Book Reviews Analysis with Text Clustering

Wendy Mo
CS 571: NLP
Emory University
Atlanta, GA, 30322, USA
wendy.mo@emory.edu

Yao Ge
CS 571: NLP
Emory University
Atlanta, GA, 30322, USA
ge.yao@emory.edu

Abstract

A book review is a description, critical analysis, and evaluation of the quality, meaning, and significance of a book. In this project, we conducted the content analysis on the book reviews written by both the professional book reviewers and the Amazon book consumers. We clustered the content of the book reviews at sentence level into eight clusters and compared the content and the distribution of the sentences written by different groups of people. We considered Fine-tuned BERT plus K-Means as the best model for text clustering on our dataset, and then conducted the content analysis for corresponding books based on the cluster generated by Fine-tuned BERT and K-Means. After the experiment, we concluded that the content of reviews for the same book is very similar among the reviews, which leads to the performance on the text clustering tasks. Meanwhile, the content between the professional book reviews and the consumer book reviews is also similar to each other.

1 Introduction

As we entered the Internet era, especially during the tough time of the Covid-19 pandemic, online shopping has become an inevitable part of life. As people browse products online, one important decision factor is the reviews written by previous consumers. We believe that the quality of the reviews would significantly affect the overall online purchasing experience. In this project, we particularly focused on book reviews. Between the two datasets, New York Times most popular book reviews from 2013 to 2018, and the Amazon book reviews, we consider the previous one as the long and professional book reviews that are good guide the consumers, and the latter one as the Amazon reviews which we would like to question the quality of the content. After the experiment, we are able to tell which category of the content is the most important to a high quality review, and how are the consumers reviews comparing to the professional reviews.

Text clustering has been well studied in the Natural Language Processing field. However, to the best of our knowledge, rare previous work was conducted on book reviews and aimed at analyzing book review content. In our work, we looked into two datasets, New York Times book reviews and Amazon book reviews, where the previous one contains 116 book reviews written by professional book reviewers, and the latter one consists 9.46 gigabytes book reviews written by consumers till 2014. The main challenges in our work are mostly related with dealing with the datasets. Since the two raw datasets are not studied and unlabeled, we made considerable effort on annotation, matching, and figuring out the reasonable approaches for comparing.

We hypothesized that the professional book review is the standard of a high-quality book review. And we assumed that the quality of a book review is directly related with the contents it mentions. Therefore, the higher ratio of overlapping with the content of the professional review leads to the consumer reviews with better quality. To make sure the comparison between reviews is reasonable and convincing, we paird up the reviews from two datasets by the name of the book. Among 116 books in the New York Times popular book reviews, we found 30 of them also being mentioned in Amazon book reviews. And meanwhile, the size of the dataset for Amazon book reviews also shrinks to a handle-able size.

We manually annotated 20% of the professional reviews at sentence level into eight categories – Information of the book, Theme of the book, Author's writing style, Author's life, Character Description, Summary of the story, Reviewer's own feeling, and Reviewer's criticism/praise. And then we fine-tuned BERT model based on these anno-

tated data, and implemented K-Means, Birch and DBSCAN as the clustering methods on the rest of the professional reviews dataset. Finally, we clustered the Amazon reviews based on the cosine similarity with each professional reviews cluster.

This work implements the existing natural language processing technique to solve emerging questions – how reliable the online book reviews are, and how to write a high-quality online book review. The approach we proposed in this work for book reviews can also be implemented to other online reviews with moderate modification. With a scientific guideline, we will be able to have higher review quality and better online purchasing experience.

In summary, the contributions of the paper are as following:

- This is one of few papers to study the book review content with the text clustering technology.
- We are the first to match up the New York Times book review data with the Amazon book review data for comparing and contrast intention.
- We compared the performance of several Transformers-based models on different clustering algorithms.
- And we claim that consumer book reviews share large portion similar content with the professional book reviews.

2 Related Work

2.1 Text Representation

Before amount of Transformers-based models (Wolf et al., 2019) was proposed, previous work mainly focused on the use of neural networks for text clustering. For example, embed words into a convolutional neural network to learn deep feature representation (Jiaming Xu and Hao, 2015), or introduce a BiLSTM-CNN layer to grasp text semantics (Yang Fan and Zhaoying, 2018). However, when Transformers-based models such as BERT (Reimers et al., 2019) and Sentence-BERT (Reimers and Gurevych, 2019) were proposed, it turns out that the embeddings of the word integrates the information of all words, and can better express its own semantics. Therefore, the method of using neural networks to extract text embeddings is gradually abandoned.

Making use of BERT, the embeddings to further feed into text clustering algorithms can be extracted from different layers of BERT. While comparing to the previous BERT Embeddings, Sentence-BERT greatly boost the efficiency by enabling the fixed-sized vectors for input sentences can be derived. At the same time, the use of pre-trained Transformers-based models for fine-tuning has also achieved state-of-the-art results in many language understanding tasks under supervised settings (such as text classification). However, relatively few researches have focused on applying pre-trained models, such as text clustering, in unsupervised environments. Inspired by these, we conducted BERT, Sentence-BERT, and Fine-tuned BERT to implement text clustering, and compared their performance.

2.2 Text Clustering

Two main text clustering methods are agglomerative hierarchical clustering and K-means respectively (Steinbach et al., 2000). The Hierarchical clustering presents the result in a multi-level tree-like structure(Zhao and Karypis, 2005) while the K-means (the partition clustering) clusters the document in a single level(Larsen B, 1999). Hierarchical clustering method generally provides the result with better quality while the K-means can cluster in the linear time which much faster than the quadratic time complexity by Hierarchical clustering.

Throughout these years, many researches implemented K-means methods for document clustering. So we tried to compared K-means with Birch (Zhang, 1996), an unsupervised data mining algorithm used to perform hierarchical clustering, and DBSCAN (Density-based spatial clustering of applications with noise) (Kamran Khan, 2014), thus trying to find the advantages of K-means in text clustering.

In addition, although many researches based on text clustering, very few studies are based on book reviews, especially book reviews written by professional book reviewers. This type of reviews' language is very beautiful, and the meanings expressed between sentences complement each other, which means the boundaries between text sentences are relatively unclear. It not only states the plot of the story, but also expresses the reviewer's own feelings, perceptions, and understanding of this book. However, a good book review will attract countless readers to be interested in the book. Therefore, we consider to find which aspects of the book the reviewers pay more attention to, and whether these

are consistent with the concerns of readers in the market. This is an innovation of our work and one of our big challenges.

3 Approach

3.1 Annotation

Book reviews written by professional book reviewers are unlabeled data, which has caused great difficulties for our research. We hope to use the labeled data for some important follow-up experiments, such as fine-tuning on BERT, or using them as true labels in the evaluation process to measure each model. As a result, we started our research from annotation.

We annotated 20 book reviews written by professional book reviewers in sentence level, with a total of 1083 sentences. It is worth mentioning that since all the books are from the New York Times most popular books, most of the books are in the genre of novel, and this common point provides some help for our annotating.

The annotation process is roughly as follows: First, each annotator was given 5 identical book reviews, and the annotators began to summarize the meaning and purpose of each sentence in the form of phrases, such as "tell the origin of the book title", "compare the author's earlier novels ", "Introduced the main character in the book", "expressed the reviewer's own evaluation of the book", etc. Then through discussion, we reached a consensus and determined to divide the sentences in the professional book reviews into 8 categories, namely:

- Information of the book (including book title, publisher, and price)
- Theme of the book (including the background of book, what the author wants to convey through the book)
- Author's writing style (including the sentences mentioned the comparison to other authors, the comparison with author's previous works, and the author's writing techniques)
- Author's life (the author's own life experience)
- Characters or description of characters
- Summary of the story or experience of characters
- Reviewer's own feeling, analysis, and own thoughts on this book
- Reviewer's criticism or praise on this book

Then, based on these 8 categories, we labeled all 20 book reviews (including the 5 book reviews that have been preliminary analyzed). After annotating, we calculated our IAA (inter-annotator agreement) score based on Cohen's kappa (Fleiss and Cohen, 1973), a statistic that measures IAA. The reason for choosing it is that Cohen's kappa has just two annotators, each annotator annotates every item, which is consistent with our annotation situation. The calculation formula of Cohen's kappa is as follows:

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \quad (1)$$

where p_o is the relative observed agreement among raters, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. If the raters are in complete agreement then $\kappa = 1$. After calculation, our initial IAA score was only 0.535. After three rounds of discussion and modification, we finally reached a consensus, making $\kappa = 1$. The distribution of 1083 sentences of 20 book reviews in 8 categories is shown in Table 1.

3.2 Baseline Model

We tried to use the combination of TF-IDF methods for weighting and K-Means algorithm for clustering as our baseline model. TF-IDF is a numerical statistic that demonstrates how important a word is to a corpus. Starting from the professional reviews written by professional book reviewers, the Vector Space Model was used to store the word frequency and weight of each document, and TF-IDF was used to calculate the weight of the term in the vector.

Since each professional review has only a few paragraphs, we implemented K-Means clustering algorithm on the sentences of all book reviews. However, TF-IDF simply calculates word frequency without considering semantic information, and the similarity between words in our professional reviews dataset is very high, thus using it for clustering is indeed not a good choice, and we will discuss it in Section 4.5.1.

Information of the book	Theme of the book	Author's writing style	Author's life	Characters	Summary of the story	Reviewer's own feeling / thoughts	Reviewer's criticism / praise
25	37	59	19	74	**386**	131	50

Table 1: The distribution for 20 book reviews in 8 categories. The most distributed categories have been marked in bold.

3.3 Transformers-based Models for Sentence Embeddings

3.3.1 BERT and Fine-tuned BERT

Encouraged by the success of Transformers-based Models, we tried to use the embeddings extracted by BERT to implement text clustering, then fed the sentence vectors into different clustering algorithms.

In this paper, the following two methods were used as the representations of sentence embeddings:

- CLS: Take the vector at the first position of the feature as the sentence embeddings, in other words, use the vector of [CLS] as the sentence embeddings.
- Mean: Take the average of all features as the sentence embeddings.

In order to receive better results, we fine-tuned on BERT using half of the labeled data (500 sentences), and used the fine-tuned model to extract sentence vectors. As with the original BERT model, we also considered two methods (CLS and Mean) to extract features.

3.3.2 Sentence-BERT

However, similar to other sentence embeddings methods, BERT does not consider the similarity between sentences; therefore, we tried Sentence-BERT, which uses twin network and triplets network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity.

Three pooling strategies were adopted in the experiment of Sentence-BERT for comparison: 1.Directly use the output vector of the CLS position to represent the vector representation of the entire sentence; 2. MEAN strategy, calculate the average value of each token output vector to represent the sentence vector; 3. MAX strategy, take the maximum value of each dimension of all output vectors to represent the sentence vector. In conclusion, the MEAN strategy is the best one.

For the professional reviews, sentences are not independent, but with certain connections, either a parallel relationship, a progressive relationship, or a transition relationship. Therefore, we hope that Sentence-BERT can better grasp the relevance and similarity between sentences, so as to better cluster the sentences. In addition, because Sentence-BERT itself has considered three ways of generating sentence vectors, CLS, MEAN, and Max, we no longer need to test and select redundantly.

3.4 Clustering Algorithm

We implemented three clustering algorithms, K-Means clustering, Birch clustering and DBSCAN clustering. The advantage of K-Means is that it is very fast, but the shortcomings are that, we need to specify the number of clusters to be classified, and the selection of the cluster centroids are random at the beginning. In contrast, both Birch and DBSCAN do not require one to specify the number of clusters in the data a priori.

The advantage of Birch is its ability to incrementally and dynamically cluster incoming, multi-dimensional metric data points in an attempt to produce the best quality clustering for a given set of resources (memory and time constraints) [1].

The advantages of DBSCAN are, first it can find arbitrarily-shaped clusters, even find a cluster completely surrounded by (but not connected to) a different cluster. Second, DBSCAN requires just two parameters and is mostly insensitive to the ordering of the points in the database [2].

3.5 Mapping Amazon Book Reviews to Professional Book Reviews

After getting the classification, we wanted to determine that each short review belongs to one or several categories. Therefore, we tested each review in the Amazon reviews dataset as an input to calculate its embeddings distance to the cluster centroids. We simply chose the cluster with the closest distance to the review as its cluster. To calculate

[1] Birch page from Wikipedia.
[2] DBSCAN page from Wikipedia.

the distance between cluster centroids and the review, we implemented both Jaccard Similarity and Cosine Similarity based on TF-IDF vectors.
Jaccard Similarity:

$$J(A,B) = \frac{A \bigcap B}{A \bigcup B} \tag{2}$$

Cosine Similarity:

$$cos(\theta) = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sum_{i=1}^{n} B_i^2} \tag{3}$$

3.6 Clustering or Classification

While clustering, we also adjusted the number of clusters to 8, hoping to match of the 8 classes we annotated to the maximum extent. If the text is distinguished clearly, and the clustering algorithm itself is very robust, then naturally each cluster should correspond to only one real class. However, the reality is not the case. Taking into account the elegance of text language from professional reviews, the mutual inheritance between sentences, and the difficulty of clustering, the 8 clusters obtained after implementing clustering algorithms cannot correspond to the real 8 classes, instead, each cluster contains scattered sentences of every class, which is almost an inevitable problem.

Faced with this challenge, we used the Fine-tuned BERT model to directly perform classification tests on unlabeled data. We divided the labeled data into training set, validation set, and test set, and only used the training set for fine-tuning, then tested on the test set. The classification accuracy is as high as 98%. Although over-fitting may occur, we still believe that the Fine-tuned BERT model can complete the classification problem well. Therefore, although we did not label the remaining data and could not give results based on evaluation metrics, we still analyzed and compared the results of the classification based on the empirical experience of annotation. However, due the time constraint, we are unable to perform Amazon reviews clustering based on the classification result. And we will leave it in the future work.

4 Experiments

4.1 Data

The two datasets being studied in this paper are New York Times most popular book reviews from 2013 to 2018, and Amazon book reviews till 2014. In New York Times most popular book reviews (professional reviews), there are 116 documents, 136,036 words. Each of the document in this dataset corresponds to one different book. Therefore, there are 116 books being covered by New York Times most popular book reviews. The Amazon book review was released till 2014, and there are 9.46 gigabytes data, including 8.9 million reviews. Despite the review text, it also includes other features, for instance "asin", "helpfulness", and "summary", etc. By looking for the 116 book titles in the Amazon metadata, we located 36 books that appear in the metadata with the "asin" number. And then, among 36 "asin" numbers, we successfully paired up 30 books in Amazon book reviews, and obtained 29,395 reviews in total.

Since the New York Times book reviews were written as documents with more than 1,000 words in each document on average, we processed these professional reviews by sentence level. While for the Amazon book reviews, they are generally shorter and on average 3 sentences for each review. Therefore, we kept the Amazon reviews as one review per unit.

4.2 Evaluation Metrics

1. Intra-cluster Metrics

We used Calinski-Harabasz, Silhouette Coefficient, Davies-Bouldin Index to evaluate our clustering models, since they are the common metrics to evaluate the unlabeled clustering algorithms.

Calinski-Harabasz mainly calculates the ratio of the distance between clusters to the distance within clusters. When the clusters are dense and the separation between clusters is better, the higher the Caliniski-Harabaz score, the better the clustering performance. The Silhouette Coefficient is in the range of [-1,1], where -1 means wrong clustering, 1 means high-density clustering, and near 0 means overlapping clustering. When the cluster density is higher and the separation is larger, the Silhouette Coefficient of the cluster is also larger. DB index calculation is similar but simpler than Silhouette Coefficient.

2. Inter-cluster Metrics

We also used ARI (Adjusted Rand Index), AMI (Adjusted Mutual Information), NMI (Normalized Mutual Information), MI (Standardized Mutual Information), Homogeneity, Completeness, Purity and Entropy to assist evaluating our clustering models. These inter-cluster metrics mainly detect

whether pairs of sentences whose true labels are in the same class are also clustered in the same cluster.

The characteristics of the above inter-cluster metrics are similar. The larger the value, the higher the similarity between the predicted cluster embeddings and the real class embeddings. The values close to 0 indicate that clusters are randomly allocated, and the negative values indicate a very poor predicted clustering.

4.3 Hyper-parameters

For K-Means clustering, after annotation, we decided to divide the professional reviews into 8 classes. In order to maximize matching, we also set the number of K-Means clusters to 8, and other parameters keeping the default settings. For Birch clustering, we set branching factor to 10 and threshold to 0.5, where branching fact is maximum number of CF subclusters in each node, and threshold is the radius of the subcluster obtained by merging a new sample. For DBSCAN clustering, we selected Euclidean metric to calculate distances between instances in a feature array.

When using BERT and Sentence-BERT to extract sentence vectors, we kept their original parameter settings. When performing fine-tuning, we chose to use half of the professional reviews (500 sentences) for fine-tuning, and divided the data into train, validation, and test sets according to the ratio of 7:1.5:1.5. In addition, we set batch size to 32, training epoch to 10, and chose AdamW as the optimizer, LLLoss as the loss function.

4.4 Set up

After we obtained the best possible clustering model, we implemented the clustering technique to analyze two kinds of reviews by looking into 30 overlapping books.

Firstly, we clustered all 30 professional reviews one at a time with the trained Fine-tuned BERT and k-mean clustering method. Figure 1 shows the proportion of each cluster in each document, where the x axis indicates the cluster index, and the y axis indicates the percentage of the document, and each line represents one document.

With the clustered professional reviews, we clustered the Amazon reviews based on the extent of similarity. We calculated the similarity between each short review and the mean of each long review cluster. And then, we clustered that short review into the same cluster id with the highest similarity score. To calculate the similarity score, we implemented both simple Jaccard Similarity and Cosine Similarity based on TF-IDF vector.

For both the Jaccard Similarity and Cosine Similarity clustering result, we grouped by the cluster id to obtain the cluster proportions on Amazon reviews, and also the average similarity score for both similarity calculation methods on each cluster.

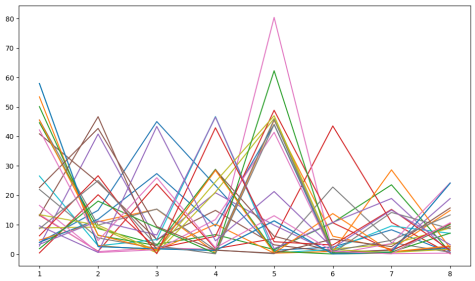

Figure 1: Cluster proportion of the New York Times book reviews

4.5 Results

4.5.1 Results on Clustering

Table 3 in Appendix indicates the intra-cluster results on professional reviews. We compared the performance on six methods of extracting embeddings with three clustering algorithms. Not surprisedly, the performance of baseline is terrible, since it only extracts word frequency information and lacks semantic information. Two methods of Fine-tuned BERT outperformed others, whether CLS or MEAN for extracting sentence vectors. Also, performance of these two methods is similar, with MEAN strategy to be slightly better. Sentence-BERT is not as good as we expected, and it is even worse than using BERT directly. In addition, although the values of DBSCAN seem good, there are only two clusters with sentences, other clusters are empty, which means this algorithm is not suitable for our professional reviews dataset.

Table 4 and Table 5 in Appendix both indicate the inter-cluster results on professional reviews, the difference is that Table 5 shows the results of Purity and Entropy. As mentioned in Section 4.2, for inter-cluster metrics, the value close to 0 indicates that clusters are randomly allocated. The results in Table4 show that values of almost all models are near 0, which fully proves that our clustering methods on professional reviews dataset are not good.

Cluster Index	Professional Reviews (%)	Amazon Reviews Jaccard Similarity Cluster	Amazon Reviews Jaccard Similarity	Amazon Reviews Cosine Similarity Cluster	Amazon Reviews Cosine Similarity
0	22.66	8.35	0.05	21.48	0.48
1	13.18	9.26	0.06	14.18	0.47
2	9.37	12.47	0.06	9.42	0.40
3	13.91	13.60	0.06	12.94	0.42
4	21.40	11.61	0.06	20.84	0.48
5	5.14	16.39	0.05	3.84	0.31
6	6.23	14.21	0.06	7.06	0.40
7	8.10	14.10	0.06	10.24	0.41

Table 2: The final comparing result between New York Times book reviews and Amazon book reviews by cluster.

In Table 5, we can see that each cluster cannot represent a certain class very well, instead, it presents a relatively scattered distribution. The purity and entropy of the 8 clusters are also very close, with no obvious difference. In addition, the noise in each cluster is relatively high, while the purity is relatively low.

4.5.2 Comparison Between Two Book Reviews

Table 2 shows the final comparing result between New York Times book reviews and the Amazon book reviews. On average, both cluster 0 and cluster 4 occupy about 20% of the entire text, and cluster1 and cluster3 occupy about 10% of the text for the professional reviews. As for the Amazon reviews, two similarity score calculation approaches generate significantly different results. While Jaccard Similarity yield relatively uniform distribution and extremely low similarity score with the long review, Cosine Similarity yields much more promising result. Same as the distribution of the professional reviews, the clustering for Amazon reviews based on Cosine Similarity scores also have cluster 0 and cluster 4 with about 20% on average and cluster1 and cluster 3 with about 10% on average. And the average similarity score for each cluster ranges from 31% to 48%. Figure 2 shows the absolute distance of the proportion for each cluster and for every document. A large portion of the clusters have the absolute distance around 0% percent, and the majority of the clusters have the absolute distance smaller than 20%. The ones that have the absolute distance larger than 20% are basically the outliers.

Therefore, we can conclude that, with a reasonable similarity calculation method, we can indeed perform pretty well clustering for Amazon reviews. And we can also project that, if a more precise word embeddings is performed on the datasets, there is the probability that we can achieve a more promising outcome.

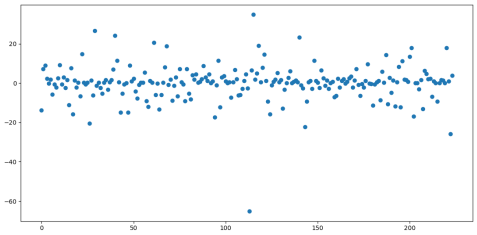

Figure 2: Absolute difference between the proportion of short review clustering and long review clustering

4.5.3 Results on Classification

Figure 3 in Appendix is a sample text of one of the classes after using Fine-tuned BERT to directly perform on classification task. This sample shows that in this class, the content of most sentences is concentrated on the book title, publisher, and price, which fits one of our categories very well. Although for the classification problem of unlabeled data, we do not have appropriate evaluation metrics to measure the performance of model, we can still use the classification results to reasonably guess that the results obtained by using the model for classification task are far better than clustering task.

4.6 Analysis

4.6.1 Analysis on Clustering and Classification

As we compared several ways to extract embeddings, the advantage of Fine-tuned BERT is obvious. Since it has gained more in-domain knowledge, it is able to complete the clustering task relatively well. However, all the clustering results show that the clusters we received have no clear boundaries between each other, and much noise

inside the clusters. In other words, the purity is relatively low.

One of the possible reason is that K-Means does not perform well in classifying very close clusters, which is precisely the case in our professional reviews dataset. Another reason is that we first determined 8 classes based on the annotations, but these categories are general. For example, the category "information of the books" includes basic information such as book title, the publisher of the book, and the price. However, semantically they are very different, and perhaps their embeddings are even not close. This kind of general classification given by us subjectively also leads to the effect of clustering that is not what we hoped.

However, BERT itself is a state-of-the-art model on many NLP tasks, including the classification tasks. Especially in the case where we have used half of the dataset for fine-tuning, so its classification results are very ideal.

4.6.2 Analysis on Book Reviews Comparison

As we look closer into the exact clusters with the highest similarity scores, there are several interesting pattern shown. Basically, there are two kinds of clusters that have the biggest chance to have a high similarity scores. The first kind of the clusters are the ones account for the highest proportion of the text. For instance, in the book "And the Mountain Echoed", cluster 3 occupy nearly 70% of the reviews, and its similarity score is also the highest among all the clusters. The other kind of the clusters is quite the opposite, which are the ones account for the lowest proportion of the text. For the book "And the Mountain Echoed", cluster 5 only accounts for less than 1%, while also has high similarity score. One reasonable guess for such situation is that, for the cluster with very few reviews, as long as the character's name and major locations are frequently mentioned, then the similarity score is very likely to be relatively high. Meanwhile, for the large clusters,

5 Conclusion

5.1 Summarize

The experiment result on book reviews shows that the content of reviews are very close to each other, which leads to relatively poor clustering result. Meanwhile, the comparison between professional reviews and Amazon reviews shows that the cluster distribution are very similar between one another, and the similarity scores based on TF-IDF Cosine Similarity are promising. The result strongly indicates that the professional reviews and Amazon reviews share a large portion of the contents. Although the result is not as we predicted ahead of the experiment, it is a good news for the book online consumers for having high quality online reviews.

5.2 Broader Impact

As this project only focuses on book reviews clustering and analysis, the same approach with few modification can be easily applied to the study on other reviews. As we notice on Amazon dataset, many other commodity reviews are available online. As long as there are professional reviews available as the standard, the method proposed is highly reproducible.

5.3 Future Work

5.3.1 Research on "Helpfulness"

The label "helpful" in Amazon dataset, represents the helpfulness rating of the review, e.g. 2/3. In other words, three people have read this review and two of them thought this was helpful.

In this project, we have already concluded that the Amazon reviews and the professional reviews, on average, have a large proportion of overlapping. However, we have not figured out the relation between the overlapping the the helpfulness. In the future work, we will conduct the experiment, for instance, the logistic regression analysis, to see whether the high proportion of the overlapping would lead to the higher rating in terms of the helpfulness. We will start with calculating the helpfulness rating for the reviews have more than one person viewed (denominator > 1) for the sake of relative objectiveness.

5.3.2 From Clustering to Classification

Since the original datasets for this project are unlabeled – there is no categorical label for either sentence or paragraph or reviews, we spent a lot of time on manual annotations, and were only able to perform Fine-Tuned BERT on classification at the very last minute. Therefore, our next step would be classifying the Amazon reviews based on the long review classification rather than clustering, and perform more meaning content analysis. Also, we would like to try other classification methods, for instance classifying by calculating the distance with the manually labeled centroid, to compare the performance.

References

Joseph Fleiss and Jacob Cohen. 1973. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. volume 33, pages 613–619.

Guanhua Tian Bo Xu Jun Zhao Fangyuan Wang Jiaming Xu, Peng Wang and Hongwei Hao. 2015. Short Text Clustering via Convolutional Neural Networks. pages 62–69.

Kamran Aziz Simon Fong S. Sarasvady Kamran Khan, Saif Ur Rehman. 2014. DBSCAN: Past, present and future. pages 232–238.

Aone C Larsen B. 1999. Fast and effective text mining using linear-time document clustering.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *CoRR*, abs/1906.09821.

M. Steinbach, G. Karypis, and V. Kumar. 2000. A Comparison of Document Clustering Techniques. In *Technical Report*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. volume 1910.

Meng Kui Yang Fan, Liu Gongshen and Sun Zhaoying. 2018. Neural Feedback Text Clustering With BiLSTM-CNN-Kmeans. pages 57460–57469.

Ramakrishnan Raghu Livny Miron Zhang, Tian. 1996. BIRCH: an efficient data clustering method for very large databases. volume 25, pages 103–114.

Y Zhao and G Karypis. 2005. Hierarchical clustering algorithms for document datasets. volume 10.

A Appendix

'THE WHISTLERBy John Grisham374 pp.',
'Doubleday.',
'$28.95.',
'DARK MATTERBy Blake Crouch342 pp.',
'Crown.', '$26.99.',
'Crouch's Wayward Pines trilogy became the basis for a Fox television series whose pilot episode was directed by M.',
'Under deeper cover, it might also be a fantasy novel shaped by C.S.',
'Le Guin, Neil Gaiman, Peter F.',
'AN AMERICAN MARRIAGE By Tayari Jones 306 pp.',
'Algonquin Books of Chapel Hill.',
'$26.95.',
'259 pp.',
'Little, Brown & Company.',
'$27.',
'8th.'",
'Try it tonight.',
'Breaded steakfish.',
'Later, Eleanor says to Alonzo, "See you next week?', 'Dalloway."',
'THE RULES OF MAGICBy Alice Hoffman367 pp.',
'Simon & Schuster.',
'$27.99.',
'Was that how the prodigal son felt?',
'Vernon, N.Y., on Fiske Place, directly above the train tracks.',
'isn't returning his calls.',
'The grandson of Robert E.',
'A HERO OF FRANCEBy Alan Furst234 pp.',
'Random House.',

Figure 3: Sample text of one of the classes after using Fune-Tuned Bert.

	K-Means			Birch			DBSCAN		
	CH	**Silhouette**	**DB**	**CH**	**Silhouette**	**DB**	**CH**	**Silhouette**	**DB**
(Baseline) TF-IDF	12.937	0.0041	10.792	5.471	-0.0038	14.783	1.898	0.0107	5.105
BERT_CLS	132.48	0.0324	3.9697	109.2	**0.0234**	3.8139	50.96	0.1478	2.5619
BERT_MEAN	115.67	-0.005	4.2239	86.65	-0.0145	5.6156	97.67	0.3065	2.1971
Fine_Tuned_BERT_CLS	**346.15**	0.0281	3.6245	**321.6**	0.0067	4.0500	139.8	0.3408	**1.6128**
Fine_Tuned_BERT_MEAN	328.01	**0.0400**	**3.5319**	307.6	0.0230	**3.7090**	**210.8**	**0.3640**	1.6529
Sentence_BERT	98.857	0.0144	4.4285	58.49	-0.0035	5.5368	74.85	0.0834	4.4791

Table 3: Intra-Cluster Results on professional reviews. CH: Calinski-Harabasz, Silhouette: Silhouette Coefficient, and DB: Davies-Bouldin Index. The best results of each evaluation method are in bold.

	ARI	**AMI**	**NMI**	**MI**	**homogeneity**	**completenes**
(Baseline) TF-IDF	0.00405	0.03811	0.05141	0.08899	0.05689	0.05141
BERT_CLS	0.03902	0.13670	0.14999	0.26153	0.16721	0.14999
BERT_MEAN	0.03964	0.12288	0.13661	0.23226	0.14849	0.13661
Fine_Tuned_BERT_CLS	0.00322	0.03321	0.04789	0.08429	0.05389	0.04789
Fine_Tuned_BERT_MEAN	-0.00060	0.03542	0.05049	0.08694	0.05558	0.05049
Sentence_BERT	0.02027	0.06320	0.07551	0.13580	0.08682	0.07551

Table 4: Inter-Cluster Results on professional reviews by using K-Means clustering. The best results of each evaluation method are in bold.

	Information of the book	**Theme of the book**	**Author's writing style**	**Author's life**	**Char-acters**	**Summary of the story**	**Reviewer's own feeling / thoughts**	**Reviewer's criticism / praise**	**Purity**	**Entropy**
Cluster 1	9	5	5	3	7	60	20	6	0.39	1.96
Cluster 2	0	7	14	6	23	89	22	14	0.51	2.19
Cluster 3	0	11	23	3	11	66	23	7	0.46	2.25
Cluster 4	10	0	0	0	1	6	1	0	0.56	1.46
Cluster 5	6	2	1	1	0	11	1	3	0.44	2.23
Cluster 6	1	1	4	3	5	51	10	3	0.65	1.78
Cluster 7	0	12	17	5	16	76	36	11	0.44	2.31
Cluster 8	3	2	13	0	25	80	26	11	0.50	2.09
Average									0.49	2.21

Table 5: Results of Purity and Entropy of each Cluster on professional reviews by using Fine-Tuned BERT to extract embeddings from CLS. Cluster 1 to Cluster 8 are clusters after implementing K-Means clustering.