IS 567 - Text Mining Project Report

Sentiment Analysis of Science Fiction Books

Prathamesh Patil(pppatil2)

Reading is by far the most popular pastime for most people, creating a better path to information and knowledge sharing. As a result, books continue to be an essential and vital commodity that is extensively purchased. Nonfiction books outsell fiction books in terms of sales. General fiction and biographies appear to be sold more frequently among the fiction and non-fiction genres.

The typical American reads 12 to 13 novels each year. The book's success can be determined by a variety of elements, including the author's unique idea, reader appeal, marketing approach, and internet sales. How people choose what they read has always piqued the curiosity of researchers. Although several analytical areas exist to investigate business model and profits, there is practically no method to comprehend the reader's psyche and what most entices them about a book.

1. Goals of the Project

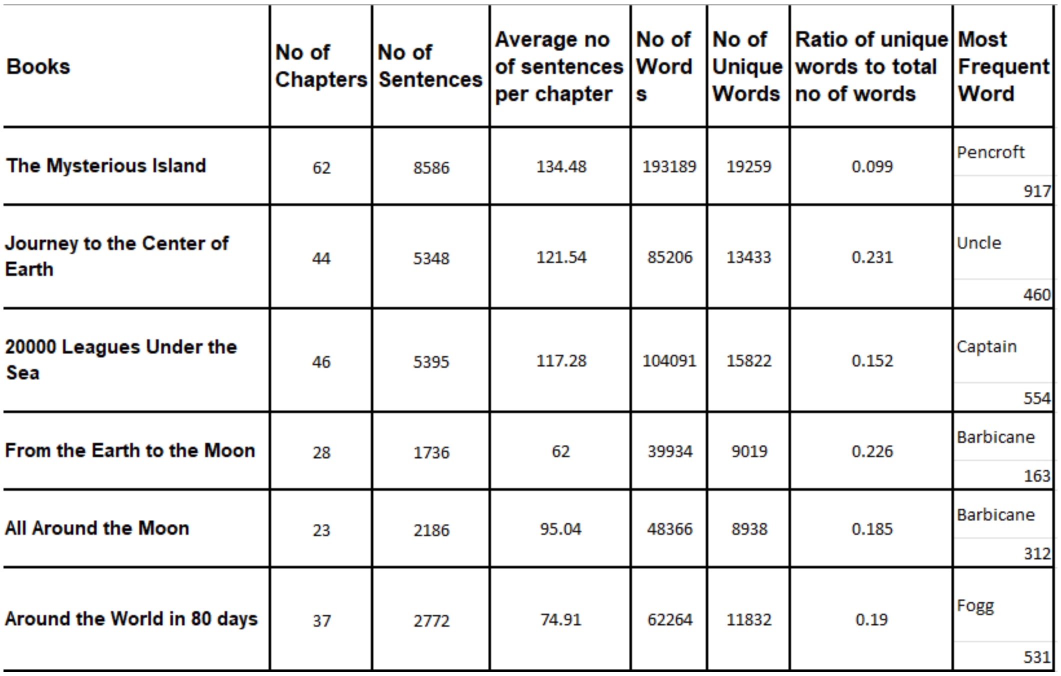
The Book Network, a well-known website for readers, states "Not only do reviews give you the lowdown of the story, genre and tone of the book, you also get a valuable impression of its quality." (*The Importance of Book Reviews*, Nov 26, 2022.) Reviews are a crucial component of the decision-making process and the first port of call for many uncertain readers. They are the most effective approach to learn how an audience thinks prior to buying or post reading completion.

This project delves deeper into the reader's mindset. It attempts to research following non-traditional questions –

> *'What is the author's writing style?'*
>
> *'How does author's writing style affect the reader while reading?'*
>
> *'Do positive and negative sentences in book affect reader's judgement while writing review about a book?'*

2. Descriptive statistics of data

The dataset contains six books - *Journey to the Center of the Earth, Twenty Thousand Leagues, Under the Seas, Around the World in Eighty Days, From the Earth to the Moon, Around the Moon* and *The Mysterious Island* by Jules Verne. These books are in standard *text(.txt)* format and have no labels for classification. A summary of all the six books in table 2.1.



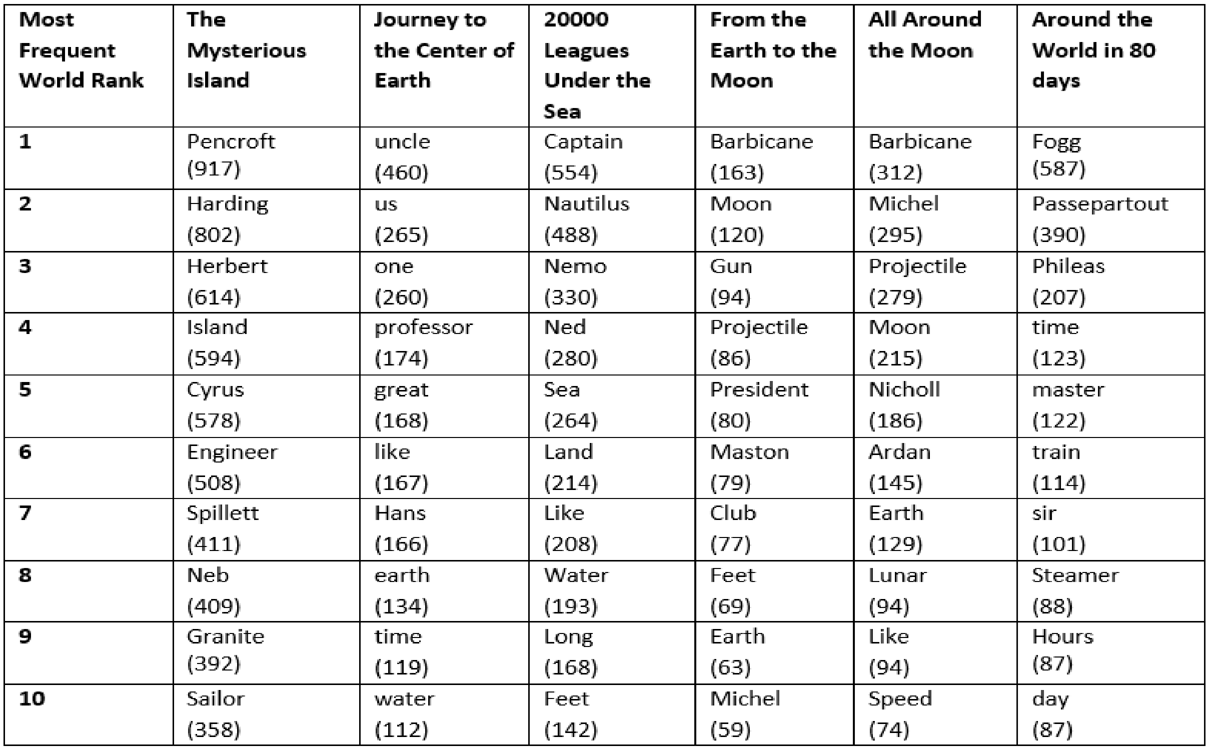
Books	No of Chapters	No of Sentences	Average no of sentences per chapter	No of Word s	No of Unique Words	Ratio of unique words to total no of words	Most Frequent Word
The Mysterious Island	62	8586	134.48	193189	19259	0.099	Pencroft 917
Journey to the Center of Earth	44	5348	121.54	85206	13433	0.231	Uncle 460
20000 Leagues Under the Sea	46	5395	117.28	104091	15822	0.152	Captain 554
From the Earth to the Moon	28	1736	62	39934	9019	0.226	Barbicane 163
All Around the Moon	23	2186	95.04	48366	8938	0.185	Barbicane 312
Around the World in 80 days	37	2772	74.91	62264	11832	0.19	Fogg 531

Table 2.1 – Popular books by Jules Verne and their statistics

As seen in the above table, The Mysterious Island is unique because it has highest average no of sentences per chapter but also lowest ratio of unique words to total words. On the contrary, although *From the Earth to the Moon* has lowest average no of sentences per chapter, it has second highest ratio of unique words to total words. It's not significant proof that the words will be reused in longer books. But we cannot overlook the fact it becomes challenging to come up with fresh concepts once the premise has been set up in a book. This could be the reason that there are less unique terms in larger novels. Because synonyms are ignored here, unique words will count words with the same meaning differently.

In addition, ideally most frequent word should be one of the three articles – *a, an,* or *the*. But since stopwords were removed, we can clearly see that the most frequent word in a particular book is the name of main character. Since the plot of any book revolves around the main character, this insight is justified.

Next table (Table 2.2) shows the top 10 words with their respective count. It is evident that nouns are the most repeated words rather than verbs in all the six books. The only verb that makes the chart is 'like' which is either used for comparison or to show affection/admiration. Logically speaking, all these books fall under the sci-fiction category, which describes surroundings to give readers a livelier scene and uses comparison to relate surroundings with day-to-day objects. Hence, 'like' making it to top 10 can be accounted for.



Most Frequent World Rank	The Mysterious Island	Journey to the Center of Earth	20000 Leagues Under the Sea	From the Earth to the Moon	All Around the Moon	Around the World in 80 days
1	Pencroft (917)	uncle (460)	Captain (554)	Barbicane (163)	Barbicane (312)	Fogg (587)
2	Harding (802)	us (265)	Nautilus (488)	Moon (120)	Michel (295)	Passepartout (390)
3	Herbert (614)	one (260)	Nemo (330)	Gun (94)	Projectile (279)	Phileas (207)
4	Island (594)	professor (174)	Ned (280)	Projectile (86)	Moon (215)	time (123)
5	Cyrus (578)	great (168)	Sea (264)	President (80)	Nicholl (186)	master (122)
6	Engineer (508)	like (167)	Land (214)	Maston (79)	Ardan (145)	train (114)
7	Spillett (411)	Hans (166)	Like (208)	Club (77)	Earth (129)	sir (101)
8	Neb (409)	earth (134)	Water (193)	Feet (69)	Lunar (94)	Steamer (88)
9	Granite (392)	time (119)	Long (168)	Earth (63)	Like (94)	Hours (87)
10	Sailor (358)	water (112)	Feet (142)	Michel (59)	Speed (74)	day (87)

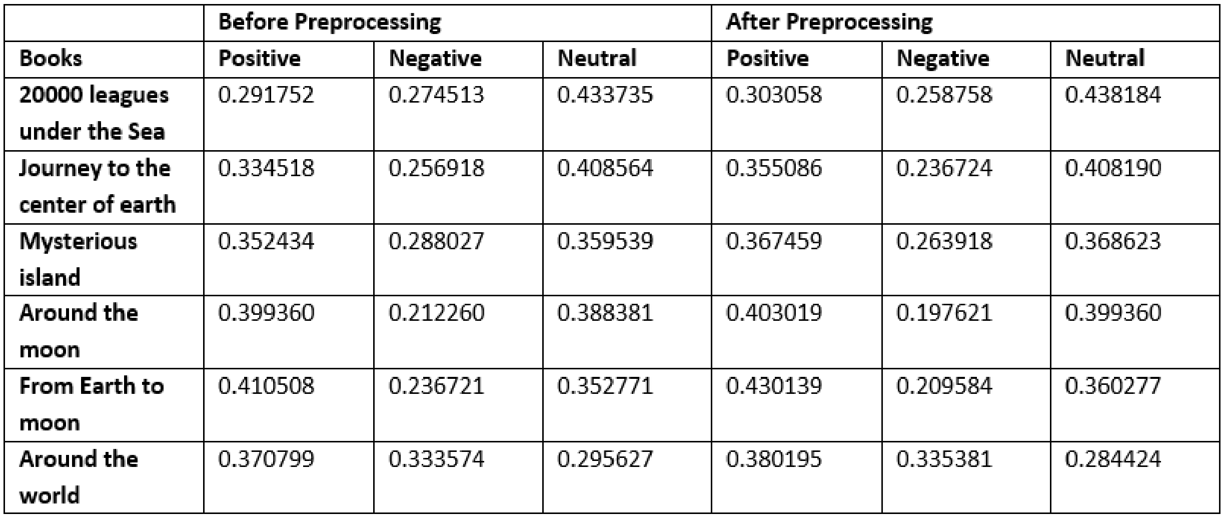
Table 2.2 – Top 10 frequently repeated words with counts in six books

In addition to these 6 books, there is a reviews dataset for the book '*Around the World in Eighty Days*'. During initial presentation, this dataset included 150 reviews manually scraped from the web. At the time of report, 350 more reviews were added making a total of 500 reviews. This was a labelled dataset with ratings from 1-5 and text of the review as two columns. No exploratory data analysis was performed on this reviews dataset.

3. Pre-processing and Transformation steps

Since the data was in text format, it required lot of cleaning before it could be loaded into pandas dataframe. To remove unnecessary headers, footers and titles, manipulation using python was done. To annotate data, it is a tardy and time-consuming way to read all the six books and then classify each sentence as positive, neutral, or negative. 3 unsupervised learning techniques - *AFINN lexicon library developed by Finn Arup Neilsen, TextBlob package which contains a pretrained sentiment polarity indicator and Sentiment Intensity Analyzer (SIA) from NLTK package* - were used to determine sentiment labels for every line in six books without any preprocessing. To take labeling to more accurate level, the usual preprocessing which included –

- Lower Case conversion
- Regular expression to remove symbols, numbers, and punctuations
- Tokenization
- Stopwords removal
- Lemmatization



	Before Preprocessing			After Preprocessing		
Books	**Positive**	**Negative**	**Neutral**	**Positive**	**Negative**	**Neutral**
20000 leagues under the Sea	0.291752	0.274513	0.433735	0.303058	0.258758	0.438184
Journey to the center of earth	0.334518	0.256918	0.408564	0.355086	0.236724	0.408190
Mysterious island	0.352434	0.288027	0.359539	0.367459	0.263918	0.368623
Around the moon	0.399360	0.212260	0.388381	0.403019	0.197621	0.399360
From Earth to moon	0.410508	0.236721	0.352771	0.430139	0.209584	0.360277
Around the world	0.370799	0.333574	0.295627	0.380195	0.335381	0.284424

Table 3.1: Comparison between positive, neutral, and negative indices before and after processing text data of 6 books

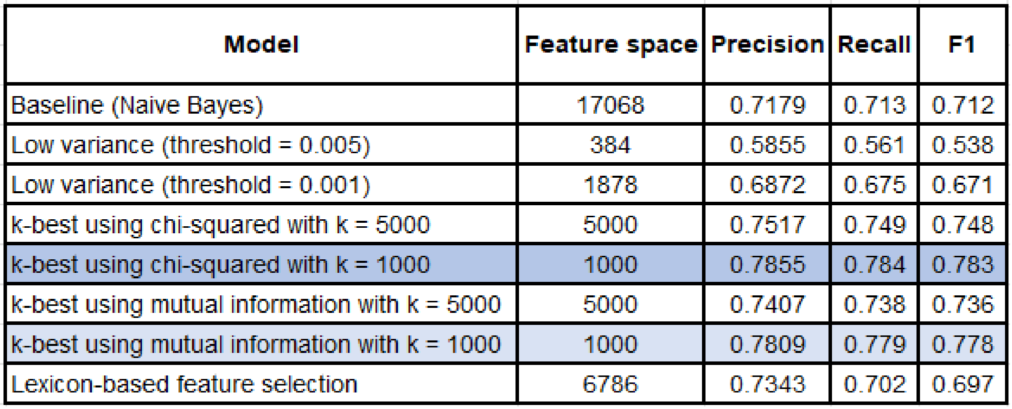
Table 3.1 shows that after performing preprocessing steps only 2% of negative sentences were misclassified for all books except ‘Around the World in 80 days’. As a conclusion, classification done by aforementioned 3 libraries is fairly accurate.

As for the reviews, same process was carried out except for an additional step where numbered rating was converted to labels. Encoding to label these reviews included following format –

1,2 – negative
3 – neutral
4,5 – positive

4. Feature extraction and Selection

Since the document here is a whole book, TF-IDF for feature extraction is preferred over Bag of Words and Count Vectorizer as it gives importance to frequency of the words. One of the most significant advantages of TF-IDF is that it is computationally inexpensive and provides a straightforward basic framework for similarity calculations using vectorization and cosine similarity. Table 4.1 shows comparison between feature space size and accuracy using Naïve Bayes as baseline model for feature selection.



Model	Feature space	Precision	Recall	F1
Baseline (Naive Bayes)	17068	0.7179	0.713	0.712
Low variance (threshold = 0.005)	384	0.5855	0.561	0.538
Low variance (threshold = 0.001)	1878	0.6872	0.675	0.671
k-best using chi-squared with k = 5000	5000	0.7517	0.749	0.748
k-best using chi-squared with k = 1000	1000	0.7855	0.784	0.783
k-best using mutual information with k = 5000	5000	0.7407	0.738	0.736
k-best using mutual information with k = 1000	1000	0.7809	0.779	0.778
Lexicon-based feature selection	6786	0.7343	0.702	0.697

Table 4.1 – Precision, Recall and F1 comparison between various feature sizes with NB classifier as baseline

As seen above, highest accuracy is achieved when feature size equals 1000. With feature size less than 1000 and low variance selection technique, Precision drops drastically. On the other hand, as feature space approaches 5000, precision drops very slowly with k-best method. Hence, k-best using chi-square for feature selection combined with TF-IDF is selected to give the best output without computationally straining the machine.

Summary:

Feature Extraction technique: TF-IDF
Feature Selection technique: K-best using Chi-squared with k=1000
No of features selected: 1000

5. Models, parameters, evaluation results and improvement strategies

Developing a predictor that reliably recognizes sentiments in sci - fi works proved to be difficult because certain manual categorization must be performed prior to applying machine learning techniques. Absence of ground truth labels made the project more complex and hence, as mentioned earlier, unsupervised techniques were used to classify and label each sentence. Instead of using a single technique to label, three techniques were used to verify. Each line had three sentiment values from each of the three techniques, which were combined to get a final sentiment value by selecting the most repeated label from the three. This ensured ground truth labels were accurate as possible without manually classifying them.

The books were classified into training and test datasets. 5 books with total of 23,251 sentences served as training set and test set included 1 book with 2772 sentences. 3 standard classification models – *Naïve Bayes, Linear SVM and Logistic Regression Classifier* - were used, and their respective weighted accuracies are reported in table 5.1

Models	Precision	Recall	F1
Naïve Bayes	0.7855	0.7842	0.7832
LinearSVM	0.8414	0.6912	0.7354
Logistic	0.6022	0.5861	0.5547

Table 5.1: Comparison between classification models

As seen above, linear SVM performs slightly better than logistic and Naïve Bayes theorem. This slight improvement can be credited to SVM being a high dimensional vector classifier. Logistic performs the worst as it is a binary classifier and will work best if there are only 2 classes.

To answer our main questions, reviews dataset was not subjected to any classification technique as it was pre-labelled and did not require data annotation.

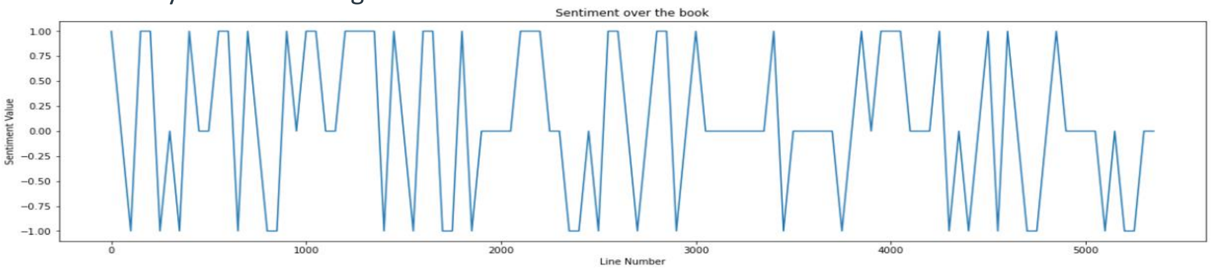
6. Error analysis, insights, and interpretations

Table 5.1 shows that the accuracies are barely within acceptable range considering test dataset is a book within same era by same author. Absence of pre trained models to correctly label sentiment in sci-fi domain make this a challenge and hence this accuracy of LinearSVM can be considered a luxury. Using neural networks might boost accuracy to 90% or more.

Comparing sentiment index of the book to sentiment of the reviews isn’t an intuitive approach. After a lot of literature review, this seemed like the easiest possible way to answer the question if sentiments in book and author’s sentiment influence reviews of the reader. The factors deciding correlation between the two are an unexplored territory and need advanced domain knowledge in human psychology. Trying to test hypothesis of the correlation with available data was indeed a behemoth task.

Taking a supervised learning approach to build a model was just an additional step to explore this uncharted territory. It didn’t in any way add value to the outcome of testing hypothesis. I learned of few drawbacks about various methods used and my point of view to answer the questions –

1. Author’s sentiments are not reflected in the book. Whatever emotion the author feels while writing the book is not congruent with the sentiment of the book. For instance, sentiment over the book ‘Twenty Thousand Leagues Under the Sea’ is shown below –



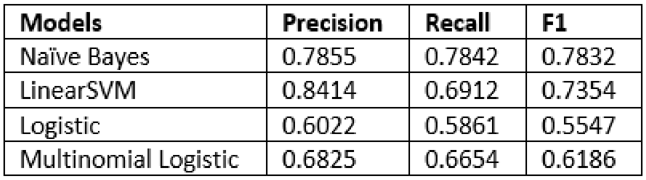
   We see fluctuations from positive at the start of the book to the neutral end. This is just the sentiment of the plot of the book and not how author actually feels. Author usually decides how a plot progresses and above diagram shows that. They may experience negative or mixed feelings while writing the plot and there is no data to know these feelings.

2. Number of reviews in dataset greatly influences the comparison between sentiment indices of the book and sentiment index of reviews. For example, when there were only 150 reviews, sentiment indices were roughly same for books and reviews with positive and negative values inverted. But when the review number increased to 500, positive reviews dominated and were a large portion of the reviews. In general, popular books will have high number of good reviews

thus skewing the opinion of readers who would want to try the same book for first time. There is no way to normalize this skewness.

Sentiment Index	Around the World	150 Reviews	500 Reviews
Positive	0.381	0.331	0.652
Negative	0.335	0.387	0.124
Neutral	0.284	0.282	0.224

3. Low accuracy for logistic regression stems from using logistic regression to classify more than 2 classes. Since there are 3 labels, using multinomial logistic regression would have been a wise choice here. After the late realization and fixing the issue, results are as follows –



Models	Precision	Recall	F1
Naïve Bayes	0.7855	0.7842	0.7832
LinearSVM	0.8414	0.6912	0.7354
Logistic	0.6022	0.5861	0.5547
Multinomial Logistic	0.6825	0.6654	0.6186

In nutshell, the project was first step towards a larger study that can be an ideal topic for thesis or Doctorate study. Having gone through multiple readings and trying ways to prove the hypothesis, finally it can be said that it is inconclusive to determine if sentiments of the books influenced the reader's point of view while writing a review.

References:

1. Yucesoy, B., Wang, X., Huang, J. *et al.* Success in books: a big data approach to bestsellers. *EPJ Data Sci.* **7**, 7 (2018). https://doi.org/10.1140/epjds/s13688-018-0135-y

2. (*The Importance of Book Reviews*, Nov 26, 2022.) *The Importance of Book Reviews*. (Nov 26, 2022.). The Book Network. Retrieved December 8, 2022, from https://www.thebooknetwork.co.uk/the-blog/the-importance-of-book-reviews

3. Koto, F., Adriani, M. (2015). A Comparative Study on Twitter Sentiment Analysis: Which Features are Good? In: Biemann, C., Handschuh, S., Freitas, A., Meziane, F., Métais, E. (eds) Natural Language Processing and Information Systems. NLDB 2015. Lecture Notes in Computer Science (), vol 9103. Springer, Cham. https://doi.org/10.1007/978-3-319-19581-0_46