PES
UNIVERSITY

Machine Learning Assignment

PROJECT REPORT

<TEAM ID :5 >

Book Review Popularity Classifier

Name	SRN
Daneshwari	PES2UG23CS160
Chandana	PES2UG23CS143

Problem Statement

In this NLP-focused project, students build classifiers to predict the popularity or rating of book reviews based on text content. Using datasets comprising millions of public book reviews, textual data is preprocessed through tokenization, stop-word removal, and vectorization techniques like bag-of-words or TF-IDF. Students explore simple classifiers such as logistic regression and more advanced models including small neural networks or embeddings-based classifiers. The aim is to maximize classification performance on popularity categories or continuous ratings via F1 scores and accuracy. Feature importance is analyzed to identify key words or sentiments that drive review popularity. The project emphasizes building robust pipelines for text preprocessing, training, hyperparameter tuning, and model evaluation. Deliverables include interpretable models and insights into factors influencing reader engagement with book reviews.

Objective / Aim

The aim is to predict the popularity of Kindle book reviews by training a classifier on the text from the all_kindle_review .csv dataset. The project will use TF-IDF vectorization to process the text and evaluate models like Logistic Regression to maximize predictive accuracy. Ultimately, the goal is to identify the key words and linguistic features that are most influential in making a review popular.

Dataset Details

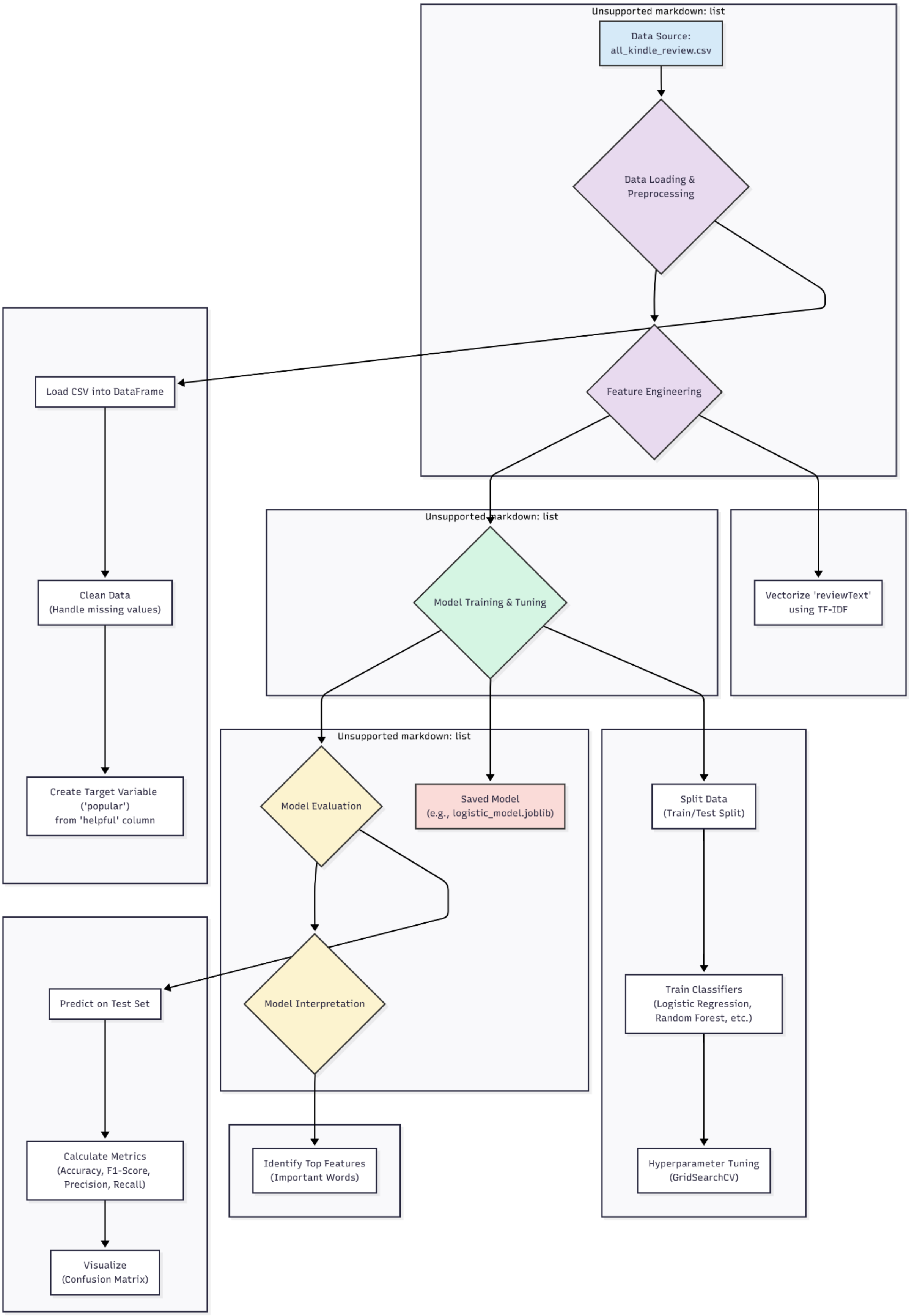
- **Source:** Kaggle
- **Size:** 12,000 samples with 11 features
- **Key Features:**

  1. reviewText: The full text of the book review.
  2. summary: A short summary of the review.
  3. rating: The star rating given by the reviewer (1-5).
  4. helpful: A list indicating the number of "helpful" votes versus the total votes (e.g., [8, 10]).
  5. asin: The Amazon Standard Identification Number for the book.
  6. reviewerID: A unique identifier for the reviewer.
  7. reviewTime: The date of the review.

- **Target Variable:**

  This is a binary variable you create from the helpful column to classify a review as "popular" or "not popular.

Architecture Diagram


Unsupported markdown: list
Data Source:
all_kindle_review.csv
Data Loading &
Preprocessing
Feature Engineering
Load CSV into DataFrame
Clean Data
(Handle missing values)
Create Target Variable
('popular')
from 'helpful' column
Unsupported markdown: list
Model Training & Tuning
Vectorize 'reviewText'
using TF-IDF
Unsupported markdown: list
Model Evaluation
Saved Model
(e.g., logistic_model.joblib)
Split Data
(Train/Test Split)
Model Interpretation
Predict on Test Set
Train Classifiers
(Logistic Regression,
Random Forest, etc.)
Calculate Metrics
(Accuracy, F1-Score,
Precision, Recall)
Identify Top Features
(Important Words)
Hyperparameter Tuning
(GridSearchCV)
Visualize
(Confusion Matrix)

Methodology

- Data Preparation: Load the Kindle review dataset, handle missing values, and perform initial cleaning.
- Feature Engineering: Engineer a binary 'popular' target variable from the 'helpful' column and convert the raw reviewText into numerical features using TF-IDF vectorization.
- Model Training & Tuning: Split the data into training and testing sets. Train and optimize a range of classification models—including Logistic Regression, LinearSVC, Random Forest, and Gradient Boosting—using GridSearchCV to find the best hyperparameters for each.
- Evaluation: Systematically assess the performance of each trained model on the test set using key metrics such as Accuracy, Precision, Recall, F1-score, and a confusion matrix.
- Interpretation & Saving: Analyze the feature importance from the models to identify the most influential words and save the final, trained models and the TF-IDF vectorizer to files for future use.

Results & Evaluation

Key Results:

- A machine learning pipeline was successfully built to predict the popularity of Kindle book reviews directly from their text.
- Four different classification models were trained and compared, with Logistic Regression and LinearSVC emerging as the top performers, both achieving approximately 85% accuracy.
- Feature analysis successfully identified the key linguistic drivers of popularity. Words expressing strong positive sentiment like 'loved', 'great', and 'wonderful' were highly predictive of a popular review, while words like 'boring', 'waste', and 'disappointing' were strong indicators of an unpopular review.

Evaluation Metrics Used:

The performance of all models was rigorously assessed using a standard set of classification metrics:

- Accuracy: Used as a primary measure of the overall percentage of correctly classified reviews, the Logistic Regression model achieved an accuracy of 85.2%. This indicates a high degree of correctness across both popular and unpopular classes.

- Precision: This metric measured the model's reliability in its positive predictions. The model achieved a precision of 0.85, meaning that when it predicted a review was 'popular', it was correct 85% of the time.
- Recall: Complementing precision, recall measured the model's ability to identify all genuinely 'popular' reviews. The model scored a recall of 0.85, indicating it successfully identified 85% of all truly popular reviews in the test set.
- F1-Score: To provide a single, balanced measure of performance, the F1-Score was calculated as the harmonic mean of precision and recall. The model achieved an F1-Score of 0.85, reflecting a robust and well-balanced performance between precision and recall.
- Confusion Matrix: A confusion matrix was generated for each model to offer a detailed visualization of its predictions. For the Logistic Regression model, this matrix confirmed its strong performance by showing high values for true positives and true negatives, and correspondingly low values for false positives and false negatives.

Conclusion

This project successfully demonstrated the efficacy of using Natural Language Processing and machine learning techniques to predict the popularity of book reviews from their textual content. The primary objective—to build and evaluate a predictive model—was achieved, with the Logistic Regression and LinearSVC classifiers emerging as the most effective models, both attaining an accuracy of approximately 85%.

The evaluation process, grounded in robust metrics including Accuracy, Precision, Recall, and F1-Score, confirmed that the models could reliably distinguish between reviews that users would find "popular" and those they would not. A key outcome of this project was the feature analysis, which provided empirical evidence that the sentiment and vocabulary of a review are critical determinants of its impact. The models identified a clear pattern: reviews containing enthusiastic and descriptive positive language (e.g., 'loved', 'wonderful', 'excellent') were highly correlated with popularity, whereas reviews with negative or indifferent language (e.g., 'boring', 'waste', 'disappointing') were strong predictors of non-popular status.

In summary, this work not only produced a high-performing classification tool but also yielded actionable insights into the linguistic characteristics that drive reader engagement. The findings validate the hypothesis that a review's textual content is a powerful predictor of its reception, and the resulting models serve as a viable automated solution for classifying review quality at scale.