Title: Analyzing Sentiment in Customer Reviews using Advanced Machine Learning Techniques and Feature Engineering

Team Members: Harshaditya Kumar; Parinay Karande ; Hemanth Mohan

Group 5

I. Project Goal/Background/Motivation

Our project aims to develop a machine learning model to predict sentiment in customer reviews using historical data and explore patterns and insights in the data using unsupervised methods. The motivation behind this study is to provide businesses with valuable insights to better understand their customers and enhance their products and services based on the analysis of customer reviews.

II. Research Questions and Hypotheses

1. How does the performance of the supervised learning classifier vary across different time periods within the dataset?   
Hypothesis: The performance of the classifier will vary significantly across different time periods due to changes in review trends and other external factors that impact customer sentiment.

2. Are there significant differences in sentiment values between different time periods within the dataset?  
Hypothesis: There are significant differences in sentiment values between different time periods, due to changes in customer preferences, market trends, and other external factors that impact customer reviews.

3. Which scaling, or transformation techniques are most effective for improving model performance?  
Hypothesis: The most effective scaling or transformation technique for improving model performance will depend on the specific features and characteristics of the dataset, as well as the choice of machine learning algorithm.

4. Are there any instances within the dataset that significantly impact the accuracy of the machine learning models?   
Hypothesis: There are some instances within the dataset that have a significant impact on the accuracy of the machine learning models, such as outliers or periods of extreme sentiment.

III. Dataset

Dataset link: [Link to a suitable customer reviews dataset, e.g., <https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products>]

The features/variables included in the dataset are:

* Product ID: Unique identifier of the product
* Reviewer ID: Unique identifier of the reviewer
* Review Text: The text of the customer review
* Rating: The rating given by the customer (1-5 stars)
* Review Date: The date of the review

The primary significance of this dataset is that it provides a rich source of historical data on customer reviews, which can be used to develop machine learning models for predicting sentiment. The sentiment, derived from the rating and review text, is the key target variable in the dataset, which can be predicted using a variety of features such as previous ratings, review text, and product information.

IV. Methodology

1. Data Collection and Preprocessing: We will use the customer reviews dataset from the provided link. The dataset will be preprocessed, including tokenization, stop word removal, and stemming.

2. Feature Engineering: We will employ TF-IDF and word embeddings for vector representation. Additionally, we will implement one-two statistical tests (e.g., Chi-Square or ANOVA) to select the most relevant features and one-two scaling/transformation techniques (e.g., Min-Max Scaling or Standardization) to normalize the data.

3. Supervised Models: We will implement a classifier (e.g., Logistic Regression) and a regressor (e.g., Support Vector Regression) for predictive modeling.

4.Unsupervised Models: We will apply a clustering algorithm (e.g., K-Means) for data reduction or pattern discovery.

5.Outlier/Anomaly Detection: We will use one-two outlier/anomaly detection methods (e.g., Isolation Forest or DBSCAN) to identify and handle anomalies in the data.

6. Model Evaluation: We will split the data into train-test sets using k-fold cross-validation and assess the models' performance using metrics such as accuracy, precision, recall, and F1 score.

7. Advanced Techniques: We will incorporate an advanced method (e.g., boosting, KNN imputation, or using deep learning with Keras) to improve the results of the models.

8. Visualization: We will use visualization techniques such as bar plots, confusion matrices, and word clouds to analyze and present our findings.

V. Technical Tools and Languages

We will use Python as the primary programming language for implementing the data mining methods and models, along with various Python libraries such as scikit-learn, Pandas, NumPy, Matplotlib, Seaborn, TensorFlow, and Keras. We will also use Jupyter notebooks as the primary tool for data analysis and visualization.

VI. Resources

1. Customer reviews dataset: The dataset available at the given URL ([Link to the chosen dataset]) will be required to perform the analysis.

2. Python: The Python programming language and its associated libraries will be used to implement the various data mining methods and models.

3. Jupyter notebooks: Jupyter notebooks will be the primary tool for data analysis and visualization.

VIII. Conclusion

By analyzing sentiment in customer reviews using machine learning techniques and feature engineering, we aim to provide businesses with valuable insights into their customers' preferences and opinions. This project will not only help improve products and services based on the analysis of customer reviews, but also contribute to the advancement of sentiment analysis techniques and applications.