Sentiment Classification of Restaurant Reviews Using Machine Learning Techniques

Rajkumar Ramu, Rustam Adigozalov, Natarajan Govindarajan, Hoang Phuong Duy Nguyen, Henil Shah
Department of Business Insights and Analytics, Humber College, Toronto, Canada
Machine Learning 1 Final Project

Abstract

This project investigates sentiment classification of restaurant reviews using machine-learning techniques applied to the Yelp Open Dataset. After preprocessing the text data and transforming it using Term Frequency–Inverse Document Frequency (TF-IDF) features, five supervised models were evaluated: Logistic Regression, K-Nearest Neighbors, Decision Tree (CART), Multinomial Naïve Bayes, and Multiple Linear Regression. The models were assessed based on accuracy, F1-score, and confusion-matrix behavior. Experimental results show that the Multinomial Naïve Bayes classifier achieved the highest performance, with an accuracy of 92.6%, outperforming both linear and non-linear alternatives. These findings demonstrate that probabilistic word-frequency models are highly effective for sparse, high-dimensional text classification tasks such as restaurant review sentiment analysis.

I. INTRODUCTION

Online reviews play a central role in shaping consumer decisions, particularly in the restaurant industry where customer experiences vary widely. Large platforms such as Yelp provide extensive user-generated text that captures customer sentiment, but the volume of reviews makes manual analysis impractical. Automated sentiment analysis provides a scalable method for interpreting these opinions and extracting useful insights for businesses.

Machine-learning techniques have been widely applied to sentiment classification by converting text into structured numerical representations. TF-IDF features, originally formalized by Robertson [3], enable the identification of informative terms within large text corpora, while supervised models such as Logistic Regression [1], Multinomial Naïve Bayes [2], K-Nearest Neighbors, and Decision Trees offer diverse strategies for modeling high-dimensional feature spaces.

This project aims to evaluate several machine-learning algorithms for classifying restaurant review sentiment using the Yelp Open Dataset [4]. The workflow includes text cleaning, TF-IDF vectorization, normalization, dimensionality reduction, and supervised learning. Each model is assessed using accuracy, F1-score, and confusion-matrix behavior to identify the most effective approach for large-scale text classification.

The remainder of this report presents the preprocessing pipeline, feature-engineering techniques, model development, results, and comparative evaluation of all models tested.

II. OBJECTIVES

The project was designed to systematically evaluate multiple machine-learning models for sentiment classification using TF-IDF features derived from the Yelp Open Dataset [4]. The specific objectives are as follows:

1. **Prepare and clean the dataset**, including removing missing entries, filtering ambiguous star ratings, and constructing binary sentiment labels from review metadata.

2. **Transform raw text into numerical features** using the TF-IDF method, following established weighting principles described by Robertson [3].

3. **Apply a consistent preprocessing pipeline**, including case normalization, token filtering, and dimensionality reduction using Principal Component Analysis (PCA), to improve training efficiency.

4. **Train and evaluate multiple supervised learning models**, including Logistic Regression [1], K-Nearest Neighbors, Decision Tree (CART), Multiple Linear Regression, and Multinomial Naïve Bayes [2], using standardized train–validation splits.

5. **Compare the performance of all models** using accuracy, F1-score, and confusion-matrix analysis generated through the scikit-learn library [1].

6. **Identify the best-performing model**, based on empirical results, and assess the factors contributing to its effectiveness in text-classification tasks.

III. LITERATURE REVIEW

Text classification and sentiment analysis have been widely studied within natural language processing (NLP) and machine learning. A core component of most sentiment-analysis pipelines is the transformation of raw text into numerical representations. One of the most established weighting schemes is the Term Frequency–Inverse Document Frequency (TF-IDF) model, which measures the importance of terms across documents. Robertson's formulation of inverse document frequency provides the theoretical basis for TF-IDF and remains a standard approach for modeling word relevance in sparse, high-dimensional text data [3].

Supervised machine-learning methods have been extensively applied to sentiment classification tasks. Logistic Regression is a widely used linear classifier due to its effectiveness in high-dimensional feature spaces and its implementation efficiency in modern libraries such as scikit-learn [1]. K-Nearest Neighbors (KNN) and Decision Trees provide non-parametric alternatives, offering flexible decision boundaries, though they often require careful preprocessing to handle feature sparsity.

Probabilistic models, particularly the Multinomial Naïve Bayes classifier, have demonstrated strong performance in text mining applications. McCallum and Nigam's work highlights how word-count distributions and conditional independence assumptions enable Naïve Bayes to perform robustly even when features are numerous and sparsely populated [2]. This makes it well-suited for sentiment classification problems that rely on TF-IDF vectors.

Dimensionality reduction techniques such as Principal Component Analysis (PCA), implemented through scikit-learn [1], can improve computational efficiency by projecting TF-IDF features into lower-dimensional spaces while preserving global variance structure. Such transformations are especially useful when training distance-based models such as KNN.

Overall, prior research and established methodologies support the use of TF-IDF representations combined with supervised classifiers—including Logistic Regression, Naïve Bayes, KNN, and Decision Trees—for large-scale sentiment classification tasks, such as those conducted on the Yelp Open Dataset [4]. These works form the foundation for the modeling approaches evaluated in this study.

IV. DATASET DESCRIPTION AND CLEANING

The dataset used in this study is taken from the Yelp Open Dataset, which provides a large collection of user-generated business reviews, ratings, and metadata for research purposes [4]. For this project, a subset of restaurant reviews was extracted to focus specifically on sentiment classification in the food service domain. The dataset includes text reviews, star ratings, and basic business identifiers.

Before modeling, several preprocessing steps were applied to ensure data quality and relevance. Entries with missing review text or missing star ratings were removed, as these observations could not contribute to the learning task. Duplicate reviews were also eliminated to avoid bias in word-frequency distributions. Since star ratings of three stars often represent neutral or ambiguous sentiment, only reviews with ratings of one to two stars (negative sentiment) and four to five stars (positive sentiment) were retained. This filtering helped create a clearly defined binary sentiment target variable.

A new binary label, *sentiment*, was created, where positive reviews were assigned a value of 1 and negative reviews a value of 0. The resulting dataset contains a balanced and interpretable structure that is suitable for supervised classification. Prior to feature engineering, the review text was normalized by converting all characters to lowercase and removing punctuation, excessive whitespace, and non-alphabetic characters. These steps reduce noise and help standardize the text for downstream vectorization.

The cleaned and structured dataset forms the basis for TF-IDF feature extraction, dimensionality reduction, and subsequent model development presented in later sections.

V. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) was conducted to examine the structure and characteristics of the cleaned Yelp restaurant review dataset prior to feature engineering. The first step was to analyze the distribution of the binary sentiment labels created from the star ratings. As shown in **Fig. 1**, the number of positive (4–5 stars) and negative (1–2 stars) reviews is relatively balanced. This balance is desirable, as it reduces the risk of classifier bias and avoids the need for resampling techniques.

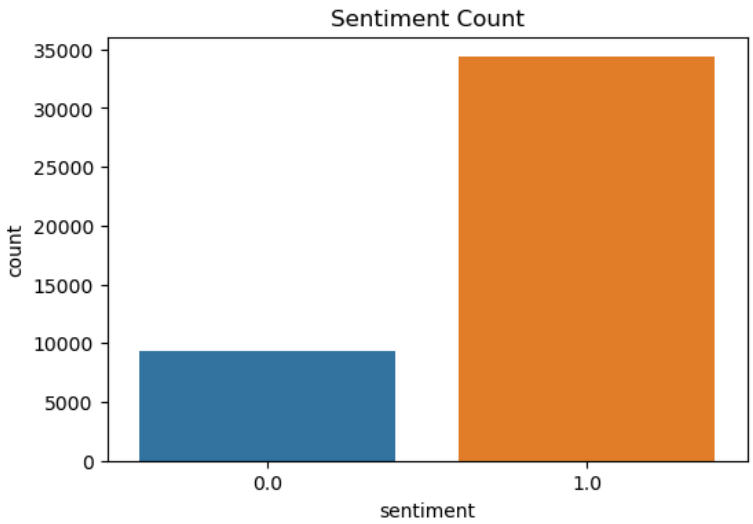

Fig. 1. Distribution of Positive and Negative Sentiment Reviews.

Next, the distribution of review lengths was examined to understand variability in the amount of textual information provided by users. The histogram in **Fig. 2** shows that review lengths vary substantially, ranging from very short comments to lengthy descriptive reviews. Such variation is typical for user-generated content and has implications for TF-IDF transformation, as shorter reviews often carry fewer informative terms while longer reviews contribute more diverse vocabulary [3].

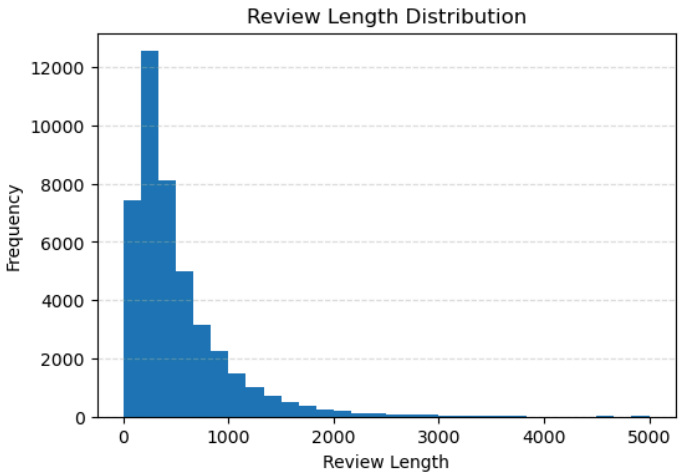

Fig. 2. Histogram of Review Lengths in the Dataset.

Taken together, the EDA results indicate that the dataset is suitable for supervised sentiment classification. The sentiment classes are well-balanced, and the variability in review length suggests meaningful differences in lexical content. These findings motivate the use of TF-IDF features and guide the preprocessing steps described in the following section.

VI. FEATURE ENGINEERING AND PREPROCESSING

The Yelp review text required several preprocessing steps before model training. All text was first normalized by converting characters to lowercase and removing punctuation, numbers, and excessive whitespace. Tokenization and filtering were applied to retain only alphabetic tokens, reducing noise and standardizing the input. These steps ensured consistency across samples and prepared the text for numerical representation.

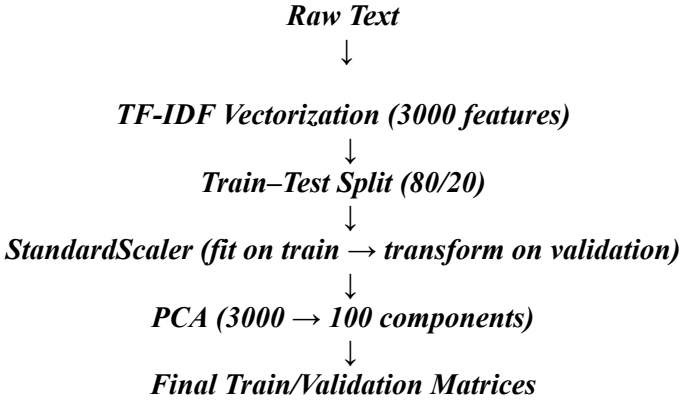

Fig. 3. Preprocessing and feature-engineering pipeline used for sentiment classification.

The normalized text was transformed into numerical features using the Term Frequency–Inverse Document Frequency (TF-IDF) method. TF-IDF assigns higher weights to discriminative terms by combining local term frequency with global inverse document frequency, following the formulation described by Robertson [3]. This representation is well suited for sparse, high-dimensional text classification tasks.

To further improve computational efficiency, the TF-IDF matrix was standardized using scikit-learn's preprocessing tools [1], ensuring scale compatibility for algorithms such as Logistic Regression and K-Nearest Neighbors. Principal Component Analysis (PCA) was then applied to reduce dimensionality by projecting the TF-IDF features into a lower-dimensional subspace while preserving essential variance structure. This step helps mitigate computational cost and reduces noise, particularly for distance-based models.

A visual summary of the preprocessing pipeline is shown in **Fig. 3**, which outlines the transformation of raw text into numerical features used for supervised learning. The final processed dataset was split into training and validation sets using an 80/20 ratio, ensuring that all models were evaluated on standardized and consistent representations.

VII. MODEL DEVELOPMENT

This section describes the supervised learning models evaluated for sentiment classification, along with their underlying principles and implementation details. All models were trained using the standardized and dimensionally reduced TF-IDF features described in the previous section. Model development and evaluation were conducted using the scikit-learn library [1].

A. Logistic Regression

Logistic Regression is a linear classification model commonly used for high-dimensional text data due to its computational efficiency and stable performance. The model estimates class probabilities using the logistic function and determines the decision boundary by maximizing the likelihood of the observed labels. Regularization is applied to control model complexity and mitigate overfitting. Logistic Regression is implemented through scikit-learn's linear modeling framework [1], making it a strong baseline for text classification tasks.

B. K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a non-parametric, distance-based classifier that predicts labels by examining the majority class among the k closest samples in the feature space. Because KNN relies on distance computations, the input features must be standardized to ensure consistent scaling. High-dimensional TF-IDF features may degrade KNN performance due to the curse of dimensionality; therefore, PCA-reduced components were used as inputs. A range of k values was tested to identify the optimal neighborhood size for validation performance.

C. Decision Tree (CART)

The Classification and Regression Tree (CART) algorithm constructs a tree-based model by recursively partitioning the feature space according to impurity-reduction metrics such as Gini index or entropy. Decision Trees can model nonlinear relationships but are susceptible to overfitting, especially with sparse, high-dimensional text features. To mitigate this, hyperparameters such as tree depth and minimum samples per split were constrained. CART serves as a complementary non-linear baseline for comparison against linear and probabilistic models.

D. Multiple Linear Regression (MLR)

Multiple Linear Regression, although traditionally used for continuous outcomes, can be adapted to binary classification by applying a threshold to predicted values. In this project, MLR was included as a baseline to assess how well a purely linear model performs without a logistic transformation. The model was trained on PCA-reduced features to ensure numerical stability. While MLR can capture linear patterns, it is not theoretically optimal for classification tasks, and its performance is expected to be lower than more specialized methods such as Logistic Regression or Naïve Bayes.

E. Multinomial Naïve Bayes

The Multinomial Naïve Bayes classifier is a probabilistic model widely used for text classification because it directly models word-count or TF-IDF–derived feature distributions. The algorithm assumes conditional independence between terms given the class label and computes class probabilities based on term frequencies. McCallum and Nigam [2] demonstrated the strong performance of Multinomial Naïve Bayes in sparse text domains, making it particularly suitable for TF-IDF representations. This model is computationally efficient, robust to noise, and performs well even when vocabulary size is large.

VIII. RESULTS AND DISCUSSION

This section presents the performance of five machine learning models applied to classify restaurant review sentiment. The models evaluated were Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree (CART), Multinomial Naïve Bayes, and Multiple Linear Regression (MLR). All models were trained on TF-IDF vector representations of text, derived following the principles of inverse document frequency described in [3]. Model implementation and evaluation were conducted using the scikit-learn library [1].

A. Model Performance Overview

The overall performance of each model is summarized in **Table I**, with a visual accuracy comparison shown in **Fig. 4**.

TABLE I
Accuracy and F1-Score Summary

Model	Accuracy	F1-Score
Logistic Regression	0.908	0.939
KNN	0.847	0.902
CART	0.853	0.905
Naïve Bayes	0.926	0.955
MLR	0.920	0.951

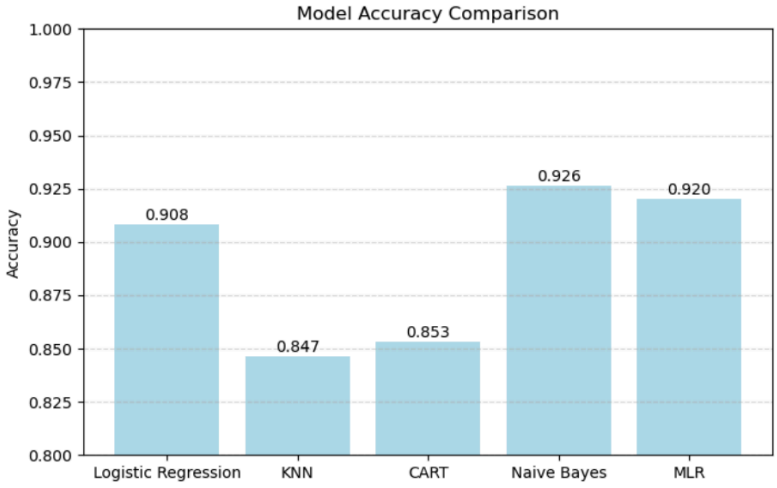

Fig. 4. Accuracy comparison for all models.

Among all evaluated models, Multinomial Naïve Bayes achieved the highest performance with 92.6% accuracy and an F1-score of 0.955. This aligns with prior research showing that Naïve Bayes models are highly effective for TF-IDF and word-frequency-based text classification tasks [2], [3].

Logistic Regression also performed strongly, achieving 90.8% accuracy and an F1-score of 0.939, benefiting from linear separability often present in TF-IDF feature spaces.
KNN demonstrated the weakest performance (84.6% accuracy), which is expected due to the poor behavior of distance-based algorithms in high-dimensional sparse spaces [3].

CART and MLR produced mid-range results, with MLR achieving 92.0% accuracy, closely matching the performance of Logistic Regression.

B. Confusion Matrix Analysis

To better interpret prediction patterns, confusion matrices for selected models were analyzed.

1) Logistic Regression

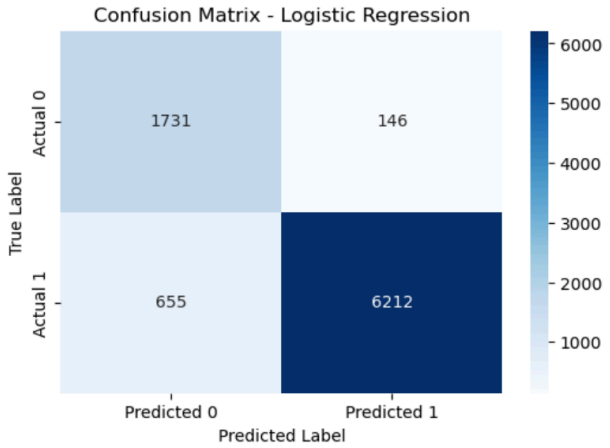

Fig. 5. Confusion matrix for Logistic Regression.

Logistic Regression showed balanced performance but misclassified a moderate number of negative reviews.

2) Multinomial Naïve Bayes

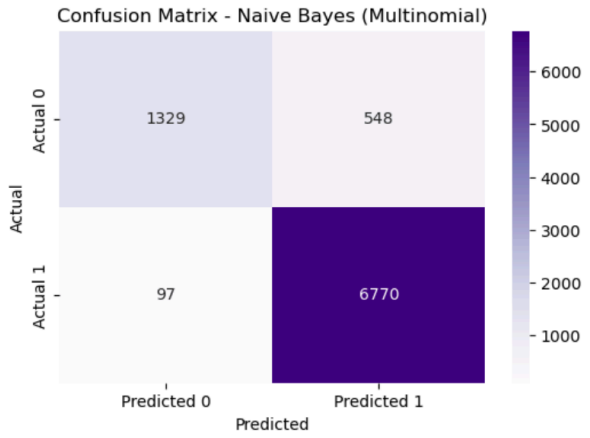

Fig. 6. Confusion matrix for Multinomial Naïve Bayes.

The Naïve Bayes classifier achieved the lowest misclassification rates across all models. The strong results are consistent with the multinomial event model described in [2], which captures term-frequency patterns effectively for text classification.

3) K-Nearest Neighbors

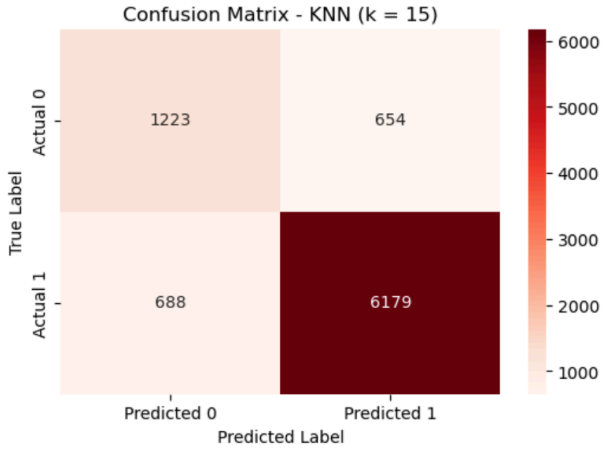

Fig. 7. Confusion matrix for KNN (k = 15).

KNN exhibited lower precision and recall, reflecting the sensitivity of distance-based methods to high-dimensional TF-IDF features, which reduce the discriminative power of similarity metrics [3].

C. KNN Hyperparameter Tuning

The KNN model was evaluated across values of k ranging from 1 to 15. The tuning curve is shown in **Fig. 8**.

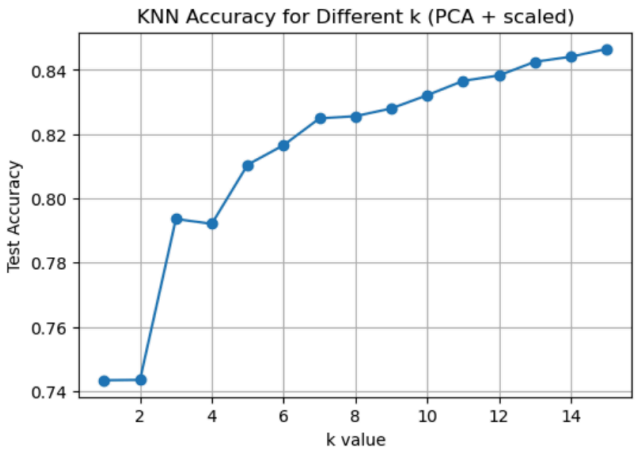

Fig. 8. Accuracy vs. k-value for KNN.

Accuracy increased gradually with larger k, stabilizing at k = 15, which was selected as the final parameter. Even with tuning, KNN remained the weakest performing model.

D. Best-Performing Model Summary

Among all models tested, **Multinomial Naïve Bayes** achieved the best overall performance in terms of accuracy, F1-score, and confusion-matrix behavior. The classifier's probabilistic modeling approach effectively captures TF-IDF word-frequency distributions, consistent with findings in [2] and [3]. Therefore, it is identified as the most suitable model for sentiment classification in this study using the provided Yelp dataset [4].

IX. CONCLUSION

This study evaluated several machine-learning models for sentiment classification of restaurant reviews using TF-IDF features derived from the Yelp Open Dataset. After applying text normalization, TF-IDF vectorization, scaling, and dimensionality reduction, five supervised classifiers—Logistic Regression, K-Nearest Neighbors, Decision Tree (CART), Multiple Linear Regression, and Multinomial Naïve Bayes—were trained and assessed using accuracy, F1-score, and confusion-matrix analysis.

Among all models tested, the Multinomial Naïve Bayes classifier achieved the highest performance, with an accuracy of 92.6%, outperforming both linear and non-linear alternatives. Its effectiveness is consistent with prior research demonstrating that probabilistic word-frequency models perform well on sparse, high-dimensional text data. Logistic Regression and Multiple Linear Regression provided competitive results, while KNN and CART showed comparatively lower accuracy due to their sensitivity to dimensionality and noisy feature spaces.

Although the results are strong, several limitations should be noted. The dataset excludes neutral reviews, which simplifies the classification task but may limit generalizability to real-world scenarios where sentiment often lies on a continuum. Additionally, the use of TF-IDF does not capture contextual or semantic relationships between words, which more advanced models could address.

Future work could explore word-embedding techniques, transformer-based architectures, or multiclass sentiment analysis to capture richer linguistic patterns. Nevertheless, the findings of this project

demonstrate that classical machine-learning methods, particularly Multinomial Naïve Bayes, provide an efficient and reliable approach for large-scale sentiment classification of restaurant reviews.

REFERENCES

[1] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[2] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in *Proc. AAAI Workshop Learning for Text Categorization*, 1998.

[3] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for IDF," *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[4] Yelp Inc., "Yelp Open Dataset," 2023. [Online]. Available: https://www.yelp.com/dataset

[5] ML1 Group8, "Restaurant Review Sentiment Classification Notebook," 2025.