Sentiment Analysis of Science Fiction Book

Prathamesh Patil
IS 567 - Text Mining Project

AGENDA


1.
Project Background
2.
Data Exploration
3.
Implementation Steps
4.
Methodology & Challenges
5.
Results
6.
Future Scope

PROJECT BACKGROUND

Comprehending the granularity of a reader's mindset and the factors that draws them most to a book using sentiment analysis

Why:
While reviews have been the most used data to understand the book's performance, analyzing the sentiment of the book and its impact on a reader before he even writes a review is yet a relatively unexplored area.

DATA EXPLORATION

Books	No of Chapters	No of Sentences	No of Words	No of Unique Words	Most Frequent Word
The Mysterious Island	62	8586	193189	19259	Pencroft (917)
Journey to the Center of Earth	44	5348	85206	13433	Uncle (460)
20000 Leagues Under the Sea	46	5395	104091	15822	Captain (554)
From the Earth to the Moon	28	1736	39934	9019	Barbicane (163)
All Around the Moon	23	2186	48366	8938	Barbicane (312)
Around the World in 80 days	37	2772	62264	11832	Fogg (531)

DATA EXPLORATION

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)

IMPLEMENTATION STEPS

- Label each line in the books

- Find overall Positive Index, Negative Index and Neutral Index of the book using following formulae -

```
pos_index =(total positive labels / total no sentences)
neg_index =(total negative labels / total no sentences)
neu_index =(total neutral labels / total no sentences)
```

- Compare it to reviews on the Internet to see if these match sentiment analysis of the book

METHODOLOGY (UNSUPERVISED)

- Each sentence needed to be labelled
- Following libraries were used to label each sentence
  - AFINN
  - TextBlob
  - NLTK

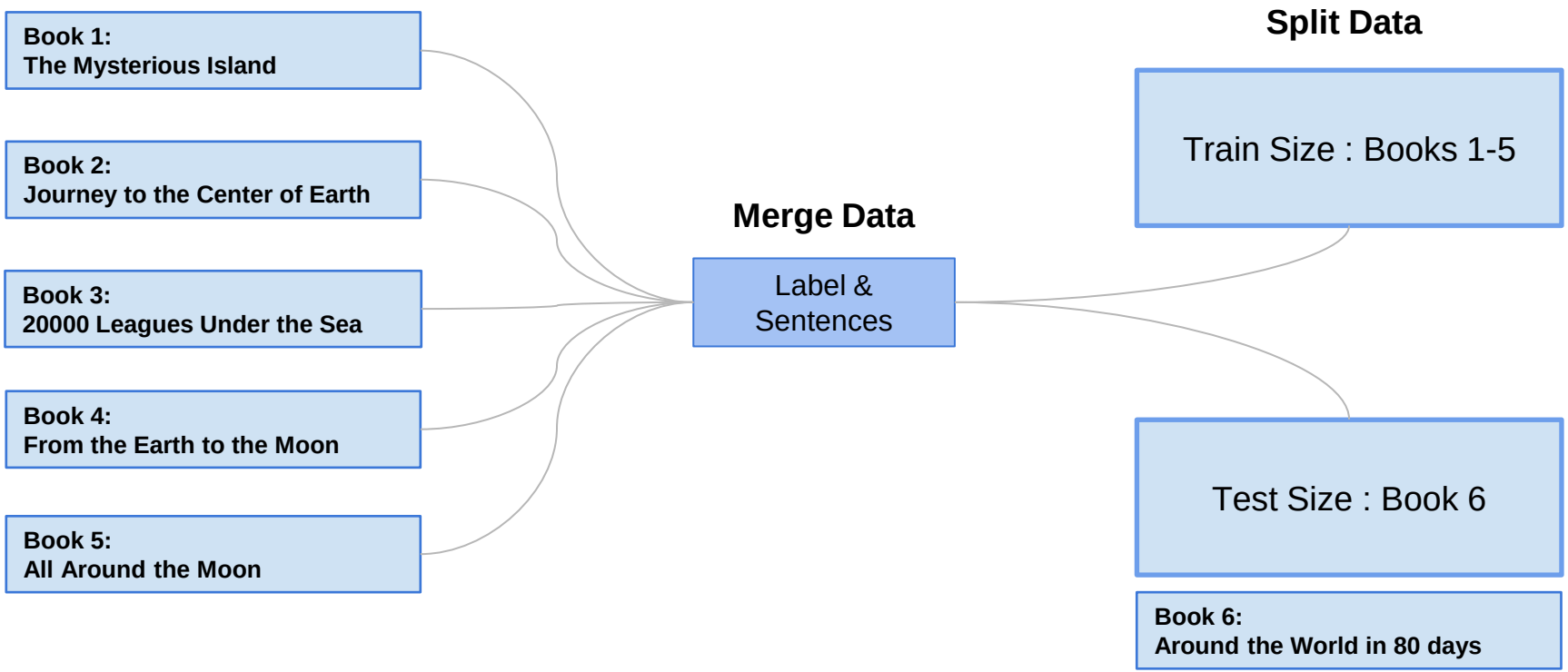
	Sentences	Afinn_Score	Afinn_Sentiment	TextBlob_Polarity	TextBlob_Sentiment	NLTK_Scores	NLTK_Sentiment	Sentiment_list	Label	Value
0	The year 1866 was signalised by a remarkable i...	0.0	neutral	0.375000	positive	0.3400	positive	(neutral, positive, positive)	positive	1
1	Not to mention rumours which agitated the mari...	6.0	positive	0.250000	positive	0.2023	positive	(positive, positive, positive)	positive	1
2	Merchants, common sailors, captains of vessels...	3.0	positive	-0.016667	negative	0.4754	positive	(positive, negative, positive)	positive	1
3	For some time past, vessels had been met by "a...	0.0	neutral	0.033333	positive	0.0000	neutral	(neutral, positive, neutral)	neutral	0
4	The facts relating to this apparition (entered...	3.0	positive	0.400000	positive	0.7501	positive	(positive, positive, positive)	positive	1
...	...	...	...	...	...	...	...	...	...	...
5390	May the judge disappear, and the philosopher c...	2.0	positive	0.312500	positive	0.5411	positive	(positive, positive, positive)	positive	1
5391	If his destiny be strange, it is also sublime.	-1.0	negative	-0.050000	negative	-0.2023	negative	(negative, negative, negative)	negative	-1

METHODOLOGY (SUPERVISED)

- Following methods were used for feature selection to train Naive Bayes Model -
  - Naive Bayes (Baseline)
  - Low Variance
  - K-best using Chi Square
  - K-best using Mutual Information
  - Lexicon Based

- 5 books were joined to use as training dataset and 1 book was used as testing dataset

METHODOLOGY (SUPERVISED)

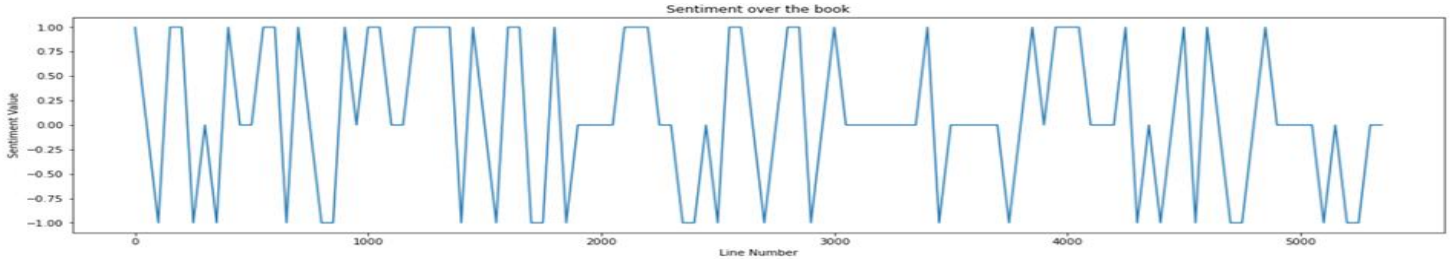


CHALLENGES

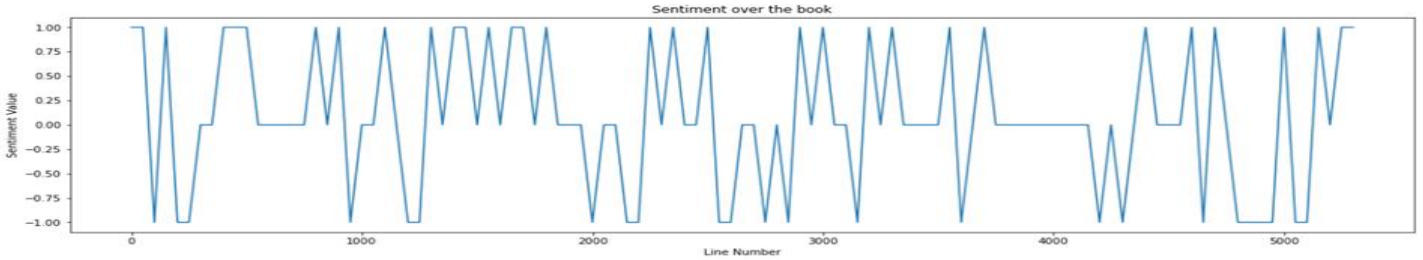
- Books in .txt format needed to be cleaned before loading in dataframe
- No models trained specifically for Sci - Fi books to label each sentence correctly
- Reading and labelling each sentence manually was not feasible
- Collecting reviews for the required book had to be done manually

Sentiment over the book (line by line)

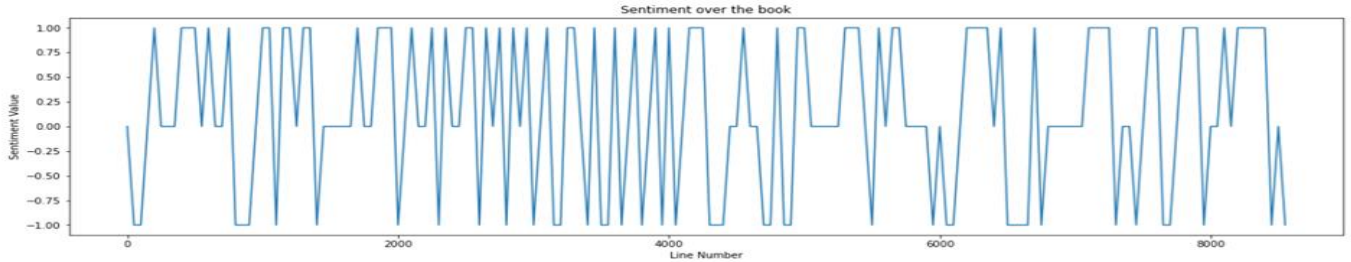
1. Twenty Thousand Leagues Under the Sea



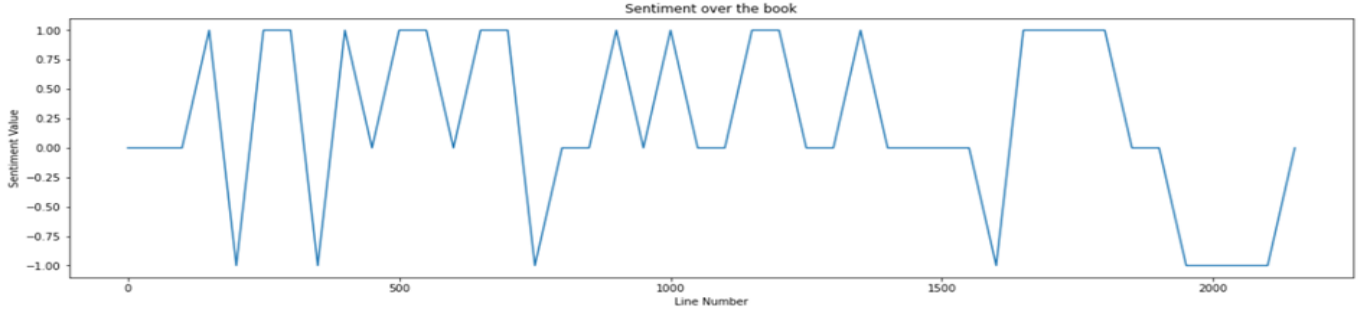
2. Journey to the Center of the Earth



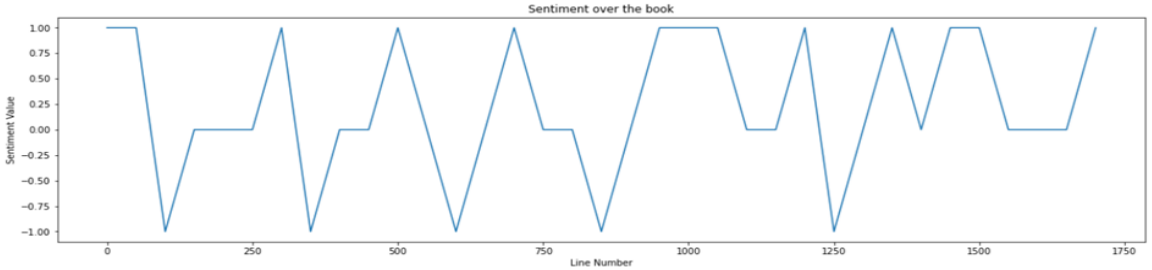
3. The Mysterious Island



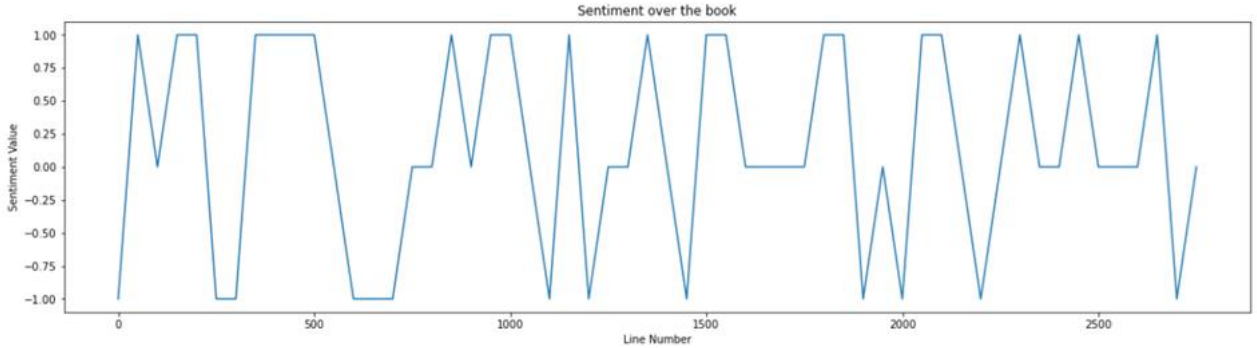
4. Around the Moon



5. From Earth to the Moon



6. Around the World in 80 days



Sentiment Index of books (overall)

	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

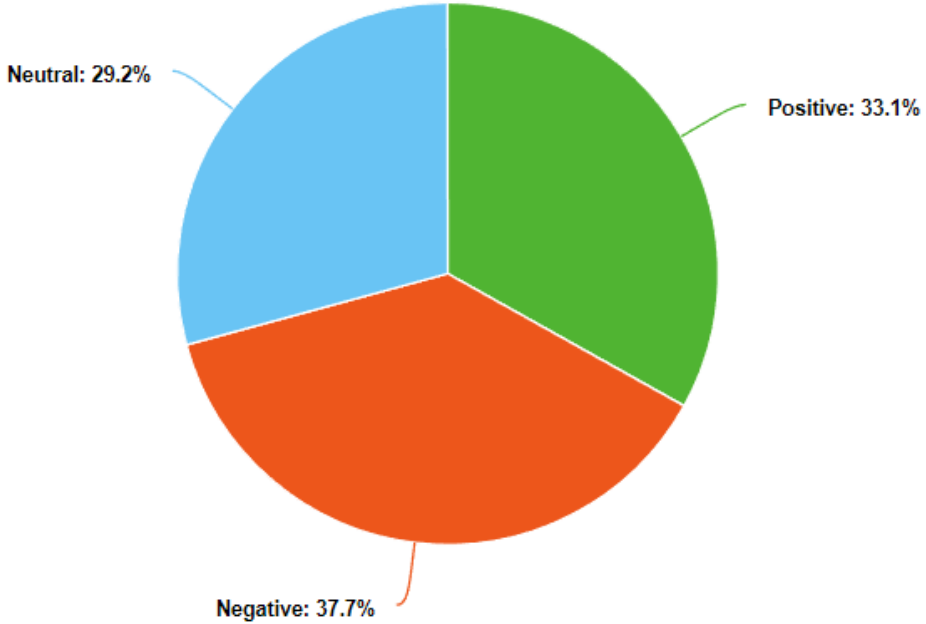
Model	Feature space	Precision	Recall	F1
Baseline (Naive Bayes)	17068	0.7179	0.7134	0.7119
Low variance (threshold = 0.005)	384	0.5855	0.5608	0.5382
Low variance (threshold = 0.001)	1878	0.6872	0.6749	0.6706
k-best using chi-squared with k = 5000	5000	0.7517	0.7490	0.7477
k-best using chi-squared with k = 1000	1000	0.7855	0.7842	0.7832
k-best using mutual information with k = 5000	5000	0.7407	0.7377	0.7364
k-best using mutual information with k = 1000	1000	0.7809	0.7791	0.7776
Lexicon-based feature selection	6786	0.7343	0.7021	0.6970

Comparison between Sentiment Indices

Sentiment Index of the Book

	Positive	Negative	Neutral
Around the world	0.380195	0.335381	0.284424

Sentiment Index of Reviews



Limitations and Future Scope

- In conclusion, the project forms a basis for the implementation of a larger idea. It can be used for detailed sentiments like happiness, anger, sadness or confusion.

- Going forward, this could be implemented for multiple genres and authors giving online bookstores a stronger direction to engage with their readers and drive sales.

- An idea like this can also be coupled with demographic aspects of the reader to create individual reader personas.

Thank you!

Any Questions?