Text Summarization of Book Reviews on Goodreads

Jessica Fang
Department of Electrical and Computer Engineering
New York University
Brooklyn, NY 11201
jf3135@nyu.edu

Abstract

Goodreads is the largest community of book-readers online. As a review aggregation site in a time when the amount of consumable content available is overwhelming, reception on Goodreads is a significant predictor of a reader's likeliness to read a given book. The task is unsupervised, multi-document, domain-specific, and extractive. The system proposed in this paper aims to use natural language processing to deliver to the user a summary of the reviews for a single book on Goodreads. It creates a feature matrix out of all of the review sentences, then uses a Restricted Boltzmann Machine to score the sentences and use the highest scoring ones for the summary.

Introduction

In the present day, though more written content is being created than ever, readership of books is on the decline. One reason for this is that compared to other industries such as music or television, the book industry has not adopted use of technology nearly as well. Compared to apps such as Spotify or television subscription services, the largest hub for the book industry is the website Goodreads, which allows users to browse books, leave reviews, and socialize with communities.

The reviews on Goodreads are a significant predictor of a reader's likeliness to read a given book. A study in 2015 comparing the reviewer base of the biography genre on both Goodreads and Amazon recorded the total number of reviews on Goodreads at 1.6 million, compared to 0.9 million on Amazon.

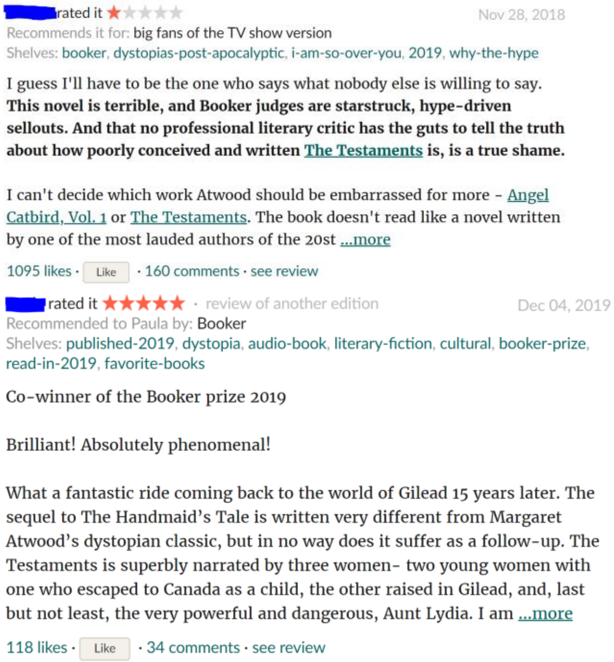
However, Goodreads is widely criticized by its userbase, with an essay by Angela Lashbrook on Medium titled "Almost Everything on Goodreads is Broken" garnering over 10,000 claps. One of the reasons she cites is the site's review system, which has a number of flaws.

1. Prioritization of Reviews

While Goodreads' current algorithm for sorting reviews is undisclosed, based on observation it seems to prioritizes showing reviews with the most likes or comments on them. This often causes the reviews shown to be the most polarizing ones or the ones by reviewers with many followers. This is both a cause and effect of many users using Goodreads as a kind of social media platform, in addition to the site's built-in friend feature and messaging system. While many users enjoy this social element to the website, it can be a detractor to the site's function as a review aggregation website. One of the most popular books of 2019, *The*

Testaments by Margaret Atwood, has a very high average rating of 4.2, yet four out of five of the top reviews are negative, despite only 3% of reviewers rating the book as 1 or 2 stars.

Figure 1



[redacted] rated it ★☆☆☆☆ Nov 28, 2018
Recommends it for: big fans of the TV show version
Shelves: booker, dystopias-post-apocalyptic, i-am-so-over-you, 2019, why-the-hype

I guess I'll have to be the one who says what nobody else is willing to say. **This novel is terrible, and Booker judges are starstruck, hype-driven sellouts. And that no professional literary critic has the guts to tell the truth about how poorly conceived and written The Testaments is, is a true shame.**

I can't decide which work Atwood should be embarrassed for more – Angel Catbird, Vol. 1 or The Testaments. The book doesn't read like a novel written by one of the most lauded authors of the 20st ...more

1095 likes · Like · 160 comments · see review

[redacted] rated it ★★★★★ · review of another edition Dec 04, 2019
Recommended to Paula by: Booker
Shelves: published-2019, dystopia, audio-book, literary-fiction, cultural, booker-prize, read-in-2019, favorite-books

Co-winner of the Booker prize 2019

Brilliant! Absolutely phenomenal!

What a fantastic ride coming back to the world of Gilead 15 years later. The sequel to The Handmaid's Tale is written very different from Margaret Atwood's dystopian classic, but in no way does it suffer as a follow-up. The Testaments is superbly narrated by three women- two young women with one who escaped to Canada as a child, the other raised in Gilead, and, last but not least, the very powerful and dangerous, Aunt Lydia. I am ...more

118 likes · Like · 34 comments · see review

Figure 1: Top 5 star and top 1 star reviews on Goodreads for Margaret Atwood's "The Testaments." The top 1 star review has nearly 10 times the number of likes despite less than 1% of reviewers rating the book with 1 star

2. Lack of Moderation

Goodreads has historically lacked moderation for its reviews. In its own review guidelines they include the statement: "*Harsh critical statements that apply to the book or the writing in it, such as "This guy can't write a lick," or "This book is absolute trash." Again, honest opinions about books are always going to be welcome and encouraged on Goodreads.*"

However, many users have stated that reviewers take this statement too far and they are uncomfortable using the site due to the highly negatively charged reviews. Due to the prioritization algorithm stated above, these reviews are often the most visible. Moreover, reviewers who wish to attract a larger following may be inclined to leave a more polarized review in order to get more likes and comments on their review. This is especially applicable to books that have just come out, or have not even come out, as Goodreads allows users to post reviews before a book's release. Multiple cases have been reported upon by news sites of books being "brigaded" with 1-star reviews before its release as a form of bullying an author. While this issue is separate from the problem of review summarization, it shows how Goodreads chooses not to moderate its reviews.

3. Quality of Reviews

As Goodreads consists of non-professional reviewers, its reviews have a different style and quality from professional book reviews. While professional reviewers use formal language and, most likely having a background in writing or journalism, are able to convey their thoughts succinctly and eloquently, there is a much larger degree of variance with reviewers on Goodreads, who do not have the pressure of writing for a general audience or write to cater to their specific audience. Though there is a sheer quantity of Goodreads review content, much of it is unstructured and has low information density, meaning more text for a prospective reader to comb through in order to make a decision on whether to read a book

Goals

To address this problem, this project aims to use natural language processing to create a summary of the corpus of reviews for a given book. Based on Karen Ball's article "6

Elements of a Good Book Review" four relevant features have been selected:

1. Representation

A review summary should be representative of common experiences among the reviewers. Opinions should be weighed as an aggregate of the reviewer base as a whole, rather than simply the most visible reviewers.

2. Balance

A review summary should be balanced and comment on both positive and critical aspects of the book, to allow the user to make an unbiased judgment.

3. Specificity

A review summary should inform the user of what distinguishes the book from other books beyond simply quality. For example, "a character with complex motivations" is more specific than "a good character."

4. Information Density

A review summary should deliver information succinctly. Moreover, relevant to the task of text summarization, information should not be redundant. That is, sentences chosen should not repeat the same information.

Possible Methodology

Text summarization problems are often defined in three dimensions: input, context, and output. For this problem, the chosen input is multi-document (that is, multiple reviews are being used as input), context is domain-specific (the domain being the context of book reviews), and output is extractive (the summary is generated by pulling sentences from the original text, as opposed to abstractive summarization, which generates new sentences).

The task is unsupervised, as there are no appropriate human-generated summaries that can be used as training data. The following are some of the methods considered for this task.

1. Feature Matrix using Restricted Boltzmann Machines

Using this method, the corpus of reviews is split into sentences and a feature matrix is created out of those sentences, each sentence translating to a vector of features relevant to the task of review summarization, such as inclusion of keywords and position in the review. The feature matrix is then passed through a Restricted Boltzmann Machine (RBM) to abstract it and produce a refined matrix. The sentences can then be ranked based on their total score after going through the RBM, and the desired number of top sentences chosen for the summarization. This method does not account for redundancy. (Verma, Nidhi 2019)

2. Sentence Embedding Clusters

This method involves using sentence embedding to encode the general semantics of each sentence in the review corpus, such as by using Skip-Thought Vectors. Once the sentence embeddings have been generated, the problem can be approached as a classic unsupervised clustering problem. The number of clusters is set to the number of desired sentences using something like a K-nearest neighbors algorithm, and the "central" sentence of each cluster taken to be used in the summary, essentially capturing the main ideas or themes in the review corpus to convey to the user and ensuring low redundancy of information. (Chauhan 2018)

3. TextRank

TextRank is an algorithm that uses a graphing method to generate a similarity matrix. The sentences are embedded into vectors, and then cosine similarity between vectors used to create a similarity matrix. The similarity matrix is used to construct a graph, each node being a sentence, and the TextRank algorithm is run on the graph, generating the top sentences based on the probability of transitioning to that sentence through the graph. This method is able to choose the most important sentences, but does not account for redundancy. (Mihalcea, Tarau 2004)

4. Latent Semantic Analysis

Using Latent Semantic Analysis (LSA) a sentence term matrix is constructed as a term-frequency based representation of the sentences in the corpus. Next, singular value decomposition (SVD) can be done on the document term matrix to model the matrix in lower-dimension vector space. From this lower-dimension space, latent variables can be picked out and the top-scoring sentence for each latent variable chosen. This method accounts for redundancy, but its interpretive power is limited because it does not take in any information about word order or syntactic relations, and it also has performance issues with larger data. (Ozsoy, Nur, Alpaslan 2011)

Method

For this project, the Restricted Boltzmann Machine method was selected. The model has five steps as follows:

1. Text Mining
2. Pre-processing
3. Feature Matrix Generation
4. RBM Training
5. Summarization Generation

Text Mining

The reviews for this project was taken directly from Goodreads.com via a scraping program that retrieves reviews by book ID. The scraper removes all HTML tags, such as formatting and links, but keeps line breaks. Due to the large amount of reviews on Goodreads, only the first 300 reviews are scraped. The scraper then formats the all of the reviews for one book in a .txt file. This will be referred to as the review corpus.

Text Preprocessing

Python’s Natural Language Toolkit (NLTK) was used to preprocess the data.

Regular expressions were used to clean the data, removing images, symbols like emoji or line breaks that do not add to the review, block quotes (many reviews contain samples of quotes from the books or summary; these may be misleading for the summarizer), and text within Goodreads’ spoiler tags (as a review summary should not spoil the book for potential readers).

The review corpus was tokenized first into paragraphs, then into sentences. Stop words were removed using NLTK’s standard English stopwords list.

Word normalization was done by stemming words to their root component with lemmatization (as to simplify different verb tenses, etc.). The Porter’s algorithm was used.

Words were tagged by Part of Speech using NLTK's POS tagger.

Feature Matrix Selection

Thematic Word Frequency: Words in the blurb dataset are weighted by frequency and given a value normalized to the length of the sentence. Words such as "book," "characters," and "author" are expected to have higher values.

Book-specific Word Frequency: Words in the review corpus not in the list of blurb thematic words are weighted by frequency and given a value normalized to the length of the sentence. Words such as the title, author's name, and characters' name are expected to have higher values.

Length: Longer sentences tend to convey more information. For this feature, sentences shorter than five words are given a value of 0, and the length of the sentence otherwise.

First or Last Sentence: A binary value to indicate if a sentence is the first/last sentence of a paragraph or not. Important sentences are more likely to occur at either the beginning or end of a paragraph.

Sentence Position: More important sentences are more likely to occur at the beginning or end of a review. The first and last sentences of the review have a sentence position value of 1, while sentences inbetween have values based on cosine function, with depreciating value toward the middle.

TF-IDF Sum: The Term Frequency Inverse Document Frequency (TF-IDF) values for each word are calculated, with each sentence serving as a document, and summed over the sentence length. TF-IDF is calculated taking the term frequency (frequency of a word in the sentence) multiplied by the log of inverse document frequency (frequency of a word across all sentences). The TF-IDF value is a statistical measure to determine the importance of words in a document.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Restricted Boltzmann Machine

After the feature matrix is generated, it is enhanced and abstracted using a Restricted Boltzmann Machine (RBM). The feature matrix is passed as input through a visible layer and goes through a number of hidden layers. In this case, one hidden layer was chosen. The hidden layer has a bias vector named b with random initialization.

Figure 2

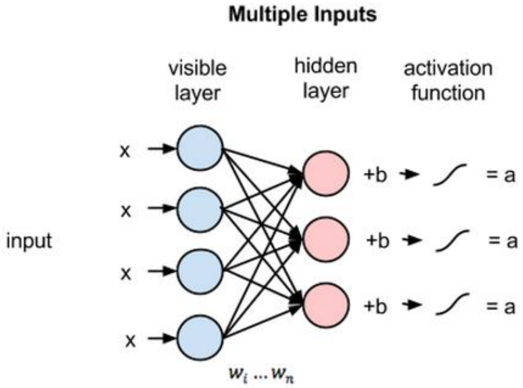


Figure 2: Model of the Restricted Boltzmann Machine

During the training phase, each row of the feature matrix is passed in as input x. The size of x is six, for each feature selected.

$$x = \{x_1, x_2 \dots x_6\}$$

Unlike normal feed-forward networks, the weights of RBMs are not adjusted through backpropagation, instead using something called contrastive divergence. The weight matrix W is randomly initialized and used to generate the hidden nodes h. Then during the reconstruction phase, the hidden layer is used to “predict” the visible layer, and weights are adjusted based on the margin of error.

$$\Delta w_{i,j} = \alpha(\langle x_i h_j \rangle_{data} - \langle x_i h_j \rangle_{model})$$

Where α is the learning rate (set to 0.1) and angle brackets denote the positive contrastive divergence and negative contrastive divergence. That is, the probability that the energy level is over a random threshold.

$$\langle x_i h_j \rangle_{data} = logistic\left(x_i \times w_{i,j}\right) > randn$$

For this RBM, a sigmoid activation function is used. The RBM is trained for 50 epochs.

The final abstracted matrix is obtained from the dot product of the original feature matrix and the trained RBM matrix.

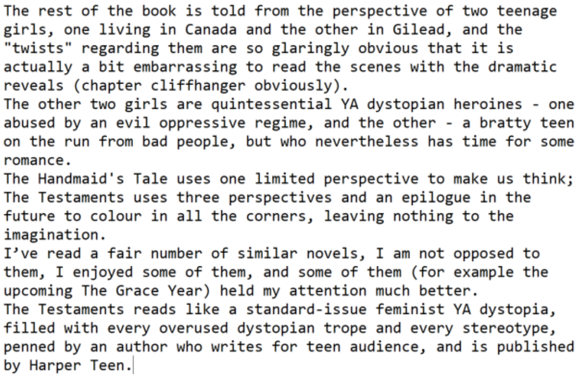
Summarization

To generate the summary, each vector in the enhanced feature matrix is summed up to get a final score. The sentences are sorted by score and the top N sentences are chosen for the summary. In this case, five sentences were chosen. They are arranged in order of their appearance in the review text.

Evaluation

Samples of summaries generated using this model are shown as follows.

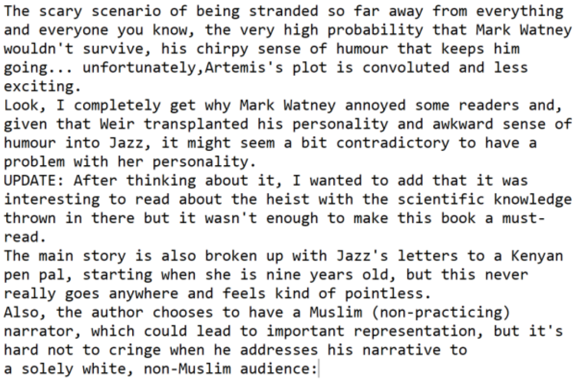
Figure 3



```
The rest of the book is told from the perspective of two teenage
girls, one living in Canada and the other in Gilead, and the
"twists" regarding them are so glaringly obvious that it is
actually a bit embarrassing to read the scenes with the dramatic
reveals (chapter cliffhanger obviously).
The other two girls are quintessential YA dystopian heroines - one
abused by an evil oppressive regime, and the other - a bratty teen
on the run from bad people, but who nevertheless has time for some
romance.
The Handmaid's Tale uses one limited perspective to make us think;
The Testaments uses three perspectives and an epilogue in the
future to colour in all the corners, leaving nothing to the
imagination.
I’ve read a fair number of similar novels, I am not opposed to
them, I enjoyed some of them, and some of them (for example the
upcoming The Grace Year) held my attention much better.
The Testaments reads like a standard-issue feminist YA dystopia,
filled with every overused dystopian trope and every stereotype,
penned by an author who writes for teen audience, and is published
by Harper Teen.
```

Figure 3: Summary generated for Margaret Atwood’s *The Testaments*

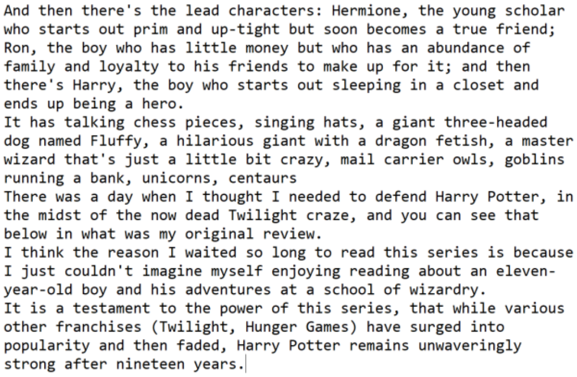
Figure 4



```
The scary scenario of being stranded so far away from everything
and everyone you know, the very high probability that Mark Watney
wouldn't survive, his chirpy sense of humour that keeps him
going... unfortunately,Artemis's plot is convoluted and less
exciting.
Look, I completely get why Mark Watney annoyed some readers and,
given that Weir transplanted his personality and awkward sense of
humour into Jazz, it might seem a bit contradictory to have a
problem with her personality.
UPDATE: After thinking about it, I wanted to add that it was
interesting to read about the heist with the scientific knowledge
thrown in there but it wasn't enough to make this book a must-
read.
The main story is also broken up with Jazz's letters to a Kenyan
pen pal, starting when she is nine years old, but this never
really goes anywhere and feels kind of pointless.
Also, the author chooses to have a Muslim (non-practicing)
narrator, which could lead to important representation, but it's
hard not to cringe when he addresses his narrative to
a solely white, non-Muslim audience:
```

Figure 4: Summary Generated for Andy Weir’s *Artemis*

Figure 5



```
And then there's the lead characters: Hermione, the young scholar
who starts out prim and up-tight but soon becomes a true friend;
Ron, the boy who has little money but who has an abundance of
family and loyalty to his friends to make up for it; and then
there's Harry, the boy who starts out sleeping in a closet and
ends up being a hero.
It has talking chess pieces, singing hats, a giant three-headed
dog named Fluffy, a hilarious giant with a dragon fetish, a master
wizard that's just a little bit crazy, mail carrier owls, goblins
running a bank, unicorns, centaurs
There was a day when I thought I needed to defend Harry Potter, in
the midst of the now dead Twilight craze, and you can see that
below in what was my original review.
I think the reason I waited so long to read this series is because
I just couldn't imagine myself enjoying reading about an eleven-
year-old boy and his adventures at a school of wizardry.
It is a testament to the power of this series, that while various
other franchises (Twilight, Hunger Games) have surged into
popularity and then faded, Harry Potter remains unwaveringly
strong after nineteen years.
```

Figure 5: Summary Generated for J.K. Rowling’s *Harry Potter and the Sorcerer’s Stone*

Ten generated summaries were given to a test audience who were asked, from zero to five out of five sentences, how many sentences they considered helpful to making a decision on whether to read or not read a book, assuming they were undecided before. The books chosen were from a variety of genres.

From ten respondents, the following evaluation metrics were obtained.

$$Recall = \frac{S_{Ret} - S_{Rel}}{S_{Ret}}$$

$$Precision = \frac{S_{Ret} - S_{Rel}}{S_{Rel}}$$

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

Where S_{Ret} is the total number of sentences returned and S_{Rel} is the number of relevant sentences as judged by the respondents. Average S_{Rel} from the respondents was 3.7.

Recall	Precision	F-measure
0.26	0.35	0.30

A qualitative look shows that the summaries that performed the best were summaries the thematic word list was trained on, typical genre fiction such as romance or fantasy, while the model performed poorly on genres such as classic literature and nonfiction.

One notable problem is that often, sentences in the summary appear to lack context. That is, some sentences reference a missing previous sentence or have pronouns that are ambiguous to the user. Additionally, many reviews provide summaries of the book or quotes from the book which are superfluous to the review. These two cases account for most of the false positives reported.

Performance is an issue, and the program takes several minutes to process a large review file. Due to these performance issues, this model may be better suited for summarizing individual reviews rather than a set of multiple reviews.

Future considerations include pronoun replacement to address ambiguity in sentences, and choosing an equal number of positive and negative reviews to pull sentences from in order to ensure a more balanced summary.

Resources

K. Blagec, H. Xu, and A. Agibetov "Neural sentence embedding models for semantic similarity estimation in the biomedical domain." *BMC Bioinformatics*, Is. 20, 2019

K. Chauhan, "Unsupervised Text Summarization using Sentence Embeddings" Medium.com, 2018

S. Dimitrov, F. Zamal, A. Piper, D. Ruths "Goodreads vs. Amazon: The Effect of Decoupling Book Reviewing and Book Selling" 9th International AAAI Conference on Web and Social Media, 2015

M. Farouk, "Measuring Sentences Similarity: A Survey" *Indian Journal of Science and Technology*, Vol. 12, 2019

P. Kouris, G. Alexandridis, A. Stafylopatis "Abstractive Text Summarization Based Semantic Content Generalization" *Proceeding of the 57th Annual Meeting of the Association for Computational Linguistics*, pgs. 5082-5092, 2019

P. Liu, S. Joty, and H. Meng "Fine-grained Opinion Mining with Recurrent Neural Networks and Word Embeddings" *Proceedings of the 2015 Conference on*

Empirical Methods in Natural Language Processing, pgs. 1433-1443, 2015

Y. Liu and M. Lapata "Text Summarization with Pretrained Encoders" *Conference on Empirical Methods in Natural Language Processing*, pgs. 3721-3731, 2019

A. Londhe, Y. Deshpande, G. Ghongde, P. Deshmukh, H. Gate "Product Review Summarization With Feature Extraction and Opinion Mining," *International Journal of Scientific Research Engineering & Technology*, Vol. 8, Is. 3, pgs 174-177, 2019

R. Mihalcea, P. Tarau "TextRank: Bringing Order into Text" *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004

B. Ojokoh and O. Kayode "A Feature-Opinion Extraction Approach to Opinion Mining" *Journal of Web Engineering*, Vol. 11, No. 1, 2012

M. Ozsoy, F. Nur, I. Cicekli "Text Summarization using Latent Semantic Analysis" *Journal of Information Science 37(4)*, pg. 405-417, 2011

B. Pang and L. Lee "Opinion Mining and Sentiment Analysis" *Foundations and Trends in Information Retrieval*, Vol. 2, No. 1-2, 2008

A. Popescu and O. Etzioni "Extracting Product Features and Opinions from Reviews," *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, 2005

J. Steinberger, K. Jezek "Evaluation Measures for Text Summarization," *Computing and Informatics*, Vol. 28, 2009

S. Verma and V. Nidhi "Extractive Summarization using Machine Learning" *18th International Conference on Computational Linguistics and Intelligent Text Processing*, 2017