Women's Clothing Review Analysis with Extraction-Based using Clustering and Topic Modelling Techniques

Nathaniel Takeshi, Nyoman Wikananda Santana, Muhammad Jalu Herlambang

June 3, 2024

1 Introduction

Consumer review on e-commerce product hold essential insights for business owner to induce business strategy. Consumer review shown the user satisfaction and dissatisfaction, giving valuable insight for future business strategy and marketing. Topic modeling technique can be utilize to capture common topics from thousands of product reviews. The topics generated could be consumer's liking, or consumer's feedback and complaints.

In this opportunity, we are performing topic modeling on **Women's Clothing E-commerce Review** dataset using different techniques, which is K-Means and HDBSCAN as the clustering; *Latent Dirichlet Allocation* (LDA) and *Bidirectional Encoder Representations from Transformer* (BERT) as the topic modeling. We will show that the combination of K-Means and BERT model excel in average coherence score for the extracted topic.

2 Dataset

The dataset used is **Women's Clothing E-commerce Review** from Kaggle. It consists of 23.486 data, which further truncated into 22.628 data, due to some missing attributes within, and will be discussed in later parts of this document. The dataset itself is highly imbalance, it contains more positive reviews and very little negative reviews. This will highly affect the topic modeling result, as probably the model won't see the negative review's topic as much worthy to be extracted. But, in this Study, we will try to utilize sentiment score and clustering to overcome such issues.

3 Methodology

We applied Exploratory Data Analysis (EDA) in the firsthand. Our main focus is to delete all data that does not have text reviews and combine the title with text reviews into one attribute. Then we add the word count and also the Vader sentiment value from the text review as one of the attributes. Furthermore, ever since we were looking for the negative reviews, we discovered that positive reviews dominated the negative ones, with an addition to people tended to recommended the clothings rather than not. As a result, our searching in finding appropriate data is quite of a challenge.

Next, we applied the text preprocessing, namely: lowercasing, stopwords removal, POS tagging. We also added business stopwords exclusion to enhance text filtration that would result into meaningful words within each data. Lastly, we also implement dimensional reduction by using a combination pipeline of Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE).

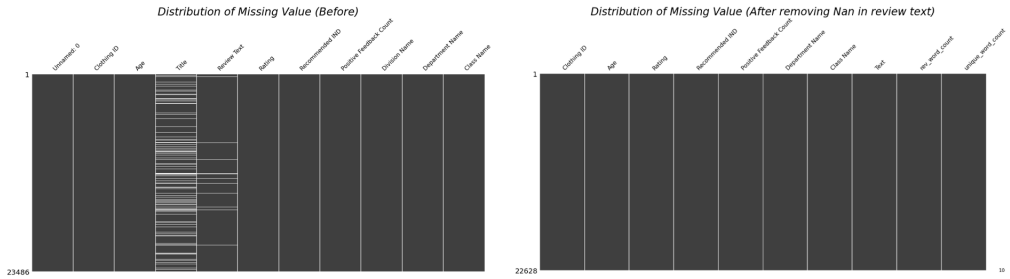

Figure 1: EDA and Feature Engineering

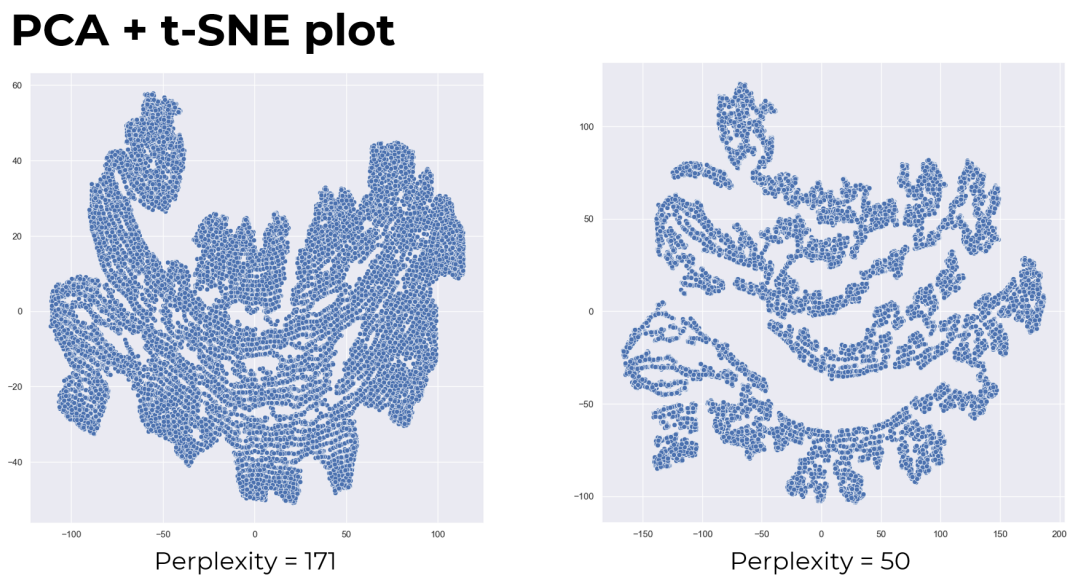

Figure 2: PCA + t-SNE result

After preprocessing, we attempted to create clusters using two approaches: K-means and Hierarchical Density-Based Spatial Clustering of Application with Noise (HDBSCAN), as a baseline and state-of-the-art methods, respectively.

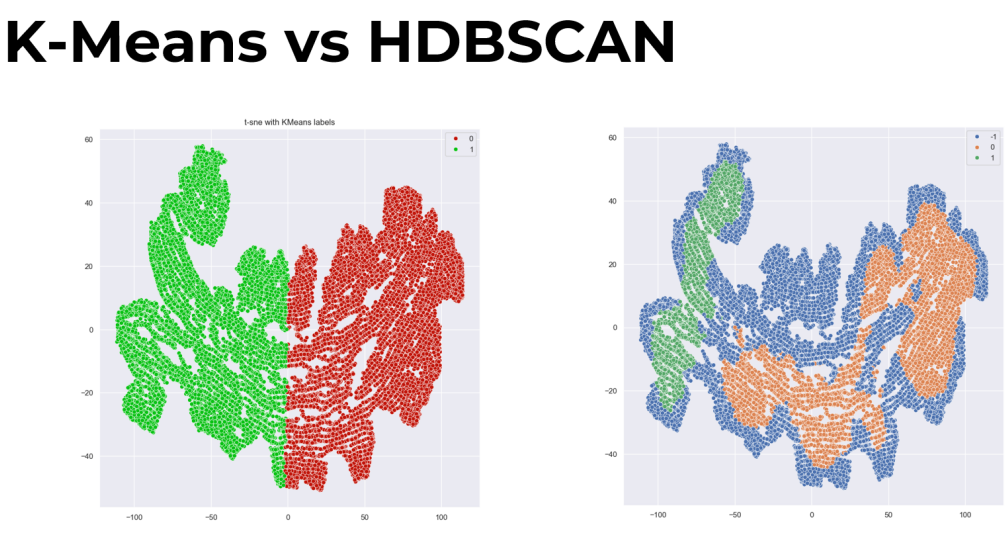

Figure 3: K-means vs. HDBSCAN

Surprisingly, K-means produced a better result than HDBSCAN. We could infer that from the silhouette score and clustering algorithm from HDBSCAN, where it would cluster data regarding to distance between and minimum number of data per cluster.

Proceeding with topic modelling, where we used LDA and BERT, specifically using BERTopic. During the process, we also used the coherence score to determine optimal number of topics, and we got the optimal answer is three.

In later sections, we provide the result in table that showed K-means and BERTopic produce the best result among all combination of tested clustering algorithm and topic modelling techniques.

4 Result

4.1 Clustering

The result for clustering with K-means and HDBSCAN is shown by the clustering result with silhouette score shown in Table 4.1. Based on optimal silhouette score, we produce 2 cluster for K-Means while HDBSCAN generated their own number of cluster, which in this case, generated 3 clusters. But color blue (label -1) in HDBSCAN is considered as a noise, which means it doesn't belong to any cluster. Hence, we can consider that HDBSCAN produce 2 clusters only. K-Means create more intuitive cluster based on the plot position while HDBSCAN produce cluster based on the hierarchy which could produce a different cluster inside a cluster. Based on this, it makes sense that HDBSCAN produce less silhouette score compared to K-Means as it is also less intuitive.

Silhouette Score	**K-Means**	**HDBSCAN**
	0.530	−0.006

Table 1: Silhouette score for different clustering techniques

After clustering, we performed topic modeling on each cluster. We use combination of the two cluster methods with two other topic modeling methods, which is, LDA and BERT model. This produce 4 different combinations of methods. The result of each combination can be shown by Table 4.1.

K-means Result				**HDBSCAN** Result			
Coherence Score	Cluster-#	LDA	BERT	Coherence Score	Cluster-#	LDA	BERT
	Cluster 0	0.251	**0.787**		Cluster 0	0.273	**0.832**
	Cluster 1	0.235	**0.801**		Cluster 1	0.252	**0.719**
Average Coherence Score		0.243	**0.794**	Average Coherence Score		0.263	**0.776**

Table 2: K-means and HDBSCAN result with topic modeling

We see that the combination of K-Means for clustering and BERT for topic modeling produce the highest average coherence score. In general, BERT model highly outperform LDA technique, mainly because BERT model is an LLM based technique which able to capture semantic dependence more accurately. While K-means and BERT produce the highest average coherence score, it is not so far from HDBSCAN and BERT. For further research, we would highly suggest different experiment of parameter for HDBSCAN and the BERT model.

The topic extracted using combination of K-Means and BERT model is shown in Table 4.1. Most user love the clothing design aesthetic and they often buy it as a gift and for their daughter. But, it seems most user have problem with the size of the cloth which often too small. Further clarification on clothing size would certainly helpful.

Topics	Meaning
Cluster 0	
['great', 'size', 'top', 'dress', 'fit', 'love', 'comfortable', 'soft', 'small', 'shirt']	Regarding the fit, probably some issue with the size (small), the material is comfortable
['daughter', 'gift', 'loves', 'size', 'top', 'beautiful', 'cute', 'birthday', 'petite', 'im']	Gift for daughter, beautiful and cute design, gift for birthday
['photos', 'measurements', 'reference', 'medium', 'size', 'top', 'person', 'online', 'much', 'mauve']	Customer may insist availability of photos and measurements. Probably some problem between online reference and actual product. Concern about specific color (mauve)
Cluster 1	
['dress', 'size', 'fit', 'top', 'im', 'great', 'small', 'fabric', 'little', 'love']	Regarding the fit, probably some issue with the size (small), the material is comfortable
['joggers', 'jogger', 'fit', 'perfect', 'pair', 'comfy', 'little', 'pants', 'comfortable', 'small']	Jogger outfit with comfortable material, potential problem with size (small)
['green', 'store', 'saw', 'try', 'great', 'love', 'little', 'online', 'fall', 'catalog']	Regarding color (green), the great online catalogue which make the customer try the cloth

Table 3: Extracted topic and the potential meanings

Finally, these also are the insights we could provide to the company based on the extracted topic from Table 4.1.

1. Most of customer are around 30-39 and 40-49. Advertising can be more directed towards those group of ages.
2. Super popular products are blouses, dresses, and fine gauge. Advertise and innovate in those products more.
3. Make clothing size information clearer.
4. Available design is good, keep it. Or innovate using the similar design mood.
5. Utilize online catalogue, make it pretty.
6. Make Birthday Gift Event (Package or Discounts) or children cloth packages or discounts.

5 References

[1] Blei David M., Ng Andrew Y., and Jordan Michael I. "Latent Dirichlet Allocation". In: *Journal of Machine Learning 3, 2003* (). Ed. by John Lafferty, pp. 993–1022.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[3] C. Hutto and Eric Gilbert. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". In: *Proceedings of the International AAAI Conference on Web and Social Media* 8 (May 2014), pp. 216–225. DOI: 10.1609/icwsm.v8i1.14550. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

[4] Leland McInnes, John Healy, and Steve Astels. “hdbscan: Hierarchical density based clustering”. In: *Journal of Open Source Software* 2.11 (2017), p. 205. DOI: 10.21105/joss.00205. URL: https://doi.org/10.21105/joss.00205.