

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.

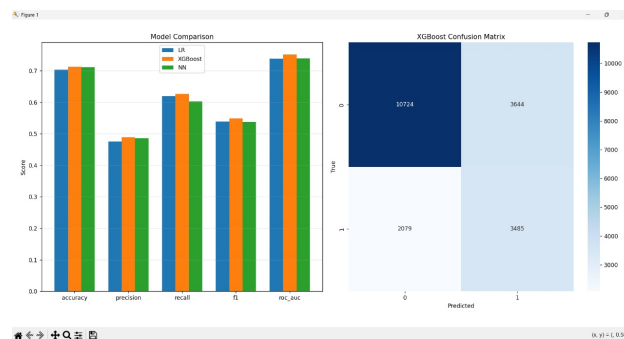


Figure 2: Confusion Matrix, BarPlots etc.

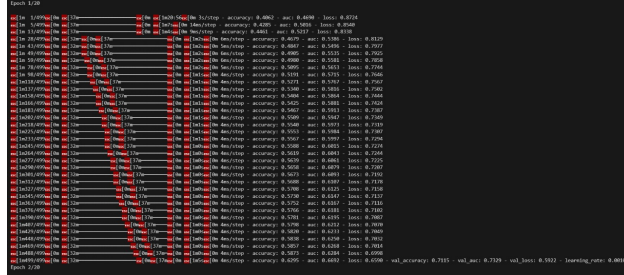


Figure 3: Epoch-Wise Training Outputs

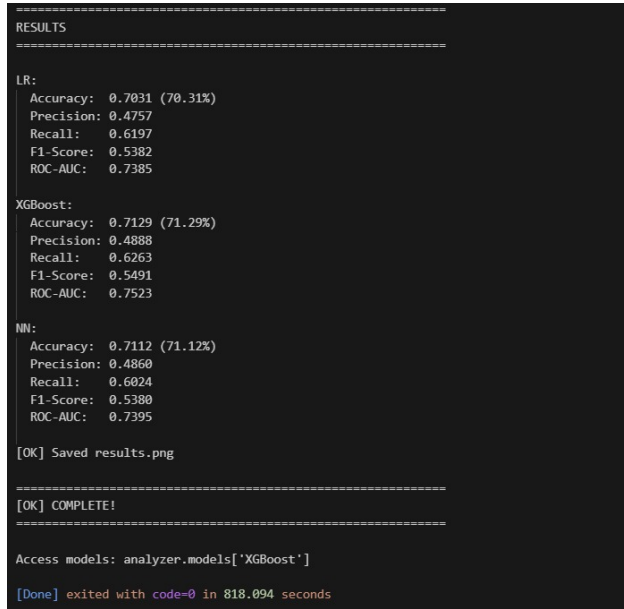


Figure 4: Performance Metrics