Opinion Summarization and Evaluation: A Survey

Tejpalsingh Siledar, Sri Raghava Muddu, Rupasai Rangaraju, Pushpak Bhattacharyya
Computer Science and Engineering, IIT Bombay India,
{tejpalsingh, sriraghava, rupasai, pb}@cse.iitb.ac.in

Abstract

Opinion summarization involves condensing opinions into concise representations that capture consensus while preserving overall sentiment. The surge in customer reviews on e-commerce platforms has spurred interest in developing robust opinion summarization and evaluation systems. This survey meticulously examines the multifaceted task of opinion summarization across various contexts: general, aspect-specific, and multi-source scenarios. Traditionally constrained by contextual limitations, recent efforts have shifted towards large-scale opinion summarization to manage large volumes of data effectively. The rise of Large Language Models (LLMs) has significantly impacted this domain, showcasing their potential in NLP tasks. The paper also delves into standard evaluation metrics and pivotal datasets used in these studies, providing a comprehensive overview for researchers and practitioners alike. By synthesizing these insights, the survey outlines progress to date and identifies future directions for advancing opinion summarization and evaluation.

1 Introduction

Opinion summarization is a crucial tool in e-commerce, offering concise representations of customer sentiments to aid in decision-making. This process involves condensing lengthy reviews into manageable insights that capture the general consensus while retaining the overall sentiment. With the rapid expansion of online shopping and the abundance of customer feedback on e-commerce platforms, there is a growing demand for efficient summarization systems. Customers often feel inundated by the multitude of reviews when making purchasing choices. A proficient summarization system that can distill these reviews into brief, informative summaries not only speeds up decision-making but also enhances user experience by allowing shoppers to grasp extensive content effortlessly. Additionally, for product manufacturers, opinion summarization provides valuable insights into customer perceptions and preferences, facilitating product enhancement and market positioning strategies.

However, traditional opinion summarization approaches often focus solely on extracting insights from customer reviews, overlooking valuable information embedded within product descriptions, specifications, and other textual sources. This limitation underscores the need for a more comprehensive approach—multi-source opinion summarization. This innovative technique integrates data from various textual sources, including product descriptions, specifications, reviews, and question-answers, to provide users with holistic summaries that encapsulate both subjective opinions and objective product attributes.

Multi-source opinion summarization represents a paradigm shift in e-commerce, offering a more nuanced understanding of products and enhancing the transparency and completeness of product information. By aggregating insights from diverse sources, these systems empower consumers to make more informed purchasing decisions by considering a broader range of factors, such as product features, functionalities, and user experiences. Moreover, by condensing vast amounts of information into concise yet comprehensive summaries, multi-source opinion summarization streamlines the shopping experience, reducing decision-making fatigue and improving user satisfaction.

Evaluating the quality and effectiveness of opinion summaries is crucial to ensure they meet user needs and expectations. Traditional evaluation methods often rely on human judgment, which can be time-consuming and subjective. Automated evaluation metrics, such as ROUGE and BLEU, are commonly used but may not fully capture the nu-

ances of opinion summarization, such as sentiment preservation and coherence. Recent advancements in natural language processing, particularly the development of large language models (LLMs), offer new possibilities for more accurate and reliable evaluation methods. These models can assess the quality of summaries by comparing them to human judgments, providing a higher correlation with user satisfaction and a deeper understanding of the summarized content.

This survey synthesizes insights from seminal papers, emphasizing the interdisciplinary essence of opinion summarization and evaluation. It stresses the significance of integrating opinions across various sources to generate succinct summaries. The aim is to offer researchers and practitioners a comprehensive overview of historical breakthroughs, current trends, and forthcoming directions in these fields.

2 Summarization Terminologies

Opinion An opinion is a subjective judgment or viewpoint regarding a specific entity, product, or topic, reflecting the individual's personal feelings or beliefs. Opinions can be found in reviews, social media posts, and other user-generated content, providing valuable insights into consumer attitudes and perceptions. For example, in the statement *The battery life of this laptop lasts really long*, the reviewer expresses a favorable opinion about the laptop's battery performance, highlighting their satisfaction with this particular feature.

Summary A summary is a condensed version of a longer text that captures its main ideas and essential points. In opinion summarization, summaries distill key opinions and sentiments from a large number of reviews, enabling users to quickly grasp the overall sentiment and major themes without reading each review individually. Summarization can be achieved through two main approaches:

1. **Extractive summarization:** This method involves selecting and stitching together important sentences or phrases directly from the original text. It maintains the original wording and structure, ensuring that the summary remains faithful to the source content. Extractive summarization is straightforward and effective, especially when the original text is well-written and coherent.
2. **Abstractive summarization:** This approach rephrases the core meaning of the original text to create a summary. Abstractive summarization involves generating new sentences that convey the same information as the source text but in a more concise and coherent manner. It is more complex than extractive summarization, as it requires natural language generation capabilities and a deep understanding of the text.

Aspect An aspect refers to a specific feature or attribute of a product or service that customers comment on in their reviews. Aspects are critical for fine-grained sentiment analysis, as they help identify which features are most frequently praised or criticized. For instance, in the statement *I love the performance of this laptop*, the aspect is *performance*, indicating that the customer is specifically commenting on how well the laptop performs.

Aspects can be categorized into two types:

1. **Implicit aspects:** These aspects are not explicitly mentioned in the review but can be inferred from the context. For example, in the sentence *The laptop is very expensive*, the implicit aspect is *price*, as the statement implies a judgment about the cost without directly mentioning the word *price*.
2. **Explicit aspects:** These aspects are directly stated in the review. For instance, in the sentence *It has a superb display*, the explicit aspect is *display*, as the feature being praised is clearly mentioned.

Sentiment Sentiment represents the qualitative polarity of the opinion expressed in a review sentence, indicating whether the sentiment is positive, negative, or neutral. Sentiments are often measured on a scale, which can range from simple classifications like *Positive*, *Negative*, and *Neutral* to more nuanced scales such as *Very Negative, Slightly Negative, Neutral, Slightly Positive,* and *Very Positive*. This granularity allows for a more detailed analysis of customer opinions, helping businesses identify areas of strength and opportunities for improvement. For example, a review stating *The laptop's battery life is fantastic* would be classified as highly positive, whereas a statement like *The battery life is just okay* might be considered neutral or slightly positive.

Sources for Opinion Summarization The four key sources for opinion summarization predominantly found on e-commerce websites are: (*a*) product description, (*b*) specifications, (*c*) reviews, and (*d*) question-answers. These sources collectively provide a comprehensive view of consumer sentiment, aiding both buyers in making informed decisions and sellers in improving their products and services.

1. **Product Description** In e-commerce, product description is vital for providing detailed information about the features and benefits of a product. Such product description play a crucial role in aiding consumers to make informed purchasing decisions by clearly conveying what the product is, what it does, and why it is worth buying. Typically, product description is presented in either a single paragraph or a headline-paragraph format, highlighting key product attributes and benefits.

2. **Product Specifications** In e-commerce, specifications provide a detailed enumeration of the technical details and functionalities of a product. These specifications are crucial for giving consumers a comprehensive overview of a product's capabilities, enabling them to make informed decisions based on their specific needs and preferences. The information included in specifications typically covers a wide range of attributes. For instance, in the context of electronic devices such as laptops, specifications might detail the following: screen size, resolution, sound properties, etc.

3. **Customer Reviews**: In e-commerce, reviews offer valuable insights into the quality, performance, and user experience of a product. These reviews provide firsthand accounts from individuals who have purchased or used the product, offering a range of perspectives that aid prospective buyers in their decision-making process. Now, reviews are also instrumental in identifying common issues or defects, as multiple reviews highlighting the same problem can signal a potential flaw that prospective buyers should be aware of. Conversely, consistently positive reviews can reassure buyers of the product's quality and reliability.

4. **Question-Answers** In e-commerce, question-answers is a crucial feature allowing consumers to inquire about specific aspects or features of a product before making a purchase. These sections provide a dynamic forum for prospective buyers to seek additional insights and clarifications directly from the seller or other customers who have already purchased the product, thus addressing potential concerns or queries that may arise during the decision-making process.

3 Foundations and Background

Transformers: Transformers, introduced by Vaswani et al. (2017), revolutionized natural language processing with their attention mechanism. Unlike traditional sequence-to-sequence models, Transformers leverage self-attention mechanisms that allow them to capture dependencies between different words in a sentence more effectively. This architecture enables Transformers to model long-range dependencies and contextual information, making them highly effective for tasks like language translation, summarization, and text generation. Transformers have since become the backbone of many state-of-the-art models such as BERT, GPT, and T5, showcasing their versatility and robust performance across various NLP tasks.

BART: BART (Bidirectional and Auto-Regressive Transformers), introduced by Lewis et al. (2020), is a state-of-the-art sequence-to-sequence model pre-trained for various natural language processing tasks, including summarization. BART utilizes a bidirectional transformer encoder-decoder architecture with a masked language modeling objective during pretraining, allowing it to effectively capture bidirectional contexts and generate coherent summaries. It leverages denoising autoencoding and token masking strategies to ensure robust representation learning and generation capabilities, making it highly suitable for abstractive summarization tasks where capturing semantic meaning and linguistic fluency are crucial.

T5: T5 (Text-To-Text Transfer Transformer), introduced by Raffel et al. (2020), is a versatile pre-trained model that excels in various natural language processing tasks, including summarization. T5 adopts a unified text-to-text framework where all tasks, including summarization, are reformulated as text generation tasks. This approach allows

T5 to achieve state-of-the-art performance across different domains by fine-tuning on specific tasks like summarization. By leveraging large-scale data and transformer architecture, T5 generates summaries by predicting target text conditioned on input text, demonstrating strong capabilities in understanding and generating human-like summaries across diverse datasets and languages.

GPT: GPT (Generative Pre-trained Transformer), introduced by Radford et al. (2018), is a widely recognized model in natural language processing that employs a transformer architecture for language modeling tasks. GPT is trained using unsupervised learning on large text corpora, enabling it to generate coherent and contextually relevant text based on given prompts. The model's autoregressive nature allows it to predict the next word in a sequence based on previous words, making it suitable for tasks such as text completion, dialog generation, and summarization. GPT's success has led to subsequent versions and adaptations, solidifying its position as a cornerstone in NLP research and applications.

LLMs: Large Language Models (LLMs) have transformed natural language processing by leveraging massive datasets and advanced transformer architectures. GPT-3 (Brown et al., 2020) stands out with its vast scale of 175 billion parameters, excelling in generating coherent and contextually relevant text. T5 (Raffel et al., 2020) adopts a text-to-text framework, achieving state-of-the-art results across various tasks. More recent models include Mistral (Jiang et al., 2023a) and LLaMA (Touvron et al., 2023), which continue to push the boundaries with enhanced efficiency and performance. Models like GPT-4 further explore larger scales, novel architectures, and improved capabilities, advancing the field by setting new benchmarks in natural language understanding and generation tasks.

4 Evaluation Metrics

In this section we discuss the reference-based and reference-free evaluation metrics in the context of assessing the opinion summary quality.

4.1 Reference-based Evaluation

Reference-based evaluations in summarization include automatic evaluation, human evaluation, and faithfulness evaluation. Automatic metrics such as ROUGE, BERTScore, METEOR, and BLEU quantify content overlap, fluency, and structural alignment between machine-generated and human-written summaries. Human evaluation relies on human assessors to judge readability, coherence, and overall quality. Faithfulness evaluation examines how accurately summaries convey the original content's meaning and nuances. Together, these evaluations provide a thorough assessment of summarization systems.

4.1.1 Automatic Evaluation

ROUGE (Lin, 2004) The ROUGE score is a set of metrics used to evaluate the quality of automatic summarization and machine translation systems in natural language processing. It compares an automatically generated summary or translation with a reference or a set of reference summaries (typically human-produced). The ROUGE score ranges from 0 to 1, with higher scores indicating higher similarity between the generated summary and the reference. The most common ROUGE metrics to evaluate the summaries are:

1. ROUGE-1: This metric measures the overlap of unigrams (single words) between the system and reference summaries. It is defined as:

$$\text{ROUGE-1} = \frac{\sum_{i=1}^{|R|} \min(\text{C}(w_i, S), \text{C}(w_i, R))}{\sum_{i=1}^{|R|} \text{C}(w_i, R)}$$

   where w_i is the i-th word in the reference summary R, S is the system-generated summary, and $\text{C}(w_i, X)$ is the number of times w_i appears in summary X.

2. ROUGE-2: This metric measures the overlap of bigrams (sequences of two words) between the system and reference summaries. It is defined as:

$$\text{ROUGE-2} = \frac{\sum_{i=1}^{|R|} \min(\text{C}(\text{bi}_i, S), \text{C}(\text{bi}_i, R))}{\sum_{i=1}^{|R|} \text{C}(\text{bi}_i, R)}$$

   where bi_i is the i-th bigram in the reference summary R, and the counts are similar to those in ROUGE-1.

3. ROUGE-L: This metric measures the longest common subsequence (LCS) between the system and reference summaries. It is based on

sentence-level structure similarity and identifies the longest co-occurring in sequence n-grams automatically. The ROUGE-L score is computed as:

$$\text{ROUGE-L} = \frac{LCS(R, S)}{|R|}$$

where $LCS(R, S)$ is the length of the longest common subsequence between the reference summary R and the system summary S, and $|R|$ is the length of the reference summary.

The ROUGE metrics provide a robust measure of the overlap between the generated summaries and the reference summaries, with ROUGE-1 and ROUGE-2 focusing on n-gram overlap and ROUGE-L emphasizing sequence similarity.

BLEU BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002), evaluates the similarity between candidate and reference summaries based on n-gram precision. It measures how many n-grams in the candidate summary match those in the ground truth summary. BLEU-1 assesses word-by-word matches, while BLEU-2 and higher consider matching pairs and longer sequences, respectively. Unigram scores gauge summary adequacy, indicating whether the model captures essential features, while higher n-grams assess fluency.

Despite its popularity in Natural Language Generation (NLG) systems, BLEU has limitations. Techniques like clipped precision address issues such as artificially inflated scores from repeated words, where each word is counted only up to its occurrence in the reference summary. Additionally, a brevity penalty discourages overly short summaries with mainly stop words, calculated based on the lengths of the predicted and reference sentences:

$$\text{Brevity Penalty} = \begin{cases} 1, & \text{if } c > r \\ e^{(1-\frac{r}{c})}, & \text{if } c \leq r \end{cases}$$

BERTScore (Zhang et al., 2020b) BERTScore is a metric used to evaluate the quality of machine-generated text, particularly in tasks like summarization and machine translation. It leverages BERT (Bidirectional Encoder Representations from Transformers) embeddings to measure the similarity between the generated summary and the reference summary.

The BERTScore ranges from 0 to 1, with higher scores indicating better quality and greater similarity to the reference summary. The BERTScore metric consists of the following components:

1. BERT EMBEDDINGS: BERT embeddings are computed for both the generated summary S and the reference summary R.

2. COSINE SIMILARITY: The cosine similarity between the BERT embeddings of S and R is calculated to measure their similarity:

$$\text{Cosine Similarity} = \frac{\text{emb}(S) \cdot \text{emb}(R)}{\|\text{emb}(S)\| \cdot \|\text{emb}(R)\|}$$

where $\text{emb}(X)$ represents the BERT embedding of summary X.

3. PRECISION: BERTScore also computes precision by comparing how well the generated summary captures important tokens from the reference summary:

$$\text{Precision} = \frac{\sum_{i \in R} \max_{j \in S} \text{emb}(r_i) \cdot \text{emb}(s_j)}{\sum_{i \in R} \text{emb}(r_i)}$$

where r_i and s_j are tokens in R and S, respectively.

4. RECALL: BERTScore evaluates recall by measuring how well the reference summary tokens are captured by the generated summary:

$$\text{Recall} = \frac{\sum_{j \in S} \max_{i \in R} \text{emb}(r_i) \cdot \text{emb}(s_j)}{\sum_{j \in S} \text{emb}(s_j)}$$

5. F1 SCORE: The harmonic mean of precision and recall provides the overall BERTScore:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The BERTScore metric integrates BERT embeddings to assess both the content overlap and the quality of summary generation, making it a robust evaluation measure for tasks requiring semantic understanding and linguistic fluency.

METEOR (Banerjee and Lavie, 2005) The METEOR (Metric for Evaluation of Translation with Explicit ORdering) score is a metric used to evaluate the quality of machine-generated text, particularly in the context of summarization and machine translation. It compares the generated summary

to a reference summary by considering synonyms, stemming, and paraphrasing, aiming to improve correlation with human judgment.

The METEOR score ranges from 0 to 1, with higher scores indicating better quality and greater similarity to the reference summary. The METEOR metric consists of the following components:

1. PRECISION (P): The fraction of words in the generated summary that are also present in the reference summary.

2. RECALL (R): The fraction of words in the reference summary that are also present in the generated summary.

3. HARMONIC MEAN (F_{mean}): The harmonic mean of precision and recall, giving a balanced measure of both.

$$F_{\text{mean}} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P}$$

4. FRAGMENTATION PENALTY (F_{penalty}): A penalty for the number of chunks or fragments in the matching, which addresses the fluency and order of the generated summary.

$$F_{\text{penalty}} = 0.5 \left(\frac{\text{chunks}}{\text{matches}} \right)$$

5. METEOR SCORE (M): The final METEOR score incorporates the harmonic mean and the fragmentation penalty.

$$M = F_{\text{mean}} \cdot (1 - F_{\text{penalty}})$$

The METEOR metric provides a comprehensive evaluation by not only considering precision and recall but also taking into account the order and structure of the generated summary, making it well-suited for tasks that require high-quality, coherent text generation.

4.1.2 Human Evaluation

Best-Worst Scaling Best-Worst Scaling (Flynn and Marley, 2014) is a technique used to measure individuals' preferences among a set of items or options. Widely applied in opinion summarization studies, it provides more reliable results than traditional ranking systems (Kiritchenko and Mohammad, 2017). In this evaluation method, participants rate various summaries generated by different models. A score of +1 indicates the best model, -1 indicates the worst, and 0 indicates the remaining models. The final scores are calculated by averaging the ratings given by different participants, resulting in a robust and comprehensive assessment.

Likert Scale A Likert scale presents a series of statements, each accompanied by a range of response options from strongly agree to strongly disagree. The scale typically ranges from two to seven points, with the five-point scale being the most common: **Strongly agree, Agree, Neutral, Disagree, Strongly disagree**. Likert scales are extensively used in surveys and research to gather data on people's views, attitudes, and behaviors. In natural language processing, they assess the sentiment of written content, such as reviews or social media posts. For instance, a Likert scale can evaluate the positivity or negativity of a review or the extent of agreement with a statement. This tool provides valuable insights into participants' attitudes and opinions, aiding researchers in drawing conclusions and making informed decisions.

4.1.3 Faithfulness Evaluation

SummaC (Laban et al., 2022): SummaC (Summary Consistency) aims to tackle the granularity in NLI models. Specifically, SummaC focuses on identifying inconsistencies in summarization, taking into account the diverse levels of granularity that can exist between sentences and documents A higher SummaC score indicates higher faithfulness.

CTC (Deng et al., 2021): CTC (Compression Transduction Creation) presents a framework that considers various natural language generation (NLG) tasks, including compression (such as summarization), transduction (like text rewriting), and creation (such as dialog generation). The CTC metric evaluates information alignment, with a specific emphasis on gauging consistency and relevance. A higher CTC score indicates higher faithfulness.

FactCC (Kryscinski et al., 2020): FactCC is a BERT-based classification model, with the objective of ascertaining the consistency or inconsistency between a provided text or summary and its corresponding source article. A higher FactCC score

indicates higher faithfulness.

FactGraph (Ribeiro et al., 2022): FactGraph uses both the text and their structured meaning representations computed using a graph encoder with structure-aware adapters to enhance the factuality of the summaries with respect to the source document. A higher FactGraph score indicates higher faithfulness.

4.2 Reference-free Evaluation

Traditional metrics like ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020b) depend on reference summaries for evaluation. ROUGE, for example, primarily measures n-gram overlap. However, with the rise of large language models (LLMs), generated text can convey the same meaning as reference summaries but with different wording. Siledar et al. (2024b) suggests that humans might even prefer summaries generated by models like GPT over human-written ones. This highlights the need for reference-free metrics that can evaluate summaries across various dimensions. The following seven metrics were introduced by Siledar et al. (2024a) for comprehensive opinion summary evaluation:

1. `Fluency (FL)`: This metric evaluates the summary's quality in terms of grammar, spelling, punctuation, capitalization, word choice, and sentence structure. A fluent summary should be error-free, easy to read, follow, and understand. Annotators were provided with guidelines on how to penalize summaries based on their fluency.

2. `Coherence (CO)`: This assesses the overall quality of the sentences in the summary. A coherent summary should be well-structured and organized, building a cohesive body of information rather than a series of related points.

3. `Relevance (RE)`: The summary should include only significant opinions from the reviews, excluding non-consensus or unimportant opinions. Summaries are penalized for redundancies and irrelevant information.

4. `Faithfulness (FA)`: All information in the summary should be verifiable, supported, or inferred from the reviews. Summaries are penalized if any information cannot be verified, supported, or inferred from the reviews, or if they overgeneralize.

5. `Aspect Coverage (AC)`: The summary should cover all major aspects discussed in the reviews. Summaries are penalized for omitting any major aspect and rewarded for comprehensive coverage.

6. `Sentiment Consistency (SC)`: The summary should accurately reflect the consensus sentiment of the aspects discussed in the reviews. Summaries are penalized for misrepresenting the sentiment of any aspect.

7. `Specificity (SP)`: The summary should avoid generic opinions and provide detailed, specific information about the consensus opinions. Summaries are penalized for lacking detail and rewarded for specificity.

5 Datasets

In this section, we first discuss the different datasets used for testing opinion summarization models followed by evaluation datasets testing the correlation of different evaluation methods with humans for the task of opinion summary evaluation.

5.1 Opinion Summarization Datasets

Amazon (Bražinskas et al., 2020): Amazon test set contains product reviews from four domains: *electronics, home and kitchen, personal care,* and *clothing, shoes and jewelry* from the *Amazon Product Dataset* (He and McAuley, 2016). The evaluation set contains three general abstractive summaries per product. Each product has 8 reviews in the evaluation set. The training set contains $\sim$ 1M reviews over 90K products.

Oposum+ (Amplayo et al., 2021): Oposum+ contains product reviews from six different domains: *laptop bags, bluetooth headsets, boots, keyboards, televisions* from the *Amazon Product Dataset*. The evaluation set contains four summaries per product: three aspect-specific abstractive summaries and one general extractive summary. Each product has 10 reviews in the evaluation set. The training set contains $\sim$ 4.13M reviews over 95K products.

Flipkart (Siledar et al., 2023b): Flipkart dataset contains product reviews from three domains: *laptops, mobiles*, and *tablets*. The test set contains

around 147 products with one summary per product. Each summary consists of multiple aspect-specific summaries. There are around 676 aspect-specific summaries in total. The original test set contains around 1000 reviews per product on average. Siledar et al. (2023a) downsample this to 10 reviews per product to compare different models. They first remove all the reviews with less than 20 and more than 100 words. For filtering out 10 reviews they use a simple approach of first checking if the reviews contain the aspects for which summaries need to be created. After the filtering step, they randomly selected 10 reviews to form input for our test set.

GPT-R/GPT-RDQ (Siledar et al., 2024b) extended the already available Amazon, Oposum+, and Flipkart test sets by leveraging ChatGPT for annotation. GPT-R used only reviews while generating the summary whereas GPT-RDQ used reviews, description, and question-answers for generating summaries. They curated 6 new test sets: Amazon GPT-R, Amazon GPT-RDQ, Oposum+ GPT-R, Oposum+ GPT-RDQ, Flipkart GPT-R, and Flipkart GPT-RDQ containing 662 opinion summaries in total.

AmaSum (Bražinskas et al., 2021): The AmaSum dataset is a large-scale abstractive opinion summarization dataset containing over 33,000 human-written summaries for Amazon products. Each summary is paired with more than 320 customer reviews and includes three types of summaries: verdict, pros, and cons.

Space (Angelidis et al., 2021): The Space dataset is a large-scale benchmark for evaluating unsupervised opinion summarizers, built on TripAdvisor hotel reviews. It includes a training set of approximately 1.1 million reviews for over 11,000 hotels, along with 1,050 human-written summaries for 50 hotels. The dataset is designed to evaluate both general and aspect-specific opinion summarization models, with six popular aspects such as building, cleanliness, food, location, rooms, and service.

Xl-Flipkart (Muddu et al., 2024): Xl-Flipkart is a large-scale ($\sim$ 3600 reviews on average per product) test set of 25 products gathered from the Flipkart website annotated using GPT-4. It was created to test the capabilities of different models in summarizing reviews ranging in thousands.

5.2 Opinion Summary Evaluation Dataset

SummEval-Op (Siledar et al., 2024a): SummEval-Op is an opinion summary evaluation benchmark dataset, consisting of a total of 2,912 summary annotations, assessing 13 opinion summaries for 32 products from the Amazon test set. The evaluation covers 7 dimensions- `fluency`, `coherence`, `relevance`, `faithfulness`, `aspect coverage`, `sentiment consistency`, and `specificity` related to the evaluation of opinion summaries

OpinSummEval: (Shen and Wan, 2023) used the Yelp test set (Chu and Liu, 2019) to annotate for 4 dimensions: `readability`, `self-coherence`, `aspect relevance`, and `sentiment consistency`. The dataset contains a total of 100 products with 8 reviews and 14 different model summaries per product. Each summary was rated by 2 annotators on 4 dimensions.

6 Summarization and Evaluation Approaches

In this section, we discuss works related to text summarization; general, aspect-specific, and multi-source opinion summarization; large-scale opinion summarization; and summary evaluation methodologies.

6.1 Text Summarization

Text summarization has emerged as a critical task in natural language processing (NLP), aiming to condense large volumes of text into concise summaries that retain the essential information. The field has evolved significantly over the past few decades, transitioning from early extractive approaches that rely on statistical and heuristic methods to advanced neural network-based techniques capable of generating abstractive summaries. Extractive approaches select key sentences directly from the source text, ensuring high factual accuracy, while abstractive approaches generate new sentences, allowing for more coherent and fluent summaries. Recent advancements, particularly the incorporation of pre-trained language models and reinforcement learning, have pushed the boundaries of what is achievable in summarization tasks.

Extractive Approaches Extractive summarization methods aim to create summaries by identifying and selecting the most significant sentences from

the source text. Initial techniques relied on statistical measures such as term frequency-inverse document frequency (TF-IDF) to determine sentence importance. Graph-based methods like LexRank (Erkan and Radev, 2004) and TextRank (Mihalcea and Tarau, 2004) introduced the concept of representing sentences as nodes in a graph, with edges indicating similarity, and employed ranking algorithms similar to PageRank to identify central sentences. Supervised learning approaches further advanced extractive summarization by utilizing linguistic and statistical features to train classifiers for sentence ranking (Kupiec et al., 1995). The emergence of deep learning brought neural network-based models into the fold, with convolutional neural networks and recurrent neural networks being employed to score and select sentences, thereby improving the performance and coherence of extractive summaries (Yin and Pei, 2015). These methodologies have provided a robust foundation for extractive summarization, ensuring that key information from the original text is retained in the generated summaries.

Abstractive Approaches Neural abstractive summarization has gained significant traction, with sequence-to-sequence (Seq2Seq) models serving as the primary framework. Early efforts focused on RNN-based encoders (Chopra et al., 2016; Nallapati et al., 2016a) equipped with attention mechanisms (Bahdanau et al., 2016) and copying mechanisms to handle out-of-vocabulary words and enhance factual accuracy (See et al., 2017). The introduction of the Transformer model (Vaswani et al., 2017) marked a significant shift, leading to the development of pre-trained language models like BERT (Devlin et al., 2019), which have been used as powerful encoders in summarization tasks (Zhang et al., 2019; Liu and Lapata, 2019). Pre-trained Seq2Seq models such as MASS (Song et al., 2019), BART (Lewis et al., 2020), and PEGASUS (Zhang et al., 2020a) have achieved state-of-the-art results by leveraging unsupervised pre-training objectives to capture rich linguistic features and contextual information. Evaluation of these models is typically conducted on benchmark datasets like CNN/DailyMail (Nallapati et al., 2016b), Gigaword (Graff et al., 2003), and XSum (Narayan et al., 2018), which provide diverse challenges and help in benchmarking progress in the field. These advancements underscore the significant strides made in abstractive summarization, producing more fluent, coherent, and contextually accurate summaries compared to extractive methods.

6.2 General Opinion Summarization

General opinion summarization involves summarizing overall opinions or sentiments expressed about a particular entity, product, service, event, or topic. The goal is to condense and present a comprehensive overview of opinions, capturing the overall sentiment polarity (positive, negative, neutral) and the key aspects that contribute to those sentiments. General opinion summarization typically aggregates opinions from reviews to provide a consolidated view.

For instance, Ganesan et al. (2010) leveraged redundancy in reviews to generate concise summaries, while Erkan and Radev (2004) used graph-based models to identify and select the most relevant sentences. More recent approaches have shifted towards neural network-based abstractive methods. Chu and Liu (2019); Bražinskas et al. (2020) use autoencoders (Kingma and Welling, 2013) and its variants to learn a review decoder through reconstruction which is then used to generate summaries using the averaged representations of input reviews.

Another approach is to curate synthetic datasets using one of the reviews as a pseudo-summary and pair it with input reviews using different strategies. Bražinskas et al. (2020) uses random sampling, Amplayo and Lapata (2020) generates noisy version of the pseudo-summary, Elsahar et al. (2021) ranks reviews using similarity and relevance, and Amplayo and Lapata (2020) uses content plans to generate synthetic datasets. Im et al. (2021) randomly selects a review as a pseudo-summary and proposes a pipeline to generate summaries using multimodal input such as text, image, and metadata. Ke et al. (2022) captures the consistency of aspects and sentiment between reviews and summary, whereas Wang and Wan (2021) learns aspect and sentiment embeddings to generate relevant pairs. Iso et al. (2021) searches for convex combinations of latent vectors to generate summaries. Siledar et al. (2023a) uses cosine similarity and rouge scores between reviews to filter highly relevant synthetic pairs that enable models to generate summaries more faithful to input reviews compared to alternative approaches.

6.3 Aspect-specific Opinion Summarization

Aspect-specific opinion summarization focuses on extracting and summarizing opinions related to specific aspects or attributes of an entity. This approach aims to provide detailed insights into how individuals perceive different features or characteristics of the entity. For example, in the context of a smartphone, aspects could include battery life, camera quality, performance, design, etc. Aspect-specific opinion summarization helps in understanding which aspects are positively or negatively perceived by users, allowing for more targeted analysis and decision-making.

Angelidis et al. (2021) proposed the first approach to generate both aspect-specific and general summaries. They utilize a Vector Quantized Variational Autoencoder (van den Oord et al., 2017) for clustering review sentences followed by a popularity-driven extraction algorithm to summarize. (Basu Roy Chowdhury et al., 2022) utilizes dictionary learning (Dumitrescu and Irofti, 2018) to acquire representations of texts based on latent semantic units. Amplayo et al. (2021) proposed the first abstractive approach for generating aspect-specific and general summaries. They generate synthetic datasets by identifying aspect-bearing elements (words, phrases, sentences) using a multiple instance learning (MIL) (Keeler and Rumelhart, 1991) model trained on silver-labeled data obtained through seed words.

Shen et al. (2023) proposes two simple solutions for generating synthetic datasets that do not rely on complex MIL modules. The SW-LOO simply matches the aspect seed words to construct synthetic datasets, whereas NLI-LOO uses an off-the-shelf NLI model to do so using only aspects and no seed words. Mukherjee et al. (2020) takes an unsupervised approach to extract aspects and manually creates a mapping between fine-grained and coarse-grained aspects using Integer Linear Programming (ILP) based extractive subset of opinions. Siledar et al. (2023a) does not rely on any human-specified aspects or seed words for generating summaries. They use off-the-shelf aspect extraction and clustering techniques to build an automatic mapping of aspects. Their approach uses two metrics: cosine similarity and rouge scores to form synthetic datasets that achieve better performance.

6.4 Self-Supervised Opinion Summarization

The lack of supervised datasets for opinion summarization led to the use of self-supervision to create synthetic datasets for the supervised training of models. This involves selecting one review from a review corpus as a pseudo-summary and treating the remaining reviews, or a sample of them, as input, forming synthetic pairs for training. This approach allows models to learn the task of opinion summarization without requiring labeled data.

Bražinskas et al. (2020) randomly selected N reviews per entity to construct N pseudo-summary, reviews pairs. Amplayo and Lapata (2020) sampled a review randomly and generated noisy versions of it as input reviews. Amplayo et al. (2020) used aspect and sentiment distributions to sample pseudo-summaries. Elsahar et al. (2021) selected reviews similar to a randomly sampled pseudo-summary as input reviews, based on TF-IDF cosine similarity. Wang and Wan (2021) aimed at reducing opinion redundancy and constructed highly relevant reviews pseudo-summary pairs by learning aspect and sentiment embeddings to generate relevant pairs.

Im et al. (2021) used synthetic dataset creation strategy similar to Bražinskas et al. (2020) and extended it to multimodal version. Ke et al. (2022) captured the consistency of aspects and sentiment between reviews and pseudo-summary using constrained sampling. Siledar et al. (2023a) use lexical and semantic similarities for creating synthetic datasets. Siledar et al. (2024b) uses cosine similarity to select input reviews and pseudo-summary pairs, using review embeddings to compute similarity instead of TF-IDF scores. Additionally, their pseudo-summary selection considers additional sources such as product description and question-answers as well. Recent opinion summarization systems (Bhaskar et al., 2023; Hosking et al., 2023) include a large number of reviews.

6.5 Multi-source Opinion Summarization

Multi-source opinion summarization involves creating concise summaries that integrate opinions from diverse textual and non-textual sources such as reviews, product descriptions, specifications, question answers, images, metadata, and more. The objective is to synthesize viewpoints and sentiments from these varied sources into a unified summary, offering a comprehensive understanding of pub-

lic opinion or information related to a specific entity, topic, or event. This approach enables holistic insights by aggregating and analyzing opinions across multiple dimensions, enhancing decision-making and understanding in various domains such as product analysis, sentiment analysis, and market research.

Zhao and Chaturvedi (2020) used aspects identified from product description to perform extractive aspect-based opinion summarization. Li et al. (2020) proposed a supervised multimodal summarization model to effectively generate summaries using reviews, product image, product title, and product details. Im et al. (2021) proposed a self-supervised multimodal training pipeline to generate summaries using reviews, images, and metadata. Siledar et al. (2023b) did supervised opinion summarization using simple rules to generate summaries separately in the form of verdict, pros, cons, and additional information using reviews, description, specifications, and question-answers. (Siledar et al., 2024b) takes inspiration from Im et al. (2021) to utilize a multi-encoder framework to effectively fuse information from various sources. However, where additional sources are all text, their approach of forming highly relevant synthetic pairs using additional sources helps in capturing relevant information. Also, their approach differs from Siledar et al. (2023b) in training models in an end-to-end fashion without the aid of supervised summaries.

6.6 Large-scale Opinion Summarization

Large-scale opinion summarization involves summarizing sentiments and opinions from a vast number of reviews or textual sources related to a specific entity or topic. Here large scale refers to handling hundreds of thousands of reviews from diverse sources like e-commerce platforms. The goal is to distill key insights and sentiments efficiently, using scalable NLP techniques to capture overall sentiment trends and critical aspects across the dataset.

Recent opinion summarization systems such as (Bhaskar et al., 2023; Hosking et al., 2023; Jiang et al., 2023b) include a large number of reviews. Bhaskar et al. (2023) explores prompting by testing GPT-3.5 (OpenAI, 2023) and introduces various pipelines whereas Jiang et al. (2023b) introduced a review sampling strategy that uses sentiment analysis and two-stage training scheme to generate the opinion summary. Hosking et al. (2023) encodes the reviews into discrete latent space and then generates the summary by decoding the frequent encodings.

Chowdhury et al. (2024) proposed CoverSumm an algorithm to perform centroid-based extractive opinion summarization incrementally. Chang et al. (2023) uses incremental and hierarchical approaches to summarise the book-length text. Muddu et al. (2024) proposed XL-OPSUMM framework for a large-scale opinion summarization system that generates the opinion summary incrementally.

6.7 LLM-based Summary Evaluation

Summary evaluation involves assessing the quality and effectiveness of machine-generated summaries against human-generated references or predefined criteria. Evaluating summaries ensures that they are accurate, coherent, and relevant, reflecting the essential information from the source text. The recent performance of Large Language Models (LLMs) plays a crucial role in summary evaluation as they can leverage their understanding of language semantics and context to measure the similarity and fluency of generated summaries.

Fu et al. (2023) introduced GPTScore that operates on the premise that a generative pre-training model (e.g. GPT-3) is likely to assign a higher probability to the generation of high-quality text in line with provided instructions and context. Chiang and Lee (2023a) were the first to explore LLMs for evaluation. Chiang and Lee (2023b) provide concrete guidelines that improve ChatGPT's correlation with humans. Wang et al. (2023) conducted an initial survey exploring the utilization of ChatGPT as an NLG evaluator. Kocmi and Federmann (2023) used GPT models for evaluating machine learning tasks. Liu et al. (2023) introduced G-Eval, a framework for evaluation of NLG outputs using *Chain of Thought* (CoT) (Wei et al., 2023) and assigning weights to a predetermined set of integer scores based on their generation probabilities from GPT-3/4. Chen et al. (2023) were the first to investigate approaches to reference-free NLG evaluation using LLMs, finding that an explicit score generated by ChatGPT is the most effective and stable approach. Zheng et al. (2023) show that strong LLMs such as GPT-4 achieve a similar level of agreement to that of humans and hence can be used

to approximate human preferences. Siledar et al. (2024a) investigates two prompt strategies and tests the applicability of different prompt approaches on closed-source and open-source LLMs for opinion summary evaluation for 7 dimensions.

7 Summary and Conclusion

In this comprehensive survey, we have thoroughly explored the domains of opinion summarization and evaluation, providing an in-depth analysis of benchmark datasets, evaluation metrics, and influential research papers that have shaped these fields. Our journey began with an examination of text summarization, laying the groundwork for understanding its pivotal role in opinion summarization. Subsequently, we delved into the nuanced dimensions of opinion summarization, encompassing general, aspect-specific, and multi-source scenarios, each posing unique challenges and opportunities for innovation.

A significant focus of our survey was on large-scale opinion summarization, aimed at overcoming limitations related to the contextual scope of models. The burgeoning volume of online reviews necessitates systems capable of effectively processing vast amounts of data to extract meaningful insights and summarize them succinctly. Current research indicates that while strides have been made in this area, significant challenges remain, opening avenues for further exploration and advancement.

Evaluation methodologies emerged as another critical aspect discussed in our survey. Ensuring the quality and coherence of machine-generated summaries is paramount, and leveraging the capabilities of Large Language Models (LLMs) holds promise in enhancing evaluation accuracy and reliability. However, exploiting the full potential of LLMs for robust opinion summarization remains an area ripe for exploration and refinement.

By synthesizing advancements and identifying gaps in these domains, our survey underscores the progress achieved and outlines crucial directions for future research. Effective summarization systems have the potential to profoundly impact fields such as market analysis, consumer feedback processing, and beyond, facilitating informed decision-making and enhancing user experiences in various domains.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2020. Unsupervised opinion summarization with content planning. In *AAAI Conference on Artificial Intelligence*.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1934–1945, Online. Association for Computational Linguistics.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Somnath Basu Roy Chowdhury, Chao Zhao, and Snigdha Chaturvedi. 2022. Unsupervised extractive opinion summarization using sparse coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1209–1225, Dublin, Ireland. Association for Computational Linguistics.

Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with GPT-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300, Toronto, Canada. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. In *Proceedings of the 2021 Conference*

on Empirical Methods in Natural Language Processing, pages 9424–9442, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study.

Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Somnath Basu Roy Chowdhury, Nicholas Monath, Avinava Dubey, Manzil Zaheer, Andrew McCallum, Amr Ahmed, and Snigdha Chaturvedi. 2024. Incremental extractive opinion summarization using cover trees.

Eric Chu and Peter Liu. 2019. MeanSum: A neural model for unsupervised multi-document abstractive summarization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1223–1232. PMLR.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bogdan Dumitrescu and Paul Irofti. 2018. *Dictionary learning algorithms and applications*. Springer.

Hady Elsahar, Maximin Coavoux, Jos Rozen, and Matthias Gallé. 2021. Self-supervised and controlled multi-document opinion summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1646–1662, Online. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Terry N. Flynn and Anthony A. J. Marley. 2014. Best-worst scaling: theory and methods.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 340–348, Beijing, China. Coling 2010 Organizing Committee.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *Proceedings of the 25th International Conference on World Wide Web*.

Tom Hosking, Hao Tang, and Mirella Lapata. 2023. Attributable and scalable opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8488–8505, Toronto, Canada. Association for Computational Linguistics.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 388–403, Online. Association for Computational Linguistics.

Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021. Convex Aggregation for Opinion Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3885–3903, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b.

Han Jiang, Rui Wang, Zhihua Wei, Yu Li, and Xinpeng Wang. 2023b. Large-scale and multi-perspective opinion summarization with diverse review subsets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5641–5656, Singapore. Association for Computational Linguistics.

Wenjun Ke, Jinhua Gao, Huawei Shen, and Xueqi Cheng. 2022. Consistsum: Unsupervised opinion summarization with the consistency of aspect, sentiment and semantic. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 467–475.

Jim Keeler and David Rumelhart. 1991. A self-organizing integrated segmentation and recognition neural net. *Advances in neural information processing systems*, 4.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Julian Kupiec, Jan O. Pedersen, and Francine R. Chen. 1995. A trainable document summarizer. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Aspect-aware multimodal summarization for chinese e-commerce products. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8188–8195.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Sri Raghava Muddu, Rupasai Rangaraju, Tejpalsingh Siledar, Swaroop Nath, Pushpak Bhattacharyya, Swaprava Nath, Suman Banerjee, Amey Patil, Muthusamy Chelliah, Sudhanshu Shekhar Singh, and Nikesh Garera. 2024. Distilling opinions at scale: Incremental opinion summarization using xl-opsumm.

Rajdeep Mukherjee, Hari Chandana Peruri, Uppada Vishnu, Pawan Goyal, Sourangshu Bhattacharya, and Niloy Ganguly. 2020. Read what you need: Controllable aspect-based opinion summarization of tourist

reviews. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1825–1828.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016a. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016b. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. ChatGPT (August 3 Version). https://chat.openai.com.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Ming Shen, Jie Ma, Shuai Wang, Yogarshi Vyas, Kalpit Dixit, Miguel Ballesteros, and Yassine Benajiba. 2023. Simple yet effective synthetic dataset construction for unsupervised opinion summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1898–1911, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuchen Shen and Xiaojun Wan. 2023. Opinsummeval: Revisiting automated evaluation for opinion summarization. *arXiv preprint arXiv:2310.18122*.

Tejpalsingh Siledar, Suman Banerjee, Amey Patil, Sudhanshu Singh, Muthusamy Chelliah, Nikesh Garera, and Pushpak Bhattacharyya. 2023a. Synthesize, if you do not have: Effective synthetic dataset creation strategies for self-supervised opinion summarization in E-commerce. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13480–13491, Singapore. Association for Computational Linguistics.

Tejpalsingh Siledar, Jigar Makwana, and Pushpak Bhattacharyya. 2023b. Aspect-sentiment-based opinion summarization using multiple information sources. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*, pages 55–61.

Tejpalsingh Siledar, Swaroop Nath, Sankara Sri Raghava Ravindra Muddu, Rupasai Rangaraju, Swaprava Nath, Pushpak Bhattacharyya, Suman Banerjee, Amey Patil, Sudhanshu Shekhar Singh, Muthusamy Chelliah, and Nikesh Garera. 2024a. One prompt to rule them all: Llms for opinion summary evaluation.

Tejpalsingh Siledar, Rupasai Rangaraju, Sankara Sri Raghava Ravindra Muddu, Suman Banerjee, Amey Patil, Sudhanshu Shekhar Singh, Muthusamy Chelliah, Nikesh Garera, Swaprava Nath, and Pushpak Bhattacharyya. 2024b. Product description and qa assisted self-supervised opinion summarization.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Ke Wang and Xiaojun Wan. 2021. TransSum: Translating aspect and sentiment embeddings for self-supervised opinion summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 729–742, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Wenpeng Yin and Yulong Pei. 2015. Optimizing sentence modeling and selection for document summarization. In *International Joint Conference on Artificial Intelligence*.

Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 789–797, Hong Kong, China. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert.

Chao Zhao and Snigdha Chaturvedi. 2020. Weakly-supervised opinion summarization by leveraging external information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9644–9651.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.