EECS 595 Project Update

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.

This project will use the bert-base-uncased model from HuggingFace to fine-tune a sophisticated system capable of multi-label genre classification from plot summaries. I will use a hierarchical multi-class approach, where each book is assigned one or more genres that are globally arranged in a hierarchy, similar to the Dewey decimal or Library of Congress numbering system. The model will then output predictions as specific as possible while being punished for wrong, overly specific classifications.

2 Data

I was originally going to use the dataset from Open Library Data Dumps (https://openlibrary.org/developers/dumps). However, when I began working with it, only a small percentage of the books had non-empty descriptions and Dewey decimal numbers. This reduced the usable data points to under 100, so I found a new dataset, the CMU Book Summary Dataset (https://www.cs.cmu.edu/~dbamman/booksummaries.html), which contains the title, author, publication date, and a list of genres and a plot summary for each of the 16,559 books extracted from Wikipedia. For example, the data for George Orwell's *Animal Farm* is:

Wikipedia ID	620
Freebase ID	/m/0hhy
Book Title	Animal Farm
Author	George Orwell
Publication Date	1945-08-17
Genres	Roman a clef, Satire, Children's literature, Speculative fiction, Fiction

After cleaning the entries with empty genres or summaries, there are 12,841 rows in the dataset spanning 227 unique genres. I will use the top 15 most frequent genres as possible labels for the books, and since the original dataset often has many genres per book, over 90% of the rows have at least one of the top 15. Figure 1 shows the distribution of these genres in the final dataset containing 11,630 rows. Note that the frequencies will not add up to this number as there are multiple genres per book.

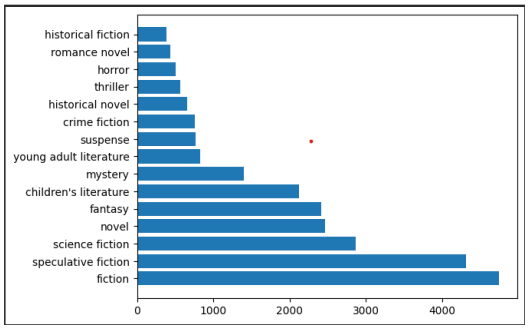

Figure 1: Genre Distribution

3 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 leverages 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.

Aly et al. (2019) introduces a novel application of capsule networks for hierarchical multi-label classification (HMC) of textual data, using the BlurbGenreCollection (BGC) and Web of Science (WOS) datasets for evaluation. This study highlights the superior performance of capsule networks in managing complex label hierarchies and combinations, particularly for rare events and diverse categories, by efficiently encoding category-specific information. The introduction of the BGC dataset and the first-time application of capsule networks to HMC fill a notable gap in the research field, offering insights that could inform the development of more sophisticated hierarchical classification methods, including those based on BERT for genre classification.

Finally, Il Kim et al. (2024) uses logistic regression, SVM, and gradient-boosting methods to predict movie genres from plot summaries, similar to predicting book genres from plot summaries. It emphasizes the importance of text representation, comparing the effectiveness of TF-IDF and bag of words algorithms. The study finds that the logistic regression model with TF-IDF outperforms other models in terms of nDCG and F1-score. They showed that in a multi-genre classification problem, traditional machine learning models, when combined with effective text vectorization techniques, can offer substantial predictive power.

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. While Aly et al. (2019) takes advantage of capsule networks for hierarchical multi-label text classification for their ability to handle complex label hierarchies, I aim to fine-tune BERT, potentially offering a more generalizable approach but without inherent design of capsule networks to directly manage hierarchical labels. Finally, Il Kim et al. (2024) tackles a similar problem in genre classification using plot summaries, with the main difference being in the methodology they used. They explored more traditional machine learning models and text representation techniques, while BERT will potentially allow me to capture a more nuanced understanding of text with its more complex architecture.

4 Methodology

The dataset was first downloaded from CMU's website at https://www.cs.cmu.edu/simdbamman/booksummaries.html. It was then cleaned to remove rows with empty genres or summaries, then other irrelevant columns were removed. Of the rows that remained, they

spanned 227 unique genres so to reduce the label space, I filtered the books to only include the 15 most frequent genres in their labels. This reduced the size of the dataset by less than 10%, since over 90% of books had at least one genre in the top 15. I then mapped each genre to an ID and manually created a genre hierarchy chart, shown in Figure 2. These were used to create a multi-hot vector for each book such that the vector contains a 1 in the corresponding index for all genres of the book and all *parents* of the genres of the book.

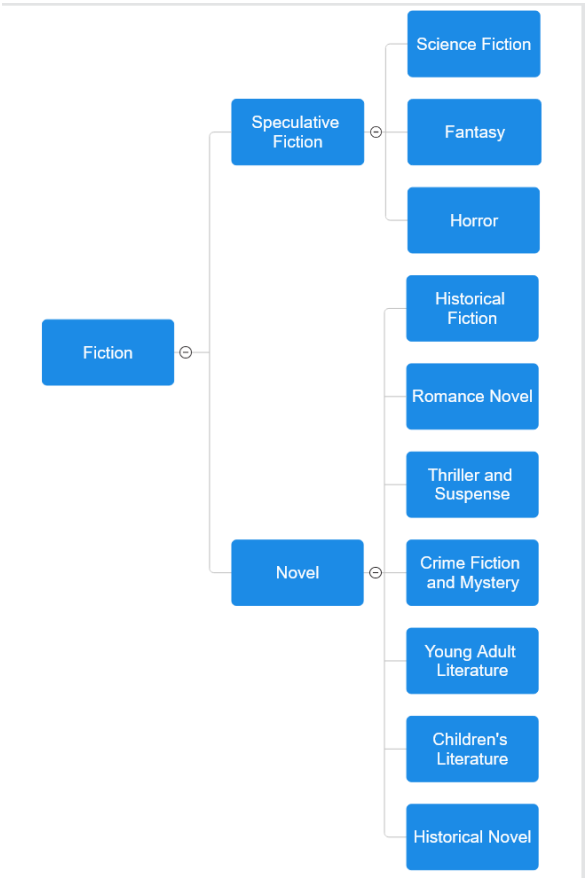

Figure 2: Genre Hierarchy

For the model, I plan to fine-tune *bert-base-uncased* from google-bert on HuggingFace. This will probably be a dense layer that matches the size of the label space with a sigmoid activation function. This will allow me to generate individual probabilities for each genre, which will be included in the final prediction if the probability is higher than a threshold. I will also adapt the loss function for my hierarchical use case so that it penalizes predictions that are “close" less than completely wrong ones.

5 Evaluation and Results

I will use micro-averaged precision, recall, and f-score as evaluation metrics for this project. My first baseline is random performance. For this, I generated a random multi-hot vector for each book and used scikit-learn to calculate the desired metrics. This was repeated 50 times and averaged across the trials. My second baseline is always predicting *Fiction* as the only label, since it is the top level genre and the most common one. The results are shown in the table below.

	Precision	Recall	F-score
Random	0.222	0.532	0.285
Fiction	1.0	0.301	0.462

6 Work Plan

So far, I have downloaded the dataset and completed pre-processing to a state where it can now be fed into the HuggingFace model to begin fine-tuning. There are 3 weeks until the project is due on April 26, 2024. My rough plan is as follows:

- Week 1: Implement hierarchical loss function and fine-tune model
- Week 2: Finish work from week 1 and evaluate using the aforementioned metrics
- Week 3: Finish evaluation and write final report

References

Rami Aly, Steffen Remus, and Chris Biemann. 2019. Hierarchical multi-label classification of text with capsule networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 323–330, Florence, Italy. Association for Computational Linguistics.

Gun Il Kim, Jae Heon Kim, Minkyoung Kim, Taekyoung Kwon, and Beakcheol Jang. 2024. An approach on improving the recommender system: Predicting movie genres based on plot summaries. In *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 270–274.

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.