Sentiment and Deception in Restaurant Reviews Through Vectorization and Naive Bayes Modeling

Introduction

Reviews are a rich source of user generated text that offer insight into consumer attitudes and experiences. While most reviews reflect genuine customer sentiment, others may be deceptive in nature, written with ulterior motives, either to promote or discredit, in this case, the restaurant. This multi layered challenge creates an opportunity for text mining to distinguish between not only positive and negative sentiment, but also between truthful and deceptive. This assignment explores both sentiment and deception by applying vectorization techniques and Naive Bayes modeling to a labeled dataset of restaurant reviews. The goal is to understand how well lexical patterns align with each classification task, and whether the same preprocessing pipeline can serve both tasks effectively.

Sentiment reflects the emotional tone, deception adds a layer of complexity, requiring models to infer intent and authenticity rather than just polarity. By comparing the performance of CountVectorizer and TF-IDF vectorization techniques across both tasks, this analysis draws on core principles of identifying not only what is said, but how it is said, and whether that reveals deeper narrative cues. Through the lens of VADER scoring, word cloud analysis, modeling, and cross validated performance metrics, this project tests the separability of truthful versus deceptive and positive versus negative reviews using a grounded framework.

Method

Data Collection, Structure and Sentiment Scoring: The dataset consisted of labeled restaurant reviews with two targets, sentiment, positive or negative, and deception, truth or lie. Each review was stored in a structured data frame containing the raw text. VADER sentiment analysis was applied to the raw review text to establish a baseline comparison between a rule-based model and the dataset's provided sentiment labels.

Exploratory Analysis and Data Preprocessing: Raw corpus, sentiment, and deception word clouds were generated to visualize lexical distributions. Initial preprocessing steps included removing outliers, cleaning text fields, removing non-alphabetic characters, lowercasing, and applying a customized stopword list

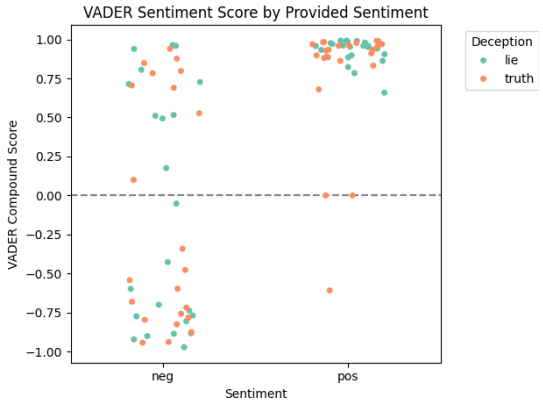
Vectorization: Two vectorization methods were used, CountVectorizer and TF-IDF. Individual approaches were configured to include unigrams and bigrams with a minimum document frequency threshold to filter out rare terms while maintaining interpretability. TF-IDF weighting was used to emphasize terms that were frequent in a document but rare across the corpus. Count-based and TF-IDF document-term matrices were created from the cleaned text for use in downstream modeling.

Train-Test Splits and Modeling: To prevent class imbalance from affecting performance metrics, stratified sampling was used for both sentiment and deception classification tasks. Multinomial

Naive Bayes was selected due to its efficiency with high dimensional sparse feature matrices. Separate models were trained on count and TF-IDF vectorizations, with optimal alpha parameters for each task. Evaluation was conducted using classification reports, confusion matrices, and 10-fold cross-validation to ensure comparable performance.

Results

Exploratory Data Analysis: Independent VADER sentiment analysis revealed a 75.8% agreement rate with the provided sentiment target feature. Notable inconsistencies included overestimated positivity, a known limitation in rule-based sentiment analysis where surface tone can mask negative content when context is overlooked.



Positive sentiment reviews contained stronger affective language like, amazing and delicious, while negative reviews emphasized dissatisfaction through words like, waiter and minutes. Truthful and deceptive reviews shared similar food centric vocabulary, but deceptive entries leaned on general descriptors like, best or place, indicating that surface-level lexical content alone may not differentiate deception as cleanly as sentiment

Model Performance Across: Sentiment classification showed stronger performance than deception classification across all vectorization methods. The TF-IDF vectorized model with alpha=1 achieved the highest sentiment classification accuracy, outperforming its count-based counterpart by a narrow but consistent margin.

Classification Report:

	precision	recall	f1-score	support
neg	0.92	0.79	0.85	14
pos	0.81	0.93	0.87	14
accuracy			0.86	28
macro avg	0.86	0.86	0.86	28
weighted avg	0.86	0.86	0.86	28

Accuracy: 0.86

Classification Report (TF-IDF):

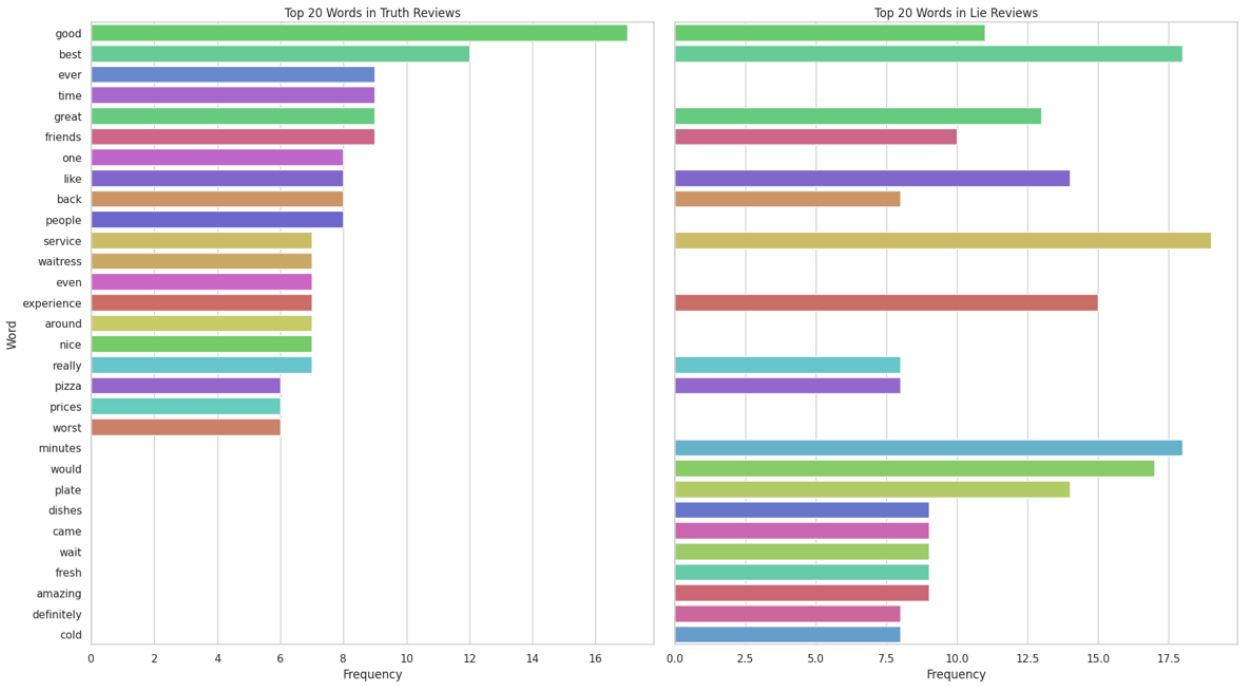
	precision	recall	f1-score	support
neg	0.87	0.93	0.90	14
pos	0.92	0.86	0.89	14
accuracy			0.89	28
macro avg	0.89	0.89	0.89	28
weighted avg	0.89	0.89	0.89	28

Accuracy: 0.89

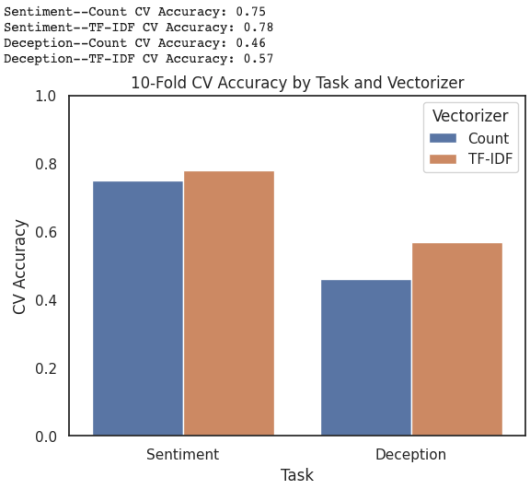
Conversely, deception classification yielded lower performance overall, Count Vectorization with an accuracy of 0.57 and TF-IDF with an accuracy of 0.54 This reflects a core distinction between the tasks: sentiment is often lexically separable, while deception relies on subtler cues that are not as easily captured by frequency based methods.

Confusion Matrix Patterns: In the deception classification task, lies were more likely to be misclassified as truths than vice versa, suggesting that deceptive reviews tend to mirror the

vocabulary and tone of authentic reviews. This aligns with linguistic theory, which suggests that deceptive language often mimics truthful expression to appear credible, reducing the effectiveness of surface-level models. In contrast, sentiment misclassifications showed clearer polarity reversals, such as positive surface tone masking negative feedback, a limitation exposed by both the VADER tool and the Naive Bayes classifier when handling context-dependent language.



Cross-Validation Summary: Cross-validated accuracy scores revealed the same trends. This suggests that emotional language tends to be more clearly differentiated by term frequency and document rarity, making both vectorization strategies effective. With cross-validated scores are notably lower than the training accuracy values, indicating that the models may have overfit to the training data and are less generalizable when applied to unseen samples.



Conclusion

This project explored the application of TF-IDF and CountVectorizer methods in classifying sentiment and deception within restaurant reviews, using Multinomial Naive Bayes as a baseline model. While sentiment classification yielded relatively strong performance, deception detection proved to be a more nuanced and difficult task. The lexical separation between positive and negative reviews made sentiment easier to model—emotional language such as delicious, awful, or perfect gave clear directional cues that both vectorization approaches were able to capture, particularly when weighted by document rarity through TF-IDF.

In contrast, deceptive reviews showed substantial lexical overlap with truthful ones, limiting the model's ability to distinguish intent. Even with different smoothing parameters and weighting schemes, classification metrics for deception remained low. This mirrors a key insight in deception research, linguistic camouflage is intentional. Deceptive reviews are written to sound realistic, often mimicking the tone and vocabulary of truthful ones, which causes confusion for models (and humans occasionally) that rely on surface level term frequencies or salience. And the cross-validation results further confirm this divide. Sentiment accuracy remained stable across folds, while deception performance dropped, highlighting the fragility of lexical indicators in more context-driven tasks.

Ultimately, this analysis reinforces that not all classification problems in text mining are created equal. For tasks like deception detection, shallow lexical models may not be sufficient, and deeper approaches, incorporating syntax, discourse structure, or neural representations, are likely necessary to capture the underlying intent.