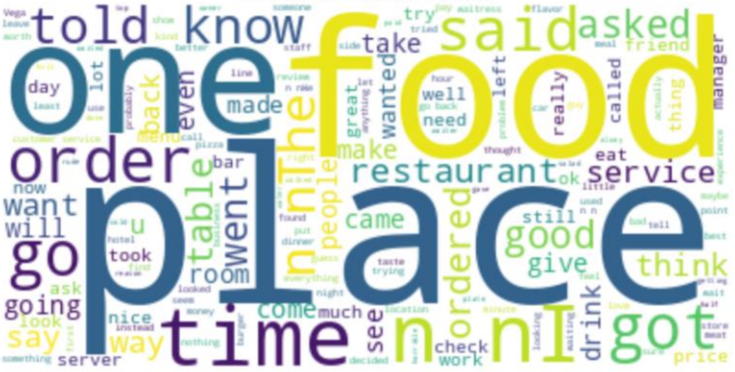
Methodology

1- **Data Preprocessing Steps**:

- **Data Loading**:
    - Load the training and testing datasets using **pandas**. The datasets are in CSV format.
- **Text Cleaning**:
    - **Lowercasing**: Convert all text to lowercase to ensure uniformity.
    - **URL Removal**: Remove any URLs present in the text using regular expressions.
    - **Number Removal**: Eliminate standalone numbers from the text.

    - **HTML Tag Removal**: Strip out any HTML tags that may be present in the text.
    - **Punctuation Removal**: Remove punctuation marks to avoid them affecting the text analysis.
    - **Newline Removal**: Remove newline characters to ensure that the text is a continuous string.
    - **Emoji Removal**: Use a regex pattern to remove emojis from the text.
    - **Contraction Expansion**: Expand common contractions (e.g., "isn't" to "is not") to standardize the text.
- **Stop Words Removal**:
    - Define a set of English stop words using NLTK's **stopwords** corpus.
    - Remove these stop words from the cleaned text to focus on more meaningful words.
- **Text Normalization**:
    - This step may include lemmatization (not explicitly shown in the TF-IDF snippet but mentioned earlier), which reduces words to their base or root form.
- **Text tokenization**
    - break down text into smaller units, such as words or phrases, which facilitates easier analysis and processing in natural language processing tasks.
- **Working with the most Frequent Words**:
    - identify key terms and themes within a dataset, allowing for focused analysis, feature selection, and improved model performance by reducing noise from less significant words.


Wordcloud for Positive Reviews


Wordcloud for Negative Reviews

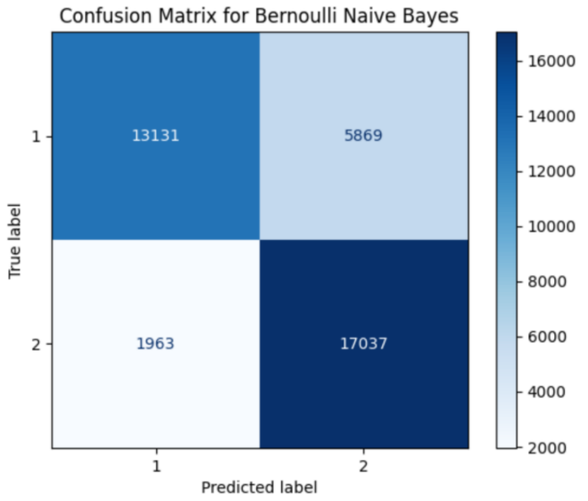
2-Model selction:

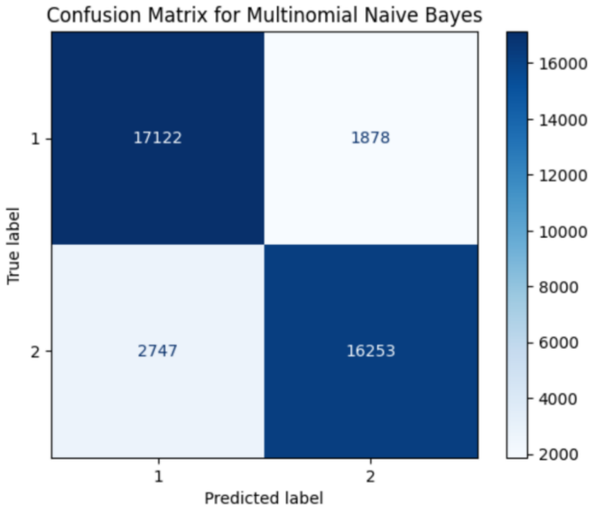
Multinomial Naive Bayes and Bernoulli Naive Bayes are both variants of the Naive Bayes algorithm used for classification tasks, particularly in text classification. Multinomial Naive Bayes is designed for datasets where features represent the frequency of words or terms in documents, making it effective for modeling the distribution of word counts. It calculates the probability of each class based on the frequencies of features, assuming that the presence of a feature is independent of others. In contrast, Bernoulli Naive Bayes is suited for binary or boolean features, where the presence or absence of a word in a document is considered, rather than its frequency. This makes it particularly useful for tasks where the goal is to determine whether specific words appear in a document, rather than how often they appear. Both algorithms leverage Bayes' theorem and the assumption of feature independence to provide efficient and effective classification, particularly in scenarios with large datasets.

3- Model Performance:

The Multinomial Naive Bayes model achieved an accuracy of 87.83%, indicating a solid ability to correctly classify reviews, with balanced precision and recall for both positive and negative classes.

the Bernoulli Naive Bayes model performed less effectively, with an accuracy of 79.39%. It showed lower precision and recall, particularly for negative reviews, which suggests challenges in correctly identifying sentiments.





4. Insights gained:

- The findings highlight the significance of choosing a model according to the particulars of the dataset and the issue at hand. The nature of the text data and the intended balance between accuracy and interpretability should be taken into account when choosing between the two Naive Bayes models, even if they are both beneficial.
- The Multinomial Naive Bayes model performed well with an accuracy of over 88%, showing high reliability in classifying reviews.
- The insights gained from these models can serve as a benchmark for future experiments with more complex algorithms, such as ensemble methods or deep learning approaches, allowing for a more comprehensive understanding of their strengths and weaknesses in sentiment analysis tasks.

4.Conclusion:

The Multinomial Naive Bayes showed superior accuracy and balanced precision, higher than the Bernoulli Naive Bayes models for sentiment classification of customer reviews. Preprocessing techniques like tokenization enhance data quality, while class imbalance can impact performance.

Names: Jori Bajahnoun (444006482)

Taghreed Alzahrani (443007366)