Fake vs Genuine Review Classification using Support Vector Machines and Natural Language Processing

Muhammad Ghulam Ali
MS Business Analytics — Machine Learning Practitioner
github.com/Muhammad-Ghulam-Ali

July 5, 2025

Project Overview

This project presents a machine learning solution to classify product reviews as either **genuine** or **fake** using a combination of behavioral signals and text features. A custom heuristic-based labeling strategy was developed to simulate real-world spam detection in the absence of verified labels.

The project leverages the *Amazon Fine Food Reviews* dataset, performs extensive preprocessing and feature engineering, and applies a Support Vector Machine classifier on TF-IDF transformed text to detect fake reviews.

Dataset Description

- **Source**: Amazon Fine Food Reviews (available on Kaggle)
- **Size**: Approximately 568,000 reviews
- **Fields Used**: `Text`, `UserId`, `HelpfulnessNumerator`, `HelpfulnessDenominator`

Labeling Strategy: Heuristic-Based Classification

Labeled as Fake (0) if any of the following conditions apply:

- User has submitted only one review.
- Review contains fewer than 20 words.
- Review received zero helpfulness votes.
- Review text is a duplicate of another entry.

Labeled as Genuine (1) if all of the following conditions are met:

- User has written multiple reviews.
- Review is long and detailed.
- At least one helpfulness vote received.
- Text is unique in the dataset.

Preprocessing and Feature Engineering

Text Cleaning

- Lowercasing
- Removal of punctuation, HTML tags, and non-alphanumeric characters
- Stopword removal using NLTK
- Lemmatization with `WordNetLemmatizer`

Feature Construction

- **review_word_count**: Number of words in a review
- **duplicate_text**: Boolean indicating if the text is a duplicate
- **user_review_count**: Number of reviews submitted by the user
- **helpfulness_ratio**: Ratio of helpful votes to total votes

Text Vectorization

The cleaned review text is vectorized using **TF-IDF** (Term Frequency-Inverse Document Frequency), transforming it into a numerical feature matrix suitable for model input.

Modeling Approach

Data Splitting

- 80/20 train-test split
- Stratified on the target class to preserve class distribution
- Fixed random seed (42) for reproducibility

Class Imbalance Handling

- Used **SMOTE** (Synthetic Minority Over-sampling Technique) to generate synthetic samples for the minority class (fake reviews)

Model

A **Linear Support Vector Classifier (LinearSVC)** was trained on the TF-IDF features.

Model Evaluation

The model was evaluated using precision, recall, F1-score, and overall accuracy.

Sample Classification Report

```
              precision    recall  f1-score   support

           0       0.84      0.88      0.86      XXXX
           1       0.89      0.85      0.87      XXXX

    accuracy                           0.86     XXXXX
   macro avg       0.87      0.86      0.86     XXXXX
weighted avg       0.87      0.86      0.86     XXXXX
```

Key Contributions

- Developed a custom labeling mechanism based on behavioral and text-based rules.
- Implemented an end-to-end NLP and classification pipeline combining TF-IDF, SMOTE, and SVM.
- Integrated domain-aware feature engineering to simulate a real-world fraud detection scenario.
- Achieved high classification accuracy with balanced performance across classes.

Future Work

- Utilize pre-trained language models (e.g., BERT) for enhanced context understanding.
- Incorporate human-annotated labels to validate and refine heuristic methods.
- Deploy the model as a web application for real-time fake review detection.
- Expand analysis to track evolving review manipulation patterns.

Author

Muhammad Ghulam Ali
MS Business Analytics
github.com/Muhammad-Ghulam-Ali
mghulamali888@gmail.com