Review Popularity Classification based on Syntactic, Semantic, and Non-Textual Heuristics

Nalin Gabriel Prabindh - PES2UG23CS360 Mithun R - PES2UG23CS341

Abstract

This project aims to classify the popularity of Goodreads book reviews using Logistic Regression, XGBoost, and Neural Networks. The model leverages syntactic, semantic, and non-textual heuristics derived from the review text and metadata to predict whether a review is popular (above the 75th percentile of total votes). The approach combines handcrafted linguistic features with TF-IDF text representations in a unified multi-model learning pipeline.

System Overview & Architecture

1. Data Handling: Reviews are loaded from the GoodReads UCSD Dataset and sampled (`sample_size`). Popularity is defined as reviews with `n_votes` above the 75th percentile.

2. Feature Engineering: Three groups of features are extracted:

- **Syntactic:** character count, word/sentence length, punctuation ratios, uppercase ratio.
- **Semantic:** VADER sentiment (pos/neg/compound), TextBlob polarity/subjectivity, lexical diversity.
- **Non-Textual:** rating, extreme rating indicator, number of comments, reading progress (`has_read`, `has_started`).

These are combined with a 100-dimensional TF-IDF vector (unigrams + bigrams).

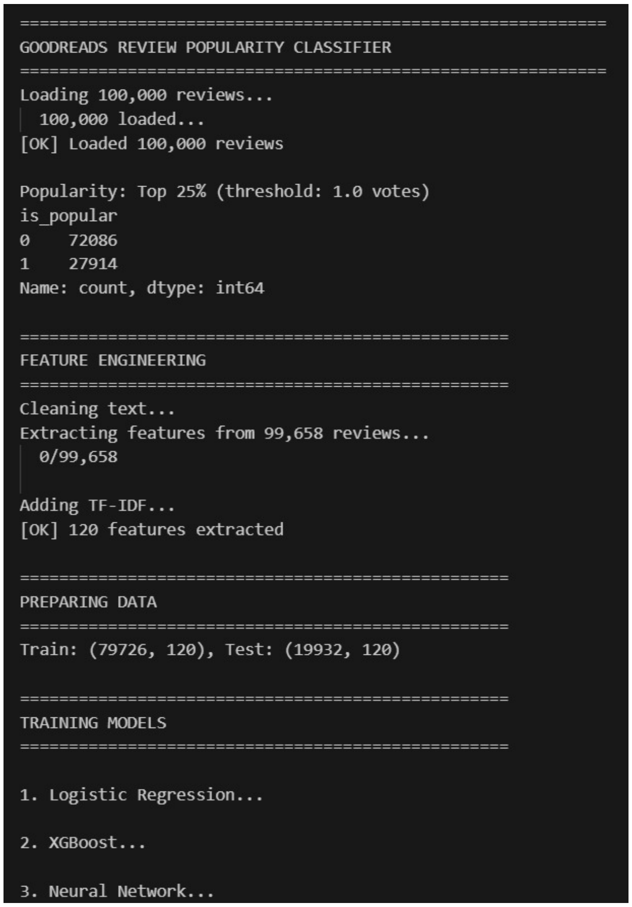
3. Data Preparation: All features are standardized using `StandardScaler`. The dataset is split 80:20 for training and testing using stratified sampling.

4. Model Training:

- **Logistic Regression:** with class balancing.
- **XGBoost:** tree-based boosting with imbalance correction.
- **Neural Network:** dense layers with ReLU, dropout, batch normalization, and adaptive learning rate scheduling.

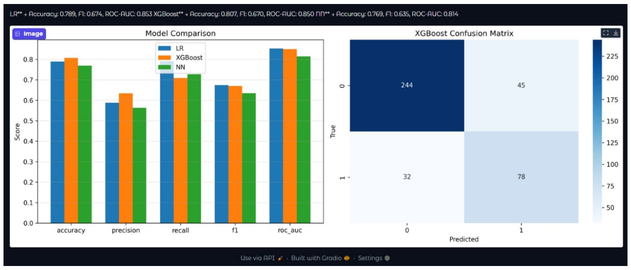
5. Evaluation: Models are compared on accuracy, precision, recall, F1, and ROC-AUC. A bar chart and confusion matrix are generated. Features are cached for fast re-runs.

Output: Performance metrics (console), `png images`, and cached feature files.



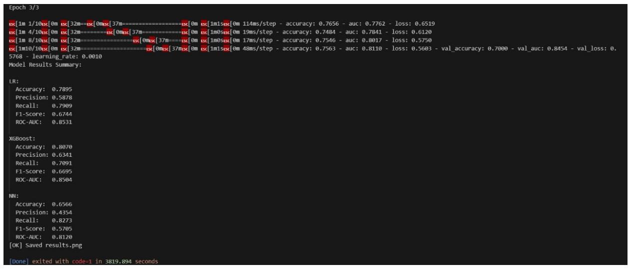
```
============================================================
GOODREADS REVIEW POPULARITY CLASSIFIER
============================================================
Loading 100,000 reviews...
  100,000 loaded...
[OK] Loaded 100,000 reviews

Popularity: Top 25% (threshold: 1.0 votes)
is_popular
0    72086
1    27914
Name: count, dtype: int64

==================================================
FEATURE ENGINEERING
==================================================
Cleaning text...
Extracting features from 99,658 reviews...
  0/99,658


Adding TF-IDF...
[OK] 120 features extracted

==================================================
PREPARING DATA
==================================================
Train: (79726, 120), Test: (19932, 120)

==================================================
TRAINING MODELS
==================================================


1. Logistic Regression...

2. XGBoost...

3. Neural Network...
```

Figure 1: Feature Engineering and Data cleaning.



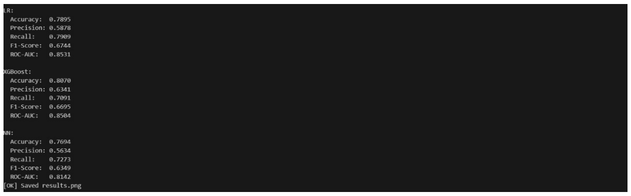
LR** → Accuracy: 0.789, F1: 0.674, ROC-AUC: 0.853 XGBoost** → Accuracy: 0.807, F1: 0.670, ROC-AUC: 0.850 NN** → Accuracy: 0.769, F1: 0.635, ROC-AUC: 0.814

Figure 2: Confusion Matrix,BarPlots etc.



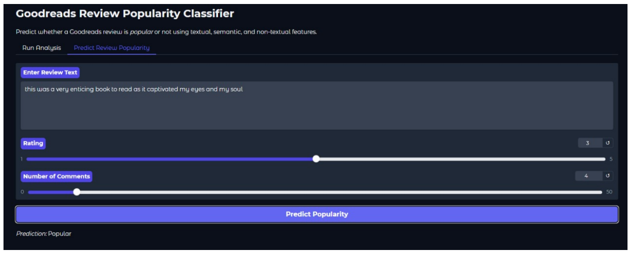
```
Epoch 3/3

ESC[1m 1/10ESC[0m ESC[32m==ESC[0mESC[37m==================ESC[0m ESC[1m1sESC[0m 114ms/step - accuracy: 0.7656 - auc: 0.7762 - loss: 0.6519
ESC[1m 4/10ESC[0m ESC[32m========ESC[0mESC[37m============ESC[0m ESC[1m0sESC[0m 19ms/step - accuracy: 0.7484 - auc: 0.7841 - loss: 0.6120
ESC[1m 8/10ESC[0m ESC[32m================ESC[0mESC[37m====ESC[0m ESC[1m0sESC[0m 17ms/step - accuracy: 0.7546 - auc: 0.8017 - loss: 0.5750
ESC[1m10/10ESC[0m ESC[32m====================ESC[0mESC[37mESC[0m ESC[1m1sESC[0m 48ms/step - accuracy: 0.7563 - auc: 0.8110 - loss: 0.5603 - val_accuracy: 0.7000 - val_auc: 0.8454 - val_loss: 0.
5768 - learning_rate: 0.0010
Model Results Summary:

LR:
  Accuracy:  0.7895
  Precision: 0.5878
  Recall:    0.7909
  F1-Score:  0.6744
  ROC-AUC:   0.8531

XGBoost:
  Accuracy:  0.8070
  Precision: 0.6341
  Recall:    0.7091
  F1-Score:  0.6695
  ROC-AUC:   0.8504

NN:
  Accuracy:  0.6566
  Precision: 0.4354
  Recall:    0.8273
  F1-Score:  0.5705
  ROC-AUC:   0.8120
[OK] Saved results.png

[Done] exited with code=1 in 3819.894 seconds
```

Figure 3: Epoch-Wise Training Outputs



```
LR:
  Accuracy:  0.7895
  Precision: 0.5878
  Recall:    0.7909
  F1-Score:  0.6744
  ROC-AUC:   0.8531

XGBoost:
  Accuracy:  0.8070
  Precision: 0.6341
  Recall:    0.7091
  F1-Score:  0.6695
  ROC-AUC:   0.8504

NN:
  Accuracy:  0.7694
  Precision: 0.5634
  Recall:    0.7273
  F1-Score:  0.6349
  ROC-AUC:   0.8142
[OK] Saved results.png
```

Figure 4: Performance Metrics



Goodreads Review Popularity Classifier

Predict whether a Goodreads review is *popular* or not using textual, semantic, and non-textual features.

Enter Review Text: this was a very enticing book to read as it captivated my eyes and my soul

Rating: 3

Number of Comments: 4

Prediction: Popular

Figure 5: Sample Output