ISSS609 Project Report

Goodreads Reviews on Children's Books

Group 4

Goh Chen Ling Beatrice (clgoh.2020@mitb.smu.edu.sg)
Kevin Magic Rialubin Sunga (kevin.sunga.2020@mitb.smu.edu.sg)
Koh Jun Jie (junjie.koh.2020@mitb.smu.edu.sg)
Wong Kian Hoong (kh.wong.2020@mitb.smu.edu.sg)

26 July 2021

1. Introduction

Parents and caretakers will go to great lengths to ensure that they obtain books of high quality for their children. As such, there is an opportunity to make use of text analysis of book reviews to recommend children's books to parents. In addition, customer reviews provide valuable information to organisations like publishers, schools and libraries.

The selected data set was scraped from Goodreads.com, a cataloging website for book recommendations and a place for users to share what books they are currently reading. Users can rate a book, give a review, and have conversations with other readers.

The aim of our analysis is to identify key topics discussed about a particular book and sentiment by analysing book reviews, thereby providing decision points to key stakeholders who would be keen to know good books to introduce to children. These would be parents who want to help their children get into a habit of reading, schools and libraries who want to add to their collection children's books of good quality, as well as publishers who would like to assess the popularity of their books.

2. Dataset

The dataset was obtained from the book review website Goodreads with information scraped in 2017 from users' public shelves, i.e., everyone can see these shelves on the web without logging in. User IDs and review IDs are anonymised. The data is contained in 2 data files as shown below. In total, there are 123,946 unique Book_id records and over 700,000 reviews.

The metadata of the 2 data files are shown in Tables 1 and 2.

S/N	Field Name	Description	Modelling type
1	Language_code	Language of the book	Nominal
2	Description	Description of the book	Text
3	Book_id	Unique Book_id	Nominal
4	Title	Title of the book	Text

Table 1: Metadata for "Goodreads_books_children.json"

S/N	Field Name	Description	Modelling type
1	user_id	ID of user who reviewed the book	Nominal
2	book_id	Unique Book_id	Nominal
3	review_id	ID of each book review	Nominal
4	rating	Rating from 1 (lowest rating) to 5 (highest rating)	Ordinal
5	review_text	Review by the user	Text
6	n_votes	Number of votes received by review	Integer
7	n_comments	Number of comments received by review	Integer

Table 2: Metadata for "Goodreads_reviews_children.json"

There are several data issues identified that posed potential challenges to the analysis. These are identified as follows:

- Special characters. The dataset contained special characters, URLs, hashtags and mentions. These were removed during pre-processing for a more meaningful analysis.

- Size of data sets. The reviews data set contains more than 700,000 reviews which do not have sentiment labels. The large size may also be computationally intensive for the analytical tasks. Hence, we have chosen to filter the dataset to a more manageable size. Details are discussed in Section 3.

- Skewness of data. Most user rating scores were in the higher range and may result in an unbalanced dataset. The distribution of reviews by user rating are shown in Figure 1. In our analysis, this was not a major issue as topic modelling, document clustering and association mining did not use the user rating scores. For sentiment analysis, user rating scores were used only as a source of "golden truth" and compared to the rating from the algorithms.

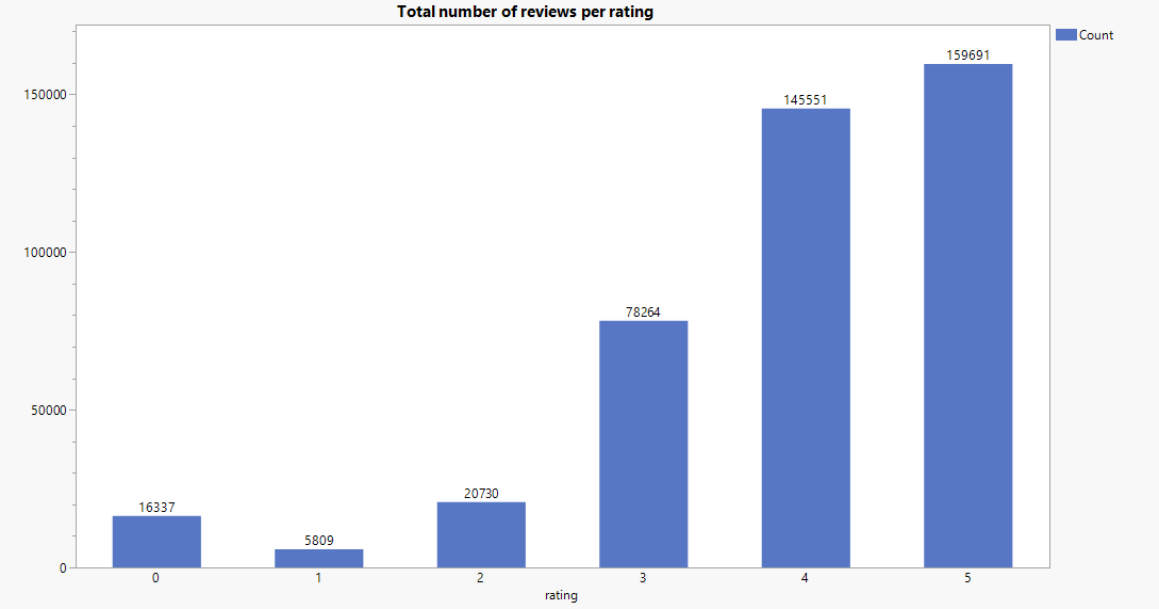

Figure 1: Distribution of reviews by user rating

3. Solution Overview

An overview of the solution and tasks can be found in Figure 2. Details of the solution are provided in Section 5.

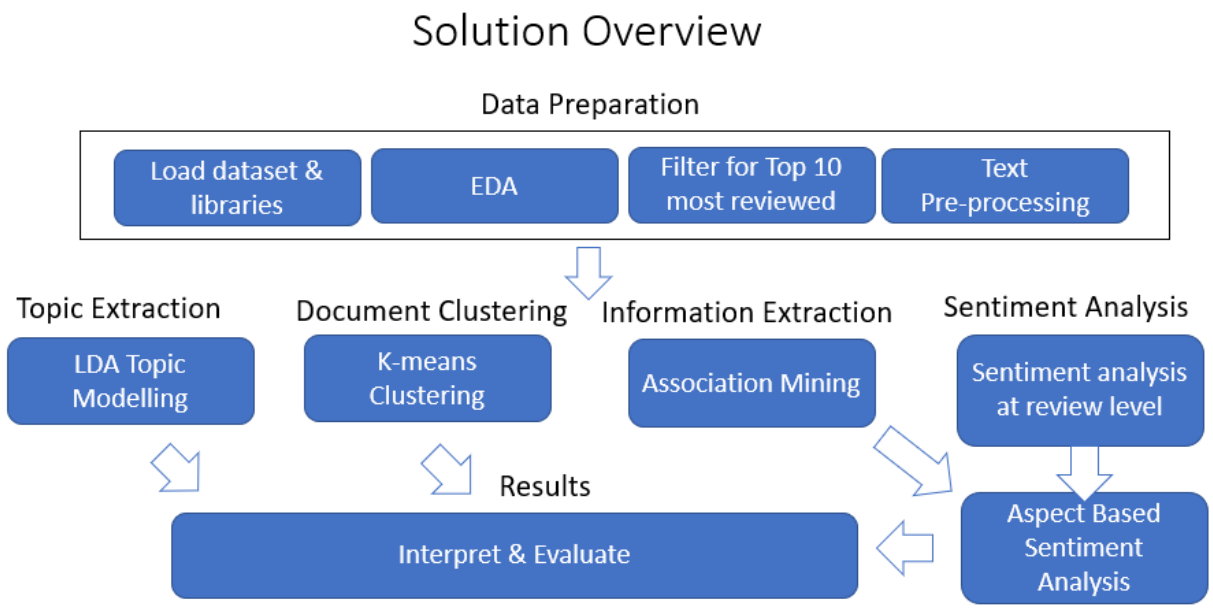

Figure 2: Overview of project solution and tasks

Given the large number of reviews as mentioned earlier, the team focused our analysis on the 10 books with the largest number of reviews. These books are shown in Table 3. The filtered dataset contained 28,928 unique reviews.

S/n	Book_id	Title	Number of reviews
1	3636	The Giver (The Giver, #1)	6,156
2	5	Harry Potter and the Prisoner of Azkaban (Harry Potter, #3)	4,696
3	11387515	Wonder (Wonder #1)	4,499
4	157993	The Little Prince	3,114
5	78411	The Bad Beginning (A Series of Unfortunate Events, #1)	1,999
6	9673436	The Invention of Hugo Cabret	1,947
7	8127	Anne of Green Gables (Anne of Green Gables, #1)	1,721
8	370493	The Giving Tree	1,697
9	11594337	The One and Only Ivan	1,575
10	2998	The Secret Garden	1,524

Table 3: The top 10 books based on number of reviews

With this corpus, LDA topic modelling and document clustering were conducted to gain insights into the topics being discussed in the reviews. Association mining and information extraction were then conducted to extract aspects and sentiments. Following that, sentiment analysis was performed, first on the review in its entirety and then on the aspects and sentiments.

4. Business Use-cases

By analysing textual reviews, we can extract insights to better understand the common topics raised by readers, what aspects of the book's readers mentioned and what they feel about those aspects. We selected one book as an example and the results are discussed in Section 6.

The analysis can then be replicated to the wider corpus to obtain further insights and apply to these three use cases:

1. The results of the analysis can uncover what type of children's books are most popular with parents and what aspects of the book are liked by readers, for example, quality of writing, high educational value or covering a particular theme or topic that is deemed more suitable for children. Rather than looking into reviews individually, parents can use the summarized results to make decisions on selecting suitable books for their children.

2. When making book procurement decisions, organizational users like schools, libraries or bookstores can apply the results of our analysis to consider relevant books to purchase for their student collection.

3. Publishers would also be interested to evaluate the performance of their books and understand user preferences so that books published would be relevant and well received by users, which in turn drives book sales and revenue.

5. Solution Details

The team first attempted to perform topic modelling and clustering across all 10 books. This resulted in a high degree of overlap between clusters and topics, pointing to the likelihood of these clusters and topics being too generic and similar. To sharpen the analysis and make distinctions between clusters and topics more meaningful, the team decided to run the analysis individually for each of the 10 books.

5.1. Topic Extraction using LDA topic modelling

Text pre-processing

The team applied LDA using Gensim LDA model on the 10 books. After loading the dataset, text pre-processing was applied to clean up the dataset. The text-preprocessing treatments applied were word tokenization, removal of stop words, numbers, dates, website URLs, hashtags, emoticons and blank lines and stemming. To enhance the clarity and distinction between topics, only nouns and verbs were included in the corpus. Manually identified stop words such as "book", "read", and "story" and foreign words were included in the pre-processing as these words did not provide much value to the analysis.

LDA analysis

The LDA analysis consisted of iterating 20 models of 1 to 20 topics per book. The coherence score for each model was calculated and output. The optimal-k for the specific book was then selected based on the model with the highest coherence score. The topics of the optimal model were then analysed in greater detail to identify the main topics/themes of the reviews.

5.2. Document Clustering

Text pre-processing

For the preparation of K-Means clustering, the team first used a regular expression function to remove all websites, emoticons, numbers, tweets and stop words like LDA, mentioned above. To further improve the quality of the results, the team selected only nouns and verbs using POS Tagging function in NLTK. Finally, we stemmed and tokenized the output for the use in our vector space model approach for K-Means Clustering.

K-Means Clustering

To better understand the different type of topics mentioned in reviews along with the keywords for each cluster, the team chose K-Means clustering along with LDA as a means of high-level extraction of keywords and classifying the reviews for each book. The K-Means algorithm is applied at the book level where all reviews for a book are used as the corpus. In essence, with the 10 books from our dataset, there are 10 corpuses and 10 K-Means cluster algorithm application (1 for each book).

Furthermore, the team also applied a single K-Means cluster run on the entire dataset itself, as a test to prove that the results when applying the algorithm on a book level will yield better insights for our business use case.

5.3. Sentiment Analysis

Text Pre-processing

After loading the reviews for all top 10 books, text pre-processing was applied to clean up the dataset, removing website URLs, hashtags and mentions. Since user reviews can contain a mixture of comments and sentiments, the reviews were retained in sentence form for application of lexicon-based sentiment packages, Vader and TextBlob to calculate sentiment polarity. The analysis was conducted at a corpus level to obtain average sentiment polarity for each book but was repeated at book level to investigate detailed reviews for each book.

Overall Sentiment Analysis

While the reviews did not contain sentiment labels, the dataset contained an attribute called 'User Rating', ranging from 1 (negative) to 5 (positive). This rating was adopted as a source of "golden truth" and a simple mapping to three sentiment labels – positive, negative, and neutral. The sentiment polarity scores for Vader and TextBlob were also divided into three groups and their mapping is shown in Table 4:

	Negative	**Neutral**	**Positive**
User rating	1, 2	3	4, 5
Vader	-1.0 to -0.31	-0.3 to +0.3	+0.31 to +1.0
TextBlob	-1.0 to -0.31	-0.3 to +0.3	+0.31 to +1.0

Table 4: Vader and TextBlob sentiment polarity scores

Two sets of classification reports were generated, one comparing user rating against Vader and another comparing user rating against TextBlob.

5.4. Aspect-Based Sentiment Analysis

Aspect-Based Sentiment Analysis which aims to identify aspects and the corresponding sentiments is a multi-step task. First, the aspects must be identified. Second, the aspects and the corresponding sentiments must be extracted. Last, the sentiments must be analyzed.

Association Mining

Aspect mining for Aspect-Based Sentiment Analysis is normally done using topic analysis. We performed LDA across the top 10 books to identify aspects. Rather than topics about aspects, however, the topics generated by LDA appear to be about the content of the books. For example, topic 2 is about *Harry Potter* while topic 5 is about *The Invention of Hugo Cabret* (see Figure 3).

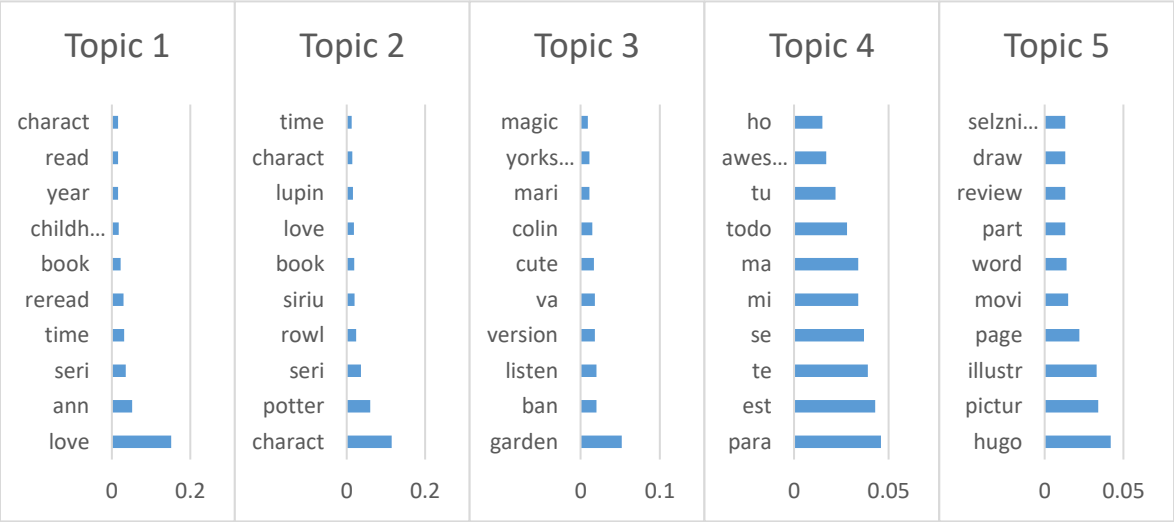

Figure 3: Sample of topics for top 10 books

Given the results from LDA did not appear to yield topics about aspects, we researched other methods for identifying aspects. Inspired by the work of Hu and Liu (2004), we use Association Mining to identify the frequently mentioned aspects in Goodreads reviews. Hu and Liu (2004) observed that "when [customers] comment on product features, the words that they use converge".

To prepare the data for Association Mining, we first apply sentence and then word tokenization to enable POS tagging which later allows us to filter out non-nouns. We then filter for words with only alphabet characters, non-stop words (including the word 'book' which occurred very frequently), and nouns.

Once prepared, we apply the Apriori algorithm using a min_support of .001, min_confidence of 0, and min_lift of 0. The rationale for these parameters is higher min_support resulted in too few itemsets, so we had to decrease min_support, and we were uninterested in order so confidence and lift are both 0.

Information Extraction

Before diving into extracting aspects and sentiments, we filter for sentences that mention at least one of the aspects identified via Association Mining. Then, we use spaCy's DependencyMatcher which enables information extraction using dependency trees. DependencyMatcher extracts information using a set of user defined rules. To define our rules, we study the sentence's grammar structure using displacy, a dependency visualization within spaCy, define and then test the rules.

Sentiment Analysis

Similar to the approach that we took in overall sentiment analysis, we apply Vader to the extracted sentiment words to understand any positive, neutral, and negative sentiments about the aspects.

6. Results and Analyses

To illustrate the results and an application of our business case, we selected the book *The Invention of Hugo Cabret* (Book_id 9673436) which contains a total of 1,947 unique reviews. We performed our analytical tasks and the results presented here can be adopted by users for their relevant business purposes.

6.1. LDA topic modelling

The coherence score output was visualized in graphical form to identify the optimal-k per book. In the case of Book_id 9673436, the optimal-k is 10 topics. See Figure 4 and Table 5.

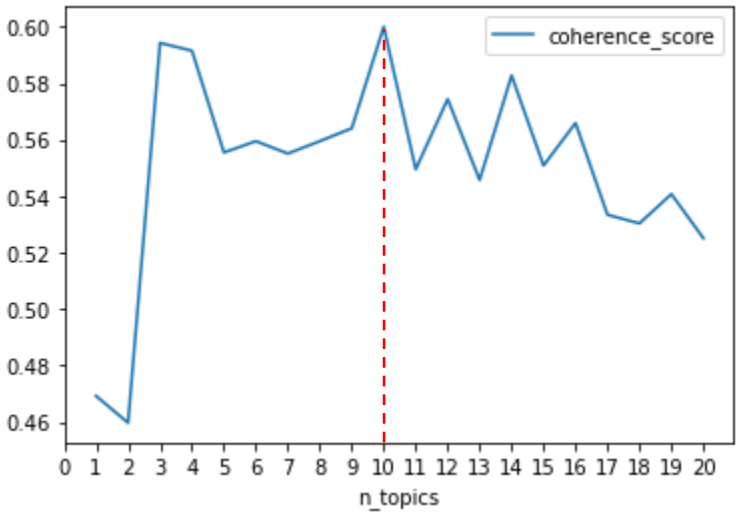

Figure 4: Identifying the optimal-k using coherence score

Model	Coherence score	Model	Coherence score
1	0.469197074	11	0.549496653
2	0.459685119	12	0.574440565
3	0.594325544	13	0.545736089
4	0.591557309	14	0.582763491
5	0.555490993	15	0.550863385
6	0.559509134	16	0.565949233
7	0.555110372	17	0.533414673
8	0.559513383	18	0.530356462
9	0.564073748	19	0.540748404
10	**0.600064742 (highest coherence score)**	20	0.525120246

Table 5: Coherence scores for the 20 models

A sample of the topics from this model is shown below (Figure 5). The topics would suggest what people liked about the book and what the book's greatest attractions are. From the three example topics, we can

see that this book teaches the value of friendship, is a book that grownups (adults) remember reading in their childhood, and that the story was made into a movie.

Figure 5: Sample of topics for Book_id 9673436

One particular challenge faced in LDA was that of misclassified words that do not actually appear in the book. For example, Topic 4 of Book_id 9673436 has the words "rowl" (author J.K. Rowling), "harri" and "potter" (Harry Potter). Upon further analysis, a number of reviews contain references to the Harry Potter series and its author JK Rowling; these were mentioned as comparisons to Book_id 9673436. This could be the cause of the misclassification. A sample of reviews containing these misclassified words is shown below.

"Not reading this novel is like not having read Harry Potter*. It is a wonderful journey that will captivate and stay with you."*

"The Invention of Hugo Cabret provided the perfect antidote to my post Harry Potter *blues this summer. Selznick has joined the ranks of writers such as Lemony Snicket, J.K.* Rowling *and Marcus Zusak who have proven that juvenile literature isn't just for kids."*

"My mom's teacher-friend told me that it's like fifth-grade level and also that it is better than Harry Potter*, which is a really high compliment that I would probably agree with if I had read it when I first read* Harry Potter*, in fourth grade."*

Given the topics identified from this book, a parent may like that the book extols good virtues such as friendship. Also, as the book has been made into a movie, the parent might introduce the movie to the child first to get him/her interested in the story, and then subsequently introduce the book to the child. Schools and libraries might evaluate that this book is suitable for schoolchildren, again based on the good virtues taught in the book, and that books reviews compared it favourably with top-sellers like Harry Potter. Publishers would similarly be interested to know that this book compares favourably to popular books like Harry Potter, and that readers of Harry Potter may like this book. This may advise promotional strategies like packaging this book together with a Harry Potter book and sold at a promotional discount.

6.2. Document Clustering

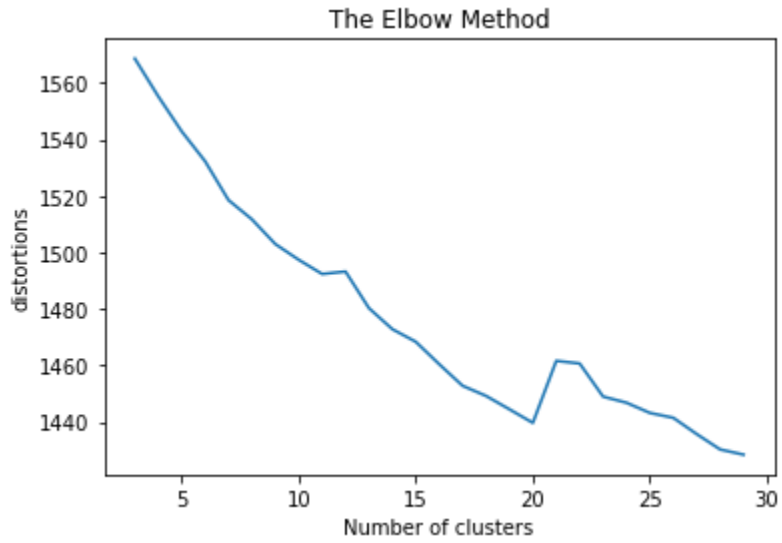

Figure 6: Result of K-Means for Book_id 9673436

The team ran the K-Means clustering algorithm for all 10 books. Figure 6 shows the result from Book_id 9673436. For the optimal number of clusters, the team opted for the elbow method and identified the optimal number of cluster2 to be 20. After selecting the number of optimal clusters, we visualized the optimal number of clusters using the YellowBrick module.

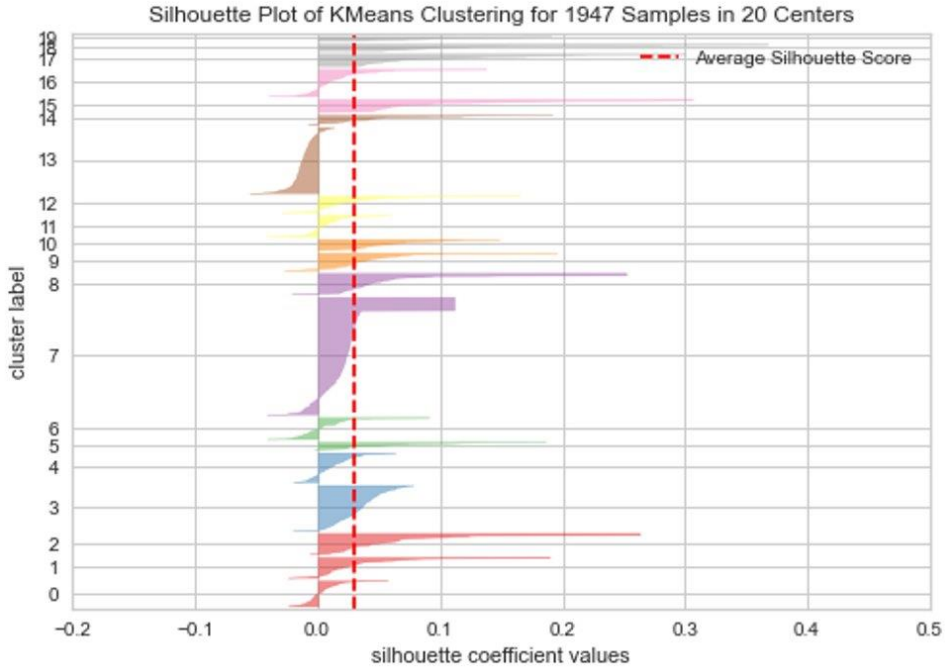

Figure 7: K-Means for Book_id 9673436 visualised in Silhouette Plot at N=20

Inspecting the silhouette plot in Figure 7, it is noted that for all clusters, a portion of the reviews have been mislabeled especially for cluster 7 and 13 where the largest number of reviews are clustered in.

Upon further inspection we noted that these mislabel reviews contained foreign languages. The algorithm is unable to classify these foreign words effectively and therefore assigned the reviews randomly.

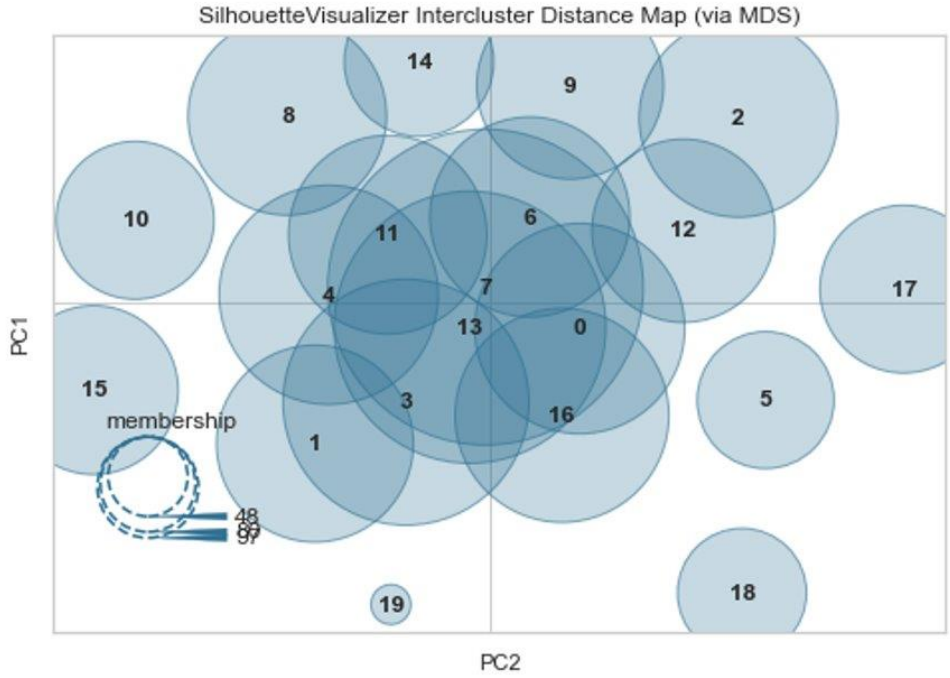

Figure 8: PCA plot for Book_id 9673436

Looking at Figure 8 above, the PCA plot shows that clusters 7 and 13 have the most overlap between clusters while clusters 4, 11 ,6, 0, 16, 3 and 1 have some overlap. To further analyse the result, the team ran the K-Means clustering algorithm at N = 20 again, this time showcasing the cluster size as well as most frequent word occurrences below. Note that the cluster number presented below is not similar to the illustrations above.

Cluster	comment_label
1	['kid', 'love', 'adult', 'think', 'pictur', 'illustr', 'children', 'enjoy', 'reader', 'draw']
2	['love', 'illustr', 'end', 'charact', 'way', 'pictur', 'plot', 'recommend', 'amaz', 'art']
3	['end', 'page', 'think', 'look', 'love', 'interest', 'charact', 'artwork', 'didnt', 'illustr']
4	['wonder', 'illustr', 'work', 'invent', 'love', 'son', 'movi', 'add', 'struck', 'film']
5	['star', 'love', 'pictur', 'illustr', 'get', 'way', 'didnt', 'part', 'word', 'day']
6	['fun', 'illustr', 'page', 'lot', 'pictur', 'see', 'way', 'age', 'movi', 'text']
7	['hugo', 'station', 'clock', 'train', 'man', 'work', 'father', 'cabret', 'part', 'automaton']
8	['children', 'adult', 'illustr', 'recommend', 'think', 'love', 'age', 'artwork', 'work', 'captiv']
9	['movi', 'love', 'see', 'watch', 'made', 'illustr', 'look', 'make', 'time', 'word']
10	['draw', 'reader', 'way', 'page', 'tell', 'work', 'time', 'recommend', 'part', 'see']
11	['pictur', 'illustr', 'word', 'tell', 'love', 'use', 'page', 'text', 'part', 'told']
12	['que', 'libro', 'hugo', 'illustr', 'time', 'experi', 'lo', 'love', 'art', 'hour']
13	['cute', 'love', 'illustr', 'quick', 'effect', 'pictur', 'get', 'divid', 'intermingl', 'fast']
14	['caldecott', 'winner', 'medal', 'pictur', 'award', 'page', 'draw', 'love', 'deserv', 'children']

15	['beauti', 'illustr', 'love', 'move', 'heart', 'everyth', 'get', 'thing', 'art', 'recommend']
16	['enjoy', 'think', 'illustr', 'pictur', 'time', 'movi', 'made', 'novel', 'text', 'everyth']
17	['amaz', 'illustr', 'time', 'pictur', 'love', 'draw', 'word', 'watch', 'thought', 'want']
18	['film', 'see', 'love', 'histori', 'illustr', 'hugo', 'movi', 'pictur', 'watch', 'part']
19	['written', 'children', 'illustr', 'part', 'draw', 'way', 'love', 'pictur', 'word', 'told']
20	['review', 'come', 'blog', 'post', 'love', 'wow', 'check', 'page', 'time', 'plea']

Table 6: Most Frequent Cluster Keywords for Book_id 9673436

Looking at Table 6 above, we note that certain clusters are distinct and mentions about topics related to the movie (cluster 4, 9 and 14) while others are about the plot of the book (cluster 7). However, we note that there are many overlapping words such as “love”, “illustration”, “enjoy” etc. This could be the cause of the overlapping clusters seen in Figure 8. The team has provided a human interpretation of each cluster in the Appendix of this report.

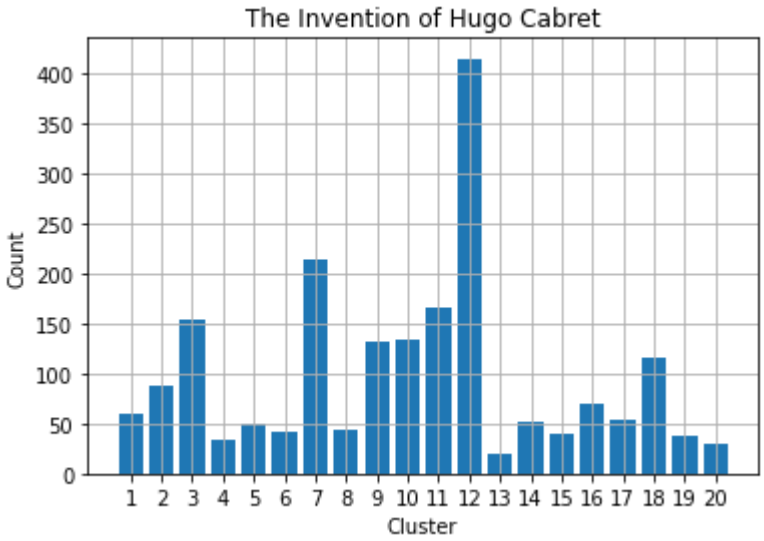

Figure 9: Cluster Frequency Distribution for Book_id 9673436

Visualising the distribution of the clusters (Figure 9), it is noted that many foreign words are represented by cluster 12, while distinct clusters mentioning separate topics have relatively small cluster size (4, 9, 14 and 20). We can interpret the results as most of the comments are positive towards the book and illustration. However, certain topics are also of interest to the readers, such as the movie adaptation and the plot of the book.

For this analysis, the use-case would be to provide a summary of the key topics to the publishers which can then be linked to further analysis such as aspect-based sentiment analysis below.

6.3. Sentiment Analysis

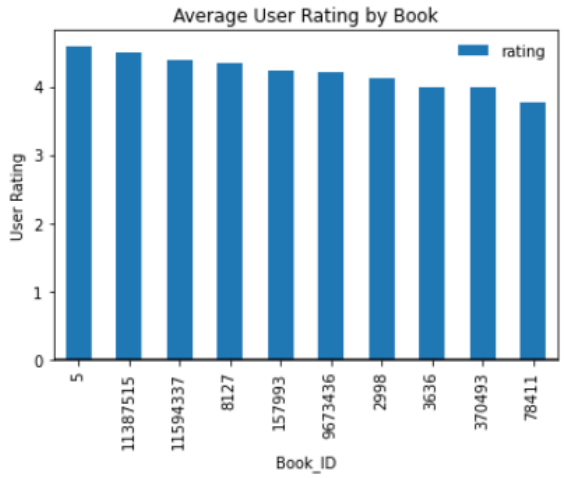

Figure 10: Average user rating by Book_ID

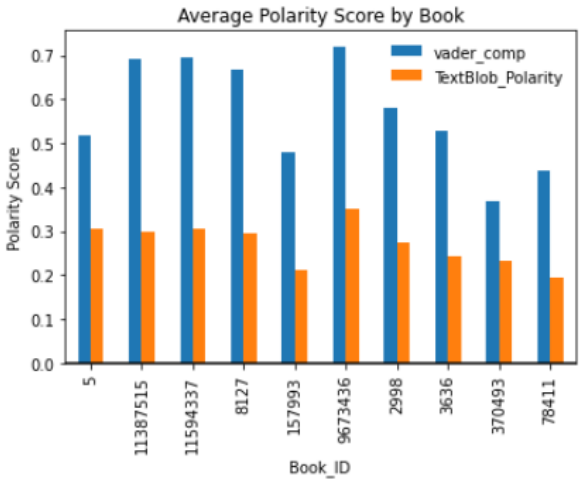

Figure 11: Average polarity score by Book_ID

Figures 10 and 11 show the average user rating and polarity scores for each book respectively, ranked by decreasing order of user rating. There are differences in polarity scoring between Vader and TextBlob.

An evaluation of the performance of both sentiment packages was performed and the classification results show that Vader has a higher accuracy rate in correctly identifying sentiment labels. The average precision and recall are 0.67 and 0.7 respectively, as shown in Figure 12. TextBlob on the other hand, has a lower average precision and recall of 0.59 and 0.45 respectively, as shown in Figure 13.

```
print(classification_report(df1['vader_sentiment'],df1['user_sentiment']))
```

	precision	recall	f1-score	support
neg	0.24	0.24	0.24	1967
neutral	0.23	0.15	0.18	4748
pos	0.80	0.86	0.83	22213
accuracy			0.70	28928
macro avg	0.42	0.42	0.42	28928
weighted avg	0.67	0.70	0.69	28928

Figure 12: Classification results from Vader

```
print(classification_report(df1['TB_sentiment'],df1['user_sentiment']))
```

	precision	recall	f1-score	support
neg	0.08	0.36	0.14	458
neutral	0.70	0.13	0.22	16531
pos	0.45	0.89	0.59	11939
accuracy			0.45	28928
macro avg	0.41	0.46	0.32	28928
weighted avg	0.59	0.45	0.37	28928

Figure 13: Classification results from TextBlob

Filtering the results to Book_id 9673436, we can see from the results that the average sentiment for this book is positive. User rating, Vader, and TextBlob polarity scores as follows (Table 7). For the three groups of users identified in our business case, the overall sentiment analysis would be able to help users obtain an overall sense of the sentiment of this book. Should the users wish to understand more about which aspects of the book were appealing, they could refer to the results from aspect-based sentiment analysis, discussed in the next section.

User Rating	Vader	TextBlob
4.2250	0.7202	0.3495

Table 7: Average sentiment from Vader and TextBlob

6.4. Aspect-Based Sentiment Analysis

Association Mining

The Apriori algorithm produced 95 itemsets. From this set, we manually skim through these to extract any children's books aspects. We extract 7 aspects: story, character, message, plot, style, writing, and plot.

These results align closely with several online articles and sources about what makes a book great. According to one website, there are 7 critical elements of a great book: plot, characters, viewpoint, dialogue, pacing, style, and beginning, middle, and end (Patterson, 2016). Table 8 shows the top 10 itemsets by support of which 5 out of 10 contain an aspect (story).

Word1	Word2	Support
('story', 'NN')	('way', 'NN')	0.00406
('story', 'NN')	('boy', 'NN')	0.00398
('spoiler', 'NN')	('alert', 'NN')	0.00397
('life', 'NN')	('story', 'NN')	0.00380
('tree', 'NN')	('boy', 'NN')	0.00366
('time', 'NN')	('school', 'NN')	0.00365
('point', 'NN')	('view', 'NN')	0.00359
('time', 'NN')	('story', 'NN')	0.00329
('school', 'NN')	('year', 'NN')	0.00279
('story', 'NN')	('school', 'NN')	0.00259

Table 8: Top 10 itemsets from Apriori algorithm

Information Extraction

Figures 14 and 15 show the top aspects and sentiments extracted from Book_id 9673436. The top aspect by far is "story," and the top aspects are "beautiful," "wonderful," and "great" which suggests positive sentiment about the story.

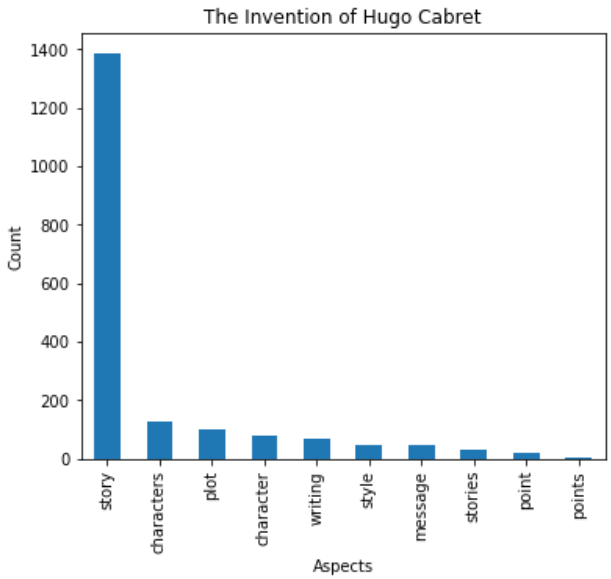

Figure 14: Top aspects from Book_id 9673436

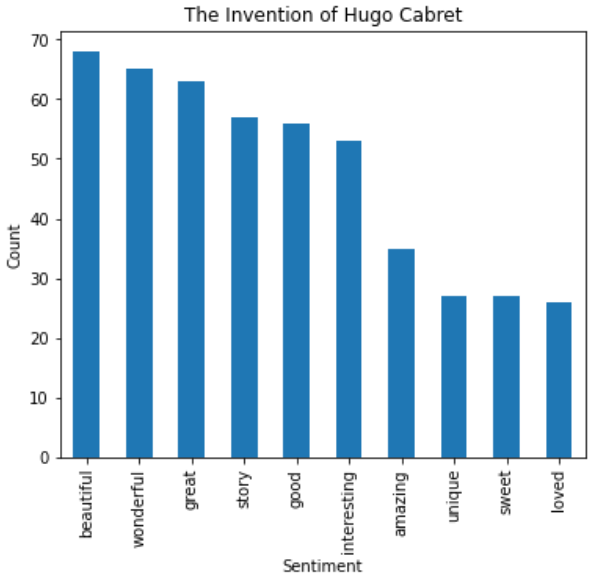

Figure 15: Top sentiments from Book_id 9673436

To evaluate the accuracy of our information extraction, we perform a manual evaluation of 50 sample sentences. We manually extract the aspect and the sentiment, if any exists, from each sentence and compare the human output to the computer output. The results from this evaluation suggest a 50% accuracy. That is, for 25 out of 50 sentences, the human and the computer aspect and sentiment pairs matched.

Table 9 shows a sample from the evaluation. The aspect and sentiment are correctly extracted only for sentence 2. For sentences 1 and 3, it extracts the incorrect sentiments; for sentences 4 and 5, it extracts sentiments where the original sentence does not contain a sentiment directed at the aspect.

S/n	Sentence	Human Aspect	Human Opinion	Computer Aspect	Computer Opinion
1	"Rowling's plots are getting tighter and tighter and this one is no exception with great surprises in the last chapters.",	plots	tighter	plots	getting
2	'It introduces Sirius Black and Remus Lupin who are some of my favorite characters.',	characters	favorite	characters	favorite
3	"The story keeps getting better and better, and one can't help but be happy to enter that world once again and keep fighting the dark wizard!",	story	better	story	keeps
4	"I know it would hurt the plot a lot, but I think Rowling shouldn't have made something so powerful.",			plot	hurt
5	'His tenderness is apparent, in the fact that unlike most of the characters, he never "shouts," only patiently inquires.'			characters	apparent

Table 9: Evaluation of computer extraction of aspects and opinions against human extraction

Sentiment Analysis

As demonstrated in Figure 16, the sentiment for the top aspects in Book_id 9673436 are all neutral (between -.3 and +.3) which contradicts our earlier hypothesis that readers had positive sentiments based on the top sentiments.

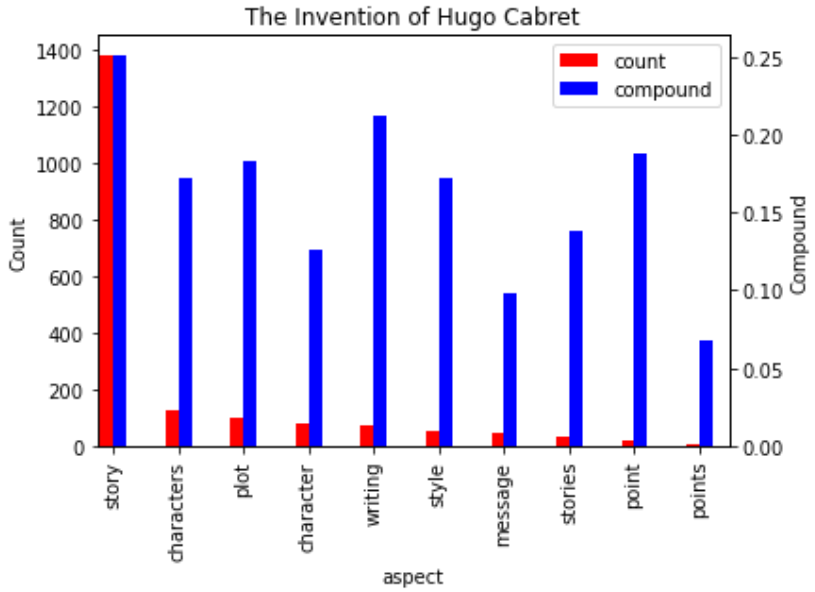

Figure 16: Sentiment for the top aspects in Book_id 9673436

To validate these results, we perform another manual evaluation on a sample of 50 sentences comparing human versus computer classification of positive, neutral, negative, or NA. NA represents the sentences with no sentiment. Tables 10 and 11 show the results from our manual evaluation. The results are the best for positive sentences. When a sentiment is identified as positive 100% of the time, it is positive. On the other hand, the results are poor for neutral, negative, and NA likely due to the inputs from aspect and sentiment extraction.

	Precision	**Recall**	**F1**
Positive	1	0.59	0.74
Neutral	0.08	1.00	0.14
Negative	0.67	0.50	0.57
NA	0.75	0.20	0.32

Table 10: Precision, recall, and F1 from manual evaluation

	Human				
Machine	**n=50**	**Pos**	**Neu**	**Neg**	**NA**
	Pos	17	0	0	0
	Neu	11	2	2	11
	Neg	0		2	1
	NA	1	0	0	3

Table 11: Observations from manual evaluation

Given the results from our manual evaluation, it does appear that we are extracting aspects and non-sentiment pairs which is causing our aggregate results to appear neutral and hiding any true positive or negative sentiments. As a result, the possible applications and use cases for this analysis are not immediately demonstrable. That said, we are still able to extract some useful insights. The results from figures 14 and 15 point to the aspect "story" as being important, and the top sentiments suggest readers think aspects of the book are "beautiful," "wonderful," and "great."

7. Discussions and Gap Analysis

7.1. Topic extraction

Some key limitations are due to the specific nature of book reviews. Words such as "book", "read", "story" which would not typically be stop words needed to be manually included into the list of stop words to ensure that the pre-processed corpus was meaningful for analysis. Secondly, there was some misclassification due to "noise". For example, a real-life author may have a similar name to a fictional character (for example, a common name like Anne), which could potentially cause this name to be misclassified. Also, book reviews sometimes contain features of other books, as the reviewer might want to compare between different books in their review. Such examples could be mentioning other titles that an author wrote or comparing the main character from another book with the book being reviewed. Such "cross-referencing" adds to the noise in the analysis.

7.2. Document Clustering

One key limitation of K-Means clustering is the inability to group words with similar meaning together. For example, "I love the book" and "I like the novel" both show positive meaning towards the title but are considered as separate after removing stop words. This causes the vector space to be large and reduces the effectiveness of the K-Means algorithm.

In addition, the team noted some reviews with non-language words in reviews. This causes mislabeling of the cluster as seen in the YellowBrick visualization tool. Finally, the keywords provide only a high-level summary of the reviews and do not go into more detail, which requires further analysis using other methods.

7.3. Sentiment Analysis

Many reviews in the corpus contained multiple statements, some of which did not express sentiment but were descriptions of certain parts of the story. These were considered to be neutral statements. Since Vader and TextBlob evaluate the review as a whole, multiple neutral statements may lower the polarity of a review, resulting in the overall review being perceived as less positive or negative than it should have been. It was also observed that there were differences in the rules applied between Vader and TextBlob (Table 12).

Sentence-level Review	Vader compound score	TextBlob polarity score
I loved the story!	0.636	0.875
I loved Hugo as a character!	0.636	0.875
I loved how the 2 storylines combined and filled each others holes!	0.636	0.6
I loved the artwork!	0.636	0.875

The pencil lines!	0.0	0.0
I loved how great the story was and how both the greatness of the story and the artwork combined made this book both incredibly enjoyable and super easy to get through in one day.	0.9599	0.553
Overall score	**0.9919**	**0.1943**
User rating for this review was **5 (positive)**.		

Table 12: Comparison of scores for Vader and TextBlob

7.4. Aspect-Based Sentiment Analysis

Association Mining

As mentioned previously, we set the min_support of the Apriori algorithm to .001 because setting it any higher resulted in too few itemsets from which to extract aspects. This is likely due to the large number of words in our corpus even after cleaning for only nouns and removing stop words and other commonly occurring non-aspect words. Lemmatizing and identifying synonyms may marginally improve the results from Association Mining because similar words will be identified.

For more marked improvement to identifying topics, expanding the analysis beyond the top 10 books or adapting the methodology by Hu and Liu (2004) for identifying infrequent itemsets should be considered.

Information Extraction

Generally, we were able to extract aspects and sentiments from simple sentences well. (Sentence 1 in Table 13). But found it significantly more challenging to extract aspects and sentiments from complex sentences or grammatically incorrect sentences. See sentences 2 and 3, respectively in Table 13.

S/n	Sentence	Aspect	Sentiment
1	'I really like their character and I do like Klaus since he enjoy reading books.',	character	like
2	'Like in the story how Frances Hodgson Burnett describes the ambiance, I am familiar with the smell of leaves after pouring down in torrents, of the river which water smells brackish wafting up in the air, with the warm welcome of the sunshine in the breaking dawn, the marvelous blossoms of flowers in a garden, the canopies of the huge trees in a forest.',	story	familiar
3	'Hope \n In the story, Mary Lennox hopes that the garden has the big potential to revive its spirit.',	story	big

Table 13: Sample of aspects and sentiments extracted

Sentiment Analysis

Vader evaluated sentiment accurately for the top sentiment words used (see Figure 17). However, because the inputs from aspect and sentiment extraction were only 50% accurate, the results from sentiment analysis were adversely impacted as well.

```
Vader result for great {'neg': 0.0, 'neu': 0.0, 'pos': 1.0, 'compound': 0.6249}
Vader result for beautiful {'neg': 0.0, 'neu': 0.0, 'pos': 1.0, 'compound': 0.5994}
Vader result for interesting {'neg': 0.0, 'neu': 0.0, 'pos': 1.0, 'compound': 0.4019}
Vader result for main {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
Vader result for favorite {'neg': 0.0, 'neu': 0.0, 'pos': 1.0, 'compound': 0.4588}
```

Figure 17: Sample result from Vader sentiment analysis

8. Limitations of the Project

For the scope of this project, emoticons which have been shown to affect scoring were not considered and removed in the pre-processing stage. This inevitably causes sentiment analysis to be affected as emoticons have been shown to affect sentiments. For example, a person might use a sad emoticon while saying he loves his work, by removing the emoticon, the sentiment changed from negative to positive. (Hogenboom et al, 2013)

For both topic extraction and document clustering, misspelled words were not corrected due to the computational requirements and complexity of the task. This inevitably affected the vector space model for document clustering and keyword distribution in topic extraction.

Our Aspect-Based Sentiment Analysis is only able to mine for known aspects. Implicit aspects contained in the text were not considered. One example would be a review that says, “I loved Hermione”, which is a positive sentiment about a character in the Harry Potter series, but it does not explicitly mention the aspect “character.” These sentences are ignored.

Another limitation with our implementation of Aspect-Based Sentiment Analysis is the sentiment is limited to one word when perhaps the sentiment is best expressed by a phrase. Perhaps the most notable example of this is the omission of negation words. If the review says, “This story was not good,” our implementation of Aspect-Based Sentiment Analysis would extract “story” and “good” as the aspect and sentiment pair, and Vader would evaluate a positive sentiment about the story. An improved implementation would extract “story” and “not good” as the aspect sentiment pair.

9. Future Work and Conclusion

To summarize, the project was helpful in providing users with an overview of the topics discussed and the sentiments relating to the book. Using the selected book “The Invention of Hugo Cabret” as an example, the overall sentiment of the book was found to be positive. Topic modelling informed us that the book taught about the value of friendship, adults remember reading this book in their childhood and that the story was made into a movie. Document clustering identified 20 clusters with similar topics discussed. Common themes identified were about the illustrations of the book, as well as the fact that the story was being made into a movie. Association mining extracted 7 features of the book that was appealing to readers - story, character, message, plot, style, writing, and plot. Extracted words describing aspects of the book (like “beautiful” and “great”) was also considered positive and aligned with the overall sentiment of the book.

The methodology of this project could be applied to another book of choice to help guide parents to select appropriate books that they wish to introduce to their children. For schools, these results would be helpful

in guiding their choice of books to introduce in the curriculum and could also guide libraries' choice of books to feature. For publishers, the results would inform them the type of books that are preferred by readers so that they can continue to select and publish relevant books to readers.

For future work, the team has identified the following areas to enhance the outcome of this analysis:

1. Depending on the analytical task, a method to reduce 'noise' before analysis should be done. This can be achieved using external packages such as polyglot which is trained to detect multiple languages with a confidence level. In addition, spelling correction should be done using spelling correction packages such as SymSpell, Bk-Tree or LinSpell. In this project, we did not utilise this due to the intensive computational runtime required.

2. A dictionary meant specifically for books and book reviews could be developed to improve classification. This is due to the unique nature of proper nouns such as author names or fictional character names. For example, the name "Anna" could be the name of an author or a main character in a story. This could potentially lead to misclassification. A specially developed dictionary would make it easier to identify popular proper nouns and associate it to the correct book.

3. We strongly recommend exploring other methods, libraries, and packages for extracting aspects and sentiments from sentences. For example, in hindsight, we should have explored the Stanford Dependency Parser introduced in class to evaluate its accuracy and ease of implementation compared to spaCy. This step is critical for improving upon the results of Aspect-Based Sentiment Analysis.

10. Project Experiences/Reflections

This was my first attempt to perform a project end-to-end on such a large corpus. Along the way, there were occasions where the results of the analysis did not perform to our expectations, and we had to tweak our steps to overcome the challenges posed by the dataset. These included having to subset the dataset to a more manageable size and exploring alternative algorithms that would provide more meaningful results. A very good learning experience indeed!
– *Goh Chen Ling Beatrice*

Working on the project gave me an appreciation for NLP. Before taking text analytics, I'd heard of NLP mentioned in conversation and in articles or papers. And during class, we learned about morphological, semantic, and syntactical analyses. But it wasn't until I had to glean meaning from unclean book reviews that I think I fully understood the "why" behind tasks like sentence or word tokenization, stop word removal, or POS tagging. I now realize that I take human speech, cognition, and understanding for granted. Trying to get a computer to identify aspects and sentiments in a sentence and extract it sounded easy in my head, but I quickly realized was a difficult and complex task when applied to "real world" text. Perhaps most importantly, I feel as if I've just seen the tip of the iceberg in terms of what's possible and what can be done with NLP. I'm eager to continue learning, and I hope to be able to apply what I've learned in my future work.
– *Kevin Sunga*

Working on this text analysis project has opened my eyes to the wide array of packages available in Python. In addition, I've also learnt on the difficulty of processing dirty data extracted online with the team

requiring to consider many edge cases when cleaning the data. Working on a large dataset has also been challenging without the computational power of enterprise systems, and with the team, I've learnt how we can prototype our analysis and link our results to business use cases. Finally, working on this project has improved my programming skills by gaining more hands-on experience.
- *Koh Jun Jie*

The experience of this project was definitely eye-opening for me. When we first started on the project, I was keen to analyse all the books contain in the data set. But it soon dawned on the team and myself in particular what a mammoth task that would have been given the time constraints of the project. In hindsight, our final approach of selecting only ten books was a wise decision, and it taught me the value of "small-wins": first developing a smaller-scale product, and then scaling it up as more skills and experience are gained. This made the task more manageable and provided much learning opportunities.
– *Wong Kian Hoong Andy*

References

Alibali, M. W. (1999). How children change their minds: Strategy change can be gradual or abrupt. *Developmental Psychology, 35,* 127-145.

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., & Kaymak, U. (2013). Exploiting emoticons in sentiment analysis. Proceedings of the 28th Annual ACM Symposium on Applied Computing - SAC '13. https://doi.org/10.1145/2480362.2480498

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04. Published. https://doi.org/10.1145/1014052.1014073

Patterson, A. (2016, August 10). The 7 Critical Elements of a Great Book. Writers Write. https://www.writerswrite.co.za/the-7-critical-elements-of-a-great-book/

Appendix

Document Clustering

To better understand the clusters, the team attempted to interpret the clusters by making sense of the most frequent keywords extracted from each cluster. By inserting an interpretation, we compared our interpretation against raw text reviews to determine if reviews were similar to our interpretation.

Cluster	comment_label	Count	Interpretation of Cluster
1	['kid', 'love', 'adult', 'think', 'pictur', 'illustr', 'children', 'enjoy', 'reader', 'draw']	60	Keywords of kid, adult, children, love, picture, illustration, enjoy. Reader and draw suggest that the illustration is enjoyable for both children and adults
2	['love', 'illustr', 'end', 'charact', 'way', 'pictur', 'plot', 'recommend', 'amaz', 'art']	87	Keywords suggest that the plot and illustration is loved by the reader, and the book is recommended
3	['end', 'page', 'think', 'look', 'love', 'interest', 'charact', 'artwork', 'didnt', 'illustr']	153	Keywords suggest that the characters and artwork is good and is loved by readers.
4	['wonder', 'illustr', 'work', 'invent', 'love', 'son', 'movi', 'add', 'struck', 'film']	33	Keywords suggest that the illustration and movie is wonderful.
5	['star', 'love', 'pictur', 'illustr', 'get', 'way', 'didnt', 'part', 'word', 'day']	50	Keywords suggest that the book is recommended.
6	['fun', 'illustr', 'page', 'lot', 'pictur', 'see', 'way', 'age', 'movi', 'text']	42	Keywords suggest that the illustration is fun to read, and there is a movie for the book.
7	['hugo', 'station', 'clock', 'train', 'man', 'work', 'father', 'cabret', 'part', 'automaton']	214	Keywords suggest that the cluster is talking about the story and characters.
8	['children', 'adult', 'illustr', 'recommend', 'think', 'love', 'age', 'artwork', 'work', 'captiv']	43	Keywords suggest the readers are captivated and love the illustration. In addition, the book is suitable for all ages.
9	['movi', 'love', 'see', 'watch', 'made', 'illustr', 'look', 'make', 'time', 'word']	132	Keywords suggest that the readers love the movie and illustration.
10	['draw', 'reader', 'way', 'page', 'tell', 'work', 'time', 'recommend', 'part', 'see']	134	Keywords suggest that the book is recommended, and the drawing is good.
11	['pictur', 'illustr', 'word', 'tell', 'love', 'use', 'page', 'text', 'part', 'told']	166	Keywords suggest that the readers love the book. In addition, 'part' suggests that certain moments in the story are noteworthy.
12	['que', 'libro', 'hugo', 'illustr', 'time', 'experi', 'lo', 'love', 'art', 'hour']	415	Keywords suggest that many reviews are written in a foreign language, and that the art is noteworthy.
13	['cute', 'love', 'illustr', 'quick', 'effect', 'pictur', 'get', 'divid', 'intermingl', 'fast']	19	Keywords suggest that the illustration is cute and loved. In addition, the word

			intermingle suggest that the illustration is intermingled with the text.
14	['caldecott', 'winner', 'medal', 'pictur', 'award', 'page', 'draw', 'love', 'deserv', 'children']	52	Keywords suggest that the film is a Caldecott award winner and is highly rated.
15	['beauti', 'illustr', 'love', 'move', 'heart', 'everyth', 'get', 'thing', 'art', 'recommend']	40	Keywords suggest that readers love and recommend the illustration of the book.
16	['enjoy', 'think', 'illustr', 'pictur', 'time', 'movi', 'made', 'novel', 'text', 'everyth']	70	Keywords suggest that readers enjoy the book, and that there is a movie for this book.
17	['amaz', 'illustr', 'time', 'pictur', 'love', 'draw', 'word', 'watch', 'thought', 'want']	54	Keywords suggest that readers love the illustration and recommend watching the film.
18	['film', 'see', 'love', 'histori', 'illustr', 'hugo', 'movi', 'pictur', 'watch', 'part']	116	Keywords suggest readers love the move adaptation as well as the book.
19	['written', 'children', 'illustr', 'part', 'draw', 'way', 'love', 'pictur', 'word', 'told']	38	Keywords suggest readers love the illustration and the way the story is written.
20	['review', 'come', 'blog', 'post', 'love', 'wow', 'check', 'page', 'time', 'plea']	29	Keywords suggest that readers love the works and came to learn of the book through a 3rd party source.