Degree Project in Technology

First Cycle, 15 credits

Abstractive Summarizations of User Reviews through Prompt Engineering

School of Electrical Engineering and Computer Science

LISA ETZELL & NORA HULTH

ABSTRACTIVE SUMMARIZATIONS OF USER REVIEWS THROUGH PROMPT ENGINEERING

Abstrakt sammanfattning av användarrecensioner genom prompt engineering

Lisa Etzell, Nora Hulth

***Abstract*—This study explores the effectiveness of using LLMs for generating summaries of customer reviews. The model GPT-3.5 was used and two different methods for formulating prompts were tested: shot-prompting and pattern prompting. When handling larger amounts of reviews that risk reaching the length of the model's context window, clustering and iterative summarization was used. The generated summaries were assessed through both human evaluation as well as automatic ROUGE and BERTScore measures. Results indicate that shot-prompting improved the quality of the generated summaries, while pattern-prompting did not show any clear improvements. Clustering reviews generally reduces summary quality. The results in combination with a literature study were used to assess what value the summaries could provide in practical application at an e-commerce company, which was conducted through a SWOT analysis. In the analysis several opportunities for companies were identified, including improved accessibility of review information for customers, leading to increased satisfaction, and internal use for business development purposes. Despite some threats such as inaccuracies and legal requirements, it was concluded that leveraging summaries of reviews can provide value to companies.**

***Sammanfattning*—Denna studie undersöker effektiviteten av att använda stora språkmodeller (LLM) för att generera sammanfattningar av kundrecensioner. Modellen GPT-3.5 användes och två olika metoder för att formulera prompts testades: shot-prompting och pattern-prompting. Vid hantering av större mängder recensioner som riskerar att överskrida modellens kontextfönster användes klustring och iterativ sammanfattning. De genererade sammanfattningarna utvärderades både genom mänsklig bedömning samt genom de automatiska måtten ROUGE och BERTScore. Resultaten visar att shot-prompting förbättrade kvaliteten på de genererade sammanfattningarna, medan pattern-prompting inte visade några tydliga förbättringar. Att klustra recensioner reducerar generellt sett kvaliteten på sammanfattningarna. Resultaten, tillsammans med en litteraturstudie, användes för att bedöma vilket värde sammanfattningarna skulle kunna ge vid praktisk tillämpning på ett e-handelsföretag, vilket genomfördes genom en SWOT-analys. I analysen identifierades flera möjligheter för företag, inklusive förbättrad tillgänglighet av informationen i recensionerna för kunder, vilket kan leda till ökad kundnöjdhet, samt intern användning för affärsutvecklingsändamål. Trots vissa hot med felaktigheter i sammanfattningarna och juridiska krav, drogs slutsatsen att användning av sammanfattningar av recensioner kan skapa värde för e-handelsföretag.**

I. INTRODUCTION

MOST reviews in today's society are based on some kind of numerical grade. Customers are asked to rate a product or service on a grade from one to five stars, and write a short comment motivating their opinion. However, when an experience is summarized in a grade, some parts of it may be lost. A number can hardly capture everything you actually experienced.

What is expressed in the written comment may also differ from the actual rating by the same customer. A review that sounds great in text may have a low rating and vice versa. There is a weak correlation between the text and the rating. Many customers tend to have a bias towards opinions already expressed regarding the product. Thus, if a product has a low average-grade, even a satisfied customer tends to give a slightly lower rating. In this case, a written comment will probably give more information regarding their actual opinion, compared to the rating [1].

From the perspective of the future customers who read the reviews, many choose a product with a higher average rating, even though the comments on a product with a lower rating are more positive [1].

Under the assumption that numerical reviews have a weak correlation to text based reviews it is likely that numerical and text based reviews do not provide the same information. Numerical reviews have the advantage that they are simple to aggregate by calculating a representative value such as mean or median. On the other hand they lack the detailed information that is provided by text reviews. A general summary of all text reviews of a single product would arguably provide customers with more information. Informed customers are more likely to find products satisfactory and according to their expectations. Hence a review summary could increase customer satisfaction. Reading many reviews is also time consuming, and thus a summary of all reviews for a given product would be advantageous for both future customers as well as for the company to take part of the feedback given.

The main issue about summarizing multiple text reviews is the time aspect. To summarize multiple texts manually is time consuming and therefore hard to scale. However, recent developments in Natural Language Processing have opened up the possibility to generate automatic summaries with generative AI by utilizing Large Language Models (LLM).

Previously fine-tuning has been a popular method to get these pre-trained LLMs to perform certain tasks but as of lately prompt engineering have become an increasingly popular alternative, which will be the main focus of this paper. A challenge in using LLMs for review summarization is their limited context window (maximum input). If a single product has too many or too long reviews, this limit risks being reached.

A. Purpose

The purpose of this study is to explore how prompt engineering can enhance the performance of Large Language

Models in generating summaries of multiple user reviews and to investigate if the standard of the resulting summaries is high enough. To be useful the model would have to guarantee that the generated summaries include all relevant information that can be scraped from the different reviews of a specific product. At the same time, the summary has to be short in order to constitute a valuable alternative to reading all the reviews, that both saves time and makes the feedback from other customers more accessible.

As a limited context window is a challenge to summarizing larger amounts of product reviews, we also aim to investigate how first summarizing clusters of reviews individually and then combining them into one unified summary affects the quality of the final summary.

Further, the study also aims to investigate if the generated summaries of customer reviews could provide value for a company. Both considering the quality of the generated summaries but also the potential impact a potential implementation of summaries of customer reviews could have for a company.

B. *Social and Ethical Aspects*

The assumption is that a summary of product reviews will make it easier for customers to engage with feedback from other customers and thereby gain more information about the products. For an e-commerce company this is especially important since customers are not able to examine the products themselves and therefore are more dependent on the information provided online. If customers make more informed decisions it is expected that they are less likely to regret their decisions [2]. For an e-commerce company this means less returned packages and increased customer satisfaction. A decreased amount of returned packages would mean decreased administrative and logistic costs. Since returned products must be considered when estimating production volumes and there is a risk of unsellable returns, reducing the number of returns could lower production costs as well. Moreover, it would also mean less transportation which would decrease the climate footprint of the company [3].

C. *Research Question*

The study aims to investigate the two research questions below. To conduct the research the listed sub-questions will be answered.

1. To what extent can prompt engineering of LLMs be used to generate a representative text summary of user reviews?

1.1. How can prompts be designed to optimize the task of opinion summarization?

1.2. Can performance be improved by subtask prompts, summarizing clusters of reviews individually and then combining these summaries into one unified summary?

2. Given the quality of the generated summaries of user reviews, as established in the findings from the research question above, could they provide value for companies?

2.1. What are the strengths and risks of the generated summaries?

2.2. What are the opportunities and threats for companies to implement and display summaries of user reviews?

The study will focus on the summarization of online reviews of products and not services. In order to investigate how prompt engineering can be used to optimize this task, we will have to find a systematic approach to vary the different prompts. This paper will handle the challenge of LLMs' limited context windows in two ways. Firstly, when investigating the formulation of prompts, the amount of reviews to be summarized will be limited in order to not reach the maximum input. Secondly, summarizing products with larger amounts of reviews is still a question of interest, for which clustering will be used. The generated summaries will have to be evaluated according to appropriate measures, to ensure quality. Finally, the strengths and weaknesses of the generated summary will be taken into consideration with the opportunities and threats of using the summaries in an e-commerce-context, in order to assess its value.

II. Theory

A. *Text Summarization*

Text summarization is the task of taking a longer text and creating a shorter summary of it. There are two main types of summarization; extractive summarization and abstractive summarization. Extractive summarization will extract key phrases and parts of a document and then concatenate these into a shorter summary. Each sentence or phrase will be ranked according to its relevance for the whole document, and the ranking will decide which parts are to be kept and which are to be removed. Thus, no new text is generated, which is the case in abstractive summarization. An abstractive summary will generate new sentences that capture the same meaning in a more concise way [4].

Multi-Document Summarization (MDS) concerns the task of generating a collective summary of several documents, usually written about the same topic. The fact that the corpus consists of multiple documents entails new difficulties for the task of summarization and in comparison to single-document summarization MDS is therefore regarded as much more complex. Since the documents of an MDS corpus cover the same topic they are likely to contain sentences or paragraphs with similar information which reinforces difficulties with redundancy. This must be managed properly by implementing anti-redundancy procedures in order to attain high informativeness of the summary. There is also a risk that the documents may contain conflicting information since they could be written by different authors, with different opinions and written during different time periods. This remains a delicate and open problem. In general MDS models are therefore more inclined to generate summaries that contain redundant, incoherent and even contradictory information [4].

Opinion summarization is a special case of multi-document summarization, focusing on the opinions and sentiments expressed. It also introduces a quantitative aspect, where the summary is to reflect which opinions are expressed more frequently than others. An opinion summary should contain different aspects of one or several topics. In contrast, a single document summarization will only extract important information, and a multi-document summarization will compare documents and remove repeated information. Thus they lack the contrastive and quantitative aspects [5].

B. Large Language Models

A large language model (LLM) is a type of language model used for general processing and understanding of natural languages. Natural Language Processing (NLP) concerns how computers can process, interpret and generate natural language and involves a range of algorithms and models with the ability to take natural language as input to generate some desirable output [6]. Thus, LLMs are suitable for a large range of NLP tasks, such as sentiment analysis, named entity recognition, and text summarization.

LLM is a type of neural networks, that uses a transformer model to learn the context and meaning of its input. They are pre-trained using very large amounts of data. This is a form of self-supervised learning, allowing the use of unlabeled datasets. After pre-training, the model is able to perform general tasks, but further tuning may be required for specific tasks. Fine-tuning is a way of customizing the model for a specific function. The model is further trained on new data for a specific task. This often involves supervised learning using labeled data [7].

Prompt engineering involves designing and finding the optimal input set of instructions (prompt) for the model to behave as wanted. Thus, it does not require any new training and is more flexible than fine-tuning. A prompt may firstly include a question or instruction for a task to be completed, which can be referred to as the base-prompt. This base-prompt can be complemented with further details such as context and examples. One common area within prompt engineering is shot-prompting, which involves providing the model with examples of how the user wants the model to respond. In zero-shot prompting, no examples are given, thus it is only the base prompt that instructs the model. In one-shot prompting, one example is provided, and in few-shot, there are multiple examples of the desirable outcomes [8].

Despite LLMs being a powerful tool, when used in summarization awareness is required regarding the risks of hallucinations. Huang et al. defines hallucinations as ”generated content that is nonsensical or unfaithful to the provided source content”. Thus, the produced summary risks inaccurate information, not representing its input documents. The risks of hallucinations may be minimized by limiting response length, provide unambiguous prompts, adjusting model parameters, as well as testing to identify vulnerabilities [9].

C. SWOT

SWOT analysis is a technique used within strategy that can be applied to summarize and identify strengths and weaknesses as well as opportunities and threats of a certain project. The internal capabilities of a project that offers a competitive advantage or in any way contribute positively to it's performance are considered strengths. The internal factors that on the contrary constitutes a limitation or a disadvantage for the project are considered weaknesses. Opportunities refers to external factors that the project have potential to benefit from meanwhile threats refer to external factors that pose potential risk or obstacles for the project. By distinguishing and mapping both internal and external factors the SWOT analysis provides a clear overview which can be used as a base for further analysis and decision making. An advantage of the SWOT analysis is that it provides visibility of internal strengths which can be utilized to diminish weaknesses and approach external opportunities and threats [10] [11].

D. Evaluation

Human judgement is naturally the best evaluation of whether a summary is representative or not, however it may be time consuming. Interrater agreement measures are also of importance, in order to assess that the raters are consistant. One such measure is Cohen's Kappa, which measures the agreement between two raters, while accounting for the possibility of the agreement occurring by chance [12].

BERTScore is one evaluation metric for text summarization. The methods compares how similar a candidate text is to a reference text, thus measuring the quality of the candidate text. To compare the texts the method uses the BERT-model to represent each token of the texts with contextual embeddings. The texts are then compared by calculating pairwise cosine similarity of the BERT embeddings computed for each token. The score is computed as recall and precision, which combined can be used to calculate an F1 measure, as shown in equation 1. The recall score is calculated by matching each token in the reference text to a token in the candidate text, and the precision score is calculated by matching each token in the candidate text to a token in the reference text. To maximize the matching similarity score the method uses greedy matching, where each token of one text is matched to the most similar token in the other text. By using contextual embeddings, BERTScore takes semantics into consideration and have been found to have a stronger correlation to human evaluation, than for example ROUGE-score [13].

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{1}$$

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) includes a set of measures that can be used to automatically evaluate the quality of a summary by comparing it to another ideal summary. The measures show the similarity between the documents by counting the number of overlapping units such as n-gram, word sequences, and word pairs between them. All measures range on a scale of zero to one, the closer the score is to one the higher similarity [14].

Three common ROUGE measures are ROUGE-1, ROUGE-2 and ROUGE-L. ROUGE-1 is calculated based on the unigrams in a candidate and reference summary. Thus, it compares the overlapping of single words in the two documents. ROUGE-2 is the corresponding metric for bigrams. The ROUGE-N precision and recall are calculated using equations 2 and **??**, from which the ROUGE-F1-score is calculated using equation 1.

$$\text{ROUGE-N-Recall} = \frac{\text{Common N-grams}}{\text{N-grams in Reference}} \quad (2)$$

$$\text{ROUGE-N-Precision} = \frac{\text{Common N-grams}}{\text{N-grams in candidate}} \quad (3)$$

ROUGE-L is based on the longest common subsequence. The words in the longest common subsequence are not required to be in consecutive positions (as in a substring).

III. Previous Studies

A. Text summarization and Prompt Engineering

Traditionally, both extractive and abstractive methods have been utilized for text summarization. Although summarization through prompt engineering of LLMs is considered abstractive, one alternative could be to integrate extractive summarization to improve performance. This approach has been tested in previous studies. For instance Benham et al. introduced a hybrid approach for summarizing user reviews, where their summarizer uses both extractive and abstractive methods. First an extractive summarization is performed by identifying key sentences and extracting them. Secondly, the extractive summary is passed onto a BERT-based abstractive summarizer. The method showed promising results and outperformed three other existing summarizers [15].

Bhaskar et al. conducted a similar study to this one where they investigated the performance of different pipeline methods to summarize user reviews by utilizing GPT-3.5 and prompt engineering. In their study they also combined extractive and abstractive summarization. They identified the model's limited maximum input length as a key issue and and wanted to investigate if this issue could be mitigated through constructing pipeline methods, where reviews were summarized iteratively in chunks. The pipeline methods consisted of a family of approaches where different extractors were used to select relevant parts of reviews. These parts were then summarized in clusters through iterative summarization, using GPT-3.5 as the summarizer. For short input reviews the study showed that basic prompted GPT-3.5 generated reasonably faithful and factual summaries. More advanced techniques did not show much improvement, thus indicating that integrating extractive summarization did not enhance performance. When used for repeated summarization of longer input reviews, GPT-3.5 tended to produce generalized and unfaithful selections of viewpoints [16].

Prompt engineering have become a popular method for utilizing pretrained LLMs to perform different tasks, including the task of text summarization. In a study conducted by Zandvoort et al. performance of transformer-based summarization of medical reporting was enhanced through prompt engineering. The study investigated both shot prompting and context pattern prompting. The shot prompting was assessed through testing zero-shot, one-shot and two-shot prompts. The context pattern prompting was then tested on the most effective of the shot-prompts through an increase of context. The added context was divided into two types of context, scope context and domain context. In the study the results showed that adding shots to the prompt was beneficial and that adding the combination of scope and domain context generated the best results. However, scope context had little effect by itself [17].

Given that adding examples and context to prompts have showed promising results in previous studies on medical reporting, this will be further investigated in this study in the context of summarizing customer reviews. The results from the study of Bhaskar et al. propsed that clustering can may be a proper method to mitigate the issue of limited context windows. The study will investigate clustering further by evaluating performance for a dataset with a large number of short reviews per product, which was not tested by Bhaskar et al.

B. Evaluation

ROUGE and BERTScore are both established evaluation metrics. ROUGE-scoring is often used in three different variants: ROUGE-L, ROUGE-1, and ROUGE-2, meanwhile BERTScore is often presented as the F1 score solely. The study conducted by Zandvoort et al., which focused on summarizing medical reports, used ROUGE-1 and ROUGE-L score as an automatic measure of the quality of the summaries. The top-performing prompt achieved a ROUGE-1 score of 0.25 and a ROUGE-L score of 0.189 [17]. Moreover, Bhaskar et al. used both BERTScore and ROUGE-1 and ROUGE-L score to measure the quality of summaries of customer reviews generated through different pipelines. When using the FewSum - Amazon dataset and GPT3.5 model, they recieved a BERTScore of 0.88-0.89. There ROUGE-1-scores were in a range of 0.26-0.27, and ROUGE-L at 0.23-0.24. Another of their pipelines involved few-shots, and had a ROUGE-1 of 0.33 [16].

Both ROUGE and BERTScore may however lack relevant measures for the task of opinion summarization. The measures are therefore usually complemented by some sort of human evaluation. Bhaskar et al. identifies the metrics factuality, faithfulness and relevance to better capture the important features of an opinion summary. Factuality considers whether the summary is based on actual statements from the data, and thus contains factual information, no hallucinations. A summary's faithfulness is how representative it is of the viewpoints from the original data. Relevance takes into consideration how relevant the aspects presented in the summary are for the task itself [16]. These measures will be adopted in this study as well.

IV. Method

A. Data and Preprocessing

Two different datasets have been used to complete this task. The first dataset consists of customer reviews from Amazon and is a subset of the dataset FewSum. The Amazon dataset consists of 480 reviews for 60 different products with 8 reviews each. The reviews are in English and each product has three corresponding human-written summaries. The products have been selected from four different categories 1) Electronics, 2) Clothing, Shoes and Jewelry 3) Home and

Kitchen, 4) Health and Personal Care. The data was divided into one big and one small subset. The small subset consisted of the data for four randomly sampled products, one from each category. This subset was used as examples when one-shot respectively few-shot prompting were tested.

The second dataset consists of reviews from a large, Swedish e-commerce brand in fashion and retail. The dataset consists of 20 products, all with over 100 reviews each written in Swedish. Each review has a creation date, product number, title and rating. The product number contains information of the size and the color of the product. As reviews can be considered helpful for the same product regardless of color or size, these were grouped together. The rating ranges from 1 to 5 stars. This dataset serves to help us evaluate the performance of LLM summarization on larger sets of reviews. In this report we will refer to this dataset as SweRev. To enable evaluation using automatic measures, a golden summary of the reviews of each product was written by the authors of this paper and added to the dataset.

B. Prompt Engineering

To investigate how prompt engineering of LLMs can be used to generate a representative summary of user reviews, OpenAI's model ”gpt-3.5-turbo” was used. Initally, an instructional base prompt was designed to instruct the model to summarize the reviews. The base prompt sets a baseline, which allows variations from this to be attributed to a specific additional prompt. In order to formulate prompts, two different methods were used: shot-prompting and pattern-prompting, which will be explained individually. The combined use of the two methods is illustrated in figure 1.

Shot-prompting was utilized as providing the model with example answers has proven to be effective in previous studies, and it is one of the most common methods of prompt engineering. Pattern-prompting was also mentioned in previous studies and allows testing different additional contexts to the base prompt in a systematic manner.

Some consideration regarding the use of GPT3.5 is that it has a limited context-window (input) of 16K tokens or characters. For this reason, the Amazon dataset will be used for the shot-prompting and pattern-prompting, as it contains fewer reviews per product.

1) Shot-prompting: Three different types of shot-prompts were tested: zero-shot, one-shot and four-shot. Four-shot prompts were chosen as the dataset contains 4 different product categories, and thus one example from each category was used. In the one-shot prompt, the given example was from the clothing-category.

In the one- and four-shot prompts, an additional instruction was added to the base prompt that one (or a few) examples would be given, this instruction was then followed by the examples. In order to highlight the examples, as well as mark where the model's response is required, we included tags for user-input and assistant-output. An example of a one-shot prompt is given in figure 2, the four-shot prompt was constructed similarly.

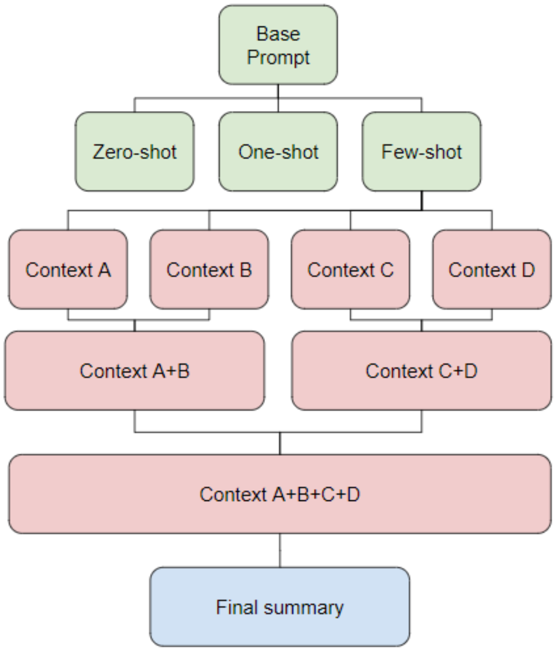
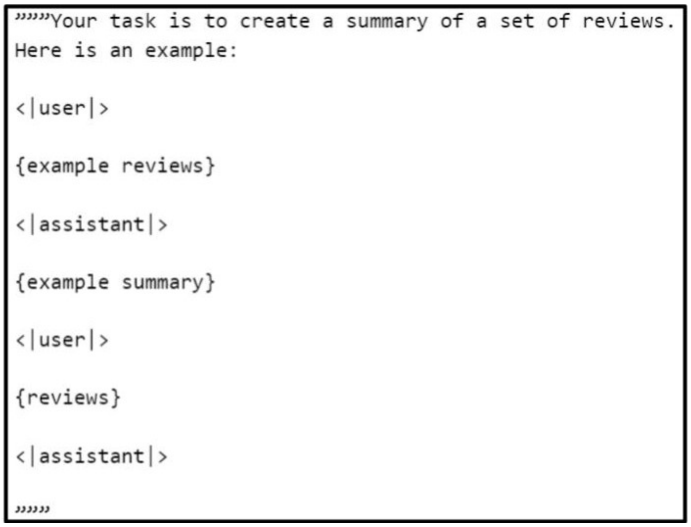


Fig. 1: The Flow of Prompt Formulation. Firstly, the base prompt is combined with different shot prompts. Secondly, the few-shot base prompt is combined with further context in different combinations, according to context pattern prompting.



```
”””Your task is to create a summary of a set of reviews.
Here is an example:

<|user|>

{example reviews}

<|assistant|>

{example summary}

<|user|>

{reviews}

<|assistant|>

”””
```

Fig. 2: Example and Structure of a One-Shot Prompt

2) Pattern-prompting: After investigating different shot-prompts, the effects of increased context were investigated. This was assessed by continuing with four-shot prompts, whilst adding information and context to the base prompt. To increase the context of the prompt, four different contexts were developed, two which concerned the background of the task, A and B, and two which concerned task related specifications, C and D.

A) Consider that you are a retail expert and that the reviews are from an e-commerce site.
B) Consider that the summary serves to assist potential

customers in determining whether to make a purchase or not.

C) The summary should be concise and based solely on the information in the reviews.

D) Include main ideas and essential information, eliminating extraneous language and focusing on critical aspects of the product

Context A serves to provide background about who is writing the summary, making sure it is from the correct perspective. Context B illuminates the target audience, allowing the summary to be adapted for customers. The purpose of context C is to minimize the risk of hallucinations, where the model makes up new, false statements. The purpose of context D is to make sure the summary focuses on the main aspects of the reviews.

When tested the contexts were added at the end of the base-prompt. Each context was initially tested independently. Then the contexts within the two categories of contexts were added to the prompt in pairs, that is AB and CD. Then finally the effect of integrating all contexts into the original prompt was examined, that is ABCD.

3) Clustering: For the initial two steps, shot-prompting and pattern-prompting, the Amazon dataset with a limited amount of reviews per product was used in order to not reach the limit of the model's context window. The purpose of clustering is to handle products with large amounts of reviews, therefore the SweRev dataset was used.

Firstly, the reviews for each product were grouped into smaller clusters, and then each cluster was summarized individually. Secondly, the summarized clusters were combined and passed on to a final step of summarization. The procedure is illustrated in figure 3. To enable more reviews in the instruction, zero-shot prompting was used, with only the base prompt. Three different types of clustering were compared:

1) Rating clustering: the reviews of a product are grouped by which rating that the user gave. Thus, reviews are grouped by customer satisfaction.
2) Random clustering: The reviews are randomly grouped together in five different clusters.
3) No clustering

C. SWOT Analysis

In preparation for the SWOT analysis a literature study was conducted to collect data about the potential opportunities and threats for companies to use summaries of user reviews. To assess the strengths and weaknesses of the generated summaries the results from this study were used. Once all data had been gathered the information was summarized in the SWOT analysis.

D. Evaluation

The summaries were first evaluated using automatic measures BERTScore and ROUGE. ROUGE evaluated word similarity between the reference and candidate summary, using its measures for unigrams (ROUGE-1), bigrams (ROUGE-2) and longest common subsequencec (ROUGE-L). BERTScore instead evalutes the semantic meaning of the summaries, calculating an F1-score. For the Amazon dataset, the model roberta-large was used. For SweRev, bert-base-multilingual was used.

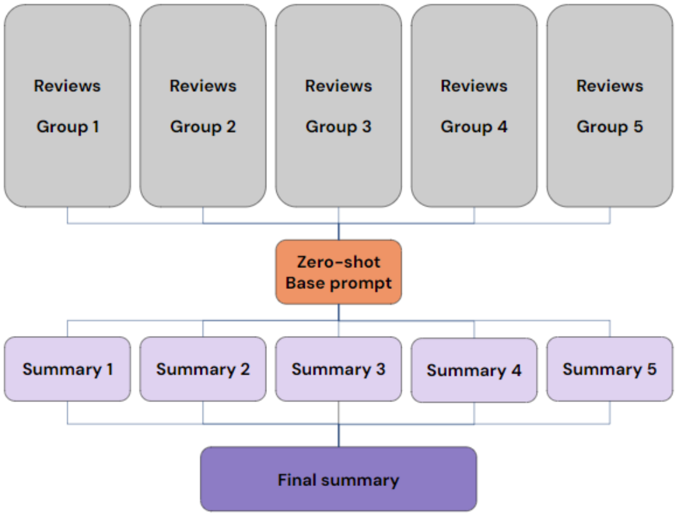


Fig. 3: Visualization of the Summarization of Clustered Reviews

To ensure quality, human evaluation was also performed on a set of summaries. In the Amazon dataset 20 of the products were chosen and their generated summaries were evaluated by the two authors of this paper. For the SweRev dataset, half of the products were evaluated. The summaries were graded according to the measures of factuality, faithfulness and relevance, using a Likert scale ranging from one to five was used, where the numbers indicated weather the criteria was fulfilled: 1) Strongly Disagree, 2) Disagree, 3) Neutral, 4) Agree, 5) Strongly Agree.

Factuality is a measure of whether the generated summary is based on the facts presented in the reviews. An unfactual summary would contain statements that are made up and not mentioned or indicated in the given reviews. To some extent, this measure represents the model's tendency not to hallucinate, which is a common problem regarding LLMs. Faithfulness means to which extent the produced summary represents the opinions expressed in the reviews. A faithful summary should bring up key points presented in the reviews and mirror the majority opinion. If contrasting opinions are expressed, a faithful summary should present this as well. The measure of relevance takes into consideration how relevant the aspects presented in the summary are for the task itself. As the summaries are to review the products, statements regarding completely separate topics, such as other products or delivery, are considered irrelevant and should be excluded.

The length of the generated summaries was also noted as it provides insight to whether the summary makes good use of the word budget.

In order to measure the interrater agreement, weighted Cohen's Kappa was used. Quadratic weighting was used because it more accurately reflects the ordinal nature of Likert scale data by assigning progressively greater penalties to larger disagreements.

V. Results

A. Shot-prompting

The results of the automatic measures from the shot-prompting are shown in table I. The general trends show that the model responds better to prompts including examples of how to respond. The few-shot prompt performs the best in all metrics, however there are very small differences compared to one-shot. All of the three different shot-prompts have an F1 BERTScore in the range of 0.87-0.89 indicating small or insignificant differences in semantics. As will be discussed further in section VI, this result is at par with similar studies covered in section III. The few-shot and one-shot prompts clearly outperforms the zero-shot prompt in both ROUGE-1 and ROUGE-2, and few-shot is also best at ROUGE-L. Compared to the studies discussed in section III, the ROUGE-1 scores are slightly higher meanwhile the ROUGE-L scores are significantly lower. This will also be discussed further in section VI.

Prompt	ROUGE-1	ROUGE-2	ROUGE-L	F1
Zero-shot	0.283	0.044	0.170	0.876
One-shot	0.323	0.064	0.192	0.885
Four-shot	0.327	0.067	0.208	0.887

TABLE I: ROUGE and BERTScore for zero-shot, one-shot and four-shot

In table II, the results from the human evaluation of shot-prompting are shown. In general, the generated summaries recieved high scores in terms of the three measures factuality, faithfulness and relevance, with all averages being above 3, and thus agreeing that the criterias are fulfilled. All three different types of prompts have very high factuality-ratings, whilst faithfulness is a bit lower. Relevance is significantly higher in one-shot and few-shot.

The majority of the generated summaries remained factual to the information presented in the reviews, without hallucinating any new information. The reviews who recieved lower scores in factuality often misinterpreted other information mentioned in the reviews. As an example, one review mentioned problems with the color of a different product, but in the summary it was mentioned as an aspect of the product being summarized.

The summaries did also in all cases present a true and generally faithful and representative summary of the products. Cases where the summaries were only considered somewhat faithful, included when singular relevant opinions were missing in the review. Another common issue was only presenting one perspective of two contrasting opinions. For example if one review said that the product was too small, and another said that the product was too large, often only one of those opinions were mentioned, without including that it was a debated topic.

In relevance, there are some more significant differences, with the zero-shot prompt having a score of 4.13, compared to one-shot and few-shot with scores 4.58 and 4.70 respectively. The main reason for a low score in relevance, was often mentioning topics out of the scope for the specific product, such as customer service or delivery.

Prompt	Factuality	Faithfulness	Relevance	Avg. word-count
Zero-shot	4.55	3.80	4.13	84
One-shot	4.4	3.80	4.58	65
Four-shot	4.65	4.03	4.70	58
Cohen's kappa	**0.24**	**0.036**	**0.33**	-

TABLE II: Human Evaluation Results for shot-prompting including Cohen's kappa for interrater agreement

Another observation during the human evaluation was the length of the generated summaries. The average word count is also presented in table II. The zero-shot prompt generated longer summaries on average, with an average word-count of 84 words. One-shot prompt summaries were about 20 words shorter, and few-shot another 7 words shorter. Thus, the zero-shot summaries often either included more aspects of the product or more commenting language.

The human evaluation show fair agreement in factuality and relevance, but chance agreement in faithfulness, which will be discussed in section VI.

B. Pattern-prompting

Table III shows the automatic measures of the results from pattern-prompting with a four-shot base prompt. The F1 BERTScore is high and very similar between all contexts. The ROUGE scores are also similar across the different contexts. Although context D performed slightly better, achieving the highest ROUGE-1 and ROUGE-2 scores, the differences are too small to definitively determine if any one context outperformed the others. This applies to both ROUGE and BERTScore. The scores are consistent with those presented in previous studies and can be considered adequate. This will be discussed further in section VI. Compared to the original four-shot base prompt shown in table I the BERTScores were more or less equivalent. The ROUGE-1 score was improved for all contexts but context AB, meanwhile no clear general improvement could be seen for ROUGE-2 and ROUGE-L scores.

Context	ROUGE-1	ROUGE-2	ROUGE-L	F1
A	0.337	0.070	0.204	0.888
B	0.332	0.072	0.206	0.888
C	0.325	0.066	0.207	0.889
D	0.346	0.074	0.206	0.889
AB	0.320	0.062	0.205	0.887
CD	0.333	0.067	0.212	0.890
ABCD	0.336	0.070	0.209	0.890

TABLE III: ROUGE and BertSCORE for different context prompts with four-shot base prompt

Table IV shows the human evaluation of the results from pattern-prompting with a four-shot base prompt. In general the results generated very high factuality and relevance scores, with factuality exceeding 4.6 and relevance above 4.4 for all contexts. The faithfulness scores were in general lower, most around a score of 4, with context D performing the best at 4.35. The average word-count for the different contexts are similar and all of them are lower than the average word-count for the different shot-prompts. Compared to the original four-shot base prompt shown in table II the majority of the factuality and faithfulness scores are equal to or higher. However, none of the contexts generated results that scored higher in relevance.

The factuality was in general high for the generated summaries indicating few hallucinations. Summaries generated with context A, which provided background on the task, scored the highest in factuality. Overall, the summaries generated with the different context had a tendency to be too generic and thereby leaving out important details from the reviews. This is the main reason why the summaries generated with the different contexts scored lower in faithfulness. However, compared to the other contexts the summaries generated with context D performed better in this category. Context D requested specifically that the summary should include the main ideas and essential information. Most of the summaries included relevant information and left out potential details such as delivery or complementing products, thus the relevance score was high.

The human evaluation only show slight agreement, which will be discussed in section VI.

Context	Factuality	Faithfulness	Relevance	Avg word-count
A	4.88	4.08	4.65	56
B	4.73	4.15	4.60	53
C	4.75	3.95	4.60	50
D	4.68	4.35	4.45	56
AB	4.65	4.13	4.68	57
CD	4.65	4.05	4.40	50
ABCD	4.63	4.00	4.65	51
Cohen's kappa	**0.15**	**0.15**	**0.14**	-

TABLE IV: Human Evaluation of different Context Prompts with Four-Shot Base Prompt including Cohen's kappa for interrater agreement

C. Clustering

The results using the SweRev dataset are shown in table V. The ROUGE-1-scores for the rating and random clustering are in the same ranges as the few-shot-prompt results previously presented. The ROUGE-2-scores are also similar, whilst the ROUGE-L-scores were a bit lower. However, the results from SweRev without any clustering show both high ROUGE-1 and ROUGE-2 scores, significantly higher than previous results using Amazon. The BERTScore is significantly lower for all three alternatives of clustering, compared to the pattern-prompting and shot-prompting sections. However, the BERT-model used was different between the SweRev dataset (clustering) and the Amazon dataset (prompt formulation).

Prompt	ROUGE-1	ROUGE-2	ROUGE-L	F1
Rating	0.327	0.062	0.165	0.685
Random	0.338	0.080	0.176	0.697
No clustering	0.380	0.094	0.191	0.716

TABLE V: ROUGE and BERTScore of Clustering

All three types of clustering produced summaries with almost perfect factuality, while faithfulness and relevance were slightly lower, as seen in table VI. In general, the summaries did contain relevant aspects of the products, which is shown by the average Relevance score exceeding 3. The common issue was the mentioning of problems with deliveries and customer service, which is out of the scope of this summarization task.

The summaries can be considered generally faithful, as seen in table VI. In the cases where a summary had a low faithfulness score, there was either a lack of specific relevant opinions, or an incorrect representation of the actual opinions presented. Another factor which may affect faithfulness is the summaries' genericity. Certain summaries contain relevant aspects and are factual, and may even be somewhat faithful in terms of representing a very general opinion. However, the information presented is generic so that it does not provide any relevant input to potential readers. Thus, there is a poor use of the word budget, which is mirrored in the faithfulness score. Genericity was in general more common in the Random and Rating clustering, which is also indicated by its lower scores. Similarly to the BERTScore and ROUGE-scores, the no clustering showed the best results also when it comes to human evaluation. An observation regarding the summaries clustered by rating is that all summaries began by presenting the negative aspects of the product, moving on to the positive and then ending with a general comment. In some cases, this caused the negative aspects to be emphasized.

The human evaluation of clustering show fair agreement in faithfulness and relevance, but chance agreement in factuality, which will be discussed in section VI.

Prompt	Factuality	Faithfulness	Relevance	Avg. word-count
No clustering	4.95	3.85	3.55	78
Random	4.95	3.45	3.60	73
Rating	5.00	3.4	3.35	111
Cohen's kappa	**0.00**	**0.30**	**0.32**	-

TABLE VI: Human Evaluation of Clustering including Cohen's kappa for interrater agreement

After performing the human evaluation, there are a few characteristics of the summaries that are not captured in the previously mentioned metrics relevance, factuality and faithfulness, yet who might bring further interesting input. Genericity was previously mentioned as it often correlated with low faithfulness, however a summary can sometimes be faithful and still have a high degree of genericity. Common in generic summaries are also larger parts or sentences that may add context to the summary, but does not actually provide any new information. For example ”There are mixed opinions of this product, which should be taken into consideration before making a purchase”. Such commenting language may help the summaries to a certain extent, however, if the summary almost solely consists of similar statements it becomes generic. The commenting language makes a poor use of the word-budget, risking unfaithfulness. Again, the clustered summaries had a higher rate of commenting language.

Another feature is the length of the summary, shown in table VI. All three types of summaries had higher average word count, compared to the Amazon summaries. Specifically the no clustering summary shows a very high word count, at 111 words.

D. The literature study for the SWOT analysis

Sharma and Kumar published a study in 2023 where they assessed the impact of online product reviews on consumer purchasing decisions through a survey. They found that the majority of consumers rely on customer reviews in their purchase decisions online. They also emphasize the importance of using customer reviews to improve products and enhance customer satisfaction, as well as their importance for brand management [18].

In their book *Customer Communities: Engage and Retain Customers to Build the Future of Your Business*, Mehta and Van Lieshout also emphasize that customers' purchasing decisions today are highly influenced by reviews of other customers. They mention that numerous studies have shown that people trust companies less than ever, meanwhile feedback from other customers are increasingly used to inform purchasing decisions. Further, they also point out the importance for the product teams to engage in a product feedback loop and use the information provided by customers to improve the products. However they argue that many companies struggle with doing this since customer feedback can be fragmented across different sources which makes it challenging to retrieve aggregated feedback. This makes it difficult for product teams to understand what limitations products currently have from the customers point of view [19]. Lowdermilk and Rich similarly suggests that customer feedback can be used for business improvement in their book *The customer-driven playbook: converting customer feedback into successful products*. They argue that customer feedback such as reviews can be used as a tool to validate that ideas create value for customers [20].

McHale and Garulay emphasize the importance of using customer reviews for marketing. In their book they highlight a study which found that the majority of customers spend ten minutes or more reading customer reviews before making purchasing decisions. By capturing and leveraging positive customer reviews they argue that companies can drive consumer demand and increase sales. However, they also point out that online reviews are not risk-free from a legal perspective. When using customer reviews for marketing there are strict legal requirements that must be complied in order to avoid liability [21].

Another publication that also argues for the importance of customer reviews is a study conducted by Lee et al. in 2006. In their study they concluded that review management is of high importance for online sellers and that especially negative reviews have a powerful impact on product attitude. They argued that sometimes reviews can be unhelpful and unreasonably negative which can have a negative influence [22].

VI. Discussion

A. Formulation of Prompts

From the results of the Shot-prompting, and pattern-prompting, it is clear that the most significant improvement comes from shot-prompting. Few-shot provides the model with further context on how the answers should be formulated and to some extent, it allows the model to be fine-tuned using chosen examples of data. This is supported both by the automatic measures ROUGE and BERTScore ,as shown in table I as well as the human evaluation in table II. Moreover, the zero-shot summaries were longer, which can be explained by the lack of a reference summary, as one-shot and few-shot will adapt their output length to match the user's examples. As mentioned in the results, the longer zero-shot summaries either resulted in more aspects of the product discussed, or more commenting language. More aspects could make the summary more faithful, but it also risks including irrelevant aspects. More commenting language can make the summary more generic and could be a sign of poor use of the word budget, thus lowering faithfulness.

All shot and context pattern prompts generated summaries which on average had a F1 BERTScore in between 0.87-0.89, shown in table I and table III. This indicates that the different modifications of the prompt resulted in small, if any, semantic differences. These results are in line with other papers on opinion summarization mentioned in section III, Bhaskar et al. also reached a BERTScore of 0.89 when generating summaries with the same dataset used in this study, Amazon (FewSum) [16]. Compared to other studies the ROUGE scores of the summaries somewhat differed. Bhaskar et al. experimented with different pipelines with zero-shot prompts. Their zero-shot ROUGE-1 scores were slightly lower, around 0.26-0.27, although the ROUGE-L scores were significantly higher, at 0.23-0.24. One of their models also involved few-shot prompts. Its ROUGE scores were slightly higher than ours, ROUGE-1 at 0.33 [16].

In the human evaluation (table II), the three different prompts all have high factuality, which can be interpreted as few hallucinations in the generated summaries. This is important for the potential application of this model. Often hallucinations are more common when the model is not provided with enough information and context, causing it to guess which response the user wants. In this case of opinion summarization, the model can rely heavier on the reviews and base the output on them.

Moving on to the context pattern prompts, there seemed to be little difference in performance between the different contexts. Context A and B, regarding the background information of the task, does not seem to improve the generated summary to any large extent. One possible reason could be that it is already clear from the reviews that they are regarding products on an e-commerce site. Context D sticks out both when it comes to ROUGE-1 in table III, as well as faithfulness in table IV. The context specifies that the summary should "Include main ideas and essential information, eliminating unnecessary language and focusing on critical aspects of the product.", which aligns well with the faithfulness metric. The improved faithfulness score therefore indicates that the additional guidance provided by the context positively influenced performance in the desired direction. However, one would also expect that this context could improve relevance, which was not the case. As low relevance often correlated to the mentioning of irrelevant aspects, if such aspects are mentioned frequently in the reviews, they may still be included by the model. To avoid this issue, one could try specifying in the prompt that such aspects should be ignored.

A general issue in the summaries generated from the Amazon dataset is how conflicting opinions are presented. As mentioned in the results, the generated summary only contained one side of the opinion, causing unfaithfulness. This was a general issue in all the given context pattern prompts.

B. Clustering

The results from the clustering showed that no clustering is to prefer, yet the rating and random clustering still show moderate performance. The no clustering had both the highest ROUGE-scores as well as the best scores in faithfulness and relevance, the reviews are longer, as shown in table V and VI. A summary of over 100 words may be too long. An interesting topic for future studies would be to investigate if the no clustering maintains the higher scores if it has a tighter word budget.

Notable is that with no clustering the SweRev dataset outperforms Amazon when it comes to ROUGE and factuality, and has similar results in faithfulness and relevance, according to tables V and VI. One theory is that the significantly larger amount of reviews (100+) creates a better representation of which opinions are relevant, as well as provides enough content to avoid hallucinations. BERTScore is significantly lower with SweRev, probably explained by the BERT-model used. Roberta-large, the model for Amazon had been trained specifically for the English language, whilst the bert-multilingual is a smaller model adapted for multiple languages, and thus it may not perform as well in a Swedish context.

The random clustering caused generic summaries, as too much information was lost in the summarization. After clustering, there are 5 summaries that probably are quite similar because they are based on a random selection of mixed reviews. When they are turned into one final summary, less common but still relevant opinions risk being excluded, resulting in lower scores in faithfulness and relevance. The same problem also occurred with the rating clustering, however not to the same extent. The main cause of the lower faithfulness for the rating clustering is more likely to be the order of which the aspects were presented. As the negative aspects were presented first, they were also emphasized more in the summary. As a reader, even though the rest of the review is positive, the first impression will affect you more. The reason for the summary starting with the negative opinions, is probably because that was the order in which it was given to the model, as the model may reinforce patterns in the beginning of the prompt. First, the 1-star rating summary was presented, then the 2-star and so on. The model may also be more sensitive to negative sentiments. Thus, an interesting future approach would be to either start from 5-star ratings, or use a random order starting from the middle, to see if this affects the produced summary.

C. Reliability of the Results

In this paper, we have tried to quantize the human evaluation by using a 5 grade scale of the measures faithfulness, factuality and relevance. As with reviews, when abstract matters are quantized into a discrete spectra of numbers, information will be lost. A summary with a score of 1 on all three measures is a poor summary, and a summary with a score of 5 in all measures is perfect, but the scale in between is not as clear. A summary can still be good enough, even though all scores are not perfect. If the scores are 3 or above, the summary can still be considered somewhat representative, as it is without irrelevant aspects, it is not unfactual and not unfaithful. In application, it is important to consider on which, if any, of these aspects a score lower than 5 is acceptable. If everything presented in the summary is true, although some aspects are missing, it may still provide good insight.

Genrally, the weighted kappa-scores were low, indicating only slight to fair agreement, according to Landis and Koch's scale [23]. The summaries generally recieved high scores, for example factuality of the clustered summaries was almost perfect. If one score is more dominant, it may cause a lower kappa due to higher chance agreement, explaining table VI. A reason for the low kappa-scores may also be that the measures are subjective and can be interpreted differently by different raters. In order to secure reliability in further studies, the criterias for each measure should be made clear, and more raters should be included in the survey.

D. SWOT Analysis

The SWOT analysis was conducted to answer the question whether the generated summaries of customer reviews provides value for companies. The internal part of the analysis was based on the results from this study and concerns the strengths and weaknesses of the generated summaries. The external part was based on the literature study presented in the results and concerns the opportunities and threats for a company to implement and use summaries of customer reviews.

1) Strengths: One of the key strengths of using generated summaries by LLMs is the efficiency. The summaries can be generated in large volumes very quickly, thereby saving time and effort in extracting insights. This provides scalability for big and growing datasets. Additionally, customization features of LLMs allows the generated summaries to be tailored to specific business needs, offering flexibility and adaptability. Compared to alternatives the summaries are also relatively cheap to produce which is beneficial for companies with a large number of products and customer reviews. Based on the results in this thesis the summaries seems to score especially well when it comes to factuality, which is of high importance. Moreover, the summaries were of high quality in general according to human evaluation.

2) Weaknesses: One of the greatest weakness of the summaries is the difficulty in portraying contrasting opinions, as well as struggles with misinterpretations or oversimplifications of customer feedback. Faithfulness had lower scores, indicating difficulties in making a perfectly representative summary To identify and correct errors in the generated summaries would require human supervision, which would be substantially more costly. Moreover, for longer and bigger quantities of reviews the summaries would have to be made through clustering. The results indicates that this would result in a lower level of faithfulness and relevance, reducing quality.

3) Opportunities: There are several opportunities with implementing summaries of customer reviews. First of all providing customers with summaries of previous customers' reviews enables them to make a more informed decision without having to take the time and effort to read each and all of the reviews on their own. Research has shown that customers value other customers' reviews highly and providing this information in a synthesized format could make the information more accessible. This could potentially increase customer satisfaction since findings from the literature study indicated that customers tend to spend substantial time reading and reflecting on customer reviews.

Further, the summaries could be used by companies internally for business development based on customer feedback. The feedback could be used for product development to improve existing products or create new ones. Publications reviewed in the literature study emphasized the importance of integrating customer feedback in the production process but also argued that many companies struggled to do so due to the fragmented nature of the feedback. Since the summaries are aggregated versions of the customer reviews they could potentially solve this problem and thereby add value in the production process. Another identified use case is marketing. Since the summaries are aggregated versions of the customer reviews they could potentially provide more value than using a singular review. Moreover, customer feedback can also be used to obtain insights and thereby enable informed decision-making and strategy formulation. The summaries can be used to validate that ideas create value for the company.

4) Threats: The research reviewed in the literature study emphasized that customers tend to trust customer reviews due to the fact that they are written by other customers and not the company. Considering this, one potential threat is that the credibility of the summaries may be perceived as low since they are provided by the company itself. Moreover, research reviewed in the literature study showed that another risk is that customer reviews can sometimes be unhelpful or unreasonably negative. Including such negative reviews in a summary risks having a negative impact on sales. The same reasoning applies if the summaries were to be used internally for business development. Misleading summaries would in this case risk that the wrong decisions and projects were prioritized. Another threat is the cost of producing the summaries. Given that e-commerce companies can have thousands of different products with multiple reviews there is a risk that producing the summaries can become costly. Another potential threat is legal constraints, for example in marketing. The regulations may vary in different countries and need to be considered before any implementation. Failure to comply with these regulations could result in legal repercussions and financial penalties.

Discussion of SWOT analysis: Based on the SWOT analysis the generated summaries of customer reviews could provide value for companies. First of all the summaries would make the information from reviews more accessible to customers. Second of all the summaries could also be used internally for business development in areas such as product development, marketing and strategy. All though the summaries did not have perfect scores their quality was in general good. By using the generated summaries the companies should be able to capture these opportunities and thus providing value for the company.

The threat of costly production of summaries is mitigated by the strength of using LLMs for summarization, allowing large volumes at low costs. Another threat is the potenital lack of credability for company-provided summaries. This threat may be reduced by explicitly stating that the summaries are automatically generated. However, there is a risk that consumers are aware of that such generated summaries can be modified, thus not increasing credibility. The weaknesses of the summaries is their varied quality of faithfulness and representativeness, which may undermine its value. It is essential to consider which requirements the summaries must follow, for example regarding factuality, faithfulness, relevance and genericity. What is good enough? Moreover, inaccurate summaries increases the threat of legal complications.

E. Future studies

For future studies the effects of summarizing reviews through clustering could certainly be investigated further, evaluating different methods for iterative summarization. With even larger numbers of reviews, one could increase the amounts of clusters as well as create further sub-clusters, however this risks increasing the iterations of summarizations significantly. A way to mitigate this may be to classify reviews as either helpful or unhelpful, and then only include the helpful reviews in the summary. However an optimal method for this classification would have to be investigated further.

Another aspect that was not considered in this study was if reviews of a product can change over time. For instance, a company might switch suppliers or make slight design changes to a product, whichcould affect customers perception of a product and be reflected in the reviews. Furthermore, events and trends in the world over time can also affect customers' perceptions of products. A topic for future studies is whether more recent reviews should be prioritized in a summary.

VII. Conclusion

In conclusion, the results of this study showed that prompt engineering has a high capability to generate a representative text summary of reviews. Although there were some variations in the summaries' quality, the results were predominantly positive according to faithfulness, factuality and relevance. Shot-prompting proved to enhance the model's performance, whilst pattern-prompting showed little difference. Using sub-task prompts and clustering did not improve performance, however this topic needs further investigation using prompt engineering.

By conducting a SWOT analysis it was concluded that the generated summaries could provide value for companies. The analysis identified several opportunities, such as making review information more accessible to customers, enhancing their purchase decisions, and using summaries for internal business development in product development, marketing, and strategy. Despite some quality imperfections, the summaries were generally considered usable. Further, the fact that the summaries can be produced in large volumes at low cost were identified as strengths, mitigating the threat of high production costs. However, risks such as potential inaccuracies and factual errors in the summaries could undermine their value and pose legal complications.

References

[1] Micael Dahlen and Helge Thorbjørnsen. *Sifferdjur: Hur siffrorna styr våra liv*. Stockholm: Volante, 2021. ISBN: 978-91-7965-154-1.

[2] Nachiketa Sahoo, Chrysanthos Dellarocas, and Shuba Srinivasan. “The Impact of Online Product Reviews on Product Returns”. In: *Information Systems Research* 29 (June 2018). DOI: 10.1287/isre.2017.0736.

[3] Regina Frei, Lisa Jack, and Stephen Brown. “Product returns: a growing problem for business, society and environment”. In: *International Journal of Operations & Production Management* 40.10 (June 2020), pp. 1613–1621. URL: https://eprints.soton.ac.uk/440529/.

[4] Juan-Manuel Torres-Moreno. *Automatic Text Summarization*. 2014. ISBN: 978-1-84821668-6. DOI: 10.1002/9781119004752.

[5] Bing Liu. “Opinion Summarization”. In: *Sentiment Analysis and Opinion Mining*. Springer International Publishing, 2012, pp. 91–97. ISBN: 978-3-031-02145-9. DOI: 10.1007/978-3-031-02145-9_7. URL: https://doi.org/10.1007/978-3-031-02145-9_7.

[6] Masato Hagiwara. *Real-World Natural Language Processing*. Manning Publications, 2021. ISBN: 978-1-61729642-0. URL: https://www.oreilly.com/library/view/real-world-natural-language/9781617296420.

[7] Dan Jurafsky and James H. Martin. *Speech and Language Processing*. [Online; accessed 7. Mar. 2024]. 2024. URL: https://web.stanford.edu/~jurafsky/slp3.

[8] Hai Dang et al. *How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models*. 2022. arXiv: 2209.01390 [cs.HC].

[9] Lei Huang et al. *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. 2023. arXiv: 2311.05232 [cs.CL].

[10] Richard Whittington et al. *Exploring Strategy, Text and Cases*. English. Ed. by Petya Koleva. Vol. 12th. Petya Koleva was responsible for writing questions/ answers for Revel – an online business simulation platform for the textbook. Pearson Education, 2019. ISBN: 978-1292282459.

[11] Bo Tonnquist. *Projektledning*. Sjunde upplagan. Stockholm: Sanoma utbildning, 2018. ISBN: 9789152354988.

[12] Jacob Cohen. “A Coefficient of Agreement for Nominal Scales”. In: *Educational and Psychological Measurement* 20 (1960), pp. 37–46. URL: https://api.semanticscholar.org/CorpusID:15926286.

[13] Tianyi Zhang et al. *BERTScore: Evaluating Text Generation with BERT*. 2020. arXiv: 1904.09675 [cs.CL].

[14] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[15] Nathaniel Benham, Siqi Gao, and Yiu-Kai Ng. “A Hybrid Approach for Summarizing User Reviews Based on KL-Divergence and Deep Learning”. In: *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*. 2022. DOI: 10.1109/ICTAI56018.2022.00054.

[16] Adithya Bhaskar, Alexander R. Fabbri, and Greg Durrett. *Prompted Opinion Summarization with GPT-3.5*. 2023. arXiv: 2211.15914 [cs.CL].

[17] Daphne van Zandvoort et al. “Enhancing Summarization Performance through Transformer-Based Prompt Engineering in Automated Medical Reporting”. In: (2023). DOI: 10.48550/arXiv.2311.13274. eprint: 2311.13274.

[18] Sunil Sharma and Satish Kumar. “Insights into the Impact of Online Product Reviews on Consumer Purchasing Decisions: A Survey-based Analysis of Brands’ Response Strategies”. In: *Scholedge International Journal of Management Development ISSN 2394-3378* 10 (Oct. 2023), p. 1. DOI: 10.19085/sijmd100101.

[19] Nick Mehta and Robin Van Lieshout. *Customer Communities : Engage and Retain Customers to Build the Future of Your Business*. eng. First edition. Hoboken, New Jersey: John Wiley Sons, Inc., 2024. ISBN: 1-394-17212-5.

[20] Travis Lowdermilk and Jessica Rich. *The customer-driven playbook : converting customer feedback into successful products*. eng. First edition. Beijing, [China: O’Reilly, 2017 - 2017. ISBN: 1-4919-8122-9.

[21] Robert. McHale and Eric. Garulay. *Navigating social media legal risks safeguarding your business*. eng. 1st edition. Indianapolis, Ind: Que, 2012.

[22] Jumin Lee, Do-Hyung Park, and Ingoo Han. “The effect of negative online consumer reviews on product attitude: An information processing view”. In: *Electronic Commerce Research and Applications* 7.3 (2008). Special Section: New Research from the 2006 International Conference on Electronic Commerce, pp. 341–352. ISSN: 1567-4223. DOI: https://doi.org/10.1016/j.elerap.2007.05.004. URL: https://www.sciencedirect.com/science/article/pii/S1567422307000415.

[23] J. Richard Landis and Gary G. Koch. “The Measurement of Observer Agreement for Categorical Data”. In: *Biometrics* 33.1 (1977), pp. 159–174. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2529310 (visited on 05/31/2024).

Lisa Etzell and Nora Hulth are both students at the Royal Institute of Technology in Stockholm, Sweden, studying Industrial Engineering and Management with a specialization in Computer Science and Machine Learning. Both students have made equal contributions to the report, and thus no part of the report can be credited to a specific author.

TRITA-EECS-EX-2024:417
Stockholm, Sverige 2024

www.kth.se