Zero-Shot Product Description Generation from Customers Reviews

Bruno Gutierrez[a], Jonatas Grosman[b], Fernando A. Correia[c] and Hélio Lopes[d]
Department of Informatics, PUC-Rio, Marquês de São Vicente, 225 RDC, 4th floor - Gávea, Rio de Janeiro, Brazil
{bgutierrez, jgrosman, fjunior, lopes}@inf.puc-rio.br

Keywords: Text Generation, Data Mining, Generative Artificial Intelligence, Large Language Model, E-Commerce.

Abstract: In e-commerce, product descriptions have a great influence on the shopping experience, informing consumers and facilitating purchases. However, creating good descriptions is labor-intensive, especially for large retailers managing daily product launches. To address this, we propose an automated method for product description generation using customer reviews and a *Large Language Model* (LLM) in a *zero-shot* approach. Our three-step process involves (i) extracting valuable sentences from reviews, (ii) selecting informative and diverse content using a graph-based strategy, and (iii) generating descriptions via prompts based on these selected sentences and the product title. For our proposal evaluation, we had the collaboration of 30 evaluators comparing the generated descriptions with the ones given by the sellers. As a result, our method produced descriptions preferred over those provided by sellers, rated as more informative, readable, and relevant. Additionally, a comparison with a literature method demonstrated that our approach, supported by statistical testing, results in more effective and preferred descriptions.

1 INTRODUCTION

Product descriptions are an important part of the customer experience in online shopping. By providing detailed information about product features, functionality, and specifications, they enable consumers to better understand their purchases and make informed decisions. Despite their importance, manual creation of product descriptions is often costly and, in many cases, prohibitive due to the constant demand for new content. Among the various machine learning applications in e-commerce (Hütsch and Wulfert, 2022), automatic product description generation stands out as a solution that not only addresses this demand but also offers significant opportunities, as it has been verified how employing a strategy for automated description generation increased sales (Novgorodov et al., 2020; Zhang et al., 2022).

Among the techniques for generating descriptions, there is considerable diversity in the data used, including attributes provided by sellers (Wang et al., 2017), product titles (Zhan et al., 2021), and even advertising slogans (Zhang et al., 2022). Our method aligns with (Novgorodov et al., 2019) leveraging customer reviews as the primary source. As noted by the authors, reviews offer a unique perspective on the interaction between the customer and the products, providing authentic insights that often go beyond the information provided by the manufacturer or seller.

Mining information from reviews has proven effective in various contexts, such as analyzing revisit intentions in the hotel industry (Christodoulou et al., 2021) and identifying negative feedback causes to assist developers in public service apps (Pedrosa. et al., 2023). In our case, considering products with abundant comments, there is enormous potential to mine valuable information, making reviews a rich source for generating descriptions.

To enhance the description generation process, we employ recent advances in natural language text generation through Large Language Models (LLM). Our method combines customer reviews and the product title with the synthesis and articulation capabilities of LLMs. Our method consists of three steps: first, extracting valuable sentences from reviews and classifying those suitable for a product description. Next, we apply a graph-based strategy to select a diverse and informative set of sentences addressing multiple product aspects. Finally, we configure and execute a prompt that uses these sentences and product title in a zero-shot way. This way, we expect the generative model to select, group, and convey part of this content

[a] https://orcid.org/0009-0001-2064-1572
[b] https://orcid.org/0000-0002-1152-5828
[c] https://orcid.org/0000-0003-0394-056X
[d] https://orcid.org/0000-0003-4584-1455

Gutierrez, B., Grosman, J., Correia, F. A. and Lopes, H.
Zero-Shot Product Description Generation from Customers Reviews.
DOI: 10.5220/0013280300003929

In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Volume 1*, pages 763-770
ISBN: 978-989-758-749-8; ISSN: 2184-4992

in an informative, concise, and readable description.

To evaluate our method, we compared the generated descriptions with those posted by sellers involving 30 evaluators in the process. The results demonstrated that our proposal produces descriptions consistently preferred over the original ones in general and in specific criteria, such as readability and informativeness. Furthermore, we replicated a method found in the recent literature (Novgorodov et al., 2019), which served as our baseline. The comparison revealed the evaluators' preference for the descriptions generated by our method, indicating that using an LLM in the description generation process is highly beneficial. Our work's main contributions are the following:

- We propose a new automatic product description generation method that was overall preferred over real descriptions posted by sellers and can be easily applied to other product categories beyond those tested.
- We demonstrate how an LLM can generate readable product descriptions articulating the information contained in reviews.
- We propose a new reproducible evaluation format, comparing the generated product descriptions with the original ones posted by sellers.

The remainder of this document is organized as follows. Section 2 reviews key works in the literature on product description generation. Section 3 details our proposed method, outlining each stage. In Section 4, we discuss sub-steps and experiments conducted to define them, starting with the two datasets that were the basis of this work. In Section 5, we evaluate the generated descriptions, aiming to assess the LLM impact. Finally, Section 6 addresses the limitations of our method and proposes directions for future research.

2 RELATED WORK

Initial techniques for product description generation relied on extractive approaches, where product information was combined with pre-existing text. Although creative, those approaches generally suffer from limitations in discourse structure, such as readability issues. After advances with the Transformer framework, new approaches were no longer restricted by discourse structure, requiring, however, larges amount of training data and also suffering from generalization issues. Following the success of pre-trained models, then, less data was need and generalization improved.

Among extractive approaches, (Wang et al., 2017) uses statistical templates and product attributes to create fluent descriptions. (Elad et al., 2019) proposes a personalized summarization method, predicting customer personality to select and condense descriptions into three tailored sentences. (Novgorodov et al., 2019) extracts sentences from customer reviews that describe a product's features, usage, or benefits, defining suitable sentences as those that could be included in a product description without modification.

Transformer-based methods have advanced product description generation. (Chen et al., 2019) proposes KOBE, a model combining product aspects, user categories, and external knowledge to create personalized descriptions. (Liang et al., 2024) merges user attributes with product titles, addressing faithfulness with a copy mechanism. (Zhan et al., 2021) develops the Adaptive Posterior Distillation Transformer, incorporating reviews, titles, and attributes to focus on user-relevant aspects. (Wang et al., 2022) improves quality by integrating auxiliary knowledge like slogans and product details, denoising content with a variational autoencoder.

Pre-trained models further enhance text generation. (Nguyen et al., 2021) adapts GPT-2 for descriptions via pre-training and fine-tuning, generating small, aspect-specific texts that combine into cohesive descriptions. Similarly, (Zhang et al., 2022) alternates between a Transformer-pointer network and a pre-trained language model, leveraging titles, attribute-value pairs, and slogans, trained on expert-created datasets to handle data scarcity.

3 METHODOLOGY

We extensively based our method on (Novgorodov et al., 2019), adopting their use of suitable sentences from reviews as a source of product information and their definition of product description proposed by the authors: a presentation of what the product is, how it can be used, and why it is worth purchasing, with its purpose being to provide details about the features so costumers are compelled to buy. We also adapted their methods for selecting and ranking sentences from reviews.

To enhance the process, we leveraged the advancements in pre-trained generative models, specifically *"gpt-3.5-turbo-0613"*, for generating fluid and informative descriptions without structural limitations. Using the model in a zero-shot manner allowed us to bypass additional training while benefiting from its generalization capabilities (Brown, 2020). Our method is divided into three macro steps, depicted in

Figure 1 and explained in the following subsections.

3.1 Sentence Extraction

To generate a high-quality description for a given product, multiple reviews covering various characteristics are needed. Given this input, the first step is to split the reviews into sentences and identify those that might be interesting for generating a description, referred to as candidate sentences. For that, we relied on the solution proposed by (Novgorodov et al., 2019), which consists of an initial filtering sub-step followed by the classification of sentences.

In the first sub-step, we applied three filters proposed by the authors, aiming to remove sentences with a very low probability of containing relevant information presented in an objective way. The targeted sentences were short (less than four words), personal, or related to advertisements sentences, with the last two filters done via unigram identification. After, the next sub-step for extracting candidate sentences is classifying the filtered sentences suitable for a product description, a concept discussed in 2. We delve deeper into the experiments to train our classifier in Section 4.

3.2 Sentence Selection

Once we have a set of suitable sentences covering different aspects of the product, the second step of our method is to select which ones will be used, which implies selecting what information to present. For this, we followed the approach proposed by (Novgorodov et al., 2019), basing our choice on the ranking and diversification of sentences.

First, to obtain the vector representation of sentences, we choose the open-source model *all-mpnet-base-v2*[1] based on the "Massive Text Embedding Benchmark" (Muennighoff et al., 2022), as it was recommended for the "Semantic Textual Similarity" task. Then, to rank the most important sentences, either because of the product aspect they address or the richness of details they provide, we adopted the same method as the authors, LexRank (Erkan and Radev, 2004).

As the next sub-step, we followed the authors' approach to sentence diversification, using cosine similarity to discard similar sentences that don't add much new information. For this, we adopted the maximum similarity threshold reported by the authors based on the 90th percentile of a set of descriptions curated by domain experts. After that, we followed their proposed greedy approach, adding sentences in the previously defined order as long as they were not similar to already selected ones.

3.3 Description Generation

We propose generating product descriptions by summarizing the selected sentences using the *"gpt-3.5-turbo-0613"* model in a *zero-shot* manner, an approach that avoids training and data collection while remaining generalizable to other domains. An immediate challenge lies in defining the prompt, which consists of an instruction and its content. The instruction is the command assigned to the model to generate a product description. The content comprises the selected sentences along with the product title as context.

Our first step in the search for our instruction was to explore a set of candidates. Also, realizing that sentence length is a relevant issue, as longer descriptions, although possibly more informative, demand a higher effort from the reader, we choose to limit ours texts, taking as reference real descriptions posted by sellers. We detail these experiments in Section 4.3.

Addressing the content aspect, we set out to examine how many sentences to generate descriptions from. Since the idea of our method is to enhance the informativeness of descriptions based on reviews, we expected to obtain richer descriptions by using more sentences. To confirm how many, we conducted an experiment detailed in Section 4.4.

4 EXPERIMENTS

Here we detail the experiments conducted to define the sub-steps of our method.

4.1 Datasets

We have based our work on two datasets. The first, the Amazon Review Data (Ni et al., 2019), is a public dataset containing reviews, user-written texts reporting costumer experience with the product, accompanied by other information such as review title and rating score. Besides products reviews, this dataset provided us their original product descriptions. We experimented with a subset of 13k products with abundance of reviews from the "*Clothing, Shoes, and Jewelry*" category. Subset details are on Table 1. Also, to limit the number of words generated in our descriptions, we established a threshold based on the observed 95th percentile of the dataset descriptions,

[1] More information at https://huggingface.co/sentence-transformers/all-mpnet-base-v2

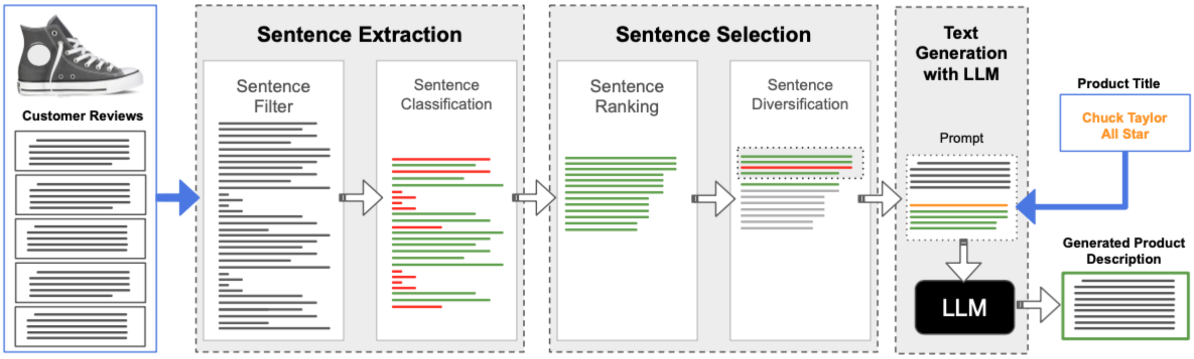

Figure 1: Overview of the proposed method.

which was 150 words. Lastly, From the initial filtering step commented in 3.1, we excluded a total of 71% of the user reviews sentences from the dataset.

Table 1: Reviews data statistics. We present the average and standard deviation for the product reviews and descriptions.

	Reviews	Descriptions
Sentences	2.9 ± 2.8	3.8 ± 3.2
Words	34.2 ± 42.9	57.1 ± 48.4
Words p. sentences	11.8 ± 8.8	15.0 ± 10.8

The second dataset was published by (Novgorodov et al., 2019), also in the public domain, and contains almost 50 thousand sentences extracted from product reviews on an e-commerce site belonging to two categories, *Fashion* and *Motors*, half of each domain. Sentences in this dataset are classified as suitable or not to be part of a product description, a concept proposed by the authors as discussed in 2. In both domains, the percentage of suitable sentences was about 8%. We used this dataset to train a sentence classifier employed in our method.

4.2 Sentence Classification

We trained a classifier to select suitable sentences using the dataset discussed in 4.1. We experimented with a few traditional models for text classification, Naive Bayes, Random Forest, and XGBoost with the TF-IDF textual representation. We also experimented the Ada model from OpenAI, a smaller variant of the GPT-3. We trained each model in each category, separating 20% of the category data for testing and also used the Area Under the ROC Curve (AUC) metric. Results are presented in Table 2.

We obtained the best results using the Ada model, which performed significantly in both categories. This result was also superior to the best model reported in (Novgorodov et al., 2019), which obtained an AUC of 0.924. Even so, we can also observe that simpler models also had a good performance, as the Naive Bayes, indicating that cheaper models can be a viable option.

Table 2: AUC in the classification of suitable sentences for a product description. The fourth column presents the generalization score (GS), where we trained the models on one category and tested on the other, and then took the average and standard deviation.

Model	Fashion	Motors	GS
NB	0,91	0,91	0,87 ± 0,002
RF	0,89	0,90	0,86 ± 0,006
XGB	0,87	0,89	0,84 ± 0,008
Ada	0,94	0,95	0,91 ± 0,001

Furthermore, to explore the generalization capacity of the models to other domains, we tried to evaluate the classifiers again but now using the test set of the other domain. That is, we selected the model trained on the *Fashion* set and measured its performance on the *Motors* test set, and vice-versa. The results can be seen in the third column of Table 2.

Overall, models seemed to generalize well, with a limited decrease in performance. For instance, the Ada model trained on a different domain had a decrease of only 0.03 and 0.04 when compared to versions trained on its own domain, indicating that, to some extent, the nature of review sentences from different categories is similar. Since this step is the only one in our method that depends on a supervised model, we verify that our method can be applicable to other domains beyond those for which we have annotated data with reasonable performance.

4.3 Model Instruction

Initially, to define our set of candidate instructions, our first task was transforming our adopted product description definition from (Novgorodov et al., 2019) into an instruction. We then proposed three leaner variations to compare how the results could vary. We arrived at the following instructions:

- **Base Instruction:** *“You will be provided with a product title and sentences extracted from reviews. Your task is to write a product description. We refer to a product description as a textual presentation of what the product is, how it can be used, and why it is worth purchasing. The purpose of a product description is to provide customers with details about the features and benefits of the product so they are compelled to buy.”*
- **Variation 1:** *“You will be provided with a product title and sentences extracted from reviews. Your task is to generate a product description based only on the title and extracted sentences. The product description should only contain objective, relevant, and positive information from the reviews”*
- **Variation 2:** *“Write an objective and informative product description based only on the product title and received sentences extracted from reviews”*.
- **Variation 3:** *“Write a product description based only on the following title and sentences extracted from reviews”*

Still, in an initial exploratory setting, we observed that all of the instructions were generating descriptions significantly longer than the ones intended and realized that additional commands to limit the text length were necessary. For that, we experimented with different commands, adding them at the end of each instruction and generating for each of the 100 descriptions.

None of the experimented commands worked as intended in all cases, so we selected the one that exceeded the stipulated limit the fewest times, which was “The product description cannot contain more than 150 words.”. Then, for each instruction, we tried reducing the specified word limit until all the generated texts contained less than 150 words, as originally intended. With this, we arrived at three candidate instructions, as one did not meet our stopping condition.

Finally, to select one we conducted a qualitative evaluation structured as follows: for a set of 80 random products, we generated trios of descriptions using the 3 instructions, and from the 80 trios, we formed 120 equally distributed pairs. We then asked one annotator to choose for each pair a description to replace the original, also giving the option of a tie.

As we observed that one instruction was consistently better, being preferred at 45% and selected as worse only 9% of the times, significantly better than the second best instruction (35% best and 19% worst). That was our selected instruction:“Write an objective and informative product description based only on the product title and received sentences extracted from reviews. The product description cannot contain more than 75 words”

Table 3: Best-worst Scaling results by varying the number of sentences in the prompt for 100 products. Annotators were presented in random order with 3 descriptions of the same product generated from different amounts of sentences. In each case, the best and worst were chosen, and the final score was calculated from their difference.

No. of Sentences	Best (%)	Worst (%)	Score
20	45	17	0.28
10	32	36	-0.04
5	23	47	-0.24

4.4 Content

Once defined the instruction, our final step was determining the number of sentences to be summarized in a description. Although including more sentences might produce more informative descriptions, due to the nature of the generative model we cannot precisely predict its effect. Beyond the cost of extra tokens, there is the possibility of issues such as hallucinations, and also doubts about how many sentences can actually be incorporated into a description.

To assess that,, we develop three scenarios in which different quantities of the ranked sentences were provided to the generative model, 5, 10 and 20. We then generated trios of descriptions for 100 products and conducted a qualitative evaluation with 10 annotators. As with the studies by (Liu and Lapata, 2019), (Puduppully et al., 2019), and (Amplayo et al., 2021), we employed the *Best-Worst Scaling* technique, asking annotators to select the best and worst descriptions among the three provided in a random order for 10 products each.

The experiment results can be observed in Table 3. We observed that the more sentences used, the better the method was evaluated, with the 20-sentence method being widely chosen as the best 45% of the time, and as the worst only 17%. In contrast, the 5-sentence method was most frequently chosen as the worst in 47% of the instances. Based on these findings, we proceeded with the 20-sentence scenario in our method, understanding that it was worth the extra tokens.

5 EVALUATION

We evaluate the descriptions generated by our method in two ways. First, we sought to understand whether the generated texts were able to leverage the content of the reviews, using the traditional summarization metric ROUGE. Second, to gain a deeper understanding of the generated descriptions across multiple criteria, we used a 7-point Likert scale (Amidei et al., 2019) with the collaboration of 30 evaluators who analyzed pairs of descriptions for a total of 150 products of the "*Clothing, Shoes, and Jewelry*" category.

5.1 Content Influence with ROUGE

To explore how the selected sentences are reflected in descriptions, we used the ROUGE metric comparing the generated text with the content that comprises the prompt (the 20 selected sentences and the product title), used as a reference. Besides, we also computed the scores from comparing the prompt content with two alternative descriptions: title-only-based descriptions, where we asked the LLM to generate a description based only on the product title; and the original descriptions posted by the sellers.

The scores obtained can be observed in Table 4 with the mean scores for 333 products, and also the standard deviation. First, analyzing the precision, we highlight the significant similarity in the vocabulary even in the case of the original descriptions that have no direct relation to the selected sentences, as 40.7% of the words used appear in the group of sentences. For the title only descriptions, precision was higher, 53%, but very distant from the score achieved by our method of 72%, indicating that ours descriptions are heavily influenced by sentences.

Regarding recall, we first observe that the original descriptions have a very low score of 12%, highlighting how reviews can be a rich source of unique information not covered by sellers. Regarding title-only-based descriptions, there is an increase in recall to 23%, partly justified by the repetition of the product title which was a constant pattern in descriptions generated by the LLM. Compared with our descriptions, recall increased by 12% to 35%, indicating that, at least in terms of vocabulary, a bigger portion of the multiple sentences is covered.

Table 4: ROUGE scores between each description and content of the prompt.

Method	Precision	Recall
Proposed	**0.72** ± 0.07	**0.35** ± 0.06
Title only	0.53 ± 0.07	0.23 ± 0.04
Original description	0.41 ± 0.15	0.12 ± 0.08

5.2 Descriptions Quality

To assess quality, we used a 7-point Likert scale comparing our descriptions with the original ones provided by sellers across multiple criteria. Additionally, to better understand how the use of an LLM contributes to this task, we replicated the extractive method proposed by (Novgorodov et al., 2019) for the same 150 products, and also compared it with the same original descriptions.

Detailing our Likert scale, each item involved comparing specific criteria between a pair of descriptions for the same product, one being the original posted by the seller and the other an alternative description either generated by our method or the replicated one. Regarding the evaluated criteria, we used the same ones as (Novgorodov et al., 2019), providing definitions for each in the form of items on the Likert scale, and asking evaluators to express their degree of agreement on a scale ranging from *"Strongly disagree"* to *"Strongly agree"*.

- **Readable:** *"When comparing to the reference description, this description is more readable, being easier to read and understand."*
- **Objective:** *"When comparing to the reference, this description presents itself in a more objective way, being more succinct and less repetitive."*
- **Informative:** *"When comparing to the reference, this description is more informative, as it presents more information and details."*
- **Relevant to the Product:** *"When comparing to the reference, the information presented in this description is more relevant to the product, as it presents more details about important features."*
- **Overall Preference**: *"Overall, I prefer this product description when comparing to the reference."*

Results are presented in Fig. 2. We first highlight how both methodologies were overall preferred against the original descriptions in significantly different proportions. For our method, 62% of the product evaluators at least agreed that the descriptions were overall preferred over the original, while only 14% at least disagreed. In contrast, for the extractive methodology evaluators agreed for only 37% of the products and disagreed for 36%. That indicates a big overall difference between both methods, but also that the original descriptions can improve a lot.

Regarding readability, both methodologies were preferred in a similar manner. For 59% of our descriptions, the evaluators considered the texts more readable, disagreeing with only 14%. The extractive descriptions were also preferred by a wide margin, with agreement being 50% and disagreement 16%. This highlights the low readability of the content made available by sellers on the platform, despite its importance.

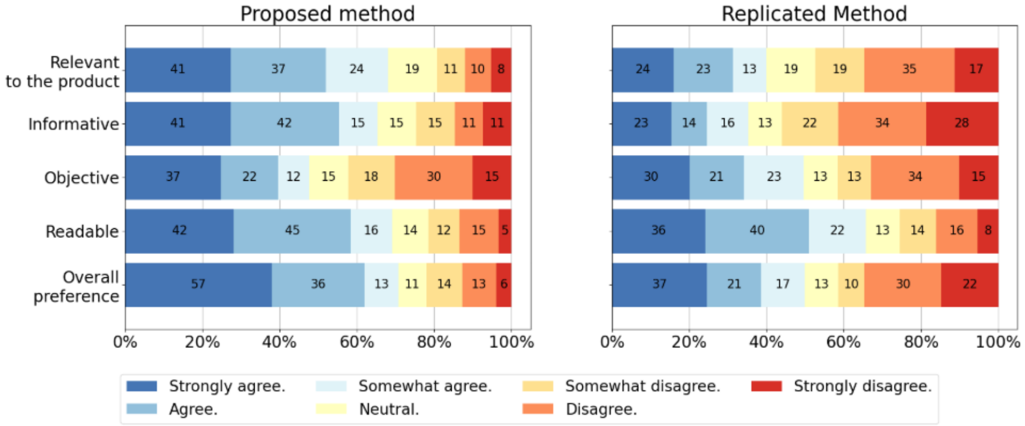

Figure 2: Comparison of each method with the original description. The numbers inside each bar indicate the number of answers given by the evaluators.

Objectivity, however, was the most penalized criterion for our method, as evaluators preferred our descriptions only 41% of the cases and disagreed in 30%. One of our suspicions for such penalization was the frequent use of generic adjectives, generating non-conclusive texts. For the extractive alternative, penalization was even worse, with agreement in only 34% of comparisons and disagreement in 33%, indicating that the few sentences selected to compose descriptions can also be generic.

Regarding informativeness and relevance to the product, we observed contrasting results for each method. While in our method there was a clear trend of preference, in 56% and 51% of the cases, respectively, and disagreement being only 14% and 12%, we found the opposite for the extractive method, with the evaluators preferring the original descriptions. That seems to indicate that five sentences extracted from the reviews are not enough to build a description with informative and relevant content, but as we add more sentences, we eventually enrich our descriptions.

Finally, to compare the two methodologies evaluations, we conducted a Mann-Whitney U nonparametric statistical test (Nachar et al., 2008). As our method showed better results in every criterion, we adopted a one-sided alternative hypothesis that the descriptions generated by our method were more preferred than those generated by the reference method, and used a significance level of 0.05. We present results in Table 5, verifying statistical significance in three criteria: informativeness, relevance to the product, and overall preference. For the other two criteria, the p-value was not small enough.

Table 5: Results of the combined Mann-Whitney U test. P-values lower than 0.05 are highlighted with a *, indicating rejection of the null hypothesis.

Criterion	Median		p-valor
	Proposed	Extractive	
Readable	6	6	0.123
Objective	4	4	0.280
Informative	6	3	0.000*
Relevance	6	4	0.000*
Overall	6	5	0.000*

6 CONCLUSION

This work presents a method for automatic product description generation that combines customer reviews with the text generation capabilities of an LLM. Our zero-shot approach allows scalability across categories without great effort, with the supervised step of sentence classification demonstrating strong generalization, as detailed in 4.2. In evaluations with 30 participants, our method was consistently preferred over original descriptions, excelling in readability, informativeness, and relevance. The comparisons with a baseline confirmed that integrating an LLM significantly improves automatic description generation.

6.1 Limitations and Future Works

Our method relies on a large number of reviews, limiting its applicability to new or low-engagement products. Additionally, it may generate hallucinations, presenting unreal attributes or inaccuracies, and can reflect false information from customer reviews, resulting in misleading descriptions.

For future work, we aim to experiment with newer LLMs and enhance prompt engineering. Exploring the inclusion of more sentences and integrating additional product information, such as manufacturer-provided technical details, could improve description accuracy and detail while helping to verify generated content.

ACKNOWLEDGMENTS

This study was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

Amidei, J., Piwek, P., and Willis, A. (2019). The use of rating and likert scales in natural language generation human evaluation tasks: A review and some recommendations.

Amplayo, R. K., Angelidis, S., and Lapata, M. (2021). Aspect-controllable opinion summarization. *http://arxiv.org/abs/2109.03171*.

Brown, T. B. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chen, Q., Lin, J., Zhang, Y., Yang, H., Zhou, J., and Tang, J. (2019). Towards knowledge-based personalized product description generation in e-commerce. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3040–3050.

Christodoulou, E., Gregoriades, A., Pampaka, M., Herodotou, H., Filipe, J., Smialek, M., Brodsky, A., and Hammoudi, S. (2021). Application of classification and word embedding techniques to evaluate tourists' hotel-revisit intention. In *ICEIS (1)*, pages 216–223.

Elad, G., Guy, I., Novgorodov, S., Kimelfeld, B., and Radinsky, K. (2019). Learning to generate personalized product descriptions. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 389–398.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Hütsch, M. and Wulfert, T. (2022). A structured literature review on the application of machine learning in retail. *ICEIS (1)*, pages 332–343.

Liang, Y.-S., Chen, C.-Y., Li, C.-T., and Chang, S.-M. (2024). Personalized product description generation with gated pointer-generator transformer. *IEEE Transactions on Computational Social Systems*.

Liu, Y. and Lapata, M. (2019). Hierarchical transformers for multi-document summarization. *http://arxiv.org/abs/1905.13164*.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2022). Mteb: Massive text embedding benchmark. *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 2006–2029.

Nachar, N. et al. (2008). The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1):13–20.

Nguyen, M.-T., Nguyen, P.-T., Nguyen, V.-V., and Nguyen, Q.-M. (2021). Generating product description with generative pre-trained transformer 2. In *2021 6th international conference on innovative technology in intelligent system and industrial applications (CITISIA)*, pages 1–7. IEEE.

Ni, J., Li, J., and McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 188–197.

Novgorodov, S., Guy, I., Elad, G., and Radinsky, K. (2019). Generating product descriptions from user reviews. In *The World Wide Web Conference*, pages 1354–1364. ACM.

Novgorodov, S., Guy, I., Elad, G., and Radinsky, K. (2020). Descriptions from the customers: comparative analysis of review-based product description generation methods. *ACM Transactions on Internet Technology (TOIT)*, 20(4):1–31.

Pedrosa., G., Gardenghi., J., Dias., P., Felix., L., Serafim., A., Horinouchi., L., and Figueiredo., R. (2023). A user-centered approach to analyze public service apps based on reviews. In *Proceedings of the 25th International Conference on Enterprise Information Systems - Volume 1: ICEIS*, pages 453–459. INSTICC, SciTePress.

Puduppully, R., Dong, L., and Lapata, M. (2019). Data-to-text generation with content selection and planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6908–6915.

Wang, J., Hou, Y., Liu, J., Cao, Y., and Lin, C.-Y. (2017). A statistical framework for product description generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 187–192.

Wang, Z., Zou, Y., Fang, Y., Chen, H., Ma, M., Ding, Z., and Long, B. (2022). Interactive latent knowledge selection for e-commerce product copywriting generation. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 8–19.

Zhan, H., Zhang, H., Chen, H., Shen, L., Ding, Z., Bao, Y., Yan, W., and Lan, Y. (2021). Probing product description generation via posterior distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:14301–14309.

Zhang, X., Zou, Y., Zhang, H., Zhou, J., Diao, S., Chen, J., Ding, Z., He, Z., He, X., Xiao, Y., Long, B., Yu, H., and Wu, L. (2022). Automatic product copywriting for e-commerce. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:12423–12431.