Book Review Analysis with Text Clustering

CS571 Final Project

Wendy Mo, Yao Ge

Motivation and Problem Statement

A book review is a description, critical analysis, and/or evaluation of the quality, meaning, and significance of a book.

In this project, our goal is conducting content analysis the book review written by professional book reviewers (long review) and Amazon book reviewers (short review) for the same book, then compare the focus categories of reviews on different types of books.

We assume the professional book reviews are the “helpful” reviews. And through these studies, we hope to know **which content readers are more concerned about for different books**, and thus **finding out which book review is more useful to readers**, instead of only a star rating.

Dataset

1. Long Review: Book Review written by professional book reviewers from New York Times' most popular book list from **2013 to 2018**. There are **101 texts.**
2. Short Review: Amazon Book Review Data. This dataset contains book reviews and metadata from Amazon, including 8.9 million reviews, which is as for **2014**.
3. Overlap between two datasets: **36 books**

Annotations ---- 8 classes

- Information of the book (title / publisher / price)
- Theme of the book (background / what the author wants to convey through the book)
- Author’s writing style (compare to other authors / compare with previous works / Writing technique)
- Author's life
- Characters / Description of characters
- Summary of the story / Experience of characters
- Reviewer’s own feeling / analysis / thoughts
- Reviewer’s criticism / praise

Annotation

1. We have annotated 20 long reviews in sentence-level, with a total of **1083 sentences**.
2. Inter-annotator agreement: **0.535** → **1.0**
   a. Cohen’s kappa: a statistic that measures inter-annotator agreement
   b. Cohen kappa has just two annotators, each annotator annotates every item.
3.

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

Preprocessing

1. Divide each long review into sentences.
   a. Annotation
   b. Input to models
2. Simple preprocessing steps:
   a. checking the repeated reviews and removing them;
   b. removing non-ascii characters and special phrases (e.g. emojis, urls, and non-alphanumeric);

Models

1. **(Baseline) TF-IDF + K-Means**
    a. Use TF-IDF to calculating the weight of the term in the vector;
    a. Implement k-means for sentence clustering.
2. **(Baseline) TF-IDF + Birch / TF-IDF + HDBSCAN**
3. **BERT**
    a. Embeddings of CLS + K-Means / Birch / HDBSCAN
    b. Embeddings of Mean + K-Means / Birch / HDBSCAN
4. **Fine-tuned BERT**
    a. use half of annotated data to fine-tune on BERT
5. **Sentence-BERT**

Evaluation for clustering ---- intra-cluster

1. Calinski-Harabasz
   a. Mainly calculates the ratio of the distance between clusters to the distance within clusters. The higher the CH score, the better the clustering performance.
2. Silhouette Coefficient
   a. When the cluster density is higher and the separation is larger, the Silhouette Coefficient of the cluster is also larger.
3. Davies-Bouldin Index
   a. The lower limit of the DB index is 0, the smaller the DB index, the better the cluster performance.

Evaluation for clustering ---- inter-cluster

1. Purity & Entropy
2. AMI (Adjusted mutual information): It corrects the effect of agreement solely due to chance between clusterings, similar to ARI.
3. NMI (Normalized Mutual Information)
4. RI (Rand index) : is a measure of the similarity between two data clusterings.
5. ARI (Adjusted Rand index) : ARI solves the problem that RI cannot describe the similarity of the randomly assigned cluster class label vectors well.
6. Homogeneity & Completeness

Result ---- intra-cluster

	K-Means			Birch			HDBSCAN		
	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_Tune_BERT_CLS	**346.15**	0.0281	3.6245	**321.6**	0.0067	4.0500	139.8	0.3408	**1.6128**
Fine_Tune_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

Result ---- inter-cluster

		ARI	AMI	NMI	MI	homogeneity	completeness
K-Means	(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_Tune_BERT_CLS	0.00322	0.03321	0.04789	0.08429	0.05389	0.04789
	Fine_Tune_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

Result ---- inter-cluster

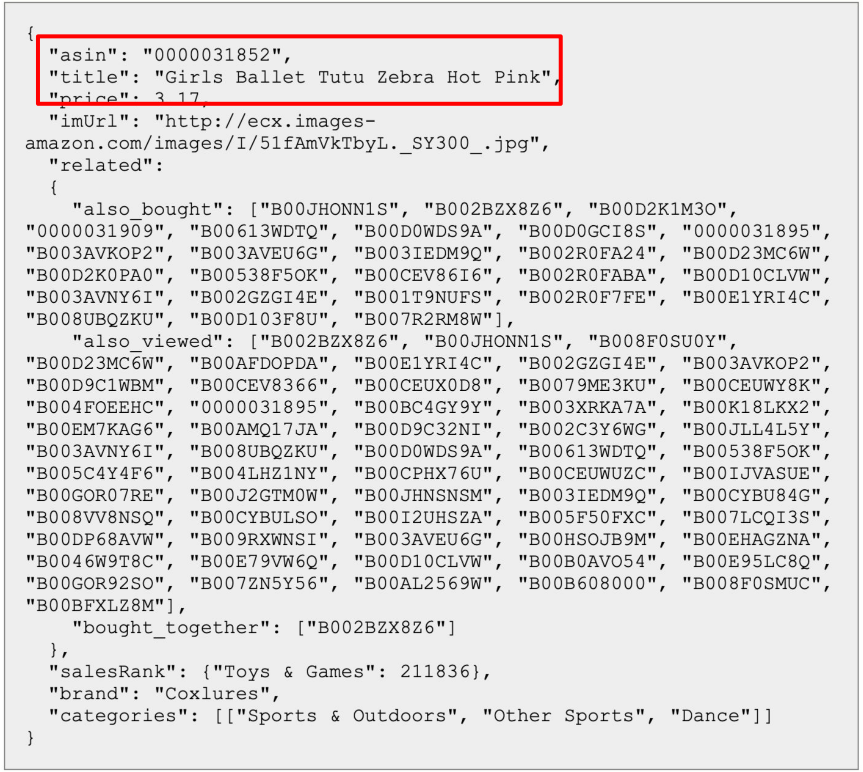
		ARI	AMI	NMI	MI	homogeneity	completeness
Birch	(Baseline) TF-IDF	-0.01501	0.02376	0.03881	0.05777	0.03693	0.03881
	BERT_CLS	-0.01112	0.11975	0.13423	0.22112	0.14137	0.13423
	BERT_MEAN	0.02186	0.11200	0.12582	0.21679	0.138610976	0.125829924
	Fine_Tune_BERT_CLS	0.00402	0.03372	0.04819	0.08577	0.05483	0.04819
	Fine_Tune_BERT_MEAN	0.00464	0.04087	0.05587	0.09601	0.06138	0.05587
	Sentence_BERT	0.03510	0.05134	0.06438	0.111969	0.07158	0.06438

Result ---- inter-cluster

BERT/ **Fine_Tune_BERT**	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	purity	entropy
Cluster 1	49 / **9**	25 / **5**	51 / **5**	0 / **3**	38 / **7**	9 / **60**	13 / **20**	0 / **6**	0.58 / **0.39**	1.90 / **1.96**
Cluster 2	51 / **0**	69 / **7**	62 / **14**	0 / **6**	38 / **23**	16 / **89**	9 / **22**	0 / **14**	0.51 / **0.51**	2.24 / **2.19**
Cluster 3	15 / **0**	16 / **11**	17 / **23**	1 / **3**	8 / **11**	9 / **66**	7 / **23**	0 / **7**	0.48 / **0.46**	2.21 / **2.25**
Cluster 4	15 / **10**	6 / **0**	8 / **0**	0 / **0**	13 / **1**	4 / **6**	2 / **1**	0 / **0**	0.58 / **0.56**	2.08 / **1.46**
Cluster 5	52 / **6**	39 / **2**	67 / **1**	0 / **1**	24 / **0**	9 / **11**	5 / **1**	0 / **3**	0.59 / **0.44**	2.07 / **2.23**
Cluster 6	69 / **1**	18 / **1**	41 / **4**	0 / **3**	62 / **5**	6 / **51**	3 / **10**	0 / **3**	0.50 / **0.65**	2.04 / **1.78**
Cluster 7	0 / **0**	2 / **12**	0 / **17**	12 / **5**	0 / **16**	8 / **76**	3 / **36**	0 / **11**	0.52 / **0.44**	1.44 / **2.31**
Cluster 8	0 / **3**	4 / **2**	0 / **13**	6 / **0**	0 / **25**	14 / **80**	44 / **26**	0 / **11**	0.46 / **0.50**	2.28 / **2.09**
									0.54 / **0.49**	2.23 / **2.21**

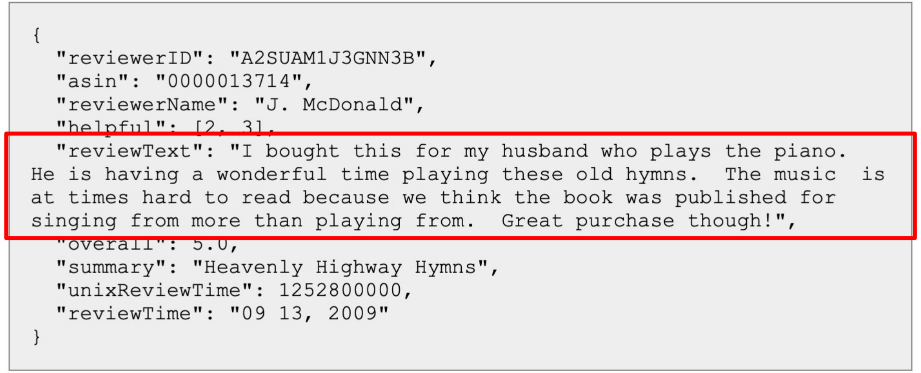
Comparison Between Short Reviews and Long Reviews

Metadata



```
{
  "asin": "0000031852",
  "title": "Girls Ballet Tutu Zebra Hot Pink",
  "price": 3.17,
  "imUrl": "http://ecx.images-
amazon.com/images/I/51fAmVkTbyL._SY300_.jpg",
  "related":
  {
    "also_bought": ["B00JHONN1S", "B002BZX8Z6", "B00D2K1M3O",
"0000031909", "B00613WDTQ", "B00D0WDS9A", "B00D0GCI8S", "0000031895",
"B003AVKOP2", "B003AVEU6G", "B003IEDM9Q", "B002R0FA24", "B00D23MC6W",
"B00D2K0PA0", "B00538F5OK", "B00CEV86I6", "B002R0FABA", "B00D10CLVW",
"B003AVNY6I", "B002GZGI4E", "B001T9NUFS", "B002R0F7FE", "B00E1YRI4C",
"B008UBQZKU", "B00D103F8U", "B007R2RM8W"],
    "also_viewed": ["B002BZX8Z6", "B00JHONN1S", "B008F0SU0Y",
"B00D23MC6W", "B00AFDOPDA", "B00E1YRI4C", "B002GZGI4E", "B003AVKOP2",
"B00D9C1WBM", "B00CEV8366", "B00CEUX0D8", "B0079ME3KU", "B00CEUWY8K",
"B004FOEEHC", "0000031895", "B00BC4GY9Y", "B003XRKA7A", "B00K18LKX2",
"B00EM7KAG6", "B00AMQ17JA", "B00D9C32NI", "B002C3Y6WG", "B00JLL4L5Y",
"B003AVNY6I", "B008UBQZKU", "B00D0WDS9A", "B00613WDTQ", "B00538F5OK",
"B005C4Y4F6", "B004LHZ1NY", "B00CPHX76U", "B00CEUWUZC", "B00IJVASUE",
"B00GOR07RE", "B00J2GTM0W", "B00JHNSNSM", "B003IEDM9Q", "B00CYBU84G",
"B008VV8NSQ", "B00CYBULSO", "B00I2UHSZA", "B005F50FXC", "B007LCQI3S",
"B00DP68AVW", "B009RXWNSI", "B003AVEU6G", "B00HSOJB9M", "B00EHAGZNA",
"B0046W9T8C", "B00E79VW6Q", "B00D10CLVW", "B00B0AVO54", "B00E95LC8Q",
"B00GOR92SO", "B007ZN5Y56", "B00AL2569W", "B00B608000", "B008F0SMUC",
"B00BFXLZ8M"],
    "bought_together": ["B002BZX8Z6"]
  },
  "salesRank": {"Toys & Games": 211836},
  "brand": "Coxlures",
  "categories": [["Sports & Outdoors", "Other Sports", "Dance"]]
}
```

Review Data



```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the piano.
He is having a wonderful time playing these old hymns.  The music  is
at times hard to read because we think the book was published for
singing from more than playing from.  Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```

Bert Clustering (And the Mountain Echoed)

Cluster	Long Review(%)	Short Review -- cosine-similarity-cluster	Short Review -- cosine-similarity	Short Review -- Jaccard-similarity-cluster	Short Review -- Jaccard-similarity
0	4.8%	18 -- 2.0385%	11.944%	208 -- 23.5561%	3.371%
1	20.3%	30 -- 3.3975%	10.645%	9 -- 1.0193%	3.356%
2	18.7%	13 -- 1.4723%	6.407%	12 -- 1.3590%	2.847%
3	19.6%	797 -- 90.2605%	17.194%	603 -- 68.2899%	3.327%
4	7.1%	13 -- 1.4723%	7.997%	16 -- 1.8120%	2.692%
5	7.1%	1 -- 0.1133%	19.901%	1 -- 0.1133%	1.695%
6	6.1%	6 -- 0.6795%	8.342%	26 -- 2.9445%	3.066%
7	16.2%	5 -- 0.5663%	9.753%	8 -- 0.9060%	1.910%

Comparison Between Short Reviews and Long Reviews

Long Review:

Cluster 3:

khaled hosseinis new novel, and **mountains echoed**, may awkward title body work, its assured emotionally gripping story yet, fluent ambitious the kite runner (2003), narratively complex a thousand splendid suns (2007).

the kite runner suns, could yield soapy, melodramatic plot twists characters very, good very, bad. mountains, too, share contrivance sentimentality, mr. hosseini narrative gifts deepened years, enabling anchor firmly maudlin aspects tale genuine emotion fine-grained details.

Short Review:

Cluster3:

Khaled Hosseini conveys life's reality in such a way that even though his characters may seem harsh, the prose is smooth, descriptive and full of emotion. It does require the reader to keep a sharp mind to follow the narrative thread.I think it helped that different characters told their story and the timeline was maintained thanks to that. And I really appreciated that the author did so well writing in English as a second language.

Comparison Between Short Reviews and Long Reviews

Short Review:

Cluster 3:

This book then is the story of &#34;why something, something like the tail end of a sad dream&#34; may sweep the human heart. Life is full of choices &#34;either you tore free or you stayed and withstood its rigor even as it squeezed you into something smaller than yourself.&#34; In his beautiful prose, Hosseini croons the stories of losses and re-discoveries in the decades of this Afghan family's lives. Starting with the folk story at the start of the book, Hosseini peers deep into the souls of parents and children struggling to find their best selves.The div is a fearsome creature that once a year knocks on the roof of one house in the wretchedly poor village. He claims one child chosen willingly, or all the children of the house. Only Baba Ayub follows the beast to claim his child, only to find she is granted a life of beauty and safety, happy under a spell of forgetting. Hana's choice resonates through the rest of the book. First we meet Pari and Abdullah, loving siblings. Each is the mirror to the other. Abdullah, especially since the death of their mother, has always protected her. Their parting starts the resonance of the travels throughout the rest of the book.Several threads of family love weave through this book suffice it to say that we see the &#34;the path behind our characters littered with all the shiny little pieces that life has ripped from him.&#34; Yet this path has great beauty, and sometimes light the way to a more complete life This is a lovely book, and it is one to savor. I was going to go to sleep but I just may read it again.

Comparison Between Short Reviews and Long Reviews

Long Review:

Cluster 5:

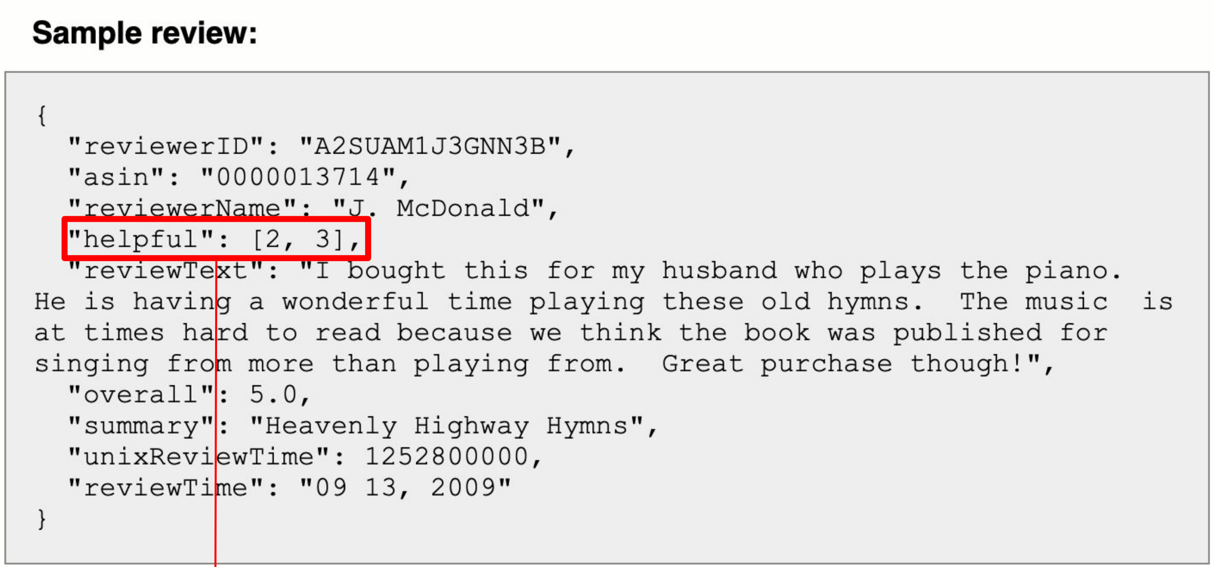
abdullah, ends california, running restaurant called abe kabob house. wife named child **pari**, long-lost **sister**, younger **pari** dream reuniting **father** missing sibling. **mother** dies, **father** begins suffer dementia, **pari** decides postpone dreams going art school take care **abdullah**.

Short Review:

Cluster 5:

It is 1952, in a small village in Afghanistan. **Abdullah** and his **sister Pari** live with their **father**, step**mother** and baby stepbrother. **Abdullah** raised **Pari** when their mother died in childbirth and the two are extremely close. In the afterword, Hosseini mentions that the title for the book was inspired by this phrase in a poem by W

Future work: Relation to “Helpful”



Sample review:

```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the piano.
He is having a wonderful time playing these old hymns.  The music  is
at times hard to read because we think the book was published for
singing from more than playing from.  Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```

“**helpful**”: helpfulness rating of the review, e.g. 2/3

We’ll first calculate the **helpfulness rating** (denominator > 1) of each review.

Reference

1. Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao,Fangyuan Wang, and Hongwei Hao. 2015. Short Text Clustering via Convolutional Neural Networks. Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, 62-69.
2. Yang Fan, Liu Gongshen, Meng Kui, and Sun Zhaoying. 2018. Neural Feedback Text Clustering With BiLSTM-CNN-Kmeans. IEEE, 57460-57469.
3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. volume 1810.
4. Nils Reimers, Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP. arXiv:1908.10084. https://arxiv.org/abs/1908.10084.
5. BERTopic: https://maartengr.github.io/BERTopic/index.html, https://github.com/MaartenGr/BERTopic.
6. Amazon Product Data: http://jmcauley.ucsd.edu/data/amazon/links.html.
7. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html.
8. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html.
9. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html.
10. https://en.wikipedia.org/wiki/Hierarchical_clustering.
11. https://blog.csdn.net/Eastmount/article/details/50473675.