Project 3

Student Names and IDs

- Zeyu Liao(zeyu9, 667691486, MCS)
- Lu,Yangliang (yl164 661963604 MCS)
- Yan, Zexi (zexiyan2, 651826615 MCS)

Project Objectives

- construct a binary classification model capable of predicting the sentiment of a review
- vocabulary size that is less than or equal to 1000
- The evaluation metric (AUC) is equal to or greater than 0.96 across all five test data sets

---

Section 1: Technical Details

Libraries used

The following libaries are used in this project.

- pandas
- nltk
- TfidfVectorizer, roc_auc_score, LogisticRegressionCV, stats from sklearn
- load from pickle
- Parallel, delayed, cpu_count from joblib (optional)

Generating Vocabulary

- Aggregate all training and test dataset into one.
- Cleanup HTML tags from the data.
- Use stopwords from `NLTK` library
- Vectorized the data with `TF-IDF` vectorizer with minimum and maximum document frequency thresholds at `0.001` and `0.5` respectively.
- A `t-test` is performed for each feature (word/phrase) to determine if there is a statistically significant difference in the usage of that feature between positive and

negative reviews.
- Features (words/phrases) with a `p-value` less than `0.05` are considered statistically significant and are selected.
- Use the `2000` most significant words/phrase to build another vecorizer and transform the training data.
- A logistic regression model with `L1 regularization` (Lasso) is trained on the significant features to further refine the feature selection.
- The top `1000` features are selected to create the final vocabulary.

Model Training and Prediction

- Read the training and test data.
- Creates a `TF-IDF` vectorizer with the predefined 1000 vocabulary.
- Configures the vectorizer to convert text to lowercase, remove stopwords, and use n-grams (`1-gram` and `2-gram`).
- Use the vectorizer to transform the training and test data.
- Fits a `Logistic Regression model` with cross-validation (LogisticRegressionCV) on the training data. Configure the model with `CV = 5` , `max_iter = 10000` .
- Predicts the probabilities of the test reviews being positive.
- Calculates and returns the ROC AUC score for the predictions.

Section 2: Results

Evaluation Metrics:

Data	AUC Score
split_1	0.9607600025218678
split_2	0.9611106214308613
split_3	0.9604750611738162
split_4	0.9613678728754387
split_5	0.9608693665237437

Run Time:

Task	Time
Generating vocabulary	over 120 mins

Task	Time
Traing model and make prediction	69 seconds

Computer configuration:

Item	Spec
Computer	MacBook Pro 2017
CPU	2.9 GHz Quad-Core Intel Core i7
GPU	Radeon Pro 560 4 GB
Memory	16 GB 2133 MHz LPDDR3