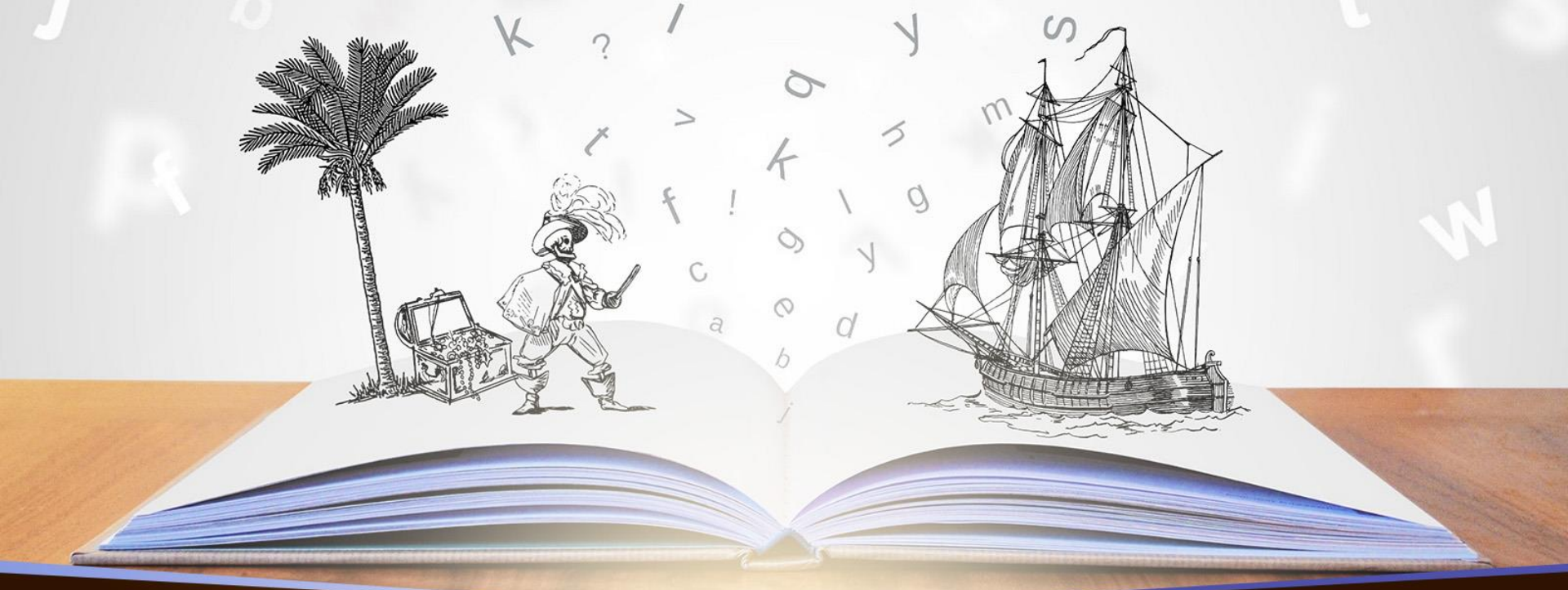


Goodreads Reviews on Children's Books

Group 4 (Goh Chen Ling Beatrice, Kevin Magic Rialubin Sunga,
Koh Jun Jie, Wong Kian Hoong Andy)

Agenda

- Introduction
- Solution Overview
- Solution Details
- Results & Analyses
- Discussion & Gap Analyses
- Future Work
- Conclusion
- Demo

Introduction

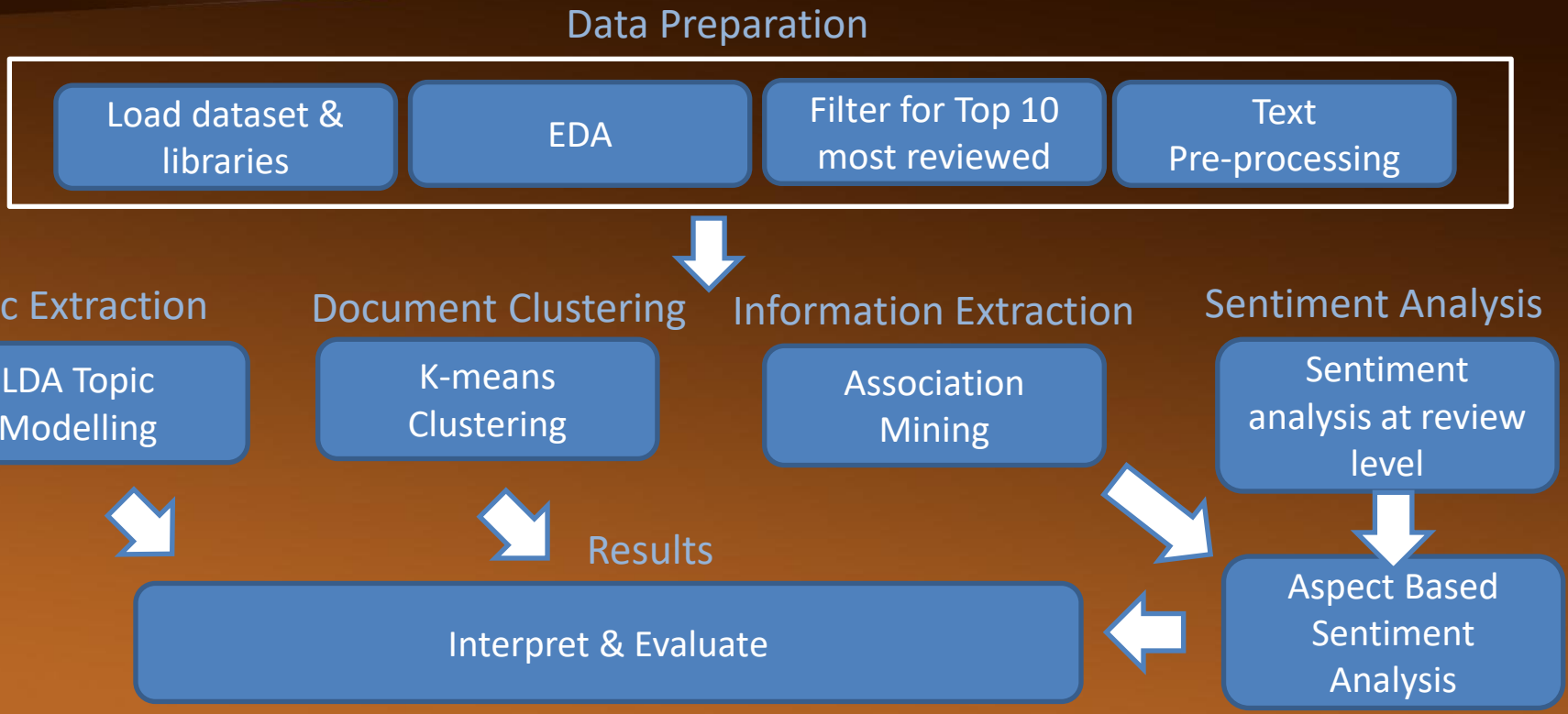
Motivation:

1. Parents want to find good books to introduce to their children
2. Schools & libraries would like to procure popular and recommended books for students
3. Publishers want to evaluate the popularity & performance of their books

Data Source:

- We analysed reader's reviews of top 10 most reviewed children's books on **Goodreads** (about 28,900 reviews)

Solution Overview



Topic Extraction

Solution Details – Topic Extraction

Text Pre-processing

- Punctuation
- Stop words removal
- POS tagging
- Word Tokenization
- Sentence Tokenization

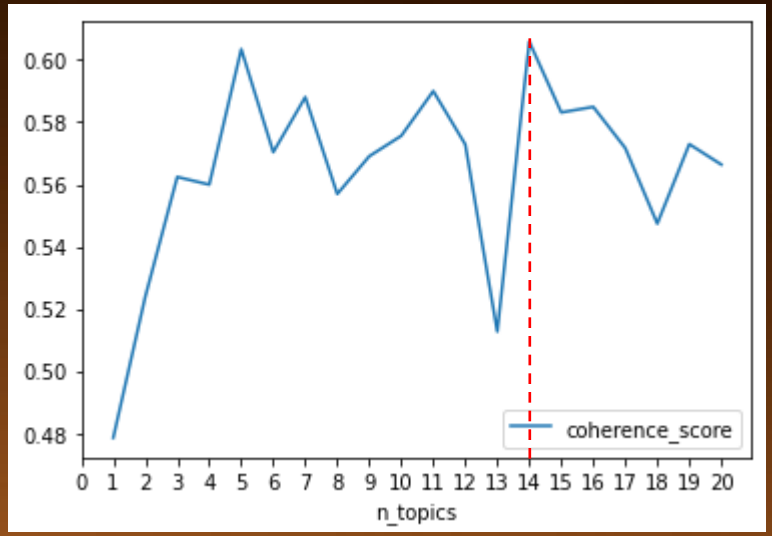
LDA Topic Modelling

- Perform 20 models per book
- Select optimal k using coherence score
- Analyze topics

Results & Analysis – Topic Extraction

Optimal *k* using Coherence Score

- The coherence score algorithm works through each of the books and generates 20 models from 1 topic to 20 topics. The graph shows the coherence score. In this case, the model with **14 topics** is the optimum.

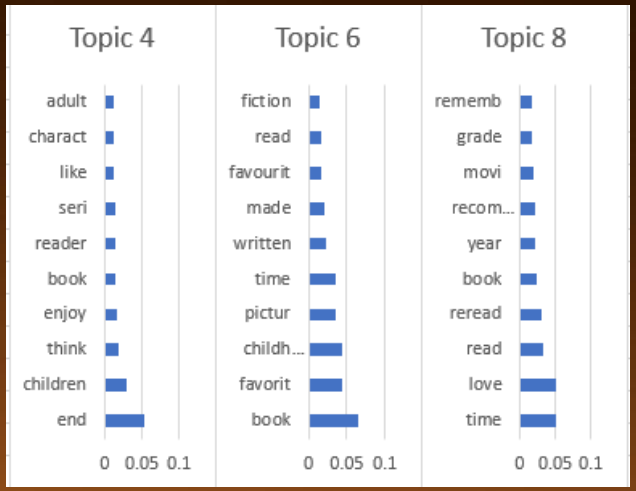



```
#Compute coherence score
coherence_model_lda = CoherenceModel(model = book_lda, texts = dataframe['Processed'],
                                     dictionary = book_dictionary, coherence = 'c_v')
coherence_score = coherence_model_lda.get_coherence()
model_topics.append(n_topics)
model_list.append(book_lda)
coherence_values.append(coherence_score)
```

Application – Topic Extraction

Schools and libraries may analyse the topics for this title to assess its suitability as a book for children.

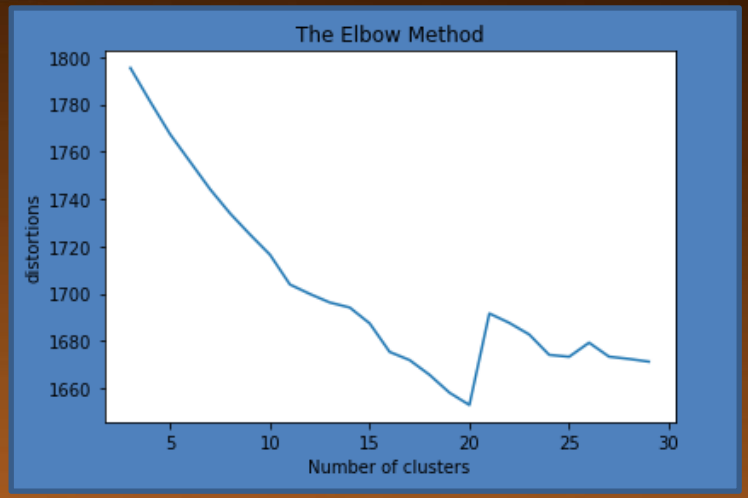
Topic 4: Children may enjoy this book.
Topic 6: A favourite childhood book.
Topic 8: Love the book and reread it. It was made into a movie.



It may be a suitable book for children, and with a movie made of it, it may increase a child's interest to read the book after watching the movie.

Document Clustering

Solution Details – Document Clustering



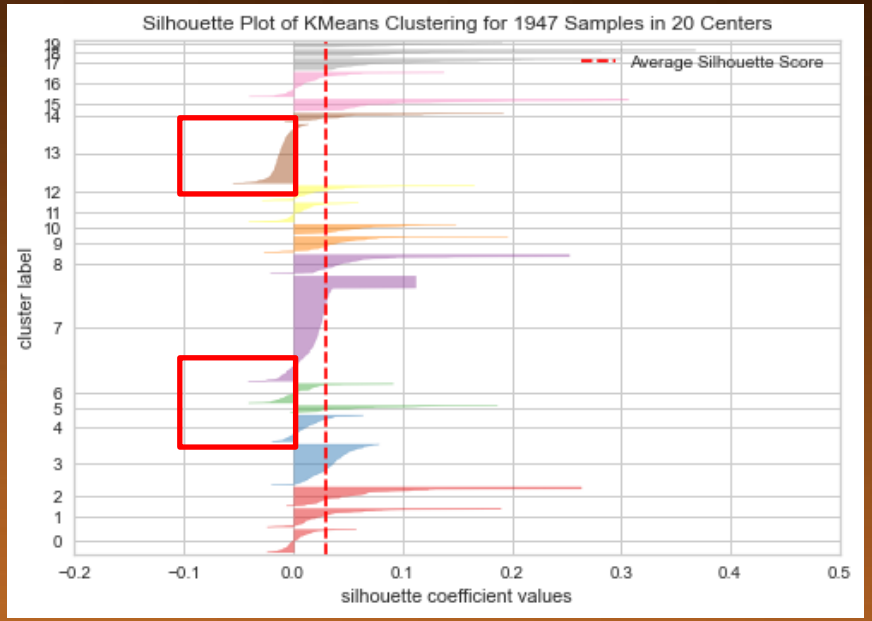
Text Pre-processing

- Text Cleaning
  - Remove websites, emoticons
- Stop words & rare words removal
- Word Tokenization
- POS tagging
  - Selection of Nouns and Verbs only
- Stemming

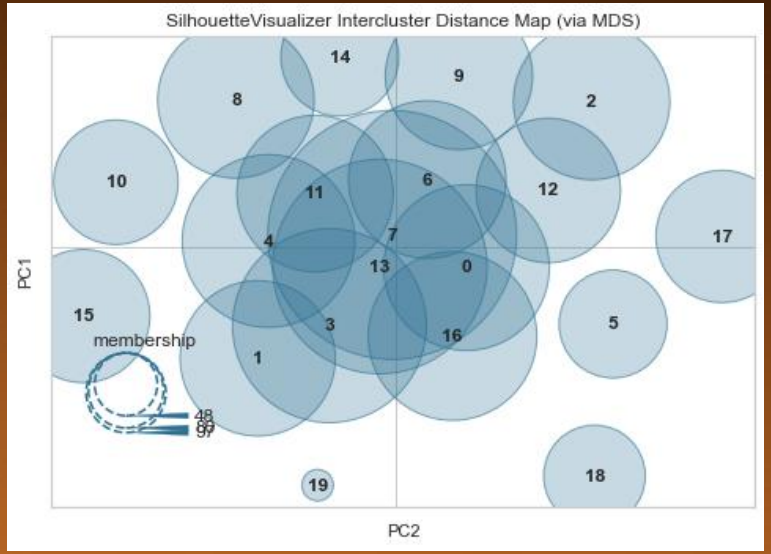
K-Means Clustering

1. Convert stemmed tokens into vectors
2. Run K-Means clustering algorithm for 3 to 30 clusters
3. Select based on elbow method and visualise output
4. View topic words and present intepret data

Results & Analysis – Document Clustering



- High level of noise
- K-Means labelling noises into incorrect clusters



Results & Analysis – Document Clustering

Summary

Some overlapping words between clusters

K-means unable to classify noise accurately

1. Reviewers love the illustration
2. The movie of the book is a critically acclaimed piece
3. The illustration is recommended for both children and adults

Application – Document Clustering

Books get reviewed on Goodreads

Goodreads provide overview for reader comments

Reduce time required by publisher to get feedback on their books

Accurate and raw feedback posted anonymously

Subscription based model where publishers pay for access to detailed information

Publishers gain feedback quickly and able to advise authors on future works

Sentiment Analysis

Results & Analysis – Sentiment Analysis

User Rating (“golden truth”) compared with Vader compound score & TextBlob polarity

	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

```
print(classification_report(df1['vader_score'],df1['user_score']))
```

	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

```
print(classification_report(df1['TB_score'],df1['user_score']))
```

	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

Results & Analysis – Sentiment Analysis

When sentiment packages perform well…..

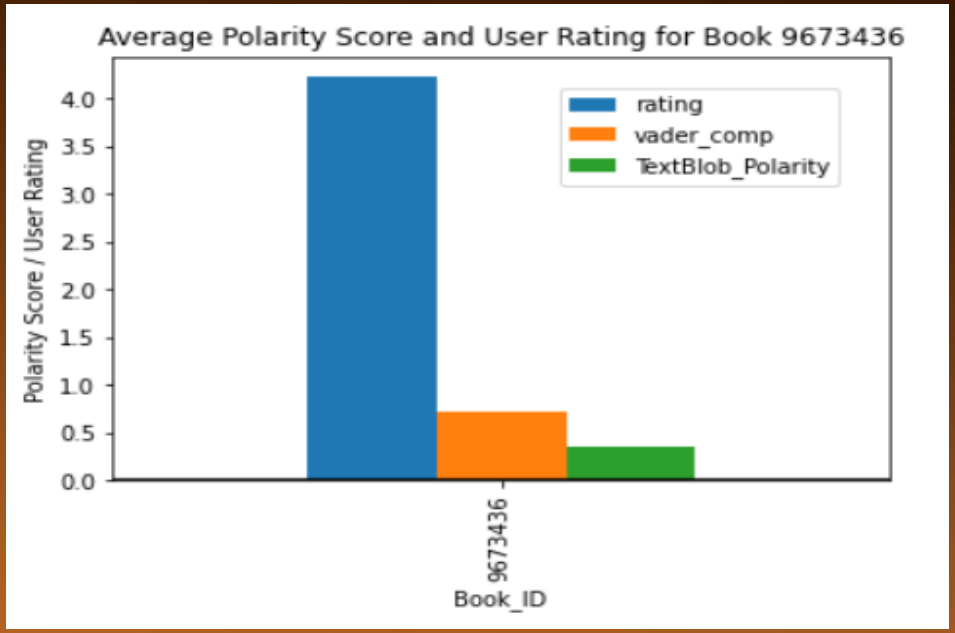
We should see with our hearts not with our eyes. This is the lesson I've learned from reading The Little Prince. I was convinced to read this book because my favorite professor, Sir Elson, told me that this was his favorite book. I've been hearing about this book and I was expecting it to be th ick and in hard bound. But to my surprise, it was just thin. You can finish reading it in a couple of hours. I love that there were illustrations! It teaches a lot of lessons. Such an awe-inspiring book.
User rating is 5
Vader polarity score is 0.9436
TextBlob polarity score is 0.055729166666666656

When sentiment packages do not perform as expected…..

A MUST READ for parents & educators, it demonstrates how a bullied child feels & how one member of the family can affect the whole that takes you on a rollercoaster of emotions.
User rating is 4
Vader polarity score is -0.6249
TextBlob polarity score is 0.2

Application – Sentiment Analysis

Overall sentiment is positive.



book_id	rating	vader_comp	TextBlob_Polarity
9673436	4.224961	0.7202	0.3495

"My three year old loved, LOVED, this book so much that afterwards we started role playing, with me as the Station Inspector and him as Hugo Cabret. Highly recommend."

User rating – 4
Vader compound score – 0.0
TextBlob polarity – 0.0533

"The juxtaposition prose/picture made this book really magical and charming… if I had a kid, this would be at the top of my list to read him/her."

User rating – 5
Vader compound score – 0.9436
TextBlob polarity – 0.0557

Association Mining

Results & Analysis – Association Mining

Example output:

```
====================================
Rule: ('adult', 'NN') -> ('child', 'NN')
Support: 0.0019346269346269347
====================================
Rule: ('story', 'NN') -> ('adult', 'NN')
Support: 0.0011068761068761068
====================================
Rule: ('story', 'NN') -> ('age', 'NN')
Support: 0.0012993762993762994
====================================
Rule: ('spoiler', 'NN') -> ('alert', 'NN')
Support: 0.0039751289751289755
====================================
Rule: ('author', 'NN') -> ('story', 'NN')
Support: 0.0020693770693770695
====================================
Rule: ('way', 'NN') -> ('author', 'NN')
Support: 0.0010683760683760685
====================================
```

- In total, the Apriori algorithm found 95 rules with a min_support of 0.001.
- From 95 rules, we extracted 7 aspects:
  - Story, writing, style, character, point, plot, message
- 5 out 10 of the top rules contained an aspect
- 1 out 10 of the bottom rules contained an aspect

Aspect and Sentiment Extraction

Results & Analysis – Aspect and Sentiment Extraction

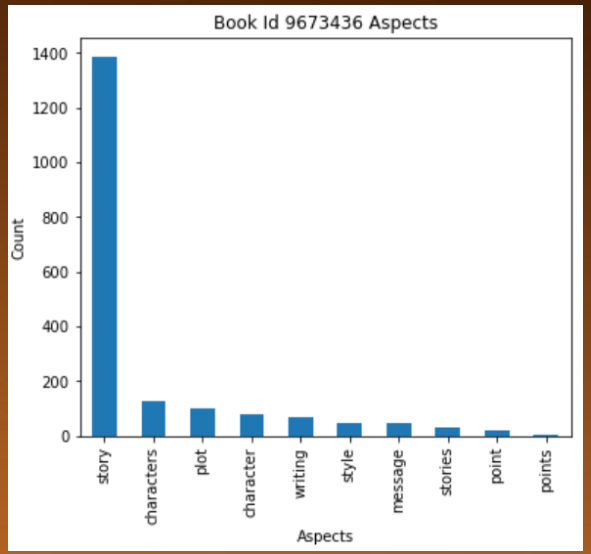


Figure: Top Aspects

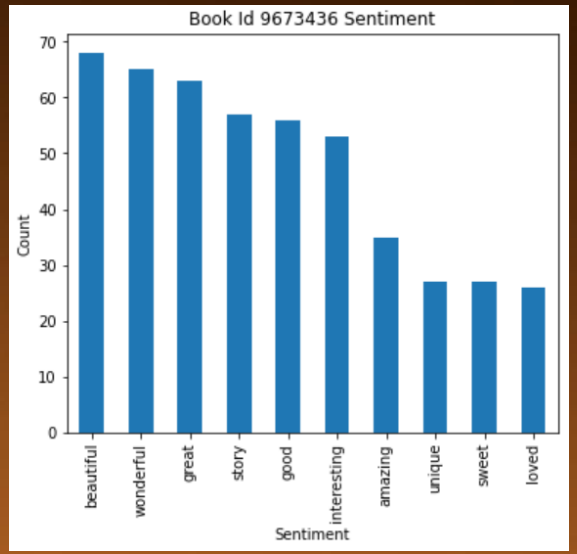


Figure: Top Sentiments

- "Story" is mentioned the most out of all aspects. This suggests that readers likely held an opinion about the book's story.
- The top extracted sentiments are "beautiful", "wonderful", and "great" which suggests that readers may have felt positively about the aspects in the book.

Results & Analysis – Aspect and Sentiment Extraction

Sentence	Human Aspect	Human Opinion	Machine Aspect	Machine Opinion
"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
'It introduces Sirius Black and Remus Lupin who are some of my favorite characters.',	characters	favorite	characters	favorite
"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
"I know it would hurt the plot a lot, but I think Rowling shouldn't have made something so powerful.",			plot	hurt
'His tenderness is apparent, in the fact that unlike most of the characters, he never "shouts," only patiently inquires.'			characters	apparent

- 50% extracted correctly (25 / 50)
- 50% extracted incorrectly (25 / 50)
- Works well on simple sentences but fails on complex sentences and sentences where no sentiment should be extracted
- Improve rule definitions for better accuracy

Sentiment Analysis

Results & Analysis – Sentiment Analysis

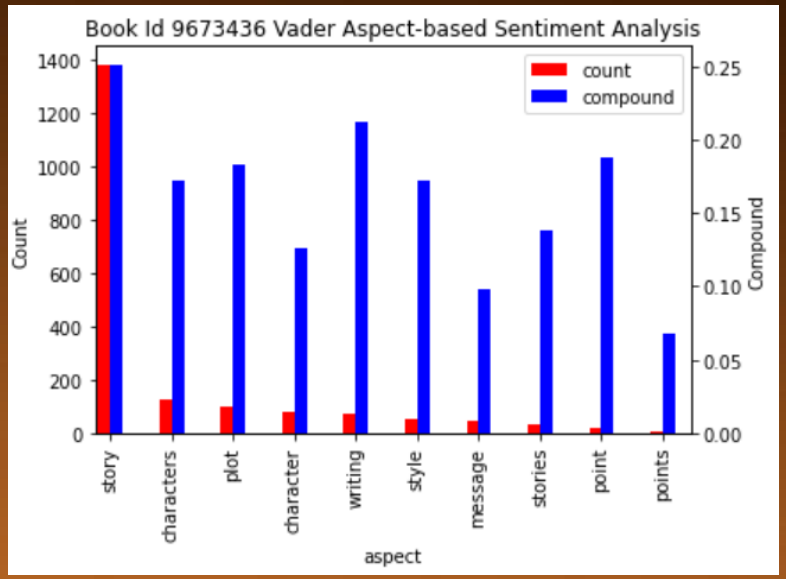


Figure: Top Aspects and Vader sentiment

- Average sentiment for each of the top 10 aspects are neutral (-0.3 to +0.3).
- A possible explanation for the neutral sentiment is the low accuracy from aspect and sentiment extraction. A large number of non-sentiment words may be decreasing the average sentiment.

Results & Analysis – Sentiment Analysis

Confusion Matrix

N=50	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

Observations

		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

- High precision on positive but lower recall. Positive sentiments are getting misclassified as neutral.
- Precision on neutral is very low. Most should be na because the extracted sentiment was erroneous.

Application – Sentiment Analysis

Although the accuracy and results of our Aspect-Based Sentiment Analysis aren't as good as we would've liked, we were able to extract some insights:

1. The most mentioned aspect in children's books reviews is "story". Therefore, publishers might further emphasize the relative importance and evaluation of stories against other aspects in signing book deals.
2. For book id 5, the most mentioned aspect is "character". Thus, for parents who want character driven book for their child, this book might be a good choice.

Challenges and Gap Analysis

Challenges & Gap Analysis

Dataset	Topic Extraction	Document Clustering	Association Mining	Sentiment Analysis
• Very large dataset, need to find a suitable way to filter • Some numbers and URLs present in the reviews • Some reviews written in foreign language • Non-language reviews (e.g. “xyzxyz”)	• Need to further refine stop words list and include frequently occurring words like "book" • Some misclassification due to “noise” (e.g. review making reference to another book)	• Very large and sparse matrix • Not meaningful results on corpus, clustering repeated at book level • High degree of overlap in the clusters • Some misclassified clusters due to “noise”	• Low support and confidence scores given large corpus and document size • Results were affected by “noise” (names of characters were mentioned often)	• Polarity scores are affected by “noise” (factual descriptions of the story) • Can mine only on known features • Difficult to mine implicit features • Difficult to extract features from long sentences

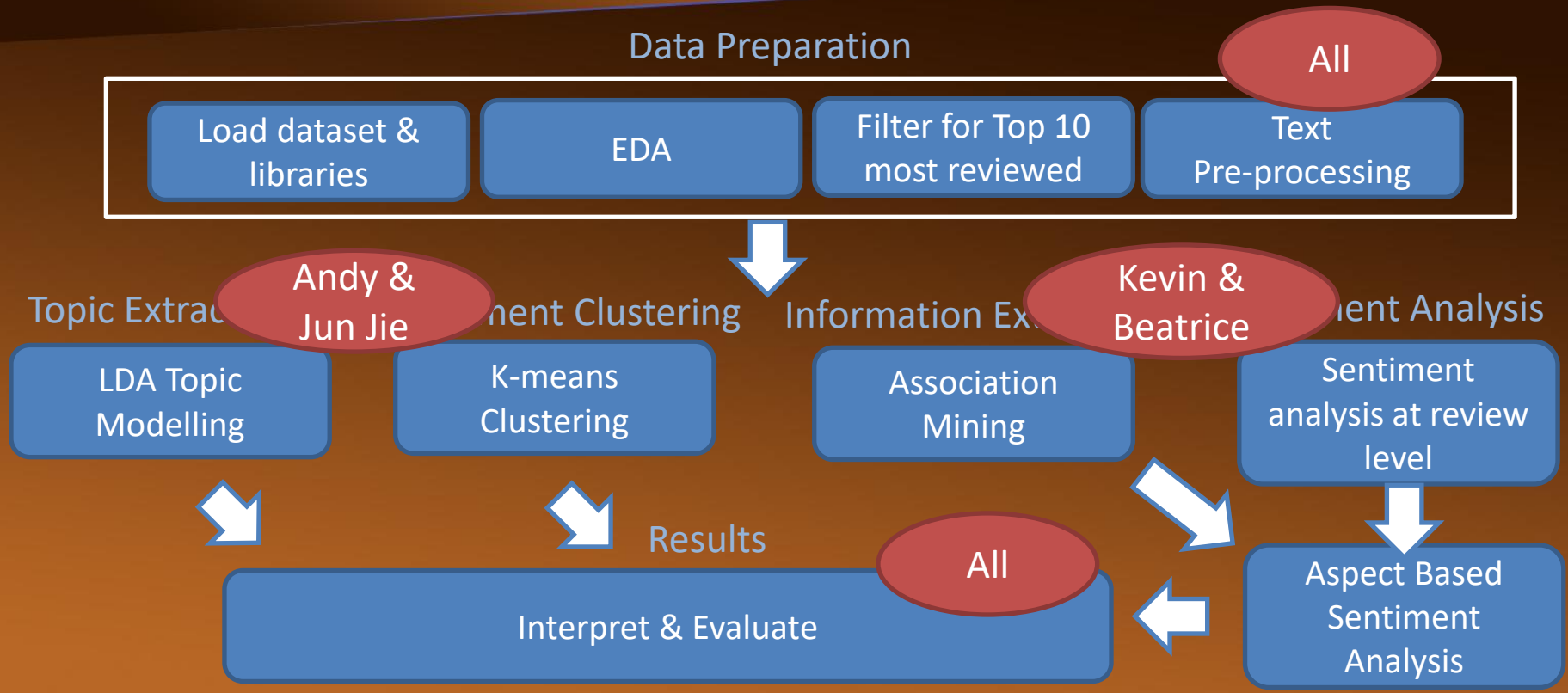
Future work and Conclustion

Future Work & Conclusion

- Long reviews contained “noise”, need a better way to filter “irrelevant” content
- Dictionary for book reviews to improve classification (e.g. author’s name and character’s name)
- Defining better rules to handle complex sentence structure – for extracting aspects for sentiment analysis
- Create a method for extracting implicit features for sentiment analysis

Demo

Team Member's contributions

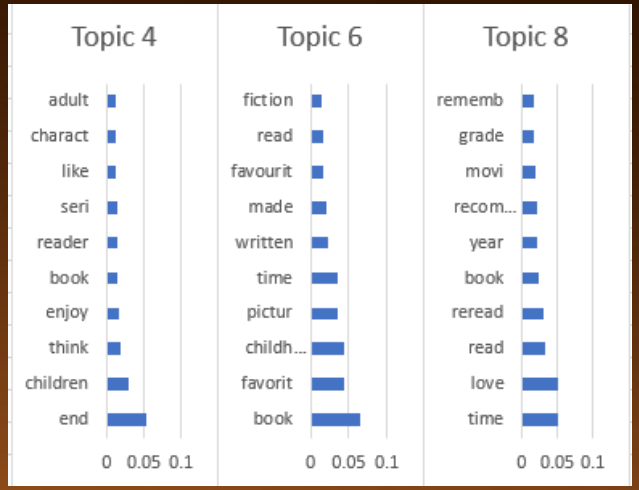


Appendix

Results & Analysis – Topic Extraction

Sample LDA output

- LDA outputs the top *n* words with the corresponding weighting, for each topic. This is visualized as a bar chart.




```
for n_topics in range(init_topic, no_of_topics, topic_step): #generate the 20 LDA models for each book_id
    book_dictionary = gensim.corpora.Dictionary(dataframe['Processed'])
    book_vecs = preprocessLDA.docs2vecs(df_top_books['Processed'], book_dictionary)
    book_lda = gensim.models.ldamodel.LdaModel(corpus=book_vecs, id2word = book_dictionary, num_topics = n_topics)
    topics = book_lda.show_topics(no_of_topics, no_of_words)
    book_topics.append(topics)
```

Solution Details – Overall Sentiment Analysis

Text Pre-processing

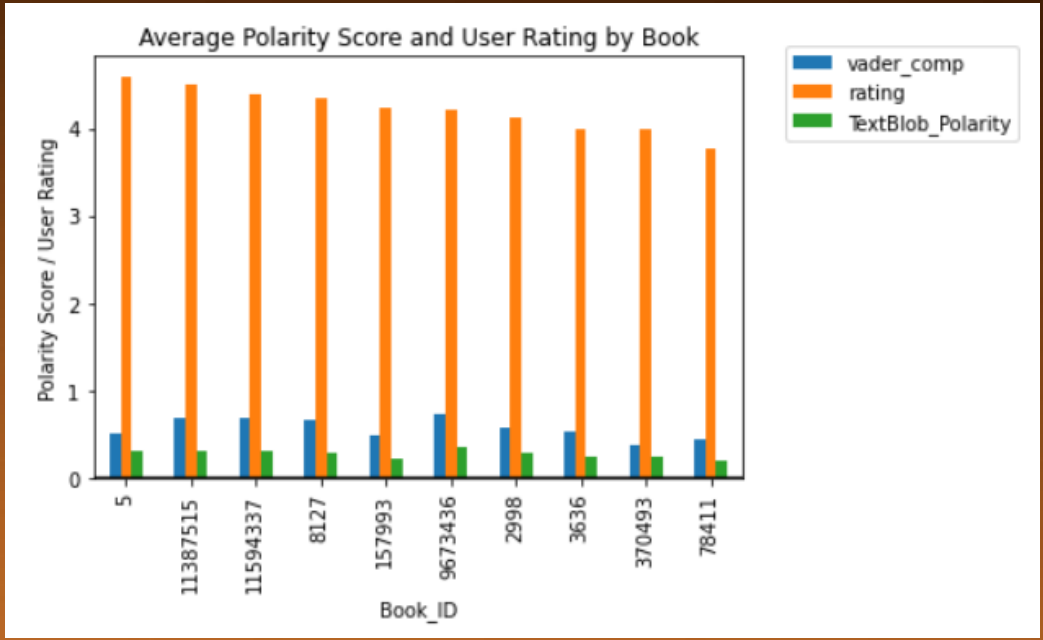
- Removal of URL, hashtags and mentions

Overall Sentiment Analysis

- Polarity Score - TextBlob
- Polarity Score - Vader
- Compare against user rating
- Confusion Matrix – precision, recall, F-score

Results & Analysis – Sentiment Analysis

Vader and TextBlob polarity scores compared with User Rating



"I read this because it's one of my friend's favourite books - it was excellent, but very upsetting as well."

User rating – 4
Vader compound score – 0.2038
TextBlob polarity – 0.5667

"I absolutely loved this book. It is not easy to write about a young person with physical disabilities and allow his voice to come through as a regular kid. This book succeeded beautifully"

User rating – 5
Vader compound score – 0.8803
TextBlob polarity – 0.2389

Solution Details – Aspect-Based Sentiment Analysis

Text Pre-processing

- Sentence and word Tokenization
- Stop word removal
- POS tagging—keep only NN words
- Lemmatization

Steps

Association Mining

↓

Aspect and sentiment extraction

↓

Sentiment Analysis

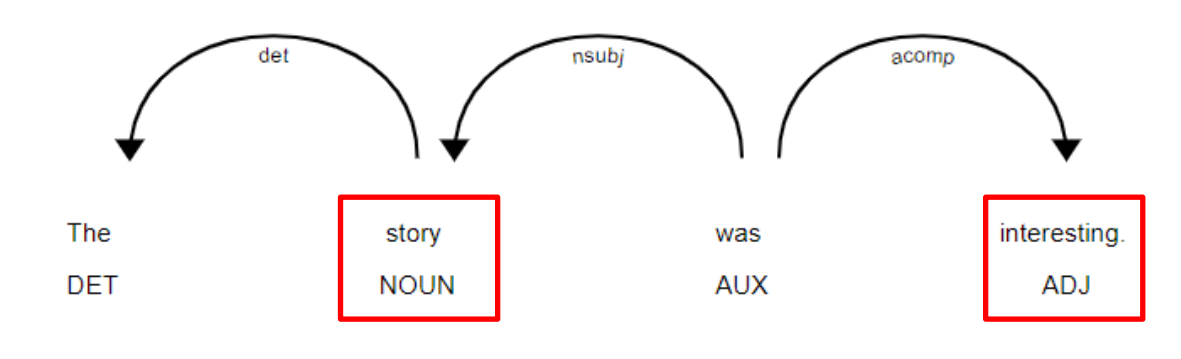
Purpose

What are the most common aspects people talk about in reviews?

What are people saying about aspects in their reviews?

Are the extracted sentiments positive, negative, or neutral?

Solution Detail – Aspect and Sentiment Extraction



- Exploit grammar rules to identify aspects and the words that modify them
- spaCy Dependency Matcher for aspect and sentiment extraction