Redefining Customer Satisfaction: A Novel Approach to Aspect-Based Sentiment Analysis

Ahmad Allaou
August 2024

Abstract

The purpose of this study is to perform topic modeling, text classification, and sentiment analysis on customer reviews. The results from the various models will be evaluated against each other, comparing not only the accuracy and coherence of the topics and sentiments generated, but evaluating their performance within the resource-constricted environment. This paper will evaluate how popular topic modeling and sentiment classification models such as LDA, BERTopic, distil-BERT, and alBERTa and leading large-language models such as ChatGPT. The Kaggle dataset explores over 14,000 reviews of a Lenovo mobile product, where topic modeling and sentiment analysis will provide some direction to what the customers are saying about the product. World-leading companies such as Amazon prioritize customer experience by constantly soliciting feedback and using this feedback to enhance their products and services. This project explores an innovative approach to customer experience using NLP techniques to extract qualitative data from customer reviews and provide insights from their reviews.

1 Introduction

In a competitive market, where product prices are nearly identical, customer experience becomes the key differentiator. Traditional metrics like Net Promoter Score (NPS) gauge customer satisfaction by asking, "how likely is it that you would recommend this product to friends or family?" and providing a rating between 0 and 10. While this metric can be useful, it often misses the nuances of customer opinions. During Amazon's early growth, Jeff Bezos states, "The most important single thing is to focus obsessively on the customer. Our goal is to be earth's most customer-centric company." Here Bezos emphasizes the importance of developing and iterating on a product based on what the customer says.

In the context of Natural Language Processing (NLP), and specifically in tasks where extracting themes from texts is important, topic modeling has been used as the standard in segmenting these topics. Topic modeling at its core is a statistical method meant to discover abstract topics that occur within a collection of documents. This is done by clustering words that tend to co-occur frequently across many documents with the purpose being to have the clusters be a representation of a distinct topic.(Rana, T. A., Cheah, Y. N., Letchmunan, S. (2016))

Following the approach of topic modeling, the objective is then to perform a sentiment analysis to not only identify some of the aspects that are often referenced, but what the sentiment towards them are. Despite sentiment analysis being the approach often referenced when discussing customer satisfaction and reviews, simply knowing if they are satisfied or unsatisfied is no longer enough. Today, with the exceptional architectures built, its possible to mine these opinions of the customer to better understand their pain points. Furthermore, by utilizing topic modeling along with sentiment analysis, its possible to capture multiple themes associated with a sentiment.(Amplayo, R. K., Lee, S., Song, M. (2018))

This study offers a comprehensive comparison between popular topic modeling approaches, such as BERTopic and Latent Dirichlet Allocation (LDA), alongside a sentiment analysis model using a pretrained transformer-based model (DistilBert and alBERTa). The results are then contrasted with those obtained from a large language model like GPT-4o, which generates themes and polarities for each topic. LDA, an unsupervised topic modeling technique, aims to identify latent themes within a document corpus through clustering algorithms.

Given a large volume of data in reviews, a topic modeling algorithm can be utilized to find the various aspects that are consistently referenced among all of the reviews. In the Kaggle dataset provided, these will be product reviews about a new mobile Lenovo phone that was released. From this, a topic modeling algorithm such as BERTopic and LDA will be used to evaluate topics, and then finally, a sentiment analysis will be performed afterwards to display the polarity of each of the topics. The main contribution of this paper will be further work into existing aspect-based sentiment analysis models. Current models such as Aspect and Sentiment Unification Models (ASUM) have not provided the greatest results, so this paper will build upon existing models and comparing the results. Amplayo, R. (K., Lee, S., Song, M. (2018))

This study will aim to contribute to current existing topic modeling and sentiment analysis works on product reviews. It will evaluate the results using F1 and accuracy scores for the sentiment analysis models, and coherence scores for the topic modeling algorithms to evaluate the quality of the topics. Coherence scores represents the coherence of the words within a topic, which gives an idea of how interpretable words within a topic are. The results of the sentiment analysis model will be limited to 5% of the due to resource constraints (this project is done in a single CPU environment).

2 Background and Method

Topic modeling is a technique used for identification of patterns in large collections of unstructured text data. In this context, a topic represents a probability distribution over a set of words, and a topic model is a statistical framework that identifies these topics across a collection of documents. Among the most common topic modeling models is the Latent Dirichlet Allocation (LDA), a generative probabilistic model for discrete data. It is the most popular and simple model for topic modeling that has improvements from prior works such as PLSI (probablistic latent semantic indexing). (K., Lee, S., Song, M. (2018))

With the advent of advanced NLP methods, sentiment analysis tasks have emerged as a response to challenges such as reviewing large amount of product reviews to mine opinions of consumers. Due to resource constraints, this study leverages distil-BERT and alBERTa, a more efficient, faster version of the original BERT model. It has 40% fewer parameters than BERT, yet it retains about 97% of BERT's performance. The sentiment analysis models utilized in this study are built off these smaller, less intensive transformer-based models.

This research endeavors to critically evaluate the efficacy and interpretability of traditional topic modeling and sentiment analysis techniques in contrast with the capabilities of modern large-language models, specifically the GPT-4o architecture. Traditional models such as LDA and BERTopic, alongside sentiment analysis through DistilBERT, will be utilized as benchmarks. The study will subsequently engage the GPT-4o model, which will be tasked with categorizing review data into pre-defined thematic categories. These categories include Battery Charging, Customer Service, Hardware Issues, Camera and Photos, Performance, Value for Money, Network and Software, Heating Issues, and Design and Aesthetics. A subsequent text-classification model will be employed to evaluate the congruence of the reviews within these predefined themes, providing insights into the comparative performance of these methodologies.

2.1 Data

The dataset used in the paper is based on a Kaggle product reviews dataset that provides a collection of reviews of a Lenovo phone from Amazon. Along with the reviews, they are each provided with a label that shows the sentiment of that review. Table 1 below depicts the distribution of the sentiment within the dataset.

Number of Rows	14,675
Number of Positive Reviews	6,963 (47.4%)
Number of Negative Reviews	7,712 (52.6%)

Table 1: Sentiment Distribution in Kaggle Dataset

During the evaluation of thematic elements within the dataset, the data was segmented by sentiment labels (Positive Negative), and n-grams were extracted to identify common patterns. This approach provides a preliminary insight into the lexical patterns associated with differing sentiments. As illustrated in Tables 2.a & 2.b, prominent negative n-grams include "battery drain fast," whereas positive sentiments are often linked to n-grams such as "battery backup good." This basic yet effective method reveals mixed sentiment towards the phone's battery performance, highlighting both strengths and weaknesses perceived by users.

N-gram	Count
lenovo k note	137
battery backup good	47
good battery backup	44
phone price range	41
camera quality good	33
best phone price	30
good camera quality	29
good battery life	27
good phone price	22
deca core processor	21

(a) Top 10 Positive N-grams

N-gram	Count
lenovo k note	377
dont buy phone	101
battery drain fast	88
please dont buy	84
worst phone ever	70
battery draining fast	53
dont buy product	50
battery backup good	42
camera quality good	40
pls dont buy	32

(b) Top 10 Negative N-grams

Table 2: Comparison of Top 10 Positive and Negative N-grams

Following this EDA, some data augmentation was performed on the text, in preparation to run the reviews through the LDA model. The reviews are initially lowercased, and all special letters are removed. The words are then lemmatized, combining the various inflected forms of a word into the same word so that they can be grouped. Finally, the words are then tokenized and all stop words are removed so that meaningless words are not represented.

2.2 LDA

Following data transformation and augmentation, an LDA model was constructed. The process began with initializing a dictionary from the corpus of pre-processed reviews. This dictionary was refined by excluding words that appeared in fewer than 10 documents and those that occurred in more than 50% of the corpus, thereby eliminating both rare and overly common terms. The LDA model was then generated using this filtered dictionary, with a baseline configuration of 12 topics, an alpha value of 0.1, and a beta value of 0.1. The figure below illustrates how the coherence score varies with an increasing number of topics.

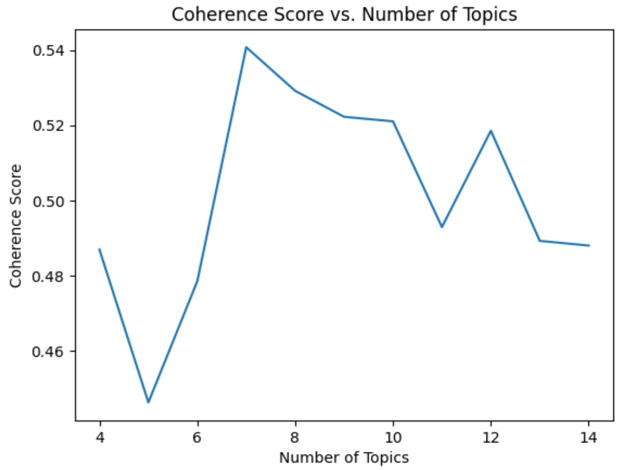

Figure 1: Line plot showing the impact of the number of topics on the coherence score

The coherence scores indicate that the optimal range lies between 6 to 9 topics, with noticeable degradation in coherence when deviating from this range. However, running the LDA model with fewer than 12 topics revealed that the topics generated were too broad, encompassing a diverse range of terms. To achieve better specificity and thematic clarity, the number of topics was increased, leading to the selection of 12 topics as the optimal configuration for the LDA model in this study.

2.3 BERTopic

The BERTopic model extends the concepts explored in the LDA model by leveraging transformer-based methods for topic modeling. It clusters topics in a way that makes them easily interpretable and offers more parameters for fine-tuning compared to LDA. Additionally, BERTopic produces more coherent and interpretable topics. In this study, the BERTopic model was configured to generate 12 topics, with a minimum topic size of 75 words. The intertopic distance map below (Figure 2) illustrates the relationships and correlations among the generated topics, highlighting their interconnections.

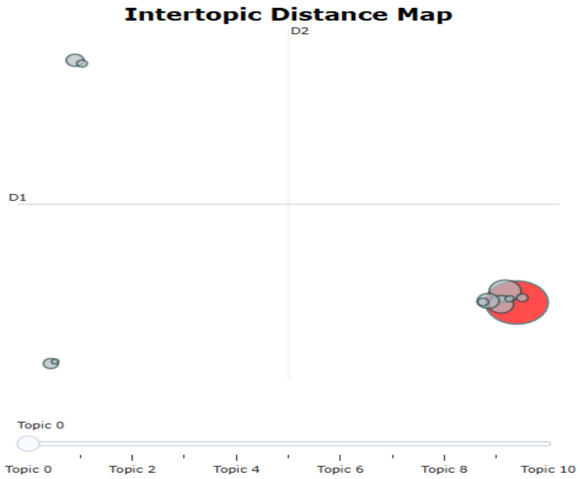

Figure 2: BERTopic Intertopic Map

2.4 DistilBERT and alBERTa

Following the topic modeling, sentiment analysis was performed on the reviews to gauge the sentiment of of the customers review. This enables the inference of the attitude of the customers towards each topic that is produced by the topic modeling techniques used. DistilBERT and alBERTa were used in place of more complex and better performing models due to resource constraints. Additionally, for the same reason the models were only trained on 300 reviews and evaluated 50 reviews (around 3% of all of the data), as the environment could not handle much more memory usage than that. The table below shows the parameters used within each of these models.

Model	**Learning Rate**	**Batch Size (Train)**	**Batch Size (Eval)**	**Number of Parameters**
DistilBERT (Model 1)	2e-5	4	4	66 million
DistilBERT (Model 2)	3e-5	8	8	66 million
ALBERT (Model 1)	2e-5	4	4	12 million
ALBERT (Model 2)	3e-5	8	8	12 million

Table 3: Parameters and Model Sizes for DistilBERT and ALBERT Models

2.5 GPT-4o

Finally, the reviews dataset is loaded into a GPT model, and prompted with the task of identifying the various themes that exist within the reviews dataset. Along with this, the LLM was also tasked with finding which of the reviews fell within each of the topics and which of them were positive or negative. The prompt that was passed into the LLM was the following: ”Find the common topics for these reviews. What i want is a csv with the topic, the count of reviews under that topic, and the ids of the reviews that are in that topic” along with the reviews dataset csv. The model then produced the following themes from the reviews: Battery Charging, Customer Service, Hardware Issues, Camera and Photos, Performance, Value for Money, Network and Software, Heating Issues, and Design and Aesthetics.

3 Results

Ultimately, the results produced by the large language model were much more effective and interpretable that of the LDA BERTopic and along with the sentiment analysis model. Unfortunately, due to the nature of the way the LLMs work, its difficult to provide a metric to contrast its performance with the other models utilized in this paper. Regarding the LDA and the BERTopic, the results were evaluated based on the coherence scores for the topics, and the number of positive and negative reviews associated with them are included.

3.1 LDA

Regarding the LDA for topic modeling, hyperparameter tuning was performed with 12 topics for the best coherence score of 67.4% with an alpha of 0.1 and beta of 10. A smaller alpha reduces the number of topics, which make the topics more concentrated. The higher the beta, the more words are encompassed in a topic.

Although the coherence for the model with these parameters were the greatest, the performance of the coherence score came at the cost of high quality. This was because one topic was encompassing so many of the themes that it made it seem as if the topics was extremely coherent, while the rest of the topics were not as coherent. The following figure displays the results of the LDA model with each of the hyperparamters tested.

Model	Alpha	Beta	Coherence
LDA_1	0.01	0.01	59.8%
LDA_2	0.01	0.1	55.1%
LDA_3	0.01	1	59.4%
LDA_4	0.01	10	64.1%
LDA_5	0.1	0.01	53.8%
LDA_6	0.1	0.1	53.8%
LDA_7	0.143	0.14	60.0%
LDA_8	0.1	1	64.2%
LDA_9	0.1	10	**67.4**%
LDA_10	1	0.01	54.8%
LDA_11	1	0.1	56.5%
LDA_12	1	1	50.6%
LDA_13	1	10	66.9%
LDA_14	10	0.01	57.2%
LDA_15	10	0.1	58.0%
LDA_16	10	1	57.3%
LDA_17	10	10	54.9%

Table 4: LDA Models with Various Alpha and Beta Values

While LDA_8 was the best performing model in terms of coherence scores, the highest quality model in terms of interpretable topics had the alpha and beta scores of 0.143 and 0.14 respectively. After the topics were created, each review was then classified into which topics they fit in the most, and then the number of positive and negative reviews associated with each topic were shown (Table 5).

Topic	Negatives	Positives
0	1817	891
1	4120	5696
2	1024	597
3	2915	2009
4	1507	670
5	1702	1173
6	1112	450
7	536	1587
8	578	2299
9	1083	751
10	2120	671
11	4621	4095

Table 5: LDA Topic Sentiments

3.2 BERTopic

Due to resource and time constraints, only 4 versions of the bertopic model were utilized, and they were evaluated based on coherence scores and manual interpretation of the scores. The table below depicts how the coherence score was impacted with the various hyperparameters tested.

Model	min_topic_size	n_grams	Coherence
bertopic1	75	1	42.3%
bertopic2	90	1	44.8%
bertopic3	75	(1,3)	45.4%
bertopic4	90	(1,3)	45.0%

Table 6: BERTopic Model Comparison

Likewise to the LDA model, the reviews were then classified within each topic and the number of positive and negative reviews were displayed for each topic generated from the BERTopic.

Topic	Negatives	Positives
0	6034	3035
1	5034	5905
2	1044	2545
3	2308	666
4	6136	4252
5	96	388
6	921	3026
7	667	519
8	23	181
9	781	191
10	92	181

Table 7: LDA Topic Sentiments

The table below depicts how each of the BERTopics and LDA topics fit into the predefined models by the GPT model:

Theme	BERTopic	LDA Topic
Battery & Charging	0, 3, 4	0
Heating Issues	4, 7	8
Hardware Issues	7	5
Network and Software		6, 7
Camera and Photos	0	9
Performance	2	1, 9
Value for Money	1, 2	2, 11, 8
Customer Service		3
Design and Aesthetics	1	7

Table 8: Categorizing themes from each topic model

Comparing just the BERTopic models and LDA models thus far in this study, the LDA model seems to outperform the BERTopic in both the coherence metrics they were evaluated on, and how well they adhere to the topics found within the reviews. Clearly, the BERTopic failed to capture the themes that discuss the network and software, and the customer service reviews. Between these two models, and the limited resources in the environment used, the LDA performed better than the BERTopic model did.

3.3 DistilBERT and alBERTa

Given that the focus of this study is on topic modeling, and due to resource constraints, limited tests were run for sentiment analysis. The sentiment analysis task in this paper was evaluated using f1 and accuracy scores since the number of positive and negative sentiments within the dataset were quite balanced. There were 4 models tested;the best model (DistilBert Model1) resulted in an accuracy score of 75% and a f1 score of 72.5%. This was the preferred model used for testing sentiment.

3.4 GPT-4o

The GPT-4 model demonstrated superior performance in topic modeling, generating coherent and meaningful themes that could be easily categorized. Not only did it produce well-defined topics, but it also successfully assessed the sentiment associated with each topic. Among the nine topics identified by GPT-4, the themes produced by the LDA model were encompassed within these, while the BERTopic model missed some of the themes identified by GPT-4. The table below illustrates how the predefined topics from the GPT model were used to classify reviews into their respective categories. This reflects the model's capability to provide comprehensive insights into customer feedback by effectively integrating topic modeling with sentiment analysis, setting a new standard in natural language processing applications.

Theme	**Positives**	**Negatives**
Battery & Charging	1050	2559
Heating Issues	110	395
Hardware Issues	280	770
Network and Software	168	889
Camera and Photos	1219	1634
Performance	600	773
Value for Money	832	390
Customer Service	108	666
Design and Aesthetics	111	151

Table 9: Positives and Negatives for Each Theme

4 Conclusion

The analysis done in this project provides valuable contribution to current topic modeling approaches, comparing that of LDA and BERTopic, and how their results differ from a LLM approach to topic modeling and sentiment analysis. LDA was tested using various hyperparameter tuning such as the alpha, beta and number of topics. Regarding the BERTopic, various features were manipulated such as the number of ngrams, the size of each topic, and the minimum number of occurence each find the most optimal implementation of the model. The study not only highlights how each model's performance differs in the objective of topic modeling but also how each of the hyperparameters impact the performance of the model in its objective. Given an increase in resources, future studies can involve more work into the sentiment analysis, and stronger fine tuning in the topic modeling approaches used in these studies. This contributes to the ongoing research in topic modeling and sentiment analysis by demonstrating the effectiveness of integrating traditional models like LDA and BERTopic with advanced large language models such as GPT-4. The findings are expected to provide valuable insights and guide future developments in the field, promoting more refined and effective techniques for extracting meaningful information from customer reviews.

References

T. A. Rana, Y. N. Cheah, and S. Letchmunan. 2016. *Topic Modeling in Sentiment Analysis: A Systematic Review.* Journal of ICT Research & Applications, 10(1).

R. K. Amplayo, S. Lee, and M. Song. 2018. *Incorporating product description to sentiment topic models for improved aspect-based sentiment analysis.* Information Sciences, 454, 200–215.

V. S. Anoop and S. Asharaf. 2019. *Aspect-oriented sentiment analysis: a topic modeling-powered approach.* Journal of Intelligent Systems, 29(1), 1166–1178.

K. Nawab, G. Ramsey, and R. Schreiber. 2020. *Natural Language Processing to Extract Meaningful Information from Patient Experience Feedback.* Applied Clinical Informatics, 11(2), 242–252. https://doi.org/10.1055/s-

0040-1708049.

C. Adams, R. Walpola, A. M. Schembri, and R. Harrison. 2022. *The ultimate question? Evaluating the use of Net Promoter Score in healthcare: A systematic review*. Health expectations: an international journal of public participation in health care and health policy, 25(5), 2328–2339. https://doi.org/10.1111/hex.13577.