Aspect-based Summaries from Online Product Reviews: A Comparative Study using various LLMs

Pratik D. Korkankar[1,2], Alvyn Abranches[1], Pradnya Bhagat[1], Jyoti D. Pawar[1]
[1]Goa Business School, Goa University, India.
[2]Dnyanprassarak Mandal's College and Research Centre, Assagao, Goa, India.
{dcst.pratik, dcst.alvyn, dcst.pradanya, jdp}@unigoa.ac.in

Abstract

In the era of online shopping, the volume of product reviews for user products on e-commerce platforms is massively increasing on a daily basis. For any given user product, it consists of a flood of reviews and manually analyzing each of these reviews to understand the important aspects or opinions associated with the products is difficult and time-consuming task. Furthermore, it becomes nearly impossible for the customer to make decision of buying the product or not. Thus, it becomes necessary to have an aspect-based summary generated from these user reviews, which can act as a guide for the interested buyer in decision-making. Recently, the use of Large Language Models (LLMs) has shown great potential for solving diverse Natural Language Processing (NLP) tasks, including the task of summarization. Our paper explores the use of various LLMs such as Llama3, GPT-4o, Gemma2, Mistral, Mixtral and Qwen2 on the publicly available domain-specific Amazon reviews dataset as a part of our experimentation work. Our study postulates an algorithm to accurately identify product aspects and the model's ability to extract relevant information and generate concise summaries. Further, we analyzed the experimental results of each of these LLMs with summary evaluation metrics such as Rouge, Meteor, BERTScore F1 and GPT-4o to evaluate the quality of the generated aspect-based summary. Our study highlights the strengths and limitations of each of these LLMs, thereby giving valuable insights for guiding researchers in harnessing LLMs for generating aspect-based summaries of user products present on these online shopping platforms.

1 Introduction

The present age of e-commerce provides online consumers with a wide range of online product reviews. These reviews contain crucial information and highlight insights about the products, thereby showcasing the strengths, quality and performance. Also, the popularity of user products on these platforms is growing at an unbelievable rate. The influx of people accessing e-commerce websites is attributed due to the presence of written product reviews by the users themselves and not by the origin of brands. This has helped prospective buyers to decide whether to buy the product based on the reviews and experiences shared by the users on these e-commerce platforms as they are considered to have more credibility and trustworthiness by the potential buyers (Maslowska et al., 2017; Watson and Wu, 2022).

Summarizing online product reviews is a daunting NLP task. It incurs a lot of computation costs in terms of time and other resources. The goal of summarization goes beyond simply identifying and extracting sentences, synthesizing reviews, and providing an overall summary. It becomes more meaningful when the summary is generated based on specific product features. Instead of a generic summary, a more focused summary on particular aspects of the product is preferred. A promising solution to this problem is aspect-based summarization, which identifies key aspects of the product and generates a summary for each extracted aspect.

The rise and power of Large Language Models (LLMs) such as GPT-4 (OpenAI, 2024), Llama (Ollama, 2023), Gemma (Team et al., 2024), Mistral (Jiang et al., 2023), Mixtral (Face, 2024a), Qwen (Face, 2024b) and others has brought a huge revolution in the field of natural language processing (NLP) including text summarization. The capability of these models is not just restricted to understanding the context, structure, and language of the input text, but it has found a huge potential and improvements in language generation or abstraction approaches. The research experiment demonstrated in this paper is inspired by (P Bhagat, 2023), which uses the Chi-square Test statistical measure to automatically calculate the aspect-based polarity of sentiment words in a given domain. The

method used in this paper also helps in discovering strong domain-specific polar adjectives that might be missing in universal sentiment lexicons.

The proposed work attempts to accurately identify product aspects and the model's ability to extract relevant information and generate concise summaries.

The remainder of the paper is organized as follows. Section 2 describes the related work studied. Section 3 is the titled proposed methodology and explains the process of generating an aspect-based summary from the given set of reviews for a specified product. Section 4 explains the implementation details and the datasets used. Section 5 elaborates the experimental details and the evaluation metrics used. Section 6 presents the results and discussions, and finally, Section 7 states the conclusion.

2 Related Work

This section provides insight into a survey of various aspect-based summarization methods and different types of large language models (LLMs).

2.1 Aspect-based Summarization

Broadly, text summarization is the process of condensing a long document consisting of text into a shorter or concise version of it while preserving the essence and overall context of the original text. Primarily, there are two types of text summarization, namely, extractive and abstractive. Extractive summarization mainly focuses on directly extracting key text phrases or sub-phrases from the original document. It often extracts exact word-to-word text phrases from the text. Extractive summarization mainly uses statistical, rule-based or probabilistic approaches. On the other hand, abstractive summarization applies machine learning or deep learning techniques to the original text, which then paraphrases or rephrases the different texts to generate a more concise summary. At times, abstractive summarization often generates new sentences which are more human-like and the flow of the generated summary is more fluid and well-connected, unlike extractive, which is often distorted (Nenkova and McKeown, 2012). Summarization tools are widely used for news articles, academic papers, and reports to make information more accessible and easier to understand. In addition to this, new summarization approaches have evolved, which use the concepts of extractive as well as abstractive summarization. These are often termed as hybrid summarization which leverages the benefits of both extraction and abstraction summarization methods. Instead of focusing on generic summarization that mainly generates a broad overall summary, our paper focuses on generating an aspect-based summary which is more concentrated on the aspect-specific summary. The aspect-based summary helps to identify and stress upon those aspects which are often important and hence find application in many NLP tasks.

(Samha et al., 2014) proposes a framework which sequentially mines product's aspects and users' opinions, groups representative aspects by similarity, and generates an output summary. This paper focuses on the task of extracting product aspects and users' opinions by extracting all possible aspects and opinions from reviews using natural language, ontology, and frequent "tag" sets. (Kamal, 2015) proposes the design of a unified opinion mining and sentiment analysis framework that facilitates subjectivity/objectivity analysis, feature and opinion extraction, anaphora resolution for feature-opinion binding, polarity determination, review summarization and visualization in an integrated manner. (Wu et al., 2016) considers Aspect-based Opinion Summarization (AOS) of reviews on particular products. Here, it addresses two core subtasks, aspect extraction and sentiment classification. Most existing approaches to aspect extraction uses linguistic analysis or topic modelling but are not precise enough or suitable for particular products. Instead, it directly maps each review sentence into pre-defined aspects. To tackle aspect mapping and sentiment classification, they propose two Convolutional Neural Network (CNN) based methods, cascaded CNN and multitask CNN. Cascaded CNN contains two levels of convolutional networks. Multiple CNNs at level 1 deal with aspect mapping task, and a single CNN at level 2 deals with sentiment classification. Multi-task CNN also contains multiple aspect CNNs and a sentiment CNN, but different networks share the same word embeddings. Experimental results indicate that both cascaded and multi-task CNNs outperform Support Vector Machines (SVM) based methods by large margins. Multitask CNN generally performs better than cascaded CNN. (Xu et al., 2020) analyzes the pre-trained hidden representations learned from reviews on BERT for tasks in Aspect-Based Sentiment Analysis (ABSA). The work is motivated by BERT-based language models for ABSA. By leveraging the annotated datasets in ABSA, they investi-

gate both the attentions and the learned representations of BERT pre-trained on reviews. They found that BERT uses very few self-attention heads to encode context words (such as prepositions or pronouns that indicate an aspect) and opinion words for an aspect. Most features in the representation of an aspect are dedicated to the fine-grained semantics of the domain (or product category) and the aspect itself, instead of carrying summarized opinions from its context. Through this investigation it can aid in improving self-supervised learning, unsupervised learning and fine-tuning for ABSA in future research aspects. (Li et al., 2020) proposes an effective new summarization method by analyzing both reviews and summaries. They first segmented reviews and summaries into individual sentiments. As the sentiments are typically short, they combine sentiments talking about the same aspect into a single document and apply topic modeling method to identify hidden topics among customer reviews and summaries. Sentiment analysis was applied to distinguish positive and negative opinions among each detected topic. A classifier was also introduced to distinguish the writing pattern of summaries and that of customer reviews. Finally, sentiments are selected to generate the summarization based on their topic relevance, sentiment analysis score and the writing pattern.

2.2 Power of LLMs in Summarization Task

For the past decade, the summarization task was very much limited and relied completely on statistical, rule-based or traditional machine learning approaches to extract product aspects from text. Recently, these approaches are now turning towards leveraging the power of deep learning and LLMs in the summarization task. Researchers have also shifted their focus on improving the extraction of fine-grained information, particularly in identifying and summarizing aspects from the large corpus of product reviews on these online shopping platforms. LLMs are now also emerging towards providing an alternative for these already existing traditional metrics and human evaluation for evaluating various NLP tasks.

Summarization models are powered by these LLMs and have gained significant momentum over the past few years. Pre-trained language models such as GPT-4, Llama, Gemma, Mistral, Mixtral, Qwen and others have displayed unmatched capabilities in the generation of extractive as well as abstractive summaries. LLMs possess a deep understanding across a vast variety of NLP tasks. They have been performing smartly in generating text across diverse domains. They have smartly outclassed many of the existing state-of-the-art summarization methods, thereby providing a strong candidate for aspect-based summarization of online product reviews.

The primary focus of our paper will be on incorporating the power of LLMs such as Gemma2 (9b, 27b), Llama3 (8b, 70b), Mistral (7b), Mixtral (8*7b, 8*22b), Qwen2 (7b, 72b) and GPT-4o on the publicly available domain-specific Amazon reviews dataset for experimentation work. The experimental results will then be analyzed with the available summary evaluation metrics to evaluate the quality of the generated aspect-based summary. This will help us to determine the best-performing LLM that generates high-quality aspect-based summaries for products belonging to different domains.

3 Proposed Methodology

The proposed work attempts to identify relevant product aspects from the given set of reviews and the LLMs ability to extract relevant information and generate concise summaries.

Our algorithm begins by identifying and extracting reviews specific to a given product. For each product, it filters reviews from a larger set by matching the product identifier called Amazon Standard Identification Number (ASIN). The resulting set contains only reviews belonging to the specified product ID. This ensures that subsequent analysis focuses only on the reviews relevant to a particular product.

After extracting the product-specific reviews, the algorithm narrows the focus to reviews with the most extreme ratings—either 1-star (negative) or 5-star (positive). These reviews are typically the most expressive in terms of customer feedback and sentiment, making them ideal for aspect-based summary and sentiment analysis. The filtered set includes only the 1-star and 5-star reviews for each product.

The next step involves cleaning the text of each review. Any numeric values or special characters (such as punctuation) are removed. This cleaning process ensures that only meaningful, interpretable words remain in the reviews. This is crucial for improving the accuracy of the subsequent language-processing tasks.

To further refine the reviews, the algorithm ap-

plies two additional filters. Firstly, it ensures that only reviews with a word count between 50 and 100 are kept. Reviews that are too short may lack meaningful content, while overly long reviews could introduce noise. Secondly, if the number of filtered reviews exceeds a predefined limit (MAX), then only the first MAX reviews are retained by randomly considering 5000 reviews.

At this point, each review is tokenized using Part-of-Speech (POS) tagging. This process breaks down the review into individual words and labels them with their grammatical roles (e.g., noun, verb). The algorithm identifies and extracts specific aspects (features) of the product from each tokenized review.

Once the aspects are identified, the algorithm applies stemming, which reduces each word to its root form (e.g., "running" becomes "run"). This helps in normalizing different forms of the same word. After stemming, similar aspects are grouped together or merged to avoid redundant or overly similar aspects for simplifying the analysis work.

Next, the algorithm calculates how frequently each aspect appears in the reviews. The resulting set contains the frequency of each aspect. The algorithm limits the number of aspects by keeping only the most frequent ones. This helps in focusing on the most commonly discussed features/aspects of the product. Using a pre-trained language model (LLM), the algorithm generates an overall summary of the reviews, thereby providing general sentiments and feedback from the customers.

In addition to the overall summary, the algorithm generates an aspect-based summary for every identified aspect. The generated summary is based only on reviews related to that aspect, thereby providing detailed insights into the feedback on specific features of the product.

After generating both the overall and aspect-based summary, the algorithm combines them to form a full summary, which includes both general sentiments as well as specific feedback on individual aspects.

Finally, the algorithm evaluates the quality of the generated summary. It compares the full summary with a reference summary (Amazon's official summary) using predefined evaluation metrics. The resulting evaluation scores indicate the effectiveness and accuracy of the generated summary.

To summarize, our proposed methodology provides a structured way to process, clean, analyze, and summarize product reviews, delivering both high-level insights and detailed feedback on specific aspects of the product.

Algorithm 1 Generating aspect-based summary of online product reviews

Input: Set of product-specific reviews.
Output: Aspect-based Summary of the given product.

Extract product-specific reviews:
for $P \in P$ **do**
    $R_P = \{r \in R \mid r.\text{asin} == P\}$
Filter 1-star and 5-star reviews:
for $P \in P$ **do**
    $R_{15} = \{r \in R_P \mid r.\text{rating} \in \{1, 5\}\}$
Remove numeric values and special characters:
$R_{clean} = \{clean(r) \mid r \in R_{15}\}$
Filter by word count and review limit:
$R_{filtered} = \{r \in R_{clean} \mid 50 \leq \text{word\_count}(r) \leq 100\}$
if len($R_{filtered}$) > MAX **then**
    $R_{filtered} = R_{filtered}[: MAX]$
Perform POS tagging and extract aspects:
$T = \{\text{token}(r) \mid r \in R_{filtered}\}$
$A = \{\text{aspect}(t) \mid t \in T\}$
Aspect stemming and reduction:
$A_{stem} = \{\text{stem}(aspect) \mid aspect \in A\}$
$A_{reduced} = \text{reduce\_similar\_aspects}(A_{stem})$
Aspect frequency and limiting:
$A_{freq} = \{\text{aspect: frequency}(aspect) \mid aspect \in A_{reduced}\}$
$A_{limited} = \text{limit\_aspects}(A_{freq})$
Generate overall summary:
$S_{overall} = \text{generate\_summary}(R_{filtered}, LLM)$
Generate aspect-based summaries:
$S_{aspect} = \{\text{aspect: generate\_summary}(R_{filtered}[r.\text{aspect} == aspect], LLM) \mid aspect \in A_{limited}\}$
Combine overall and aspect-based summaries:
$S_{full} = \text{combine\_summaries}(S_{overall}, S_{aspect})$
Evaluate the full summary:
$scores = \text{evaluate\_summary}(S_{full}, S_{amazon}, metrics)$

4 Implementation Details and Datasets used

The experiment is implemented using Python Programming Language (Sanner et al., 1999). The text processing tasks are carried out using the Natural Language Toolkit (NLTK) Library (Loper and Bird, 2002), spaCy (Honnibal et al., 2020) and Hugging Face Transformers (Ollama, 2023). The dataset used by us is a collection of reviews from Amazon.com (He and McAuley, 2016) (McAuley et al., 2015). We test our experiment on three distinct domains of data namely, *Electronics*, *Cell Phones and Accessories* and *Grocery and Gourmet Food*.

5 Experimentation Details and Evaluation Metrics

The three domains considered for experimentation in our study are 1) Electronics 2) Cell Phones and

Accessories and 3) Grocery and Gourmet Food. We have considered 5,000 reviews in each of these domains. The quality of the generated aspect-based summaries from the given set of product-specific reviews using the different LLMs was evaluated using the traditional and GPT-4 criteria evaluation metrics. Under the traditional approach of evaluation on different LLMs, we applied the likes of Meteor (Banerjee and Lavie, 2005), BertScore F1 (Zhang et al., 2019), Rouge 1, Rouge 2 and Rouge L (Lin, 2004) metrics. And, for GPT-4 criteria-based evaluation (Valmeekam et al., 2023); (Sun et al., 2024), the following parameters (Mullick and et al., 2024); (Mullick et al., 2024), such as Relevance, Coverage, Impurity, Rating and Goodness, are considered. The results of each approach are presented in the next section under results and discussions.

6 Results and Discussions

Through the experimentation process, our results are tabulated in terms of the number of parameters used for language models, namely medium-language models and large-language models. We present our results obtained using the traditional metrics and GPT-4 criteria metrics on medium-language models and large-language models.

As shown in Table 1 and the corresponding Figure 1, we can see the comparative results obtained using the traditional evaluation metrics on medium-language models. Here, the GPT-4o model achieves higher scores for metrics such as Meteor, Rouge 1, Rouge 2 and Rouge L. However, the Gemma2_9b model gets the higher score for the BertScore F1 metric. Hence, this shows that the GPT-4o model fairs well for medium-language models compared to others.

In Table 2 and the corresponding Figure 2, we see the results obtained using the traditional evaluation metrics on large-language models. Here, the combination of Mixtral_8*7b and Mixtral_8*22b models achieves higher scores for different metrics. However, Gemma2_27b gets the higher score for Rouge 2 metric. Hence, this shows that the combination of Mixtral_8*7b and Mixtral_8*22b models fairs well for large-language models.

Now, in Table 3 and Table 4 and the corresponding Figure 3 and Figure 4, we summarize the results of GPT-4 criteria evaluation metrics on medium and large-language models, respectively. Clearly, we see that the GPT-4o model achieves higher scores for Relevance, Coverage, Rating and Goodness metrics on medium as well as large-language models. These results put a high focus on the performance of the GPT-4o model's capability to generate an aspect-based summary which is relevant to the specific aspects of the product, correctly covering all the important aspects of the product, scoring well on the overall quality of the product summary and thereby also verifying how good the generated summary is. However, the Qwen2_7b model gets the best Impurity score for the medium-language model, and both Mixtral models get the best Impurity score for large-language models. The low impurity score signifies that the generated aspect-based summary does not contain any out-of-context information.

Table 1: Traditional evaluation metrics on medium-language models.

model	meteor	bert_f1	rouge1	rouge2	rougeL
gemma2_9b	14.29	**70.03**	34.07	5.93	16.39
llama_31_8b	9.73	66.08	26.94	6.82	14.98
mistral_7b	10.47	63.35	27.07	4.89	12.84
qwen2_7b	14.76	59.08	34.65	5.92	14.86
gpt_4o	**15.60**	68.81	**35.27**	**7.26**	**16.58**

Table 2: Traditional evaluation metrics on large-language models

model	meteor	bert_f1	rouge1	rouge2	rougeL
gemma2_27b	12.43	61.34	32.90	**7.49**	16.96
mixtral_8x7b	15.71	**71.45**	37.92	7.01	**18.09**
mixtral_8x22b	**19.02**	70.15	**39.25**	6.67	16.27
llama_31_70b	10.46	64.85	28.24	4.06	14.83
qwen2_72b	15.71	66.77	36.02	5.64	15.44
gpt_4o	15.60	68.81	35.27	7.26	16.58

Table 3: GPT-4 criteria evaluation metrics on medium-language models.

model	rel	cov	imp	rat	good
gemma2_9b	68.33	66.67	28.33	60.00	58.33
llama_31_8b	70.00	63.33	31.67	63.33	65.00
mistral_7b	61.67	56.67	28.33	53.33	50.00
qwen2_7b	66.67	61.67	**23.33**	56.67	51.67
gpt_4o	**75.00**	**75.00**	25.00	**65.67**	**67.67**

Table 4: GPT-4 criteria evaluation metrics on large-language models.

model	rel	cov	imp	rat	good
gemma2_27b	68.33	58.33	33.33	58.33	55.00
mixtral_8x7b	65.00	60.00	**20.00**	55.00	50.00
mixtral_8x22b	70.00	65.00	**20.00**	60.00	55.00
llama_31_70b	68.33	68.33	26.67	65.00	65.67
qwen2_72b	66.67	61.67	23.33	56.67	51.67
gpt_4o	**75.00**	**75.00**	25.00	**65.67**	**67.67**

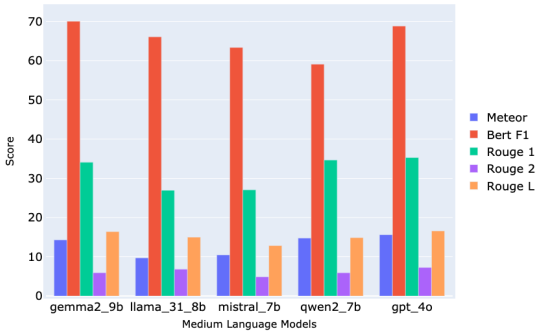

Figure 1: Comparison between results of different traditional evaluation metrics on various medium-language models.

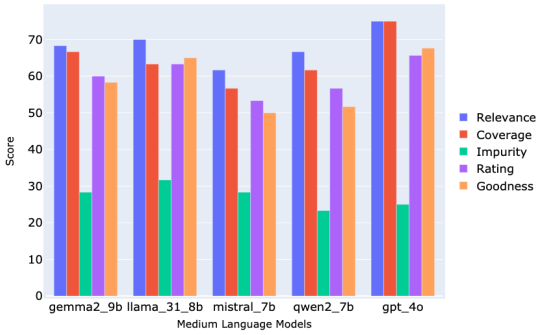

Figure 3: Comparison between results of GPT-4 criteria evaluation metrics on medium-language models

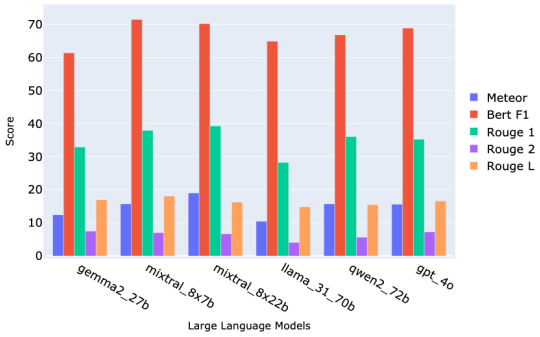

Figure 2: Comparison between results of different traditional evaluation metrics on various large-language models.

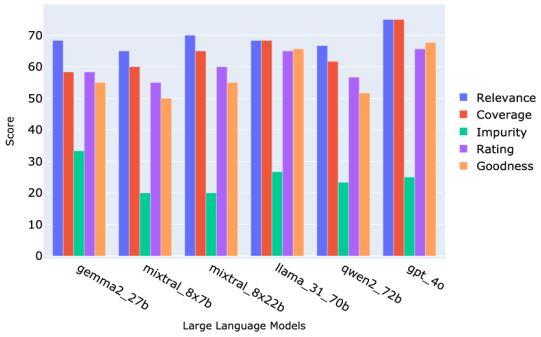

Figure 4: Comparison between results of GPT-4 criteria evaluation metrics on large-language models

7 Conclusion

In this paper, we explored the use of various LLMs such as Llama3, GPT-4o, Gemma2, Mistral, Mixtral and Qwen2 models for the summarization tasks. Our work involved experimenting with the variations of the above-mentioned models on the publicly available domain-specific Amazon reviews dataset. The experiment was tested on three different domains, and the results show that each of the LLMs was able to generate an aspect-based summary on the given set of domain-specific reviews. We postulated the algorithm used in our study accurately identifies and extracts relevant product aspects and generates a concise summary using the various models. We analyzed our experimental results produced by each of these LLMs with the summary evaluation metrics such as Rouge, Meteor, BERTScore F1 and GPT-4o metrics and evaluated the quality of the generated aspect-based summary. With this work, we were able to highlight the strengths and limitations of each of these LLMs used.

Our findings show that the GPT-4o model performs well on traditional as well as GPT-4 criteria evaluation metrics on medium as well as large-language models. However, the variations of Gemma2, Mixtral and Qwen2 models possess a lot of potential and fair well for some of the metrics on medium as well as large-language models. The introduction of Qwen2Audio and Qwen2_VL models opens the door for further exploration in multimodal summarization. So, as part of our future work, we would like to fine-tune our models to improve the performance and efficiency of our summarization task. Also, we would like to analyze the performance of small-language models with medium and large-language models in terms of computation time and efficiency.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved cor-

relation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Hugging Face. 2024a. Mixtral model documentation. Accessed: 2024-09-30.

Hugging Face. 2024b. Qwen2 model documentation. Accessed: 2024-09-30.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. https://spacy.io.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ahmad Kamal. 2015. Review mining for feature based opinion summarization and visualization. *arXiv preprint arXiv:1504.03068*.

Pengyuan Li, Lei Huang, and Guang-jie Ren. 2020. Topic detection and summarization of user reviews. *arXiv preprint arXiv:2006.00148*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Ewa Maslowska, Edward C Malthouse, and Stefan F Bernritter. 2017. The effect of online customer reviews' characteristics on sales. In *Advances in Advertising Research (Vol. VII)*, pages 87–100. Springer.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Ankan Mullick and et al. 2024. On the persona-based summarization of domain-specific documents. *arXiv preprint arXiv:2406.03986*.

Ankan Mullick et al. 2024. Leveraging the power of llms: A fine-tuning approach for high-quality aspect-based summarization. *arXiv preprint*, arXiv:2408.02584.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In *Mining text data*, pages 43–76. Springer.

Ollama. 2023. Ollama: Run large language models locally. https://ollama.com.

OpenAI. 2024. Hello, gpt-4o. https://openai.com/index/hello-gpt-4o/.

JD Pawar P Bhagat, PD Korkankar. 2023. Aspect-based sentiment words and their polarities using chi-square test. *Computación y Sistemas*, 27(June):389–399.

Amani K Samha, Yuefeng Li, and Jinglan Zhang. 2014. Aspect-based opinion extraction from customer reviews. *arXiv preprint arXiv:1404.1982*.

Michel F Sanner et al. 1999. Python: a programming language for software integration and development. *J Mol Graph Model*, 17(1):57–61.

Shichao Sun, Junlong Li, Weizhe Yuan, Ruifeng Yuan, Wenjie Li, and Pengfei Liu. 2024. The critique of critique. *arXiv preprint*, 2401(04518).

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans? *arXiv preprint arXiv:2310.08118*.

Forrest Watson and Yinglu Wu. 2022. The impact of online reviews on the information flows and outcomes of marketing systems. *Journal of Macromarketing*, 42(1):146–164.

Haibing Wu et al. 2016. Aspect-based opinion summarization with convolutional neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 2016–2023. IEEE, IEEE.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2020. Understanding pre-trained bert for aspect-based sentiment analysis. *arXiv preprint arXiv:2011.00169*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.