**Introduction**

Online platforms like Yelp and Google Maps are vital for finding trustworthy recommendations, but their value hinges on the authenticity of user reviews. Some reviews may be biased or influenced by outside factors, making it crucial to identify whether they express genuine praise or criticism. Understanding the true sentiment behind reviews helps users make better decisions and enhances their overall experience.

Equally important is ensuring that reviews are truthful. Fake reviews, designed to either unfairly boost or damage a business's reputation, can erode user trust. Platforms must differentiate between honest feedback and deceptive content to maintain their credibility and provide accurate recommendations.

Given the vast number of reviews posted daily, platforms need efficient methods to assess them. Quickly identifying the sentiment and detecting potential fakes allows these platforms to filter out misleading content, ensuring users see reliable, authentic opinions. To support this effort, the goal of this analysis is to evaluate several machine learning models that can effectively facilitate this review analysis, providing a lightweight, fast solution for classifying reviews at scale. This approach will help maintain user trust and satisfaction with both the platform and the businesses they choose.

**Analysis**

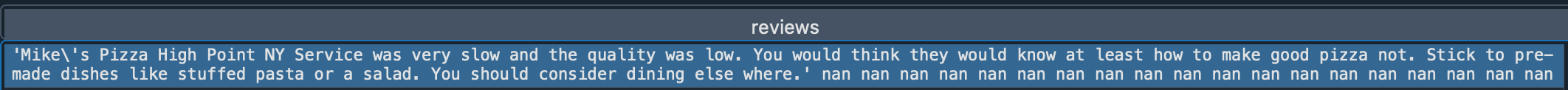
**Data Preparation**

This section focuses on the process of preparing the data for analysis. The steps include loading and cleaning the data, labeling it appropriately, and visualizing it to gain initial insights. Proper data preparation is crucial to ensure that the subsequent analysis is accurate and meaningful.

1. **Overview of the Data:** The dataset used in this analysis consists of scraped reviews of restaurants. Each restaurant review has been labeled with its sentiment, either positive (“p”) or negative (“n”), and it’s truthfulness, either false (“f”) or true (“t”). As the review data was scraped from the web, there are some html encoding characters such as line breaks (“\n”) and escape characters (“\”) included in each review.

|  |  |
| --- | --- |
| **Raw Reviews Data:** A screenshot of a computer  Description automatically generated | |
| **Lie Labels (F = False, T = True)** | **Sentiment (N=Negative, P = Positive)** |

1. **Data Loading and Cleaning**
   1. **Loading**: The data was loaded from a CSV file containing restaurant reviews, each labeled with sentiment (positive/negative) and truthfulness (true/false). During the loading process, the presence of punctuation in the reviews caused issues, leading to the text being incorrectly split across multiple columns. To resolve this, the fragmented parts of each review were re-combined into a single column, ensuring that the full text of each review was preserved for analysis.
   2. **Cleaning Data:**
      1. **Combining Review Columns:** All relevant columns containing parts of reviews were combined into a single "reviews" column.
      2. **Removing Labels from the DataFrame:** The labels (lie and sentiment) were stored separately and removed from the main DataFrame to prepare the review text for vectorization.
      3. **Handling Special Characters:** Escape characters and missing values (like 'nan') within the reviews were removed or replaced. For example, characters such as "Mike\’s" and "nan" were cleaned from the reviews, as shown in the provided example.



* + 1. **Dropping Stop Words:** During preprocessing, common words that generally do not carry significant meaning (e.g., "the," "is," "and") were removed to reduce noise and improve model performance. This was done using the 'english' stopword list available in the TfidfVectorizer from scikit-learn. This step allows the analysis to focus on more informative words that contribute meaningfully to the context of the reviews.
    2. **Vectorization:** The cleaned reviews were vectorized using the TF-IDF (Term Frequency-Inverse Document Frequency) method with L2 normalization. TF-IDF converts text into a numerical format by weighting terms based on their importance within a specific document relative to their frequency across all documents in the dataset. This helps to highlight more distinctive words in each document while down-weighting common words that are less informative. Additionally, L2 normalization was applied to the TF-IDF vectors to ensure the length of the document does not disproportionately influence the vector, allowing the model to focus on the relative importance of terms within each document rather than the document length itself.
    3. **Dropping Unnecessary Words:** Words with two or fewer characters, more than 13 characters, or containing numbers were identified and removed to refine the dataset. A total of 32 words were dropped from the vocabulary, a full list of which is below:
       - 10
       - 100
       - 15
       - 16
       - 20
       - 25
       - 2nd
       - 30
       - 50
       - 5pm
       - 6pm
       - 90
       - ad
       - aforementioned
       - disappointment
       - extravaganzaburger
       - ll
       - ny
       - pm
       - recommendation
       - recommendations
       - se
       - su
       - ve
    4. **Data After Cleaning and Vectorization:**

**A screenshot of a red table

Description automatically generated**

* 1. **Visualizing Cleaned Data-set**
     1. **Word Clouds:** To better understand the distribution of words associated with sentiment and truthfulness labels, word clouds are created. Word clouds are visual representations where the size of each word indicates its frequency in the dataset. This visualization helps to quickly identify the most prominent words associated with each label.
        + **Consistency Across Labels:** Words such as "restaurant," "food," and "service" are consistently dominant across all labels, showing that these are core elements in all reviews, regardless of the truthfulness or sentiment.
        + **Differentiation:** The differentiation between positive and negative sentiments is clearly visible through the adjectives used (e.g., "great" vs. "terrible"). Additionally, the distinction between true and false lie labels is subtle, with slight variations in the focus on timing ("minutes") and service quality.

|  |  |
| --- | --- |
| Word Cloud for Truthful (T) Reviews: | Word Cloud for Truthful (F) Reviews: |
| **Lie = t (True):**  **Dominant Words:** The most prominent words in this word cloud are "food," "restaurant," "place," "people," "great," "good," "nice," and "ordered."  **Analysis:** These words suggest that discussions around food and restaurant experiences are central in the reviews classified as "true" lies. Words like "great," "good," and "nice" indicate that these reviews tend to contain positive sentiments about the restaurant and food experiences. The presence of words like "ordered" and "went" implies that the actions of visiting and ordering at the restaurant are frequently mentioned. | **Lie = f (False):**  **Dominant Words:** The prominent words include "great," "best," "service," "minutes," "restaurant," "food," "went," and "place."  **Analysis:** Similar to the "true" lie category, "food" and "restaurant" are key topics. However, words like "minutes," "best," and "service" appear more prominently here, indicating that reviews flagged as "false" may focus more on the quality of service and the timing of the service ("minutes"). The presence of words like "cold" and "like" might suggest more mixed or specific feedback. |
| Word Cloud for Positive Sentiment Reviews | Word Cloud for Negative Sentiment Reviews |
| **Sentiment = p (Positive):**  **Dominant Words:** The most prominent words are "great," "best," "service," "restaurant," "food," "friendly," "amazing," and "good."  **Analysis:** Reviews classified as positive sentiment heavily emphasize positive adjectives such as "great," "best," and "amazing," indicating strong satisfaction with the food and service. The word "friendly" suggests that customer service and the behavior of staff are significant factors in positive reviews. "Food" and "restaurant" continue to be central themes, indicating that the core focus of the reviews remains on the dining experience. | **Sentiment = n (Negative):**  **Dominant Words:** The key words here are "went," "food," "minutes," "restaurant," "ordered," "terrible," "service," and "bad."  **Analysis:** The negative sentiment word cloud reflects dissatisfaction, with words like "terrible," "bad," and "minutes" indicating complaints about service and food quality. The emphasis on "minutes" could suggest issues with waiting times. The repetition of words like "ordered" and "went" reflects detailed descriptions of experiences that led to dissatisfaction. |

* + 1. **Word Frequency Across Reviews:** The box plot displays the distribution of TF-IDF frequencies across all words in the dataset. Most words have relatively low TF-IDF scores, which is indicated by the concentration of values in the lower quartile of the box plot. This suggests that many words in the reviews are either common across all documents (and thus receive lower TF-IDF scores) or are only slightly significant in the context of individual reviews.

Several outliers are visible above the upper whisker of the box plot. These outliers represent words with exceptionally high TF-IDF scores, indicating that these terms are particularly important in distinguishing the content of specific reviews. These high TF-IDF words will likely be important in our models to differentiate between labels.

**A diagram of a number of numbers

Description automatically generated**

* + 1. **Top 20 Tf-IDF Words Across All Reviews:** The bar chart shows the top 20 words with the highest TF-IDF scores. The most prominent words include "restaurant," "food," "place," "great," and "good." These words are heavily weighted because they are not just frequent, but also more common in certain reviews compared to others in the dataset.

The presence of positive evaluative terms like "great," "good," and "best" highlights the focus on customer experiences and satisfaction. These words are likely used to describe positive experiences, contributing to their high TF-IDF scores.

Words like "service," "ordered," and "minutes" suggest that aspects of the dining experience, such as service quality and wait times, are also heavily discussed in the reviews and may offer some discriminatory value in differentiating between sentiment and truthfulness.

A graph of a bar graph

Description automatically generated

**Model/Methods**

**Model Training:**

To evaluate the effectiveness of different models in classifying sentiment and detecting truthfulness in restaurant reviews, three models were trained: Bernoulli Naive Bayes (BNB), Multinomial Naive Bayes (MNB), and Decision Tree Classifier (DT). The training process utilized 10-fold cross-validation, a technique that splits the data into ten subsets (folds). Each model is trained on nine of these folds and validated on the remaining one. This process is repeated ten times, with each fold serving as the validation set exactly once. This approach ensures that the models are evaluated across different segments of the data, providing robust performance metrics while maximizing the use of the data for both training and testing.

The models were evaluated based on four key performance metrics: Precision, Recall, Accuracy, and F1-Score. These metrics provide a comprehensive view of how well each model performs in both identifying the correct class labels and minimizing false positives and false negatives.

**Sentiment Classification Results:**

1. **Bernoulli Naive Bayes (BNB) Sentiment Model:**
   * **Precision:** 0.7334
   * **Recall:** 0.6848
   * **Accuracy:** 0.6848
   * **F1-Score:** 0.6675
   * **Significant Words for Negative Sentiment ("n**"): The BNB model identified words such as "bad," "waiter," "wait," "asked," "terrible," and "minutes" as highly indicative of negative sentiment. These terms often reflect poor service, delays, and overall dissatisfaction, which align with the common themes in negative reviews.
   * **Significant Words for Positive Sentiment ("p"):** For positive sentiment, the model highlighted words like "quality," "favorite," "sauce," "delicious," and "friendly." These terms are typically associated with positive dining experiences, particularly emphasizing food quality and pleasant service.
   * **Analysis:** The BNB model achieved the highest precision among the three models, indicating that it was more effective in correctly identifying positive sentiment when it predicted it. However, its recall, accuracy, and F1-score were moderate, suggesting that while it was good at precision, it missed a fair number of positive sentiments (lower recall), which affected its overall balance (F1-Score).
2. **Multinomial Naive Bayes (MNB) Sentiment Model:**
   * **Precision:** 0.6106
   * **Recall:** 0.6087
   * **Accuracy:** 0.6087
   * **F1-Score:** 0.6070
   * **Significant Words for Negative Sentiment ("n"):** The MNB model similarly focused on words like "service," "waiter," "pizza," "bad," and "terrible" for negative sentiment. The repeated emphasis on service quality and specific food items like pizza suggests that these are common sources of dissatisfaction in the reviews. For a more generalizable model, we may consider removing “pizza” and other specific food groups from our vocabulary.
   * **Significant Words for Positive Sentiment ("p"):** Positive sentiment words identified by the MNB model include "favorite," "noodle," "quality," "delicious," and "amazing." These words highlight customer preferences for specific dishes and the overall quality of the dining experience. Like our conclusions from the negative sentiment, for a more generalizable model, we may consider removing specific food groups from our model vocabulary.
   * **Analysis:** The MNB model demonstrated a balanced yet lower performance across all metrics compared to the BNB and DT models. Its precision, recall, accuracy, and F1-score were close, indicating that it performed consistently but not as strongly as the BNB model, particularly in identifying positive sentiment.
3. **Decision Tree Classifier (DT) Sentiment Model:**
   * **Precision:** 0.6870
   * **Recall:** 0.6848
   * **Accuracy:** 0.6848
   * **F1-Score:** 0.6838
   * **Highly Discriminative Words:** The DT model identified a distinct set of words, including "felt," "fiance," "filled," "filet," and "tripadvisor," as key discriminators. These terms suggest that the model may have picked up on more personalized and unique experiences, as well as specific dining items like "filet," which are mentioned in reviews with strong opinions.
   * **Analysis:** The DT model performed similarly to the BNB model, with slightly lower precision but comparable recall, accuracy, and F1-score. This indicates that the DT model was relatively well-balanced in identifying both positive and negative sentiments, with a decent trade-off between precision and recall.

*A diagram of a tree

Description automatically generated*

**Lie Detection Results:**

1. **Bernoulli Naive Bayes (BNB) Lie Model:**
   * **Precision:** 0.2822
   * **Recall:** 0.2826
   * **Accuracy:** 0.2826
   * **F1-Score:** 0.2823
   * **Significant Words for False Reviews ("f"):** The BNB model identified words like "dining," "friends," "wait," "ordered," and "just" as indicative of false reviews. These words are generic and suggest that the model may struggle to find distinctive features that clearly separate false reviews from truthful ones.
   * **Significant Words for True Reviews ("t")**: For true reviews, words like "nice," "menu," "great," "waitress," and "friends" were considered significant. These terms reflect a mix of general positive sentiments and social dining experiences, which might be common in authentic reviews.
   * **Analysis:** The BNB model for lie detection showed poor performance across all metrics, with precision, recall, accuracy, and F1-score all around 0.28. This suggests that the model struggled significantly to correctly classify true versus false reviews, with a near-random performance.
2. **Multinomial Naive Bayes (MNB) Lie Model:**
   * **Precision:** 0.1496
   * **Recall:** 0.1630
   * **Accuracy:** 0.1630
   * **F1-Score:** 0.1550
   * **Significant Words for False Reviews ("f"):** Words like "really," "pizza," "menu," "cold," and "definitely" were highlighted for false reviews. These words suggest that false reviews might include exaggerated statements or focus on specific complaints, but the model’s low performance indicates it had difficulty distinguishing these reliably.
   * **Significant Words for True Reviews ("t"):** For true reviews, words like "bar," "indian," "just," "salad," and "waitress" were significant. The diversity in these terms suggests a broader range of experiences being discussed in authentic reviews, from specific cuisines to general service feedback.
   * **Analysis:** The MNB model performed even worse than the BNB model, with precision, recall, accuracy, and F1-score all below 0.17. This indicates a severe difficulty in detecting true and false reviews, making this model unreliable for lie detection in this context.
3. **Decision Tree Classifier (DT) Lie Model:**
   * **Precision:** 0.3570
   * **Recall:** 0.3587
   * **Accuracy:** 0.3587
   * **F1-Score:** 0.3568
   * **Highly Discriminative Words**: The DT model identified words like "fell," "love," "awkward," "food," "water," and "shopping" as significant in distinguishing between true and false reviews. These terms may reflect more nuanced or emotional content that could be indicative of either genuine sentiment or fabricated stories, although the model’s overall performance indicates limited reliability in lie detection.
   * **Analysis:** The DT model outperformed the Naive Bayes models in lie detection, though its performance was still low overall. With precision, recall, accuracy, and F1-score all around 0.36, the model shows some ability to distinguish true from false reviews, but it remains far from reliable.

*A diagram of a tree

Description automatically generated*

**Findings:**

The results of the sentiment classification task showed that the Bernoulli Naive Bayes (BNB) model had the highest precision, making it the best choice for applications where minimizing false positives is critical. However, the Decision Tree Classifier (DT) offered a balanced approach, with similar performance across all metrics, making it more suitable if both precision and recall are important.

For lie detection, all models struggled significantly, with the Decision Tree Classifier performing slightly better than the others but still far from satisfactory. These results suggest that the current feature set and model configurations are insufficient for accurately detecting fake reviews, and further refinement, such as feature engineering or exploring more advanced models, may be necessary.

**Conclusion:**

This analysis sought to detect user sentiment and truthfulness in online reviews. The findings reveal that while certain words, such as "great," "good," and "terrible," are effective in differentiating between positive and negative sentiments, the task of detecting whether a review is truthful is far more challenging.

The ability to clearly distinguish between positive and negative reviews based on specific keywords provides valuable insights into customer sentiment. However, when it comes to identifying deceptive content, analyzing words individually proves insufficient. The complexity of human language and the subtlety involved in crafting deceptive reviews mean that more advanced techniques are required to reliably detect fake reviews.

In conclusion, while analyzing the sentiment of reviews using individual words can offer useful insights, ensuring the truthfulness of all reviews remains a significant challenge. This highlights the ongoing need for improved methodologies that can more effectively discern between genuine and deceptive content, thereby helping users make more informed decisions and enhancing trust in the recommendations they find online.