Using Natural Language Processing to Classify Literary Genres

Brian Mann

Background

In the year 2023, over $12.5 billion worth of books were sold in the US (Milliot, 2024). Globally, that number was over $129 billion (Curcic, 2023). The share of fiction to non-fiction books was roughly 50-50. Although sales peaked across the board in 2021 due to the pandemic, sales have nearly reached that peak once again just a few years later. The number of ebooks and audiobooks have steadily been increasing, with print books still accounting for over 75% of the market share. Over half of Americans have read at least one book in the previous year, with 20% having read at least 10 books (Montgomery, 2023). The fact of the matter is, the publishing industry is still alive and thriving, as much as social media might lead us to believe otherwise.

The Problem

This project aims to use Natural Language Processing (NLP) techniques to classify a selection of open-source novels by word frequency, complexity, and sentiment analysis. By understanding the characteristics of a literary text, publishers have a better picture of how to tailor marketing to meet the demands of their readers. If a certain piece of text is more likely to produce a negative reaction in a reader, the recommendation algorithms at play may need to respond differently to be effective. A data-driven analysis such as this is more likely to keep reader engagement high, enhance recommendation systems, and efficiently optimize inventory to manage the supply chain.

Dataset

The data used in this project comes from the site www.gutenberg.org, which is the oldest open-source digital library in the world (Preston, 2023). Each of the over 70,000+ texts on the site are within the public domain, meaning that all their data is free to access. This project will specifically be looking at the 25 most popular texts from four unique genres – adventure, fantasy, horror and humor, for a total of 100 books. Although it would be more beneficial to analyze a larger number of books, the number of fiction books organized into genres is significantly limited compared with non-fiction books or books written in a foreign language. Fantasy books include Robin Hood and The Wizard of Oz; Adventure includes The Count of Monte Cristo and Moby Dick; Humor includes works by P.G. Wodehouse and Edward Lear; Horror includes such works as Metamorphosis and Dracula. Data will be extracted by using web-scraping techniques to download each book as a plain text document to a designated folder. Text data will be further extracted and processed as the project progresses.

Data Prep

Before data could be gathered from any of the target novels, it first must be scraped from Project Gutenberg's repository. First, python's BeautifulSoup library will be used to extract the links to each individual novel's plain text file. Then, once each text file has been downloaded into a directory on a local machine, the text files will then be edited to remove irrelevant tags at the beginning and end of the story that were produced by the Project Gutenberg team. Finally, preliminary information about each novel (title, author, file name, and genre) will be placed into a pandas dataframe. From there, all subsequent information used for the machine learning portion will need to be generated from exploratory data analysis and placed into our dataframe.

Methods

In the first part of the project, initial steps will be taken to determine complexity of each text (number of pages, average length of words used, difficulty level of vocabulary, etc.), word frequency (what non-filler words were used most often), and sentiment analysis (where each text lies on the spectrum of negative to positive sentiment). Once each book's text file has been scraped from Project Gutenberg using BeautifulSoup, python libraries such as TextBlob, NLTK, and TextStat will be used to extract stats such as sentiment polarity, lexical complexity, and word count. Text will first be tokenized and filtered to eliminate the influence of stop-words and non-meaningful words. This data will then be organized into a pandas DataFrame, then further comparisons will be drawn using graphical tools such as matplotlib and seaborn.

Once the initial exploratory data analysis (EDA) is complete, we will progress to using all the data we have collected for machine learning modeling. First the data will be split into training and test data in the ratio 70/30. A multinomial logistic regression algorithm will be performed initially to generate a baseline, then a decision tree will be made. Lastly, a K-means model will be used to categorize each novel into clusters using unsupervised learning, with the hope that new insights may be gained that couldn't otherwise be found using supervised learning.

Challenges

One of the challenges with this set of texts is that the overall number of samples is quite limited, particularly in the humor genre. Many of the texts categorized as 'humor' are from the same few authors, which could lead to a greater amount of overlap that does not accurately reflect the wider spectrum of humorist literature. Likewise, with each text being at least 100 years old, results gained from this analysis may not translate well to the same genres in modern literature. Additionally, if no significant difference is found between any genre when assessing the sentiment, complexity, or word frequency of each text, it may then become highly unlikely that any accurate machine learning model can be produced.

Analysis

In the first part of our exploratory analysis, we looked at word frequency, with the goal of producing word clouds separated by genre. One of the difficulties with accomplishing this was that we first needed to filter out insignificant or non-standard words from our frequency counts. For example, some of the novels included words that contained unorthodox spelling, and all the novels contained people's names, place names, and filler-words (also called stop-words). To extract only the target words of significant meaning, a text file was used containing the 25,000 most common English words. Additionally stop words were filtered using a program provided by the NLTK package.

After this initial filtering, it was still apparent that certain words with little value, like 'would', 'could' and 'become' were not getting filtered, clogging up the word clouds for each novel. It was then decided that a few more transformations would need to happen. First, if a word was shared between a given genre and any other genre, that word was taken out. Then, only words that were 6+ letters were taken into consideration, as these words are much more likely to be meaningful. The result is seen in Figure 1.

Unsurprisingly words such as 'knight' and 'battle' were common in fantasy, and 'terror' and 'darkness' in horror. For adventure, many of the words were taken directly from French, such as 'monsieur' and 'madame'. This most likely due to the influence of the novel 'The Count of Monte Cristo', which has over 1,500 pages – the longest book in the dataset by far. The humor genre seemed to have the least consistency. No words of significant consequence seemed to stick out. The only reason the word 'illustration' was so prominent is due to the nature of the plain text files substituting the word in place of any pictures that were present in the original novel. This in and of itself seems to indicate that visual media is more important than any individual words in facilitating humor.

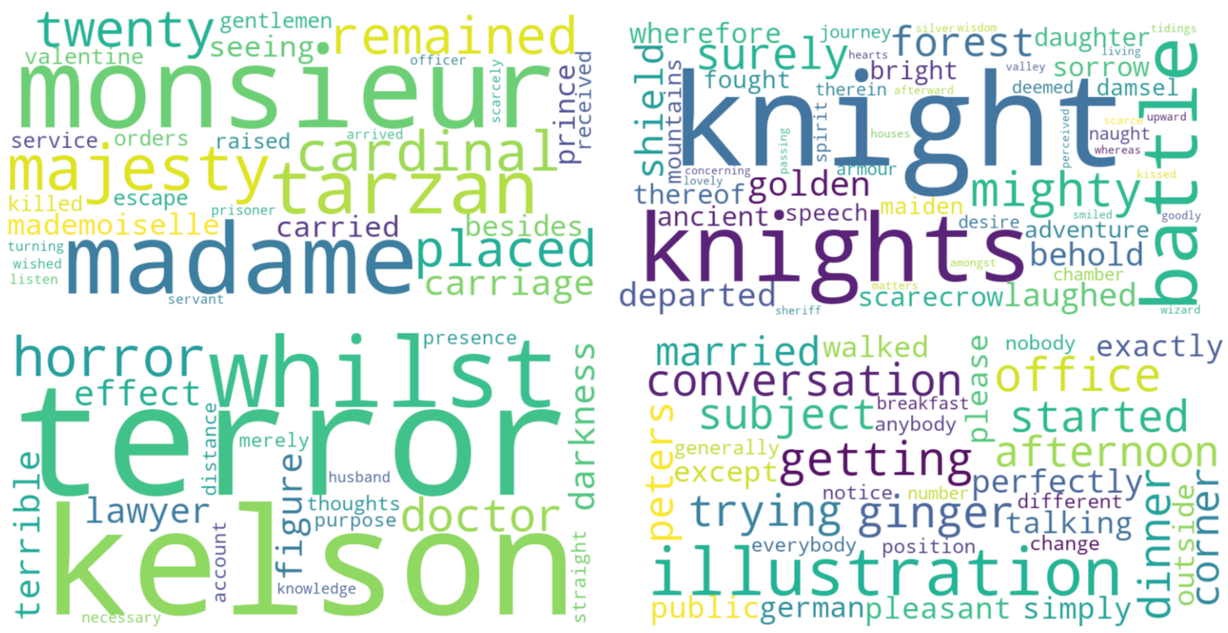


Figure 1: Word frequency by literary genre (from left to right, top to bottom: Adventure, Fantasy, Horror, Humor)

The next part of our analysis revolved around the complexity of each text. For each novel, six new variables were calculated and added to our dataframe – word count, page count, average word length, vocabulary size, vocabulary diversity, and difficult score. Word count and average word length are self-explanatory. Page count was calculated as word count divided by 300 (the typical number of words on a page). Vocabulary size was calculated as the total number of unique words used in the novel. Vocabulary diversity was derived by taking the total number of unique words (vocab size) and dividing it by the total number of words (word count). An additional difficulty score was provided for each novel by finding the percentage of more obscure/difficult words used by each novel. This was derived by finding the percentage of words with a synonym set at or below two from the NLTK synsets library.

With a few exceptions, most of these calculated variables were normally distributed, as seen in Figure 2. The average word count was about 70,000 words / 250 pages, average word length was 4.2 words, vocab size was 6,000, vocab diversity 1 unique word in 10, and difficulty level about 0.35 (out of 1). Page count was somewhat positively skewed, with much more values lower on the spectrum, with one particular outlier being

‘The Count of Monte Cristo’ as mentioned previously. Another interesting finding is that vocab size seemed to be bimodal, with a lot of books having close to 5,000 or 7,000 words, but fewer values in between.

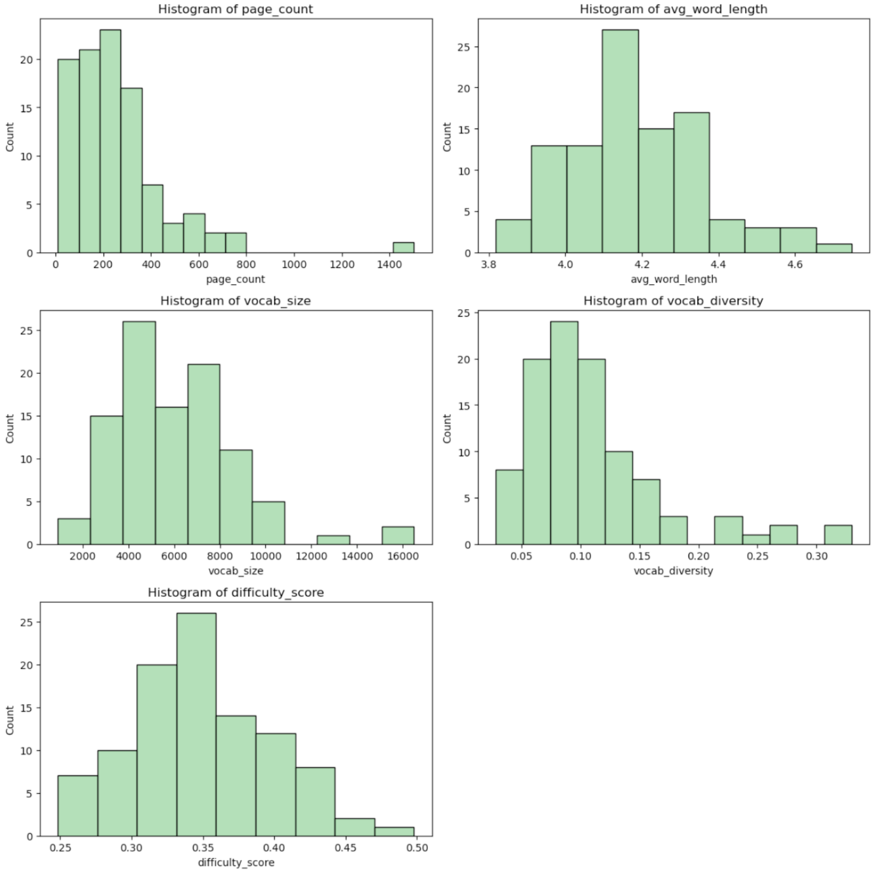


Figure 2: Histograms of complexity for overall novels

In addition to distributions for the novels overall, additional plots were made that separated each novel into the four genres – Adventure, Fantasy, Horror, and Humor. From Figure 3, we can see that in addition to having the longest novel overall, Adventure novels on average were nearly double the length of any of the other genres. Due to this fact, most likely, adventure novels also tended to have a substantially higher average vocab size. Horror and Humor both had a higher typical vocab diversity percentage, while none of the genres dominated in difficulty score or average word length.

The next part of exploratory analysis involved assessing the sentiment of the language used in each novel. Two new columns of our dataframe were calculated: average sentiment and percentage of positive sentiment. First, each novel was split into 100 equal portions. Then, a sentiment score was calculated on each of those portions using the TextBlob library. Sentiment ranges from -1 (completely negative) to 1 (completely positive), with 0 being neutral. Average sentiment is equal to the average across all 100 sections of the novel. Percentage positive sentiment is calculated as the percentage of those 100 sections that were positive. This was done to account for the effects of simple averaging masking other characteristics within the data.

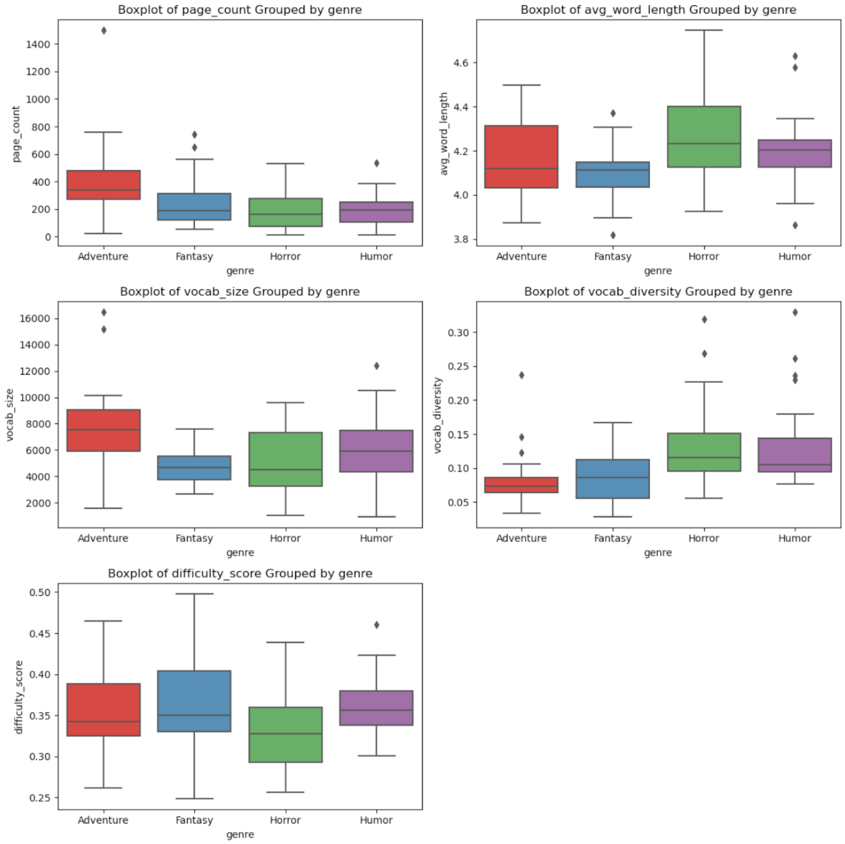


Figure 3: Boxplots of complexity by genre

In Figure 4, we can see that all 100 novels in the dataset had an overall average sentiment score above 0, and all novels except one had a higher than 50% positive sentiment. Average sentiment scores were most normally distributed, with percentage positive sentiment heavily negatively skewed. Fantasy novels tended to have the highest average sentiment and positive sentiment percentage, whereas horror novels scored significantly lower on both. However, this difference for horror novels was not quite as pronounced as one would expect, as the average percentage of positive sentiment was still nearly 80%.

In the last part of our initial analysis, we calculated the overall correlation amongst all the numerical variables in our dataframe. For better comparison, we first one-hot encoded each of the genres into a separate feature of the dataset. Figure 5 shows the results of the correlation calculations.

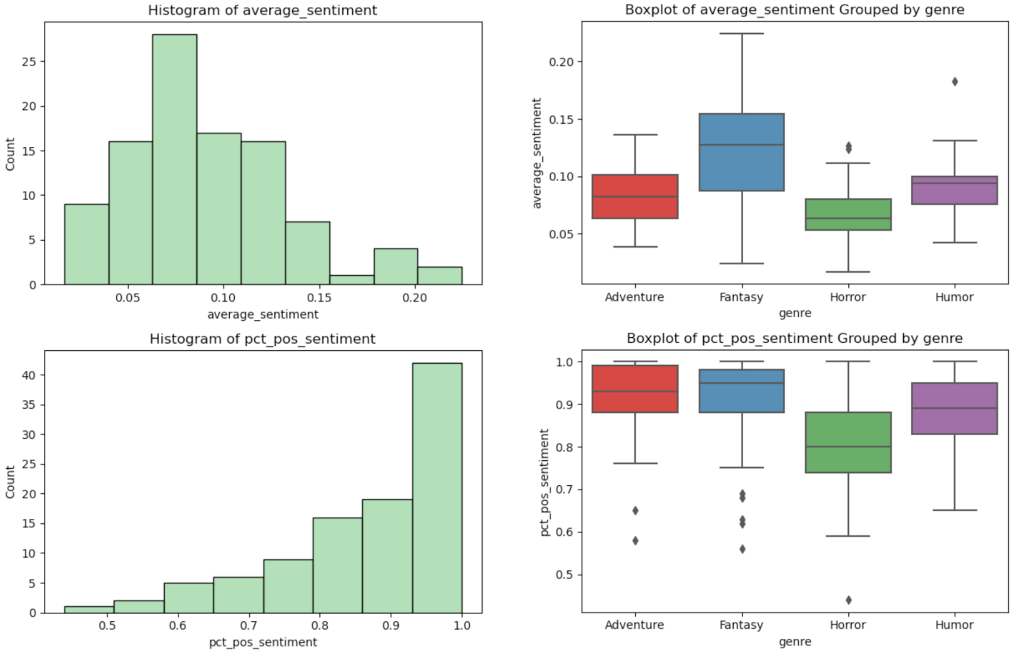


Figure 4: Histograms and boxplots for sentiment scores

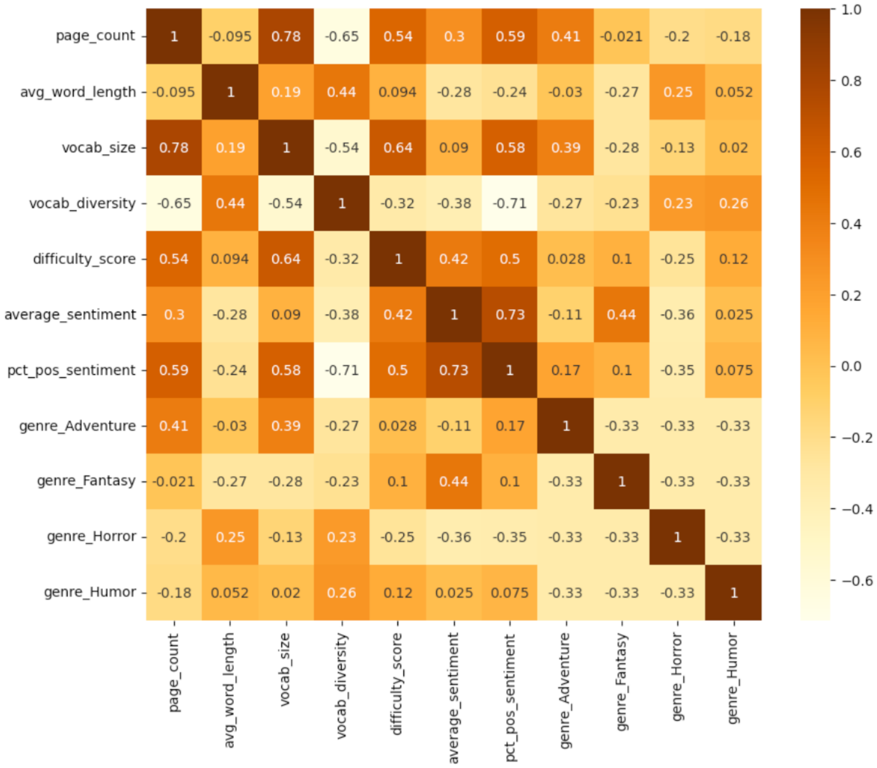


Figure 5: Correlation diagram across all novels

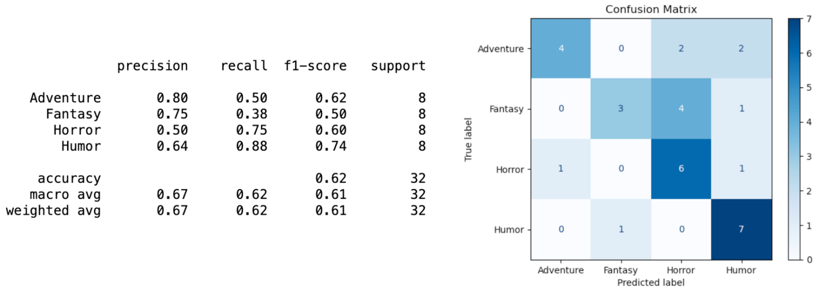
For Adventure novels, there is a strong positive correlation with page count and vocabulary size. For Fantasy novels, there is a strong positive correlation with average sentiment score, and with Horror novels, there is a moderately negative correlation with average and percentage of positive sentiment. Humor novels do not seem to have any strong indicators of correlation besides a moderately positive correlation with vocab diversity. One curious fact of this analysis is that vocab size is very strongly positively correlated while vocab diversity is strongly negatively correlated with the percentage of positive sentiment.

Modeling

A total of four different machine learning models were made during the course of this project, including a logistic regression, decision tree, ensemble and K-means model. The data used as features were each of the numerical values calculated from the exploratory phase, minus word count (for redundancy with page count), while the target variable was genre. The data was first split into training and testing sets, with a ~70/30 ratio, with exactly 8 books in each genre used in the test set and 17 in the training set.

The first model was a logistic regression model, with training and testing data scaled using a standard scaler. The model did very well in predicting the Horror and Humor genres, but struggled a bit more when it came to classifying Adventure and Fantasy (Figure 6). The overall accuracy was 62% with an F-1 score of 0.61. Subsequently, a decision tree was made using a max depth of three (Figure 7). The decision tree was more accurate at predicting Adventure, about the same at predicting Horror, but significantly worse at predicting Fantasy and Humor. This ended up producing an accuracy of only 47%, with an F-1 score of 0.44 (Figure 8). An ensemble model was then made using a weighted voting system combining both the decision tree and logistic regression models, with twice the weight given to the logistic regression model. Although this model had the highest precision (68%), it still struggled to predict Adventure and Fantasy (Figure 9). It performed better than the tree, but worse than the logistic regression model, with an accuracy of 53% and F-1 score of 0.52.

After these models were completed, a K-means clustering algorithm was performed to find out how the novels would be classified given unlabeled training data. The vast majority of Fantasy was in one cluster, a majority of Horror in another cluster, Adventure was mostly split evenly into two clusters, and Humor was spread pretty evenly across each of the clusters. One cluster was significantly larger than the other clusters and contained a fairly even amount of each genre (Figure 10).



```
              precision    recall  f1-score   support

   Adventure       0.80      0.50      0.62         8
     Fantasy       0.75      0.38      0.50         8
      Horror       0.50      0.75      0.60         8
       Humor       0.64      0.88      0.74         8

    accuracy                           0.62        32
   macro avg       0.67      0.62      0.61        32
weighted avg       0.67      0.62      0.61        32
```

Figure 6: Results of logistic regression modeling

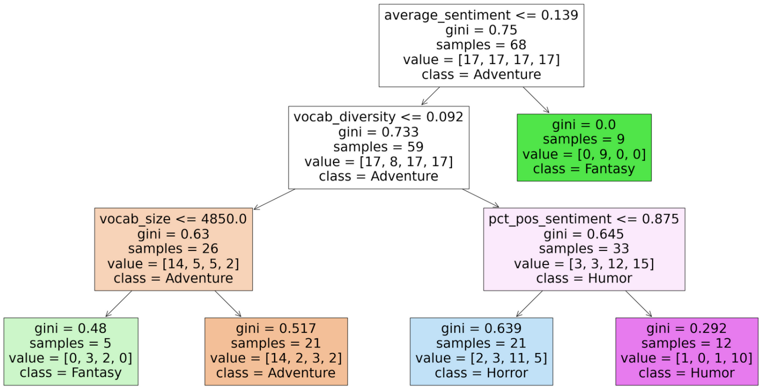
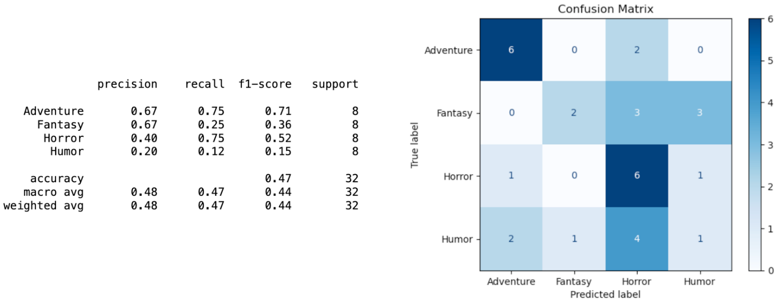


Figure 7: Decision tree model bifurcations

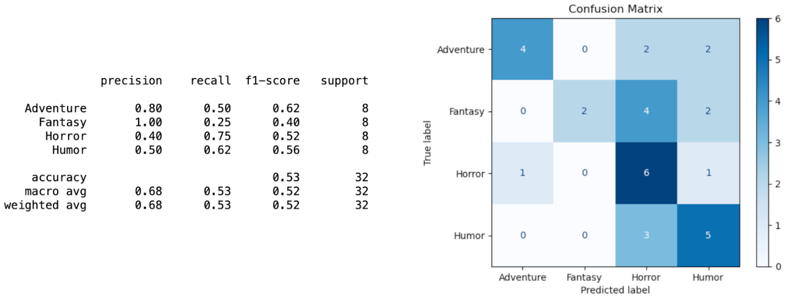


```
              precision    recall  f1-score   support

   Adventure       0.67      0.75      0.71         8
     Fantasy       0.67      0.25      0.36         8
      Horror       0.40      0.75      0.52         8
       Humor       0.20      0.12      0.15         8

    accuracy                           0.47        32
   macro avg       0.48      0.47      0.44        32
weighted avg       0.48      0.47      0.44        32
```

Figure 8: Results of decision tree modeling



```
              precision    recall  f1-score   support

   Adventure       0.80      0.50      0.62         8
     Fantasy       1.00      0.25      0.40         8
      Horror       0.40      0.75      0.52         8
       Humor       0.50      0.62      0.56         8

    accuracy                           0.53        32
   macro avg       0.68      0.53      0.52        32
weighted avg       0.68      0.53      0.52        32
```

Figure 9: Results of voting ensemble modeling

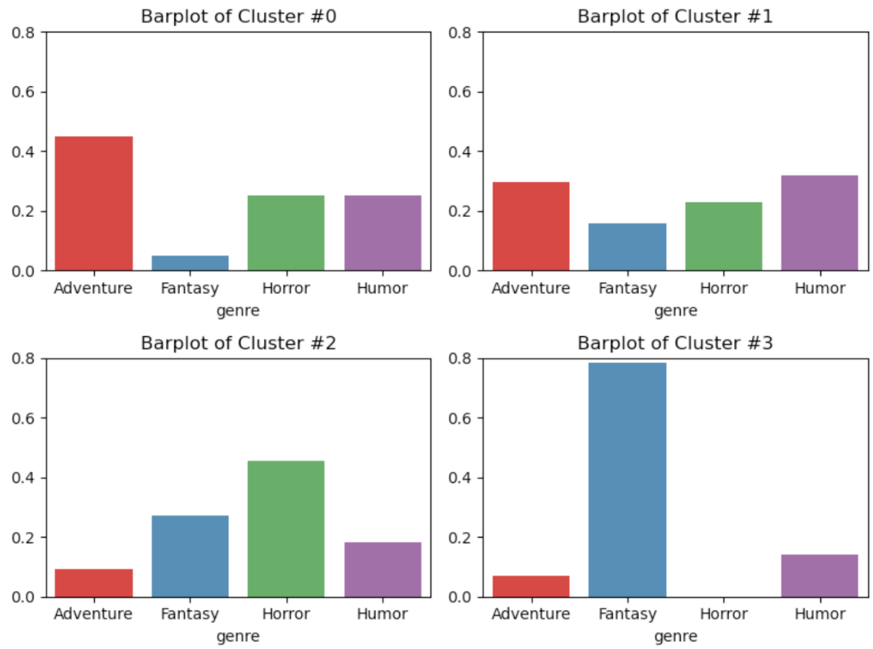


Figure 10: Percentages of each genre contained within K-Means clusters

Conclusion

In our final analysis, we conclude that the logistic regression model performed the best out of all three supervised models. Although the accuracy of this model is still not quite that high at 62%, it is about 2.5x as accurate as random chance. Although many attempts were made to tweak the hyperparameters in the ensemble modeling method, the ensemble model still seemed to perform worse than the initial logistic regression model. The decision tree seemed to struggle with precision, where it predicted each genre as Horror at a substantially higher rate. This seems to have led to a dilution in the effectiveness of the ensemble model. It is surprising that the logistic regression model was highly accurate in predicting Humor, as there were few variables that correlated much with Humor, and the decision tree model and the K-means clustering model both struggled substantially with its classification.

Assumptions and Limitations

There were a significant number of limitations over the course of this project, most notably the lack of diversity in the novels that were acquired. A select few authors seemed to dominate most of the available free literature in each of the target genres, despite some manual filtering being done to limit this bias. Therefore it is unclear whether these same results would apply to these literary genres as a whole, or perhaps they are just a reflection of the specific literary styles of particular authors. Additionally, as the number of novels available for use was quite limited, it was difficult to train and produce accurate machine learning models, as well as derive more impactful analysis and statistics.

Implementation Plan

It is not recommended that the current model be deployed, particularly due to the limitations given above. While it is still nearly 2.5x as accurate as random chance, it is unlikely that this model will be able to accurately reflect target genres on a professional level.

Recommendations and Future Applications

In the future, we recommend increasing the number of genres, books, and authors, as well as including more recently published works. Additionally, other more detailed information should be included in the analysis, such as publishing date, plot descriptions, characters, and demographics. More detailed machine learning models could also be crafted for each novel to reflect these modifications.

Ethical Considerations

Generally, when analyzing text data, protecting the intellectual property of the publisher is paramount. However, in this instance, each of these texts has been verified to be in the public domain, meaning that the copyright that was initially held no longer applies. Additionally, all data used in this project is purely for the purpose of education, and in no way will be used for profit or financial gain. Likewise, all authors will be properly credited for their respective works. Another important consideration to make is that most of these novels were written around 100-200 years in the past. The cultural norms of those times may not align with those that are present today, and as such, proper care will be taken when presenting any potentially offensive language or themes.

References

Curcic, D. (2023). *Book Sales Statistics*. Wordsrated. https://wordsrated.com/book-sales-statistics/

Milliot, J. (2024). *Publishing Sales Inch Up in 2023 Per AAP Sales Report*. Publishers Weekly. https://www.publishersweekly.com/pw/by-topic/industry-news/financial-reporting/article/94669-publishing-sales-inch-up-in-2023-per-aap-sales-report.html

Montgomery, D. (2023). *54% of Americans read a book this year*. YouGov. https://today.yougov.com/entertainment/articles/48239-54-percent-of-americans-read-a-book-this-year

Preston, S. (2023). *Cover to Cover: Access thousands of books on Project Gutenberg*. Gering Courier. https://starherald.com/news/community/gering/cover-to-cover-access-thousands-of-books-on-project-gutenberg/article_0a968c7a-6706-11ee-b8b5-1bb3f4ad78df.html