EECS 595 Project Proposal

Andrew Zhuang
University of Michigan
azhuang@umich.edu

1 Introduction

The goal of this project is to develop an NLP classifier that can accurately identify a book's genre given a short description excerpt. This project is driven by the question of whether NLP classifiers can effectively categorize books into genres based solely on their descriptions and how precise the categorization can be. An effective classifier can be used to streamline the way readers discover books, allowing them to quickly filter through books to find ones that match their preferences. It would also be an invaluable tool for publishers and library managers to streamline content organization and recommendation for their users. Finally, authors and book marketers would gain a better understanding of how well their book descriptions align with the target audience's expectations of the book's genre.

2 NLP Task Definition

The specific problem this project aims to solve is the classification of books into genres based on short description excerpts commonly found on the inside cover or back of a book. The input to our system will be text-only, consisting of short descriptions or blurbs of books, typically ranging from a few sentences to a paragraph in length. These descriptions are expected to provide enough contextual and thematic information for the model to infer the book's genre. The output of the system will be a single classification label indicating the genre the book belongs to, such as Fiction, Non-fiction, Mystery, Science Fiction, Romance, Fantasy, etc. The model will not use any other information that may be correlated to genre including author, publisher, or title.

3 Data

I will use the dataset from Open Library Data Dumps (https://openlibrary.org/developers/dumps) which contains various metadata for books, most importantly a short description and its Dewey number, which will be used to extrapolate a genre. I am currently experiencing issues with processing the dataset and do not have statistics on the number of instances, but these problems should be resolved within a few days. I will use the standard 80/20 train/test split on my data.

4 Related Work

Several previous works have laid the groundwork for classifying book genres. Sobkowicz et al. (2018) embarked on an exploration into book genre detection utilizing short descriptions, a work foundational for understanding the intricacies of short text classification in the literary domain. Their methods pivot around naive Bayes models and semantic enrichment techniques, offering a critical perspective on handling texts that are often incomplete and highly subjective. This study underscores the complexity of genre classification when faced with the brevity and bias inherent in book descriptions, setting a benchmark for semantic analysis in genre classification.

Following this, Ozsarfati et al. (2019) presented a novel approach by focusing on book titles as a primary source for genre classification. Their exploration into the effectiveness of various machine learning algorithms, particularly highlighting the success of LSTM models, marked a significant step towards understanding the potential of titles in genre prediction. While their work demonstrates the applicability of comparative machine learning techniques in genre classification, it also shrinks the scope by its exclusive focus on titles, leaving room for further exploration.

In Thakur and Patel (2021), a novel approach for book genre classification was introduced, focusing on a dictionary-based method that uses both the title and abstract of e-books. This method lever-

ages both features to identify key phrases or words that are indicative of a genre. By analyzing the occurrence and relevance of these keywords within the text, the model assigns a genre based on the highest matching criteria set by the dictionary, creating a curated, context-specific vocabulary that for a given genre.

The approaches taken by Ozsarfati et al. (2019) and Thakur and Patel (2021) differ from my project primarily in the data sources used for classification. Ozsarfati et al. (2019) focused on book titles, which contain limited information about the book's genre. Thakur and Patel (2021) used a dictionary method that combines titles and abstracts, providing a broader textual base but still relying on predetermined keyword mappings which may not capture the full nuance of a book's theme. I will only use book descriptions similar to Sobkowicz et al. (2018), but with more advanced models than naive Bayes.

5 Evaluation

The F1-score will be used as the primary evaluation metric. Since it takes both precision and recall into account, it is particularly useful for assessing performance in a skewed dataset where some genres will be much more popular than others.

The first baseline for comparison will be random performance, classifying book descriptions into genres randomly to establish the lowest expected performance benchmark. The second baseline will employ a frequency-based approach, assigning each book to a genre with a probability based on its frequency in the training set. This will simulate someone randomly guessing genres based on prior knowledge.

6 Work Plan

There are 10 weeks until the project is due on April 26, 2024. My rough plan is as follows:

- Week 1: Pre-process dataset by turning Dewey number into genre
- Week 2: Gather statistics on dataset to inform features, clean missing/incomplete data
- Week 3-5: Create features using techniques learned in class
- Week 6-8: Test and tune various models
- Week 9-10: Focus on fine tuning features and hyperparameters for best model, write report

References

Eran Ozsarfati, Egemen Sahin, Can Jozef Saul, and Alper Yilmaz. 2019. Book genre classification based on titles with comparative machine learning algorithms. In *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pages 14–20.

Antoni Sobkowicz, Marek Kozłowski, and Przemysław Buczkowski. 2018. Reading book by the cover—book genre detection using short descriptions. In *Man-Machine Interactions 5: 5th International Conference on Man-Machine Interactions, ICMMI 2017 Held at Kraków, Poland, October 3-6, 2017*, pages 439–448. Springer.

Vrunda Thakur and Ankit C Patel. 2021. An improved dictionary based genre classification based on title and abstract of e-book using machine learning algorithms. In *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020*, pages 323–337. Springer.