Classifying Books by Genre Based on Reader Reviews with Machine Learning

Jenna Bittner
bittner.87@osu.edu
Department of Physics, The Ohio State University

Motivation

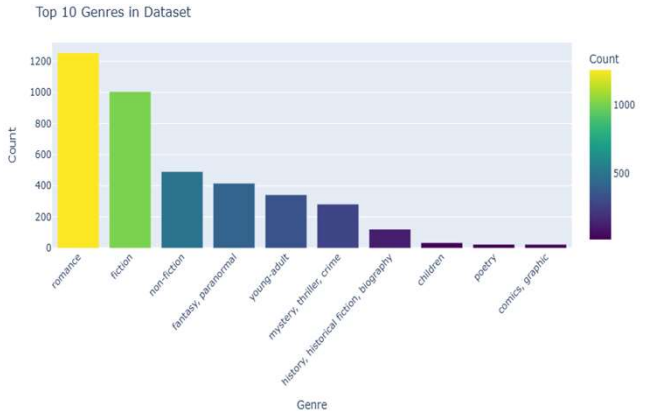
The motivation for this project is the growing need for a recommendation system to help readers navigate vast digital libraries. Existing systems can be enhanced by using classification techniques based on user-generated book reviews for more precise categorization. The primary objective is to develop a genre classification model that predicts a book's genre from its reviews. By effectively categorizing books into genres, personalized book recommendations can be enhanced, allowing readers to easily discover books that align with their individual interests.

Dataset

The dataset for this project is sourced from Mengting Wan's Goodreads dataset, containing 15 million reviews for 2 million books. For this project, 10,000 book reviews are randomly selected.

Data preparation involves using the Natural Language Toolkit (NLTK) for preprocessing

- Reviews are tokenized
- Text is converted to lowercase
- Stop words, non-alphanumeric characters, and short words are removed
- Lemmatization reduces words to their base forms



Methods

Book reviews are transformed using TF-IDF vectorization

- Converts raw text data into numerical values suitable for machine learning
- Each word in the text is assigned a TF-IDF score, reflecting its importance within the document and across the entire dataset

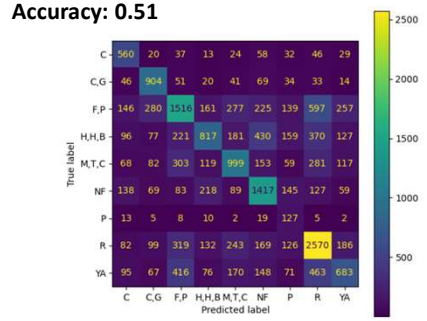
For genre classification, a Random Forest classifier is used

- Trained with bootstrap sampling, which creates subsets of training data, each constructing a decision tree
- Each tree aims to maximize genre separation
- Every tree in the forest predicts a genre
- The genre predicted by the majority trees is the final output

This approach leverages the collective decision-making of multiple trees to improve the accuracy and robustness of the genre classification.

Results

Accuracy: 0.51



The model performs effectively for well-defined categories but needs improvement for smaller classes

The model performs significantly better than random guessing

Discussion

The model's performance varied across different genres. It achieved its best results in the comics, graphics, and romance genres, with high precision and recall, reflected in f1-scores of 0.64 and 0.61. Non-fiction also performed well with an f1-score of 0.56. However, the model struggled with poetry and young-adult genres, showing low precision and recall, particularly in poetry with a precision of 0.14. Larger genres like fantasy and history had moderate performance, with f1-scores of 0.46 and 0.40. Children and mystery genres showed balanced but moderate performance with f1-scores of 0.54 and 0.47. The ROC curve demonstrated the model's strong ability to distinguish between classes, while the precision-recall curve showed a consistent balance between precision and recall across most genres. The confusion matrix contains high values in the diagonal, but the off-diagonals further emphasize the challenges in classifying overlapping genres, suggesting the need for further optimization for underrepresented categories.

Conclusions and Future Work

The classification model performs well in genres like comics, graphics, and romance but struggles with smaller or ambiguous genres such as poetry and young-adult. With an overall accuracy of 0.51 and an AUC of 0.93, the model shows moderate performance and strong class differentiation. Despite this, there is room for improvement, especially in handling class imbalances and optimizing parameters. Future work could explore using BERT to capture deeper semantic relationships and improve performance. Multi-label classification could address issues with books belonging to multiple genres, enhancing the model's accuracy and representation of complex, real-world scenarios.

References:
Wan, Mengting. "Goodreads Dataset." https://mengtingwan.github.io/data/goodreads.html, Accessed2024.[2] Dpm-a. "ML-NLP-Goodreads-Book-Classification." GitHub. https://github.com/Dpm-a/ML-NLP-Goodreads-Book-Classification/, Accessed 2024.[3] Suresh, P. and Manoharan, R. "A Survey on Book Genre Classification System using Machine Learning."Philippine Statistics Association, 2024. https://www.philstat.org/index.php/MSEA/article/view/1597,Accessed 2024.