Master's Thesis

Weakly Supervised Opinion Summarization with Focus on Implicit Aspects

NGUYEN, Khanh Vinh

Supervisor SHIRAI, Kiyoaki

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

September, 2025

Abstract

In the era of user-generated content explosion, many social media and e-commerce sites, such as Yelp and Amazon, receive millions of reviews daily. These textual data sources provide valuable insights into customer satisfaction and product quality. However, a huge volume of customer reviews poses a significant challenge for manual analysis by users or enterprises. To tackle this problem, the study of opinion summarization has emerged and has received much attention.

Opinion summarization is a task that aims to automatically generate a concise and informative summary for a given set of reviews, which reflects the overall preference of users. Unlike conventional summarization tasks focusing on factual or event-based content, opinion summarization aims to address subjective expressions and fine-grained aspect-level sentiment, which are denoted by a wide variety of linguistic expressions.

One essential challenge in opinion summarization is the prevalence of implicit opinions where users' sentiments are not directly stated through obvious aspect-opinion pairs but instead emerge from contextual cues. For instance, the sentence "I had to wave three times before getting help" implies dissatisfaction with the quality of the service, even though the word "service" is not explicitly mentioned. Most existing methods in opinion summarization tend not to focus on capturing such implicit expressions. Furthermore, the majority of current approaches rely on supervised learning, which requires large-scale datasets with manually annotated reference summaries. However, the construction of such datasets is time-consuming and expensive. A lack of labeled data severely restricts the scalability and adaptability of opinion summarization systems across new domains or languages.

This study proposes a novel weakly supervised method for opinion summarization. The main idea is to create mixed-structured data that consists of both structured data and unstructured data, which is derived from a set of reviews, and a pseudo-summary of these reviews. The structured data contains opinion-aspect (OA) pairs that explicitly associate an aspect of a product with users' sentiment, while unstructured data consists of implicit sentences that express users' opinions implicitly. In addition, a ground-truth summary is associated with the structured and unstructured data by selecting a representative review from a review set as a pseudo-summary. This approach addresses the challenge of manual annotation of reference summaries while still ensuring that the model is exposed to rich and sentiment-aware training

data. The proposed method consists of two main stages: the construction of mixed-structured data and the training of a summarization model.

In the first stage, the mixed-structured is constructed by the five steps. First, OA pairs are extracted using a large language model (LLM), i.e., the LLaMA-2 chat model. We instruct the LLM to extract OA pairs that fulfill strict syntactic requirements (e.g., OA should be a noun-adjective pair). Unlike traditional rule-based or dependency-based extraction methods, which are limited and hard to extract OA pairs from noisy reviews across diverse domains, the LLM-based approach provides greater flexibility, allowing the model to more appropriately handle ungrammatical sentences. Second, opinionated sentences that lack explicit aspect terms are extracted as implicit sentences (ISs). Third, both the OA and IS are annotated with sentiment polarity, i.e., positive or negative, to indicate the emotional orientation of the users' feedback. Fourth, a single review within the set of reviews is chosen as a pseudo-summary by a sentiment-aware selection strategy. For each time, a pseudo-summary is expected to satisfy the following four criteria: (1) it covers commonly mentioned aspects, (2) it maintains a balance of positive and negative sentiment, (3) it does not contain first-person pronouns (e.g., "I", "my"), (4) it does not contain contents other than sentences. This pseudo-summary acts as a substitute for a ground-truth summary, although it is not originally written as a summary. Fifth, the OA pairs are sampled to make a diverse mixed-structured dataset. We divide the OA pairs into two types: popular OA and unpopular OA. The popular OA appears in other reviews within the dataset. The soft matching based on cosine similarity between two embeddings of OAs is employed to check whether an OA appears in other reviews. The unpopular OAs are not included in the pseudo-summary; they are randomly sampled. Meanwhile, all implicit sentences are kept intact. The pseudo-summary selection and OA sampling steps are repeated until no more pseudo-summaries satisfy the aforementioned conditions to construct the final mixed-structured data.

After constructing the mixed-structured data, a summarization model is trained. It is a dual-encoder sequence-to-sequence model based on the pre-trained language model BART. This architecture consists of two encoders for processing opinion-aspect pairs and the implicit sentences. Each input is prefixed with a special token ([OA] or [IS]) and sentiment tag (positive or negative). In the decoder, the representations of OAs and ISs are jointly attended using attention fusion, thereby allowing the model to generate a structurally faithful and sentiment-aware summary. The use of the dual-encoder enables us to consider both explicit and implicit opinions in the reviews, which is essential for handling noisy reviews in real-world settings.

Experiments on the Yelp and Amazon datasets are carried out to evaluate

the effectiveness of the proposed method. Despite using only around 43K mixed-structured data in the Yelp dataset and 11K in the Amazon dataset, which are far fewer than the 100K+ used in the previous study, our model achieves competitive or superior results. Our model does not exceed the baselines in terms of ROUGE-1 and ROUGE-2 scores, which measure strict lexical overlap between generated and reference summaries. However, our method outperforms all baselines in terms of the ROUGE-L metric, which evaluates the longest common subsequence between generated and reference summaries. This finding suggests that our model produces more fluent and coherent summaries. Furthermore, our model achieves the lowest self-BLEU score on the Amazon dataset, indicating that it generates lexically diverse summaries.

To validate the effectiveness of each component, an extensive ablation study is conducted by removing or altering individual components. The removal of ISs results in the most significant drop in ROUGE-L, confirming the vital role of implicit content in ensuring the quality and coverage of generated summaries. The elimination of OAs from mixed-structured data also degrades performance, affirming the efficacy of both structured and unstructured data in the mixed-structured dataset. Finally, the utilization of randomly selected reviews as pseudo-summaries results in a decline in the performance of opinion summarization when compared to the use of sentiment-balanced and aspect-rich pseudo-summaries selected by the proposed criteria.

Some limitations are found during the mixed-structured data construction stage, especially in the extraction of OA pairs. Since OA pairs are generated by prompting an LLM rather than relying on gold-standard annotations, some of the extracted pairs are erroneous or semantically ambiguous. Nevertheless, our model demonstrates robustness in producing coherent summaries, suggesting that the integration of explicit opinion-aspect pairs and implicit sentences, as well as adding the sentiment of OAs and ISs to the input, can contribute to generating higher-quality summaries.

Acknowledgements

I would like to convey my extreme appreciation to my supervisor, Professor Kiyoaki Shirai for his dedicated guidance and encouragement throughout my Master's degree program. I extremely appreciate his knowledgeable advice, which has helped me make significant progress in my research. With his guidance, patience and encouragement, I have been able to achieve my thesis goals and define my future research path. Additionally, I would also like to express my sincere appreciation to all members of Shirai & Natthawut Laboratory for their advice and support.

Besides, I would like to extend my thanks to Professor NGUYEN Minh Le and Professor KERTKEIDKACHORN Natthawut for their valuable advice and suggestions for my research.

Finally, I am very grateful to my family and friends for their unconditional love, support and understanding.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter presents the background and goal of the thesis. Section 1.1 gives an overview of the study background. Section 1.2 then discusses the reason and motivations of the research. Lastly, Section 1.3 presents an overview of the overall structure of this thesis.

1.1 Background

The explosive growth of user-generated information on e-commerce platforms, social media, and review websites has generated an overwhelming amount of textual data. Among them, customer reviews are particularly valuable because of their insights into user satisfaction, product quality, and service performance. However, the huge volume of reviews often makes it impractical for potential customers or businesses to read through and extract useful information manually [1].

Opinion summarization has emerged as a solution to this problem. Unlike typical text summarization, which aims to condense content based on topic coverage or sentence salience, opinion summarization specifically targets opinion-aspect pairs, summarizing users' preferences based on certain aspects of a product or service [2]. For example, from Figure 1.1 we can see the differences in the output of typical text summarization and opinion summarization. In that, a typical summary represents the summary of factual content, which is more suitable for news, while an opinion summary focuses on depicting sentiments and aspects.

Despite growing importance, the development of effective opinion summarization models faces two primary challenges [3]. The first challenge is a lack of high-quality labeled datasets. Most summarization models require large-scale, annotated datasets for training. However, creating gold-standard

Reviews	Traditional summarization	Opinion summarization
Not good at all. The young staff basically ignores you if you have any grey hair. The food was just OK, but I won't be back. The Good Egg is much better.	This place serves breakfast in Gilbert and offers tiramisu cake, pastries, and other food. Some customers mention specific staff members like Kayla, Daniela, and Stella. The restaurant has been around for some time and has both positive and negative reviews.	Customers praised the excellent service, friendly staff, and delicious pastries, especially the tiramisu cake. Daniela and Stella were mentioned positively for their attentiveness. While many expressed strong satisfaction and loyalty, one customer criticized the staff's attitude and found the food average, preferring another restaurant instead.
Great hip breakfast in Gilbert. Food and waitress Kayla were excellent. Reviews didn't do it justice. We will be back.		
We love the service and food. The staff is always friendly. Daniela always takes care of us and remembers us by name!!! It's our favorite place.		
Awesome tiramisu cake, taste so good. Been a customer for some time and will return again and again. Greaaaaat place.		
Stella was very nice and helped us out with our birthday cake! Their customer service is impressive and their pastries are even more impressive!!		

Reviews are selected from the Yelp website. Summaries generated by GPT-4.

Figure 1.1: Comparison between traditional text summarization and opinion summarization

summaries from multiple user reviews is time-consuming and costly. As a result, few datasets exist that are domain-diverse, large-scale, and manually labeled for opinion summarization tasks. The second one is neglect of implicit aspects, which are illustrated in Figure 1.2. This figure compares explicit and implicit aspects using a review sentence. An explicit aspect refers to a specific product attribute directly mentioned in the text, such as 'pasta', which is the aspect term, and is associated with the opinion term 'bland'. In contrast, an implicit aspect refers to an attribute that is not directly stated but is implied through context. For example, the sentence 'I had to wave three times before anyone noticed I needed help' implies dissatisfaction with the service, even though the word 'service' is never mentioned. Most existing summarization methods primarily rely on the detection of explicitly mentioned aspects. However, many important opinions in reviews are conveyed implicitly. These implicit signals are often subtle and context-dependent, making them difficult to capture with rule-based or purely surface-level models.

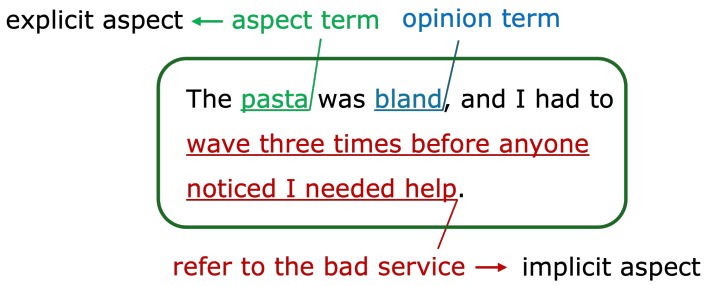


Figure 1.2: Examples of explicit and implicit aspects

1.2 Goal

The primary objective of this thesis is to propose a novel weakly supervised opinion summarization framework, with a particular focus on addressing the lack of ground-truth datasets and incorporating implicit aspects commonly overlooked by existing approaches. Specifically, this research aims to address two major challenges:

- Minimizing dependence on large-scale human-annotated summaries by constructing a pseudo-labeled corpus from user reviews using automatic techniques that account for both sentiment polarity and aspect types.
- Enabling the summarization model to capture both explicit and implicit opinion aspects, thus generating more informative and user-centric summaries.

To achieve these objectives, the thesis proposes a two-stage pipeline:

- Stage 1: Automatic Data Construction. A method is designed to construct weakly labeled training data automatically by selecting reviews as pseudo-summaries. In pseudo-summaries, there are both explicit and implicit aspects. Sentiment-aware selection is applied to guarantee consistency and quality.
- Stage 2: Opinion Summarization Modeling. A Transformer-based sequence-to-sequence model - BART [4] is fine-tuned on the constructed dataset. The model utilizes both explicit and implicit aspects guided by sentiment-informed pre-processing and aspect-aware encoding strategies.

This thesis makes the following primary contributions:

1. Proposal of a sentiment-aware weak supervision framework to construct training data for opinion summarization without totally relying on manual annotations.

2. Incorporation of implicit reviews within the process of summarization, so that the model can learn fine-grained cases of context-based opinions that are otherwise lost in existing techniques.

3. Blending sentiment polarities in data building and model-level operations, so that the quality and polarity sensitivity of generated summaries are enhanced.

4. The proposed method outperforms existing baselines on ROUGE-L metrics on benchmarks.

1.3 Structure of this thesis

The rest of the thesis is organized as follows.

- Chapter 2 discusses previous studies on opinion mining, text summarization, opinion summarization, large language models, and neural architectures relevant to this work.

- Chapter 3 introduces the overall framework, including the mix-structured data generation process and the design of the summarization model.

- Chapter 4 presents the datasets, baselines, evaluation metrics, experimental results and an ablation study.

- Chapter 5 summarizes vital contributions as well as outlines future research directions.

Chapter 2

Related Work

This chapter presents the related work that forms the foundation of this study. Section 2.1 introduces sentiment analysis, a fundamental component in opinion mining pipelines. Section 2.2 reviews the field of text summarization, including both extractive and abstractive approaches, and highlights recent advances based on Transformer architectures. Section 2.3 focuses specifically on Sentiment Summarization, discussing its objectives, system architecture, current methods, and unique challenges. Section 2.4 provides an overview of recent advances in Large Language Models (LLMs) and highlights their applications, advantages, and limitations in the context of opinion summarization. Section 2.5 introduces the Transformer model, detailing its architecture, attention mechanism, and role in modern NLP systems. Building on that, Section 2.6 provides an in-depth overview of the BART model, which serves as the backbone of our proposed framework. Finally, Section 2.7 summarizes the key characteristics of this thesis and highlights how our approach differs from existing studies.

2.1 Sentiment Analysis

Sentiment analysis is a fundamental task in natural language processing that involves identifying the emotional factors conveyed in a piece of text [5]. Due to the increasing availability of user-generated content on social media and review platforms, sentiment analysis has gained widespread attention thanks to its potential applications in a variety of domains, including marketing, healthcare and finance [6]. The primary objective of sentiment analysis is to determine the sentiment polarity of a given input, typically categorized as positive, negative, or neutral and to extract the underlying subjective opinions embedded in text [7, 8].

Over the past two decades, various techniques have been proposed to address this task. One of the earliest approaches is the lexicon-based technique, which uses predefined sentiment lexicons wherein each word is associated with a polarity score [9]. Lexicon-based methods are unsupervised and easy to implement, but they suffer from domain dependence—words like "small" may be negative in one context (e.g., "the TV is too small") but positive in another (e.g., "the camera is compact") [10, 11]. Domain-specific or adaptive lexicons have been proposed to address this limitation.

In terms of lexicon-based techniques, the corpus-based approach uses syntactic and semantic patterns in large corpora to learn the polarity of unknown words based on their co-occurrence with known sentiment words. Statistical techniques such as mutual information [12], latent semantic analysis [13], and sentiment consistency patterns [14] have been applied to infer sentiment orientation in a data-driven manner. Although corpus-based methods require significant labeled data, they provide improved adaptability and context sensitivity. Besides, the dictionary-based approach constructs a human-created seed list of opinion words and then expands it using resources such as WordNet or thesauri [15, 16]. Tools such as SentiWordNet [17], Bing Liu's Sentiment Lexicon[1], and SentiStrength[2] are popular examples. However, dictionary-based approaches often suffer from limited scalability.

Recent advances in deep learning and machine learning have revolutionized the performance of sentiment analysis. Initial models used Naïve Bayes, SVM, and rule-based methods [18, 19]. More recently, deep neural network structures such as CNNs, RNNs, LSTMs, and attention-based models have achieved state-of-the-art results on benchmarking datasets like IMDb, Yelp, and Amazon [20, 21, 22]. Hybrid models such as CNN-LSTM [23] and attention-based BiLSTM [24] have been successfully applied in multi-domain sentiment classification settings.

Realistic applications of sentiment analysis span across many domains. In the business sector, it helps identify product strengths and weaknesses [25]. In healthcare, it is used to analyze patient satisfaction and service quality [26, 27]. In finance, sentiment from news and social media is used to predict stock trends and market volatility [28]. In tourism and hospitality, it improves service personalization by analyzing customer feedback [29].

In the context of this thesis, sentiment analysis plays a crucial role in the proposed weakly supervised opinion summarization framework. It not only indicates the polarity of individual reviews but also guides pseudo-summary generation and provides sentiment-aware input to the model. This integra-

[1] `https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`

[2] `http://sentistrength.wlv.ac.uk/`

tion ensures that the generated summaries reflect both the diversity and emotional depth of user opinions, including implicit sentiments that may not be explicitly stated.

2.2 Text Summarization

Text summarization is a fundamental natural language processing task that aims at generating a condensed version of a document without compromising its crucial information and semantic meaning [30]. With the rapid growth of online text data, automatic summarization has become increasingly important in helping users extract valuable information from long documents efficiently. Text summarization techniques are generally classified into two main categories, extractive and abstractive summarization [31].

Extractive summarization is focused on concatenating the most salient sentences from the source text to form a summary. Over the years, a wide variety of extractive methods have been developed. Statistical-based approaches, such as TF-IDF, are computationally efficient; however, they often prioritize frequent words over contextual content [32]. Concept-based and topic-based methods aim to ensure thematic coverage via sentence selection covering a range of concepts or key topics in the document [33, 34]. Clustering-based approaches group semantically similar sentences and extract representatives to reduce redundancy, though they rely on predefined cluster parameters and may fragment overlapping content [35]. Graph-based techniques, such as TextRank, represent sentence relationships in a network structure and rank nodes using centrality measures; these are domain-independent but sensitive to edge weighting and semantic ambiguity [36, 37]. Semantic-based methods like Latent Semantic Analysis (LSA) uncover latent structures and sentence meanings, though they are computationally expensive and difficult to interpret [38]. Supervised machine learning-based methods use classifiers trained on engineered features to identify important sentences [39], while deep learning-based approaches such as CNNs, RNNs, and LSTMs learn representations directly from data, offering improved generalization but requiring large annotated corpora and substantial computing resources [40]. Some of the others are optimization-based methods that apply genetic algorithms for selection of sentence subsets [41], and fuzzy logic-based methods that address linguistic uncertainty in subjective or noisy texts [42].

In contrast, abstractive summarization generates new sentences that may not exist in the source text. It mimics human summarization by paraphrasing, compressing content, and repetition. Several traditional methods have been proposed. Graph-based approaches build word graphs across the docu-

ment to generate connected sentence paths, though they often struggle with semantic equivalence [43, 44]. Tree-based methods use syntactic parse trees and generation rules to produce less redundant output, but they lack contextual understanding [45]. Rule-based and template-based systems rely on predefined patterns for sentence generation and slot-filling, offering high control but requiring extensive manual rule design [46, 47]. Ontology-based approaches leverage structured domain knowledge to disambiguate meaning and generate coherent summaries, but ontology construction is costly [48]. Semantic-based techniques employ semantic role labeling to better detect relations between sentence elements, although these are heavily dependent on the accuracy of SRL tools [49].

Deep learning-based abstractive methods have experienced the most growth over the past few years. Sequence-to-sequence (Seq2Seq) models with attention mechanisms, and more recently, Transformer-based models such as BART and T5, have attained state-of-the-art performance on benchmark datasets. These models are able to capture long-range relationships and generate coherent, context-aware summaries. However, they require large amounts of annotated data, are computationally intensive, and remain susceptible to factual mistakes and repetitive generation in applications with noisy or casual user-created text [50, 51].

Although existing summarization methods have been effective in structured and factual text domains such as news and scientific articles, they often fall short when handling sentiment texts like customer reviews. These texts are highly subjective, linguistically diverse, and usually implicit with parts unstated but crucial to the identification of sentiment. Furthermore, traditional summarization models barely account for sentiment polarity, aspect-opinion structures, or user perspective diversity, which are all essential in opinion summarization.

2.3 Opinion Summarization

Opinion summarization is a task of sentiment analysis that aims to generate concise summaries, which capture the sentiment conveyed in user-generated opinionated content such as product reviews [1]. As opposed to traditional summarization, which summarises the general input information, opinion summarization aims to extract subjective information, such as sentiment polarity (positive, negative and neutral) and emotional intensity. This would be particularly helpful in real applications, like a market study, customer experience analysis, where a quick understanding of general sentiment trends from large volumes of unstructured text is required.

A typical pipeline for opinion summarization has a number of stages as shown in Figure 2.1. Opinion retrieval initiates the process by fetching the user-generated content from the web. This is followed by a subjectivity filtering step where subjective or opinionated texts are separated from neutral or objective texts. Then, the subjective comments are fed into the sentiment classification module, which labels every review or sentence as positive, negative, or neutral. In certain systems, the data is divided into positive reviews and negative reviews explicitly to perform sentiment-specific analysis or to perform targeted summarization. The opinion summation component then processes these labeled inputs to compose a summary which is coherent, concise and reflects the emotional polarity and diversity of users' opinions. The result is a sentiment-aware summary, which can be used for stakeholders to rapidly comprehend the sentiment trends towards a product, service or topic.

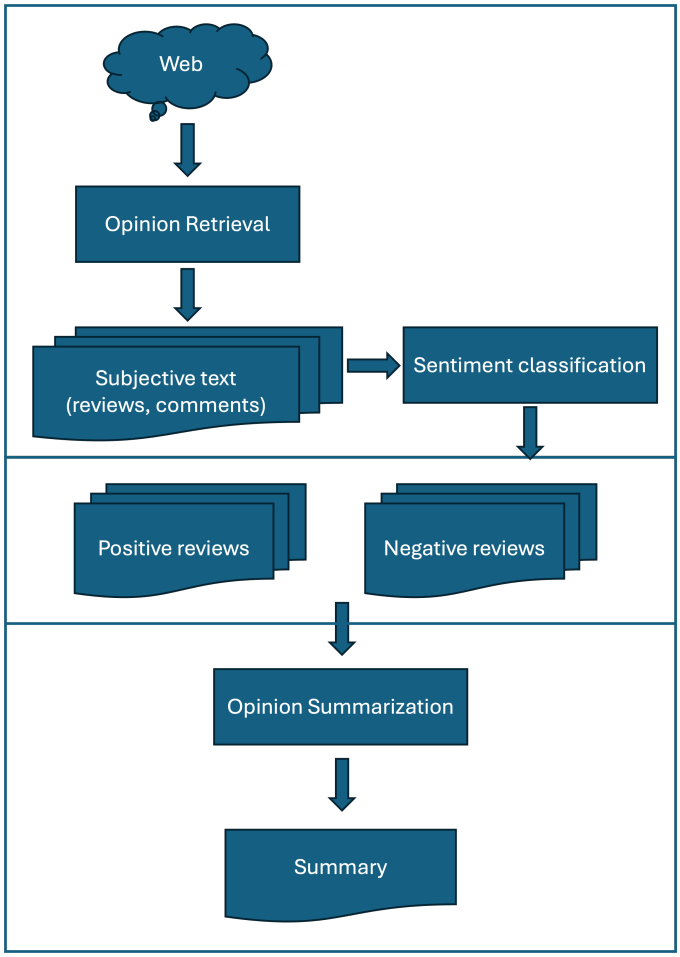


Figure 2.1: Architecture of opinion mining and summarization [1]

There are three main directions in opinion summarization: extractive methods, abstractive methods and weakly supervised or unsupervised approaches. Early extractive methods rely on utilizing only complete sentences that include explicit aspect–opinion pairs from the original text. However, they often struggle to produce coherent summaries due to redundancy and lack of fluency. With the spread of neural networks, especially transform-based models, abstractive methods have shown great promise. These models are trained to synthesize and write fluent summaries based on the information of the source text. For example, models like BART fine-tuned on review datasets can produce summaries that are not a set of selected sentences. Nevertheless, training these models requires large-scale annotated datasets, which are expensive to obtain.

Abstractive opinion summarization has been explored through a wide range of techniques, including template-based, graph-based, semantic, data-driven, and neural methods. Template-based approaches utilize predefined sentence structures guided by communicative intents such as speech acts, where classifiers like SVMs or Naive Bayes identify the type of topics, and key phrases are embedded into templated summaries [52]. Graph-based methods model reviews as word or phrase graphs, compressing and merging redundant opinions; for instance, the Opinosis framework [53] constructs summary paths over redundancy-aware graphs, while subsequent work enhanced fluency and sentiment fusion by integrating sentiment analysis and optimization techniques [54, 55]. Semantic-based approaches abstract content through deep representations such as semantic role labeling and predicate–argument structures, often employing optimization algorithms like genetic search or integer linear programming to select the most informative concepts [56]. Data-driven techniques like NAMAS [57] adopt sequence-to-sequence neural architectures with attention mechanisms, mapping input reviews to abstractive summaries, and follow-ups from IBM Watson [58] further improve generation using pointer networks and hierarchical attention. Finally, hybrid models combine extractive and generative techniques by first selecting salient quotes and then incorporating them into automatically generated summaries to provide supporting evidence [59].

Among various tasks in opinion summarization, aspect-based opinion summarization has received notable attention for its ability to organize sentiments according to specific aspects or features of a product. A typical framework involves three key steps: aspect extraction, sentiment polarity classification, and summary generation [60] as shown in Figure 2.2. One of the early frameworks [61] leveraged sentiment, feature frequency, and review scores to generate structured summaries. Hybrid approaches combining supervised and unsupervised polarity detection [60] have improved

robustness across domains by using topic detection and domain-specific lexicons. Some systems further enhance semantic organization through weakly supervised topic modeling using hashtags [62] or by evaluating feature relevance via user feedback. Despite promising results, current methods still face challenges related to domain generalization, mockery detection, and semantic understanding, motivating continued exploration of more flexible and context-aware summarization models.

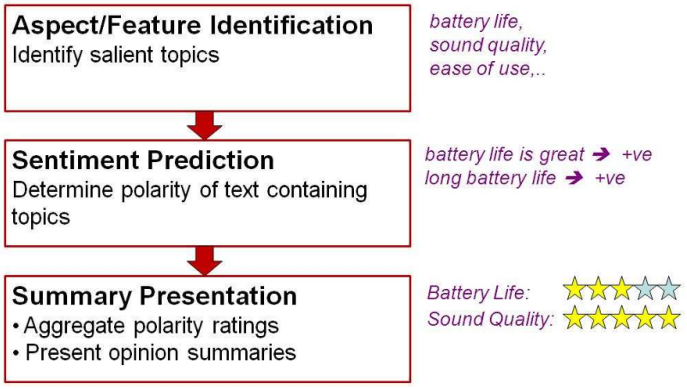


Figure 2.2: General three steps of aspect-based opinion summarization [1]

A notable work in weakly supervised opinion summarization is proposed by Liu et al. [63], where the authors address the shortage of reference summaries by introducing a novel method to synthesize training pairs composed of mix-structured input and textual output. Their method constructs the input by extracting two complementary forms of data: explicit opinion–aspect (OA) pairs and implicit sentences (ISs). The IS captures useful subjective content not formalized into OA pairs. They introduce a dual-encoder summarization model, where separate encoders are used to process OAs and ISs independently, and their outputs are fused during decoding to generate abstractive summaries. Compared to previous methods relying solely on either textual or structured input, this mixed-structured approach significantly improves the alignment between training input and target summaries. Inspired by this framework, our study builds upon their method and addresses several remaining challenges in the synthetic training process.

2.4 Large Language Models

Large Language Models (LLMs) such as BART [4], T5 [64], and GPT [65] models have significant advances in the domain of natural language pro-

cessing, particularly in generation-oriented tasks like summarization. These models undergo pretraining on extensive corpora utilizing self-supervised learning objectives, which enables them to encapsulate intricate syntactic and semantic structures. In the field of opinion summarization, LLMs have been extensively utilized owing to their capacity to produce coherent, fluent, and informative summaries, even in scenarios characterized by limited task-specific guidance [66].

Recent developments in large language models (LLMs) have catalyzed the formulation of diverse methodologies for sentiment summarization, particularly in instances where the data manifests as long-form, noisy, or weakly supervised. A number of contemporary studies investigate the application of LLMs to tackle various challenges that arise within this particular domain.

A representative approach is LFOSum [67], which addresses long-form opinion summarization by combining LLMs with extractive preselection, followed by controlled generation. The model identifies salient content and guides the summarizer using a contrastive ranking objective, improving factuality and user-centeredness in the generated summaries. In another line of work, XL-OpSumm [66] proposes an incremental summarization framework that generates summaries in small steps, guided by LLMs and memory-based fusion mechanisms, to maintain coherence over long inputs while capturing diverse user opinions.

For controllable and unsupervised generation, iteratively calibrated prompting [68] introduces a method to automatically adjust prompt templates for LLMs to better satisfy diversity and relevance goals without fine-tuning. This method is especially suitable for unsupervised settings and demonstrates strong performance across multiple domains.

Focusing on structured summarization, Korkankar et al. evaluate the capability of multiple LLMs (including GPT-4o, LLaMA3, Gemma2, and Mixtral) to generate aspect-specific summaries from Amazon reviews [69]. This study presents a pipeline combining aspect extraction, sentiment filtering, and model-driven generation, and compares LLM outputs using both traditional metrics (ROUGE, METEOR, and BERTScore) and GPT-4-based criteria (relevance, coverage, impurity, and goodness). Their results suggest GPT-4o performs best overall, while Mixtral and Qwen2 variants show competitive results for specific evaluation dimensions.

Additionally, Siledar et al. propose a unified prompting framework for multiple summarization tasks, demonstrating that the prompting technique can generalize across different summarization formats—including opinion, dialogue, and instructional summarization—without requiring specific fine-tuning for each task [70]. They highlight the increasing versatility of prompt-based LLM for domain-specific summarization.

In summary, these studies highlight the growing flexibility and potential of LLMs in handling the unique challenges of opinion summarization, including aspect alignment, implicit content, and prompt controllability. Our thesis builds upon these insights by integrating structured input modeling and polarity-aware sampling into a unified LLM-based summarization framework under weak supervision.

2.5 Transformer

The Transformer architecture [71] has emerged as a foundational model in natural language processing due to its ability to focus on long-range dependencies, support parallel computation, and achieve superior performance across various sequence modeling tasks. Unlike traditional recurrent neural networks, which handle the entire input sequentially, the Transformer uses a fully attention-based mechanism that enables each token to attend to all other tokens in a sequence simultaneously. This structure facilitates simple training and enhanced global context perception.

Figure 2.3 illustrates the architecture of the Transformer model. It follows an encoder–decoder structure, with the encoder and decoder comprising multiple stacked layers. Each layer consists of a multi-head self-attention mechanism and a position-wise feed-forward network, wrapped in residual connections and layer normalization. The attention mechanism is the core of the model, allowing the network to weigh the influence of different tokens when encoding or decoding information. Formally, given query (Q), key (K), and value (V) matrices, the scaled dot-product attention as shown on the left of Figure 2.4 is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{2.1}$$

where d_k is the dimensionality of the key vectors. This captures similarity between tokens and assigns higher weights to more relevant positions.

To enhance representational capacity, the Transformer uses multi-head attention as shown on the right of the **Figure 2.4**, where several independent attention mechanisms (or 'heads') are computed in parallel:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.2}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.3}$$

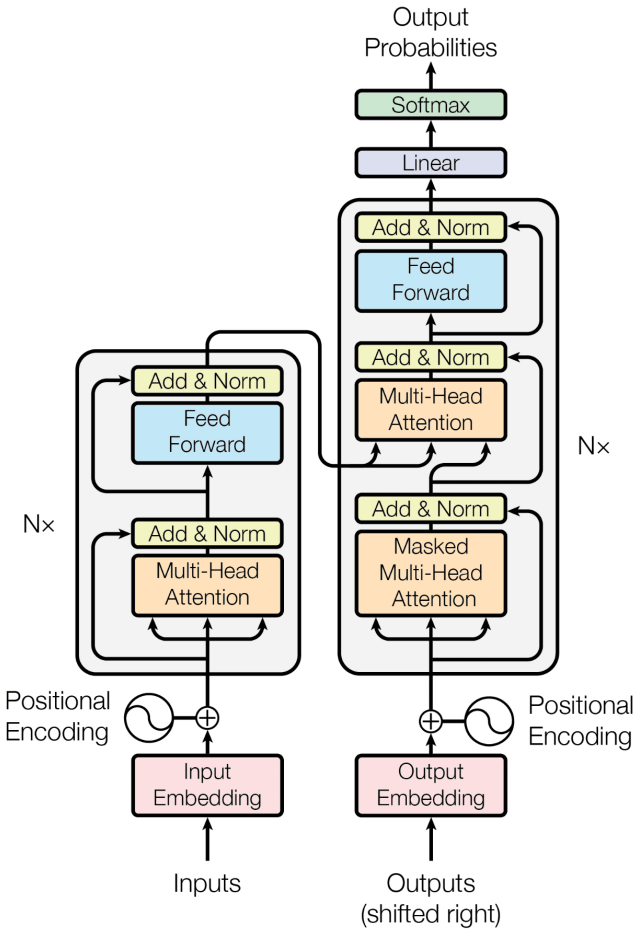


Figure 2.3: Architecture of Transformer [71]

Each head projects the inputs into different subspaces, enabling the model to capture diverse types of relations across tokens.

Since the model lacks recurrence, it employs positional encodings to incorporate information about word order. The original Transformer uses sinusoidal positional encodings defined as:

$$\mathrm{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{2.4}$$

$$\mathrm{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{2.5}$$

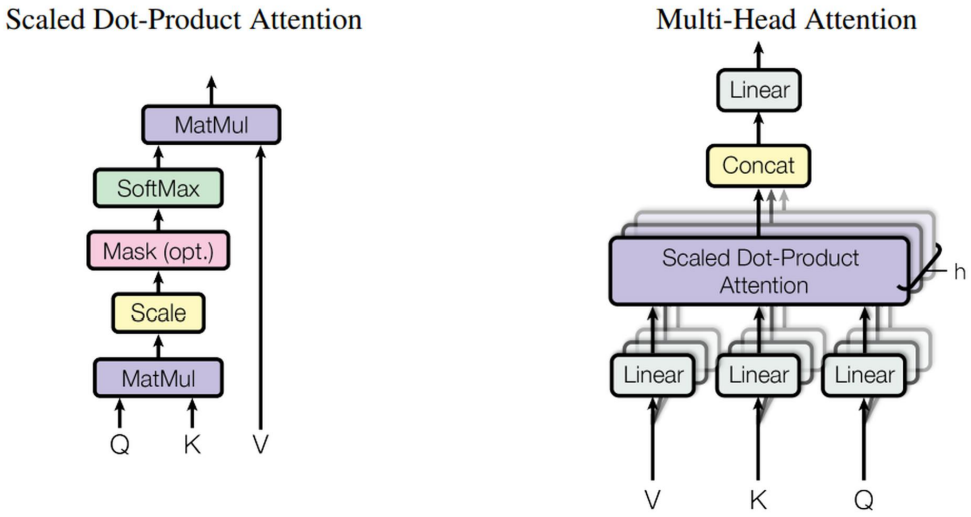


Figure 2.4: Scaled Dot-Product and Multi-Head Attention [71]

These encodings are added to the input embeddings at the bottom of the encoder and decoder stacks.

In addition to attention, each layer includes a position-wise feed-forward network applied independently to each token position:

$$\mathrm{FFN}(x) = \mathrm{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{2.6}$$

Residual connections and layer normalization are applied after both attention and feed-forward sub-layers to improve training stability and gradient flow. In the encoder, each layer consists of self-attention followed by a feed-forward network. The decoder adds an additional masked self-attention mechanism to prevent the model from accessing future tokens during training. The decoder has also cross-attention layers that attend to the encoder output to enable it to condition generation on the input context.

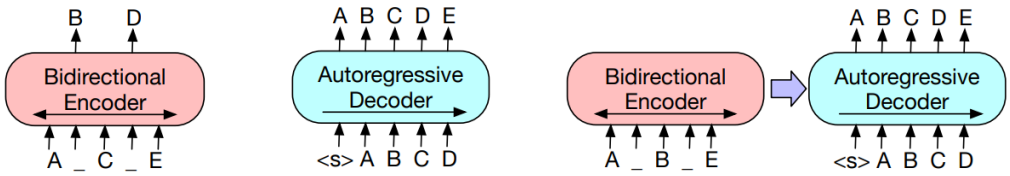
Thanks to its flexibility and performance, the Transformer has become the backbone of numerous large-scale language models such as BERT, GPT, T5, and BART. These models, which are typically pre-trained over large text datasets and fine-tuned on specific downstream tasks, have demonstrated strong performance in summarization, translation, and question answering. In this thesis, the proposed summarization framework leverages the Transformer architecture due to its ability to model rich contextual relationships and support generative tasks.

2.6 BART

BART (Bidirectional and Auto-Regressive Transformers) [4] represents a denoising sequence-to-sequence model developed by merging the capabilities of BERT [72] bidirectional encoder models with GPT [65] autoregressive decoder models. The model employs a Transformer-based encoder–decoder structure while denoising autoencoding training makes it excel at sequence generation tasks, including text summarization and translation and question answering.

The architecture of BART, which is described in Figure 2.5, consists of a Transformer encoder to process the entire corrupted input sequence in both directions while its Transformer decoder generates output through autoregressive prediction. The model combines BERT's understanding capabilities with GPT's generation capabilities because it uses complete sequence understanding during encoding and left-to-right prediction during decoding.

The training process of BART relies on a noising–denoising pre-training objective. The pre-training process involves applying different noise functions, including token masking, token deletion, text infilling, sentence permutation, and document rotation, to input sequences. The model learns to reconstruct the original text from these noisy versions, which helps develop strong semantic relationships and syntactic structures.



(a) BERT (b) GPT (c) BART

Figure 2.5: A schematic comparison of BART with BERT [72] and GPT [65]

After completing the pre-training step, the model is fine-tuned for each task. Regarding summarization, BART is usually trained via fine-tuning with document-reference summary pairs. The model encodes the document and then learns to generate the corresponding summary from scratch. Due to its denoising objective and powerful generative capabilities, BART consistently achieves state-of-the-art results on standard benchmark summarization datasets.

In the context of opinion summarization, BART has several advantages. Its encoder effectively captures the diverse input from user reviews, while its decoder generates coherent and fluent summaries that can incorporate paraphrasing and abstraction. Moreover, BART has generative capability to deal

with the implicit content, which is important for summarizing user opinions that may not always be explicitly stated. We believe that, due to its pre-training on large corpora and fine-tuning on limited task-specific examples, BART is especially effective in scenarios where high-quality human-written summaries are limited.

In this thesis, we use BART as the backbone model for abstract summarization. We improve it by incorporating additional information such as sentiment polarity, aspect-sentiment alignment, and implicit aspect signals into the input representation. This allows the generated summarization to be closer to the nuance of the user review, making it an effective tool for sentiment-aware summarization tasks.

2.7 Characteristics of this thesis

This thesis addresses key limitations in existing opinion summarization systems by proposing a sentiment-aware framework capable of handling both explicit and implicit opinions in user-generated reviews. While weak supervision has been previously explored, our work introduces several novel components that enhance the quality, coverage, and contextual understanding of opinion summaries.

First, we enhance the existing weakly-supervised opinion summarization framework to cover a broader range of sentiment expressions. By incorporating implicit reviews into the summarization pipeline, the proposed method is able to capture fine-grained opinions, allowing for a more faithful representation of user intent.

Second, our framework explicitly integrates sentiment polarity during both data construction and summarization. This polarity-aware design improves the consistency between the expressed sentiments in the input and the generated summaries, enabling better handling of subjective nuances in user feedback.

Third, we develop a weakly supervised data construction strategy that selects high-quality candidate summaries from unlabeled review sets based on semantic relevance and sentiment alignment. This allows our system to operate effectively in the absence of annotated summaries.

Our method demonstrates superior performance compared to existing baselines on benchmark datasets, achieving higher ROUGE-L scores and confirming the effectiveness of our contributions.

Chapter 3

Proposed Method

This chapter presents our proposed weakly supervised framework for opinion summarization using large language models. Section 3.1 presents an overview of the overall approach. Section 3.2 introduces the core component of the pipeline, including four submodules opinion-aspect pair extraction, implicit sentence extraction, sentiment polarity estimation, and pseudosummary selection. These components aim to construct diverse and semantically rich training inputs under weak supervision. Finally, Section 3.3 describes the architecture of the summarization model used to generate the final opinion summaries based on the constructed inputs.

3.1 Overview

The core idea of our proposed method is to construct a mix-structured training dataset composed of explicit and implicit opinion content and to train a summarization model that can effectively generate aspect- and sentiment-aware summaries without relying on manually annotated references.

The opinion summarization task is defined as follows. Given a set of user reviews $R = \{r_1, r_2, ...r_n\}$ for a particular product or entity, the goal of opinion summarization is to generate a concise and coherent textual summary S that captures the main aspects discussed in the reviews along with their associated sentiments. Figure 3.1 shows an overview of the opinion summarization task. Unlike extractive methods, we focus on generating abstractive summaries that may include rephrased or novel expressions. Under the weak supervision setting, we assume no access to gold-standard reference summaries during training.

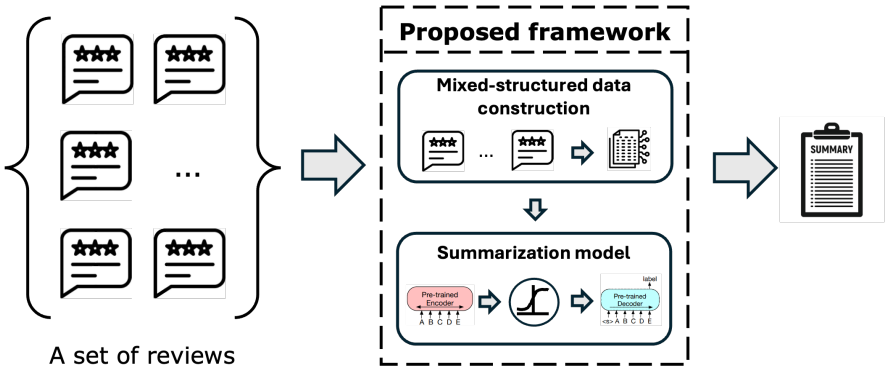


Figure 3.1: Overview of opinion summarization task

As illustrated in Figure 3.2, the proposed method is composed of two main stages: mix-structured data construction and summarization model training.

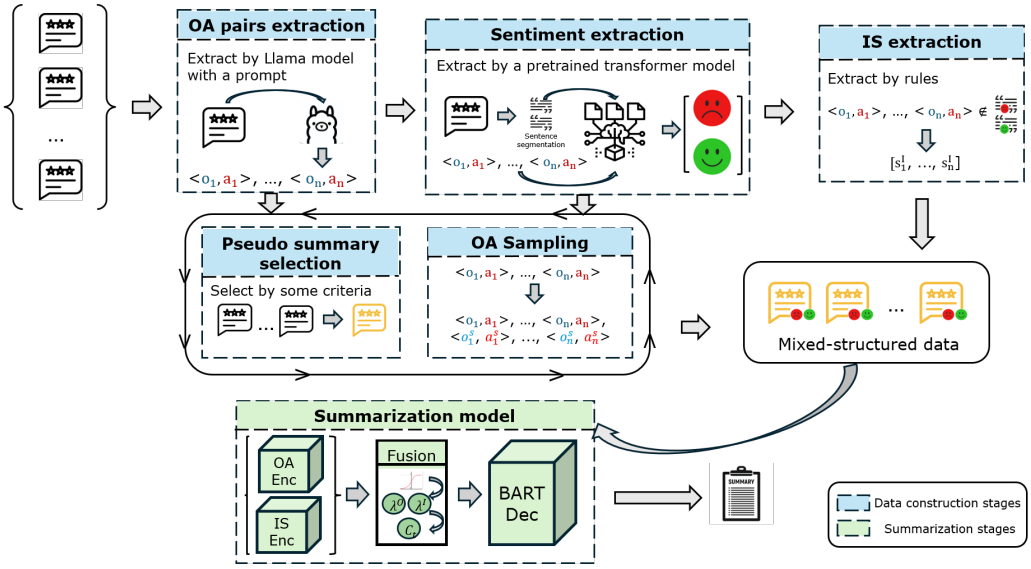


Figure 3.2: Overall structure of proposed framework

The first stage begins with a collection of user reviews, from which we extract three types of information: (i) opinion–aspect (OA) pairs, (ii) implicit opinion-bearing sentences (IS), and (iii) sentiment polarity. These components are obtained via dedicated modules for OA pair extraction, IS extraction, and sentiment analysis. To enable training without human-labeled references, we adopt a sentiment-aware pseudo-summary selection strategy, which chooses representative reviews that maintain sentiment balance and aspect diversity. The selected reviews are paired with the corresponding

structured content to form mixed-structured data, which serves as pseudo-labeled training pairs.

In the second stage, we fine-tune a summarization model on this constructed data. The model adopts a dual-encoder architecture, where OA pairs and IS sentences are encoded separately, followed by self-attention within each stream. The encoded outputs from both the OA and IS encoders are then attended by the decoder through encoder–decoder attention mechanisms to generate abstractive summaries. The architecture ensures that both explicit and implicit opinion information contribute to the generation process in a balanced and interpretable manner.

This two-stage framework enables effective training without human-written reference summaries, while promoting content coverage, sentiment diversity, and structural alignment between input and output.

3.2 Mix-structured Data Construction

This section details the process of constructing weakly supervised training data by extracting and combining different types of review content. The goal is to form a set of training instances that include explicit and implicit opinion information along with sentiment polarity. These mixed-structured representations are paired with pseudo summaries to serve as inputs for model training. The data generation process consists of four components, described in Figure 3.3.

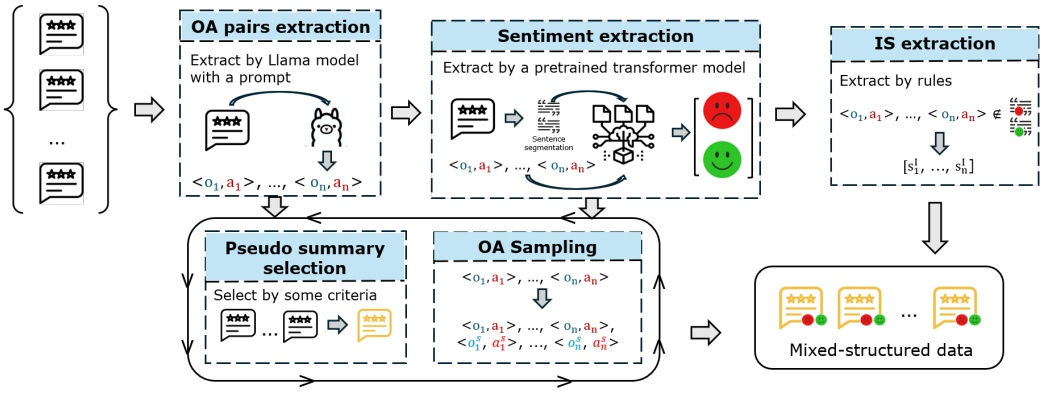


Figure 3.3: Mix-structured data construction architecture

3.2.1 Opinion-Aspect Pairs Extraction

To extract structured information from user reviews, we identify opinion–aspect (OA) pairs, where each pair consists of an aspect (typically a noun or noun phrase referring to a product feature) and an opinion (typically an adjective expressing sentiment). These pairs form the basis of the structured input used in our mix-structured training pipeline.

Instead of relying on traditional rule-based or dependency-parsing methods, we leverage a Large Language Model to perform OA extraction in a weakly supervised setting. Specifically, we utilize the LLaMA model and prompt it with a carefully designed instruction that enforces strict syntactic constraints on the output. We adopt LLaMA-2-7b-chat due to its competitive performance on instruction-following tasks and its open-source availability, which allows reproducibility and efficient deployment. The model is instructed to extract only pairs where the aspect is a noun and the opinion is an adjective, and to return results in a clean tuple-based format such as texttt [(camera, fantastic), (battery life, disappointing)].

Figure 3.4 shows how OA pairs are extracted using the LLM. To operationalize this process, we designed a prompt that clearly defines the extraction rules and avoids verbose responses. The prompt requires the model not to extract verbs, adverbs, or full sentences as aspects and strictly adhere to the output format. This approach provides several advantages, like it generalizing well to informal review text, maintaining output consistency, and eliminating the need for annotated data.

The prompt specifies strict syntactic constraints, requiring aspect terms to be nouns and opinion terms to be adjectives. It enforces structured output in a tuple format and avoids extra explanations, enabling efficient and consistent pair extraction from user reviews. These extracted OA pairs are later paired with sentiment polarity and combined with implicit content to construct mix-structured training inputs.

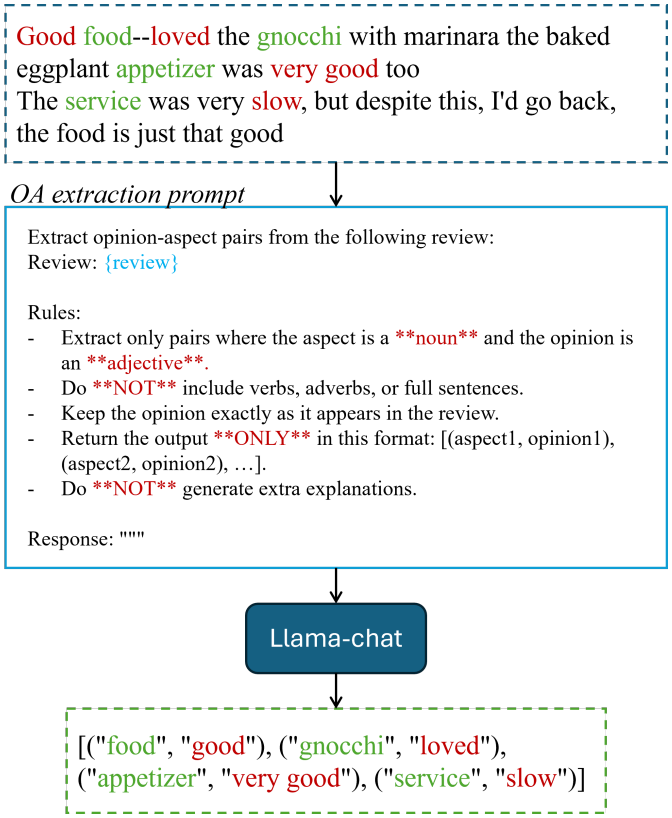


Figure 3.4: Extraction of opinion–aspect pairs using LLaMA model

3.2.2 Implicit Sentences Extraction

While opinion–aspect pairs capture structured opinions tied to specific features, many user reviews contain subjective statements that express sentiment in a more implicit or general form. These sentences, which often lack explicit aspect mentions, are still valuable in summarizing user impressions and thus are included as part of the mix-structured training data.

In our framework, we process each review by first splitting it into individual sentences. If a sentence does not yield any OA pairs, it is considered an IS candidate. To determine whether the candidate truly expresses an opinion, we apply a sentiment classifier. We utilize a pre-trained transformer-based sentiment analysis model `siebert/sentiment-roberta-large-english`[1] to identify whether a sentence conveys a positive or negative sentiment. Only those sentences that exhibit clear positive or negative sentiment are retained as valid ISs.

[1] `https://huggingface.co/siebert/sentiment-roberta-large-english`

Algorithm 1: Implicit Sentence Extraction Procedure

Input: A review $r_i = \{s_{i1}, s_{i2}, \cdots, s_{in}\}$
Output: An implicit sentence IS for r_i

$IS\_candidates \leftarrow \emptyset$;
$no\_OA\_sentences \leftarrow \emptyset$;
foreach $s_{ij} \in r_i$ **do**
    if $OA\_extraction(s_{ij}) = \emptyset$ **then**
        $no\_OA\_sentences \leftarrow no\_OA\_sentences \cup \{s_{ij}\}$;
        if $SentimentClassifier(s_{ij}) \in \{positive, negative\}$ **then**
            $IS\_candidates \leftarrow IS\_candidates \cup \{s_{ij}\}$;
if $IS\_candidates \neq \emptyset$ **then**
    return a list of valid $IS\_candidates$;
else
    return a random sentence from $no\_OA\_sentences$;

To ensure that every review contributes at least one IS to the training set, we apply a fallback strategy: if no IS with sentiment is found in a given review, we randomly select one sentence from that review as the implicit sentence. This guarantees coverage while maintaining diversity in sentence types and review contexts. Algorithm 1 outlines the procedure for extracting implicit sentences from a review.

This method ensures that ISs represent opinionated, unstructured content that complements the structured OA pairs. All extracted ISs are passed to the sentiment extraction step (subsection 3.2.3) and used to form the mix-structured input for training.

3.2.3 Sentiment Extraction

We estimate the sentiment polarity of both the extracted opinion–aspect (OA) pairs and implicit sentences (ISs). When the OAs and ISs are fed into the summarization model, the extracted sentiment is added to enhance the information of the input. Furthermore, the sentiment of OAs is also used for pseudo summary selection and sampling of OAs, which will be explained in subsection 3.2.4. Sentiment annotations allow the training pipeline to maintain a balanced representation of positive and negative opinions, which is essential for generating diverse and realistic summaries.

We employ the pretrained transformer-based model to return a sentiment label from the set POSITIVE, NEGATIVE, NEUTRAL along with a

confidence score. For each OA pair, we extract the opinion term (e.g., "disappointing") and pass it as input to the sentiment classifier. This lightweight formulation assumes the sentiment is sufficiently encoded in the opinion word. Each pair is then extended with the predicted sentiment and confidence score.

For implicit sentences, we pass the full sentence text directly to the classifier without modification. Sentences predicted as POSITIVE or NEGATIVE are retained, while those classified as NEUTRAL are discarded to reduce ambiguity. The remaining sentences are stored along with their polarity metadata, e.g.:

```
{
"text": "Totally worth it!",
"sentiment": {"label": "POSITIVE", "score": 0.9924}
}
```

The sentiment-enriched OA pairs and ISs are preserved for downstream use in pseudo summary selection (subsection 3.2.4), where sentiment balance is explicitly considered. This step enables the summarization model to learn from a mixed-structured input that reflects both the structural and emotional dimensions of user reviews.

3.2.4 Pseudo Summary Selection and Sampling of Opinion-Aspect Pairs

This section outlines the final steps for constructing mix-structured training instances by (i) selecting a pseudo summary from the review pool and (ii) sampling a set of OA pairs to simulate the input content typically summarized by that pseudo summary.

Pseudo Summary Selection To simulate a human-written summary in a weakly supervised setting, we select one review from training data as a pseudo summary. This selection is not random, but guided by semantic and structural constraints that aim to ensure the chosen review reflects common aspects, diverse sentiments, and a summary-like writing style. The candidate review must demonstrate content alignment with other reviews in the cluster and satisfy several heuristics to ensure its quality and representativeness. Specifically, we apply the following criteria to select a review as a pseudo summary:

1. **Aspect coverage constraint**: Let A be the set of aspects extracted from a candidate review, and A' be the set of aspects from all other

reviews in the training data. The candidate is only selected if $A \subseteq A'$, meaning it summarizes commonly discussed aspects.

2. **Sentiment balance**: Reviews should contain both negative and positive aspects, with the aim of covering the emotional spectrum in the summary.

3. **First-person avoidance**: Reviews that include first-person singular pronouns such as "I" and "my" are omitted to avoid overly subjective or personal descriptions.

4. **Noise reduction**: Reviews containing non-alphanumeric symbols are excluded to preserve textual clarity.

This strategy ensures that the pseudo summary approximates the structure and function of a gold summary—covering shared opinions in a concise and balanced manner.

Sampling of OA pairs After selecting a pseudo summary for a review cluster, we construct the corresponding mix-structured input by sampling a set of OA pairs from the remaining reviews. These pairs serve as the input content that the model will learn to summarize. To mimic the variability of natural reviews and promote a balanced representation, we organize OA pairs into two categories (popular and unpopular OA pairs) and apply a targeted sampling strategy for each.

- **Popular OA** pairs are pairs that have their aspect terms also appearing in the selected pseudo summary. Specifically, we use cosine similarity between the candidate opinion term and the corresponding opinion within the pseudo summary to select semantically aligned opinions using soft-matching. Each opinion word is represented using a 300-dimensional pre-trained GloVe embedding, with zero vectors assigned to out-of-vocabulary words. Cosine similarity is computed using the cosine_similarity function from PyTorch's module. This semantic matching strategy enhances the coherence and topical consistency between the input and the target summary.

- **Unpopular OA** pairs involve aspect terms that are not present in the summary. These are sampled randomly to promote diversity and introduce novel or less frequently mentioned aspects.

The number of OA pairs sampled from each review is not fixed, but follows a normal distribution to simulate natural variations in review length

and content density. At this stage, we do not apply any sampling to implicit sentences; all valid ISs associated with the review remain unchanged. This sampling strategy aims to preserve both relevance (through similarity-based selection) and variety, enabling the summarization model to generalize across a range of input conditions.

3.3 Summarization Model

To generate aspect- and sentiment-aware summaries from the constructed mix-structured input, we adopt a sequence-to-sequence (seq2seq) architecture with a dual-encoder design, which is illustrated in Figure 3.5. The model is conceptually straightforward yet effective in learning to map structured and unstructured data to coherent summaries. The model consists of two separate encoders:

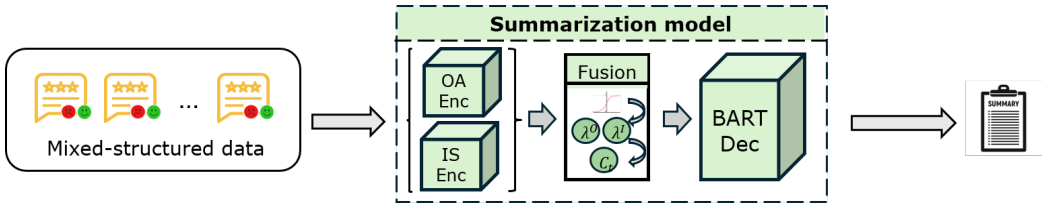


Figure 3.5: Summarization model architecture

- OA Pairs Encoder: processes the sequence of extracted opinion–aspect (OA) pairs and their sentiment.
- IS Encoder: which processes the implicit sentences (ISs) collected from the same reviews and their sentiment.

These encoders have non-shared parameters, allowing them to learn independent representations suitable for the different nature of their respective inputs. To better capture intra-type relationships, we prepend a special type token `[OA]` and `[IS]` to each input element, serving as a semantic anchor for representation aggregation.

To incorporate sentiment information, we append the phrase "*with sentiment <label>*" to each OA pair and implicit sentence before tokenization. For example, an OA pair such as "battery: great" with positive sentiment becomes "[OA] battery: great with sentiment positive". Similarly, an implicit sentence like "It lasts all day" is transformed into "[IS] It lasts all day with sentiment positive". These annotated inputs are then passed to the tokenizer and encoded by the respective BART-based encoder. This strategy

leverages BART's pre-trained language understanding to interpret sentiment cues as part of natural text, without requiring additional embedding layers or model modifications.

Both encoder outputs are passed through self-attention layers to obtain contextualized hidden states, denoted as H^O and H^I for the OA and IS encoders, respectively. These states are then used in an attention fusion mechanism during decoding. Specifically, for each decoder time step, attention distributions A^O and A^I are computed over the two encoder outputs. These are combined (via softmax and weighted sum) to form the final context vector C_t that conditions the decoder's output.

The decoder follows the standard Transformer-based structure with encoder–decoder attention, feed-forward layers, and layer normalization. It generates the summary token-by-token, attending dynamically to both OA and IS representations.

This architecture enables the model to flexibly align with both structured features and unstructured data, resulting in summaries that are more faithful and sentiment-aware than those generated from flat textual input alone.

In our implementation, we adopt the pre-trained BART model as the backbone for the sequence-to-sequence summarization architecture. Specifically, we initialize both the OA encoder and IS encoder from the BART encoder, and the decoder from the BART decoder. These two encoders are fine-tuned independently to capture domain-specific signals from structured and unstructured inputs. Leveraging pre-trained weights allows for more efficient learning and better generalization, especially given the limited size of our annotated dataset.

3.4 Difference with Liu's model

Our method is an extension of the opinion summarization framework proposed by Liu et al. [63]. While building upon the same objective of generating summaries from user reviews, our model introduces several significant improvements over prior work.

First, instead of using rule-based methods to extract opinion-aspect (OA) pairs, we leverage large language models (LLMs) to perform OA extraction. This method not only enhances flexibility and generalization but also allows the system to capture more diversity of user opinions.

Second, we introduce stricter filtering criteria for pseudo-summary selection. By applying these refined constraints, we ensure that only high-quality summaries are used for training supervision, thereby improving both training stability and output reliability.

Third, our model explicitly incorporates sentiment polarity into both the structured OA inputs and unstructured implicit sentences (ISs). This additional integration of sentiment polarity allows the decoder to better pay attention to and preserve sentiment signals during summarization, resulting in more sentiment-aware summaries.

These improvements contribute to a more robust and semantically faithful summarization model, especially in domains where sentiment and opinion structures play an important role.

Chapter 4

Evaluation

This chapter presents a comprehensive evaluation of the proposed framework. We begin by describing the datasets used for training and testing, followed by a comparison with strong baseline methods. After that, we outline the evaluation criteria and then experimental results are reported, accompanied by ablation studies to assess the contribution of each component in our pipeline. Finally, we provide in-depth discussions to highlight the strengths, limitations, and potential implications of our approach.

4.1 Datasets

To evaluate the effectiveness of our proposed framework, we conduct experiments on two publicly available opinion summarization datasets: Yelp and Amazon. These datasets cover both service and product review domains and include human-written summaries for reliable evaluation.

- **Yelp dataset** [1] consists of user reviews about local businesses and consumer services. Each sample contains a cluster of 8 reviews on the same entity (e.g., a restaurant), accompanied by one human-written summary [73]. We follow prior work and use 100 samples for development and 100 samples for test. In training, we use 43k synthetic review–summary pairs generated by our pipeline.
- **Amazon dataset** [2] comprises product reviews spanning multiple categories. Each development and test sample contains 8 reviews and three human-written summaries, providing multiple reference points for evaluation [74]. The development and test sets include 28×3 and

[1] `https://business.yelp.com/data/resources/open-dataset/`
[2] `https://jmcauley.ucsd.edu/data/amazon/`

32×3 samples, respectively. For training, we construct 11k synthetic pairs.

Table 4.1 summarizes the number of training, development, and test samples used for each dataset.

Table 4.1: Dataset statistics.

Dataset	Training	Mix-structured data	Development	Test
Yelp	1M	43,441	100	100
Amazon	100k	10,886	84	96

Figures 4.1 and 4.2 illustrate representative samples from the test sets of the Yelp and Amazon datasets, respectively. Each figure displays the full set of input reviews grouped by domain, along with the corresponding human-written summaries.

Yelp Test Sample

Input Reviews:
1. I eat here once a year or two. It is always good...
2. Great authentic Mexican food at a reasonable price...
3. We love this place! Our "date night" always consist of Margaritas...
4. This was my family's first time visiting and we had an amazing experience...
5. The service is pretty good but not extraordinary...
6. Patio seating is great. Waiters are awesome...
7. The best bar service in the area. Jose. Rocky. Drew. You rock...
8. One of the best Mexican restaurants I've been to...

Human-written Summary:
The servers are kind and knowledgeable, they will patiently answer your questions. They offer patio seating if you'd prefer to sit outside. The free chips and salsa are always a plus, and the margaritas are amazing too. The menu is full of tasty authentic Mexican dishes.

Figure 4.1: Example reviews and summary from Yelp test set

Amazon Test Sample

Input Reviews:
1. This shoe is very classy and chic. The colour is rich and...
2. This shoe was picked by the bride-to-be for her upcoming wedding...
3. I love the look of these shoes and they will forever be special to me as I wore them at my wedding...
4. These are VERY nice shoes and so pretty, but there is a rim inside...
5. These shoes run 0.5 size small...they are very comfortable and cute.
6. I bought these in navy to wear in my sister's wedding... The lower heel gave me just enough height.
7. These shoes run appropriate to size, but I felt like they pinched my toes as I have a wider foot.
8. Great fast Amazon shipping... a comfortable shoe for a wedding...

Human-written Summaries:
Summary 1: *Very pretty shoes and nice quality. The shoes run a bit small, about half a size, and there is a ridge in the shoe that rubs on your toe. Nice formal night shoe, not so much for every day.*
Summary 2: *Great fast amazon shipping. This shoe is very chic, classic and with a rich color that blends with almost any navy blue colored outfit. I love the look of the shoes and would be forever special to me. There is a rim inside the topmost part of the shoes that could cause blister to my toes.*
Summary 3: *These shoes are best fit and nice looking for parties or specially for wedding functions. The toe of the shoe can be a challenge for some feet and may make the feet sore due to the rim to keep the shape perfect. Overall, price wise and looks are great and would recommend for long term use.*

Figure 4.2: Example reviews and summaries from Amazon test set

4.2 Baselines

We compare our proposed framework against several strong baselines that represent various directions in opinion summarization. These baselines are summarized below. Each is trained on its own synthetic training data.

- **MeanSum** [73]: A reconstruction-based framework that produces summaries by decoding from the mean latent representation of input reviews, trained without access to reference summaries. The synthetic training data is implicitly created by using each review to train an auto-encoder and then generating a pseudo-summary by averaging the latent representations of the input reviews. The model learns to make this pseudo-summary semantically similar to the original reviews using a cosine similarity loss.

- **Copycat** [74]: A variational autoencoder model is used to enable unsupervised many-to-one summarization by learning latent representations of individual reviews. The synthetic training data is constructed by converting input reviews to their semantic embeddings, ensuring the model captures semantics of an entire review. At inference time, they compute the mean of the embeddings from multiple reviews, and decode it into a summary.

- **OpiDig** [75]: A weakly supervised approach that leverages structured data in the form of opinion–aspect pairs extracted from reviews. The approach is trained to reconstruct the review text from these structured representations, thereby guiding the summarization process. The synthetic training data consists of extracted opinion phrases used as model inputs, with the original reviews serving as output targets.

- **Denoise** [76]: An unsupervised approach that corrupts input data through syntactic and semantic noise to train the model in a denoising fashion, aiming to encourage robustness and generate coherent summaries. The synthetic training data is created by selecting a review as a pseudo-summary and generating noisy versions through controlled token-level and sentence-level perturbations.

- **FewSum** [77]: A conditional transformer model designed to address challenges such as content coverage, stylistic variation, and length control, especially under limited supervision scenarios. The model is first trained on large-scale unannotated reviews using a leave-one-out objective, then fine-tuned on a few manually written summaries. The

unannotated reviews and their contextual properties act as synthetic supervision during the pretraining phase.

- **Weak-Supervision Sum** [63]: This method synthesizes mix-structured training data by combining structured opinion–aspect pairs with unstructured opinionated sentences. Synthetic training data is constructed by sampling a review as a pseudo-summary and pairing it with sampled opinion–aspect pairs and implicit sentences extracted from other reviews about the same entity. It utilizes a dual-encoder model to encode each input type separately and merges their representations during decoding through attention. Among the methods compared, this is the most conceptually aligned with our proposed framework.

4.3 Evaluation Criteria

We evaluate output quality and diversity in the model by standard automatic metrics, and we report ROUGE and self-BLEU scores in particular.

ROUGE-N [78]: We use the F1 scores of ROUGE-1, ROUGE-2 to compute n-gram overlap between generated summaries and human-written reference summaries. ROUGE-1 and ROUGE-2 report unigram-level and bigram-level content overlap. The definition of ROUGE-N is given as follows.

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \tag{4.1}$$

- gram_n denotes any n-gram (e.g., unigram for ROUGE-1, bigram for ROUGE-2).
- $\text{Count}_{\text{match}}(\text{gram}_n)$ is the maximum number of n-grams co-occurring in both the generated summary and reference summary.
- $\text{Count}(\text{gram}_n)$ is the total number of n-grams in the reference summary.

ROUGE-L [78]: ROUGE-L measures the longest common subsequence (LCS) between the generated and reference summaries. It captures sentence-level structure similarity without requiring consecutive matches. The definition of ROUGE-L is given as follows.

$$\text{ROUGE-L}_{F1} = \frac{(1+\beta^2) \cdot \text{LCS}_{\text{precision}} \cdot \text{LCS}_{\text{recall}}}{\text{LCS}_{\text{precision}} + \beta^2 \cdot \text{LCS}_{\text{recall}}} \tag{4.2}$$

$$\text{LCS}_{\text{precision}} = \frac{\text{LCS}(X,Y)}{|X|}, \quad \text{LCS}_{\text{recall}} = \frac{\text{LCS}(X,Y)}{|Y|} \tag{4.3}$$

- $\text{LCS}(X, Y)$ is the length of the longest common subsequence between the generated summary X and the reference summary Y.
- $|X|$ is the length (number of words or tokens) of the generated summary.
- $|Y|$ is the length (number of words or tokens) of the reference summary.
- β is a parameter that balances recall and precision, typically set to 1.

Self-BLEU: To measure the lexical diversity of generated outputs, we compute self-BLEU [3] [79]. For each generated summary, self-BLEU treats all other generated summaries as references and computes the BLEU score. A lower self-BLEU value indicates higher diversity, indicating that the model avoids generating repetitive or generic outputs across different samples.

4.4 Experimental Results

Table 4.2 presents the automatic evaluation results of our proposed framework and a number of baselines on Amazon and Yelp datasets. In this table, R-1, R-2, R-L and S-B stand for ROUGE-1, ROUGE-2, ROUGE-L, and self-BLEU, respectively. We compare models using ROUGE-1, ROUGE-2, and ROUGE-L to assess content relevance, while self-BLEU is used to measure diversity; the latter being a metric in which lower values indicate more diverse generations.

Table 4.2: Automatic evaluation results on the Yelp and Amazon test sets. Best values are bolded.

Approach	**Yelp**				**Amazon**			
	R-1↑	R-2↑	R-L↑	S-B↓	R-1↑	R-2↑	R-L↑	S-B↓
MeanSum [73]	28.86	3.66	15.19	0.38	29.20	4.70	18.15	0.40
Copycat [74]	29.47	5.26	18.09	0.34	31.97	5.81	20.16	0.43
OpiDig [75]	29.96	5.00	17.33	0.33	29.02	5.14	17.73	0.32
Denoise [76]	30.14	4.99	17.65	0.27	31.76	5.85	19.87	0.27
FewSum [77]	31.96	5.64	17.77	0.28	32.04	5.93	20.03	0.30
Weak-Supervision Sum [63]	**36.78**	**8.66**	20.52	**0.20**	**34.50**	**7.64**	20.73	0.26
Our model	25.31	3.88	**23.09**	0.4718	31.18	5.56	**28.68**	**0.0685**

On the Yelp dataset, our model achieves the highest ROUGE-L score (23.09). In contrast, Weak-Supervision Sum attains the highest ROUGE-1 and ROUGE-2 scores (36.78 and 8.66), along with a notably low self-BLEU score (0.20), indicating strong lexical alignment and diversity. Considering

[3] `https://github.com/geek-ai/Texygen`

the limitations of ROUGE-1 and ROUGE-2, which evaluate the overlap of short fragments (one or two words) in generated and reference summaries, the fact that our method achieves the highest score on ROUGE-L indicates its superiority over the baselines.

Our proposed method on the Amazon dataset achieves the best ROUGE-L score (28.68) of all baselines, significantly outperforming FewSum (20.03), Copycat (20.16), and Denoise (19.87) by a large margin. Although Weak-Supervision Sum achieves the best ROUGE-1 (34.50) and ROUGE-2 (7.64) scores, our model exhibits much better diversity, with the lowest self-BLEU (0.0685) across all systems. This indicates that our summaries are not only informative but also less repetitive and more varied in expression.

While the self-BLEU score of 0.0685 in the Amazon dataset may seem unusually low, this result is a direct result of our model's ability to avoid word repetition. Unlike template-based models, our system generates distinct and lexically diverse summaries for different input instances, even when they belong to similar product categories. Furthermore, Amazon's current test set is relatively small and spans four diverse domains (Clothing, Footwear, and Jewelry; Electronics; Health and Personal Care; and Kitchen and Home), which leads to differences between the generated outputs. Therefore, the low self-BLEU is not an anomaly, but rather the effectiveness of our approach in generating domain-adaptive, non-duplicate summaries that prioritize diversity without sacrificing informativeness.

These results demonstrate that our method is capable of generating summaries that are both accurate and diverse, striking a desirable trade-off between content coverage and linguistic variation. We attribute this improvement to the use of sentiment-aware pseudo-summary selection and structured sampling, which together help expose the model to a wider variety of inputs while preserving sentiment and aspect alignment.

4.5 Ablation Study

To evaluate how each component in our proposed framework works well, an ablation study is carried out by removing or modifying specific components to measure performance changes. Specifically, we design five ablated variants, each modifying or removing a key element from the full pipeline. These experiments are performed on the Amazon and Yelp datasets using a mix-structured training set (1M Yelp, 100k Amazon). All models are trained under the same conditions and evaluated using ROUGE-1, ROUGE-2, ROUGE-L, and self-BLEU. Each ablated model is constructed to isolate the impact of one particular factor on the overall summarization performance. The details

of the ablated models are shown below.

- **OA Ablation**: Opinion–aspect (OA) pairs are entirely removed from the input during training. Since OA information is used in multiple stages (input construction, sampling, alignment), this ablation examines the contribution of structured opinion–aspect pairs to the overall performance of the framework.
- **OA LLM Ablation**: OA pairs are retained, but extracted using a heuristic rule-based method instead of a large language model. This tests the importance of extraction quality in the training pipeline.
- **IS Ablation**: Implicit sentences (ISs) are excluded from the input. This isolates the effect of unstructured, sentiment-rich content on summary quality.
- **Sentiment Ablation**: Sentiment scores are removed from both OA and IS components. As a result, the model receives OAs and ISs without their sentiment, which may reduce its ability to generate sentiment-aware summaries.
- **Pseudo Summary Selection Ablation**: The pseudo-summary selection strategy is replaced by uniform random sampling. That is, one review from each review group is randomly chosen to serve as the training summary, without considering polarity or aspect alignment.
- **Full Model**: We incorporate all of the above components, using OA+IS input, sentiment filtering, and guided pseudo-summary selection.

Figure 4.3 shows the ROUGE-1, ROUGE-2, ROUGE-L, and Self-BLEU of the full model and five ablated models on the Yelp dataset. Concerning ROUGE-L, the OA Ablation results in a notable decline (from 23.09 to 21.01), indicating that OAs play an important role in opinion summarization. The Pseudo Summary Selection Ablation and the IS Ablation also lead to a significant decline, suggesting that the contributions of these components are not negligible. On the other hand, the OA LLM Ablation is almost equivalent to the full model. This indicates that the extraction of OAs by an LLM and by rules is not significantly different.

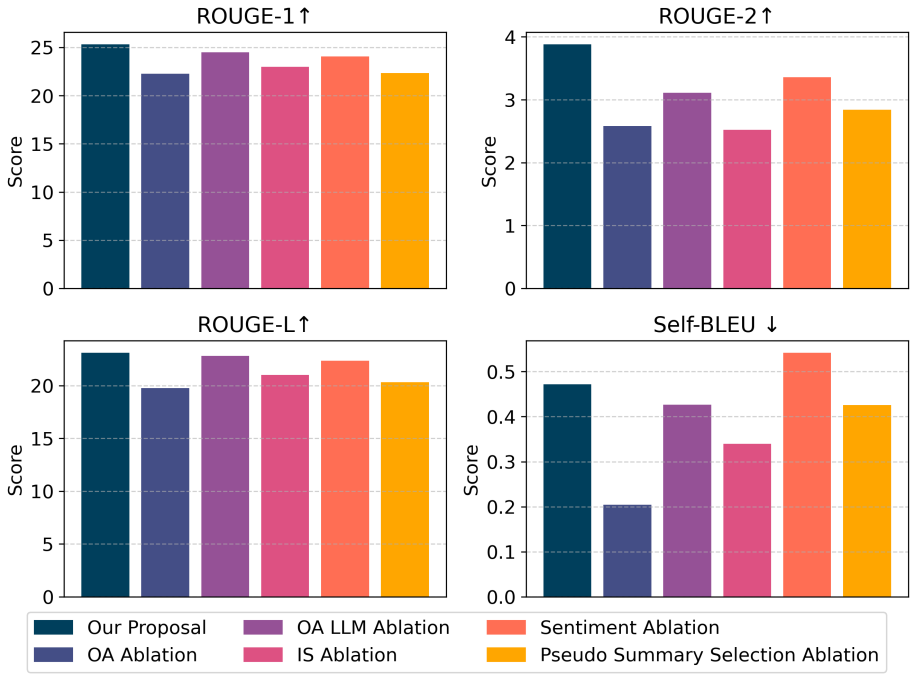


Figure 4.3: Ablation study on Yelp dataset

Figure 4.4 shows the results of the ablation study on the Amazon dataset. Among all components, IS Ablation leads to the most significant performance drop across all ROUGE metrics, particularly ROUGE-L, where the score drops sharply to around 24. This suggests that implicit sentences (IS) play a crucial role in enhancing content richness and coherence. Pseudo-summary Selection Ablation and Sentiment Ablation also degrade performance, indicating that selecting appropriate pseudo-summaries and leveraging sentiment alignment are both essential for generating effective summaries. Notably, this degradation trend contrasts with the results observed on the Yelp dataset, where OA-related ablations had a greater impact. These findings highlight the greater importance of implicit and semantic structure on the Amazon domain, possibly due to the more diverse and subjective nature of product reviews.

Overall, these results demonstrate that every model component, including implicit sentence integration, sentiment filtering, and pseudo-summary selection, significantly improves summary generation by producing coherent and diversified sentiments in the output.

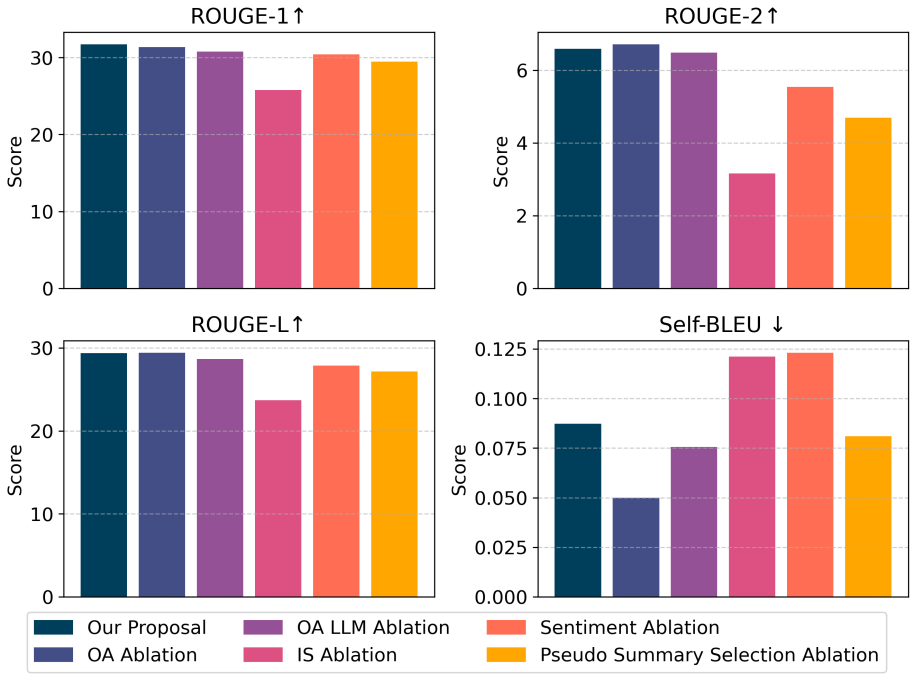


Figure 4.4: Ablation study on Amazon dataset

4.6 Evaluation of Implicit Sentence Augmentation

To assess the impact of augmenting implicit sentences on summarization performance, we conduct a series of experiments using ISs paraphrased by large language models. Specifically, we utilize the publicly available pre-trained Pegasus paraphrase model to generate one or two paraphrased variants for each implicit sentence in our input. The augmented IS are appended to the input structure before summarization.

This augmentation aims to enrich the input space with diverse yet semantically related information, potentially improving the model's ability to generalize over latent opinions not explicitly associated with aspect terms. We compare three settings below.

- Full model: uses the original IS without augmentation.
- Paraphrase 1 IS: adds one paraphrased variant per IS.
- Paraphrase 2 ISs: adds two paraphrased variants per IS.

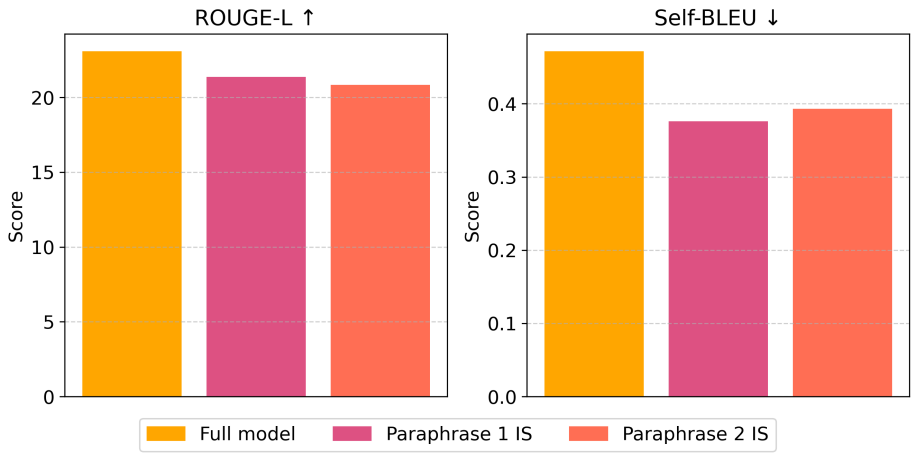


Figure 4.5: Results of Implicit Sentence Augmentation on YELP dataset

The ROUGE-L and Self-BLEU on the Yelp and Amazon datasets are shown in Figure 4.5 and Figure 4.6, respectively. The original model performs best, achieving 23.09 and 28.68 ROUGE-L, respectively. Augmenting IS with paraphrases slightly degrades performance, particularly on Yelp. These results indicate that while paraphrasing increases diversity, it may also introduce semantic drift or redundancy. This may cause a decrease in the quality of the generated summary.

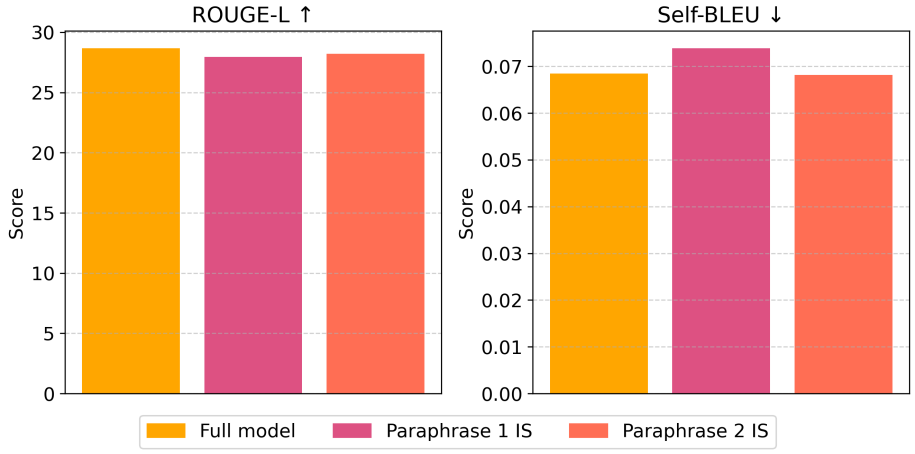


Figure 4.6: Results of Implicit Sentence Augmentation on Amazon dataset

4.7 Evaluation of Opinion-Aspects Pairs Extraction

To evaluate the effectiveness of our Opinion-Aspect (OA) extraction module, we conducted a detailed manual evaluation on 50 randomly sampled reviews from each of the Yelp and Amazon datasets. The details of 50 reviews are illustrated in Appendix A. For each review, we compared the automatically extracted OA pairs with human-annotated gold standards. The evaluation focuses on two key criteria: precision and recall. Figure 4.7 presents a comparative analysis of the OA extraction performance under exact and partial matching criteria on the Yelp and Amazon datasets, respectively. The OA pairs in which the opinion word and the aspect word are coincident with those in the ground-truth are judged as correct in the exact matching. In contrast, the OA pairs where either the opinion word or the aspect word aligns with the ground-truth are considered correct in the partial matching.

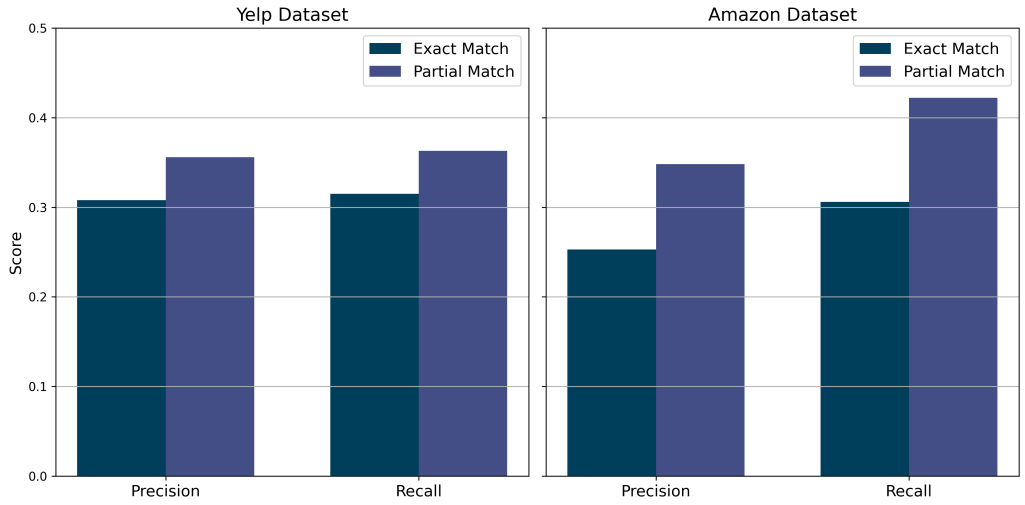


Figure 4.7: Evaluation of OA extraction on the Yelp dataset

Generally, the evaluation results indicate that while the model demonstrates a reasonable ability to extract OA pairs, its performance remains limited. Precision and recall scores lower than 40% suggest that the model still produces some irrelevant pairs and misses many valid ones. Across both datasets, we observe a consistent difference in performance when comparing the exact and partial match evaluations. For Yelp, the precision and recall are increased from 30.8% and 31.5% to 35.6% and 36.3%, respectively. A similar trend is observed in the Amazon dataset, where precision increased

from 25.3% to 34.8% and recall increased from 30.6% to 42.2%. These results suggest that while the model may not always capture OA pairs in exact lexical form, it frequently extracts semantically relevant pairs that overlap meaningfully with the human annotations.

Table 4.3: Example of extracted opinion–aspect pairs from a user review.

Review Excerpt	**Extracted OA Pairs**
Love these yoga pants. With a 36 inch inseam I love that these actually go over my heels. The fabric is perfect in between thick and thin, comfortable medium weight. And the pockets! Who doesn't love pockets? I would buy these again and recommend to anyone who is y'all and looking for a quality quality legging.	1. ⟨pants, comfortable⟩ 2. ⟨fabric, perfect⟩ 3. ⟨pockets, loved⟩ 4. ⟨inseam, goes over⟩ 5. ⟨quality, quality⟩

To better illustrate major causes of errors, we analyze a representative review and examine the output of the OA extraction module in detail. The example of the extracted opinion–aspect pairs are presented in Table 4.3. While some pairs are correct, ⟨quality, quality⟩ are redundant, and ⟨inseam, goes over⟩ is vague in expressing polarity. These types of noisy inputs, when propagated through the pipeline, can subtly degrade the training signal and the quality of the generated summary. Although our system demonstrates robustness against such imperfections, this observation highlights the potential benefit of integrating post-processing or confidence-based filtering for OA extraction.

4.8 Error Analysis

To better understand the limitations of the proposed weakly supervised opinion summarization system, we conducted error analysis on two evaluation datasets. Our analysis focused primarily on ROUGE-2 scores, which is the lowest score of our model, as well as the quality of the generated summaries.

Low ROUGE-2 Scores Across both datasets, we observed consistently low ROUGE-2 scores. On Yelp, the average ROUGE-2 was only 0.0388, with some samples scoring as low as 0.000. Similarly, on Amazon, the average

ROUGE-2 reached only 0.0556. These results suggest a notable lack of bi-gram overlap between generated summaries and references. Such low scores typically indicate such problems as poor word ordering, surface form mismatch, or weak semantic alignment. For instance, generated outputs often used alternative phrasings or omitted critical transitional bi-grams, thereby reducing ROUGE-2 score.

This phenomenon is further exemplified in specific error cases. In a sample in the Yelp dataset, the model produced a summary "The Willis Tower is very delicious, and I . . . " where Willis Tower is the name of a building that has a sushi restaurant. Compared this with the human-written summary "I'm a big sushi fan and this place did not disappoint," which conveyed a similar meaning but with different lexical choices.

In another case, although candidate opinion–aspect and implicit sentence were technically present, their alignment with the review content was weak, resulting in a generated summary ("Candance and Shehelle were amazing, ...") that failed to reflect the actual review content ("The staff and service are top notch."). These failures suggest that the model struggles to connect extracted inputs to semantic of the reviews meaningfully.

On the Amazon dataset, similar problems were observed. In one example, the generated summary ("This thing works. The sound is great, and worth the price.") shared only shallow lexical similarity with the human-written summary ("This is a great product. It has amazing sound and value for money."). While both conveyed positive sentiment, the lack of lexical and structural overlap led to low ROUGE-2 score, likely due to limited training diversity and poor alignment of word-level co-occurrence patterns.

High Variance When examining the model outputs, we also observed high variance of evaluation metrics across samples. In the Yelp data, ROUGE-1 score fluctuated between 0.109 and 0.426, while in the Amazon between 0.2017 and 0.4865. Similar significant variability was observed for ROUGE-L score, indicating inconsistent performance across reviews. This inconsistency means that while the model can perform well on some individual instances, it does not do so for all of them.

4.9 Summary of Experiment

Our experimental results reveal several important insights on the effectiveness and design choices of the proposed framework. First, while our method does not achieve the highest ROUGE-1 and ROUGE-2 scores compared to the strongest baseline (Weak-Supervision Sum), it consistently obtains the

highest ROUGE-L on both Yelp and Amazon datasets. ROUGE-L is often more indicative of sentence-level fluency and structure preservation, suggesting that our summaries are more coherent and better aligned with human-written outputs. Moreover, our model significantly outperforms baselines in diversity (self-BLEU), confirming its ability to avoid repetitive phrasing and generate more lexically varied summaries, an important property in opinion summarization tasks where richness of expression enhances user trust and readability.

Second, the ablation study highlights the importance of each component in the framework. The removal of implicit sentences (IS) or sentiment-aware filtering consistently leads to large performance drops across both relevance and diversity metrics. This validates our decision to include not only structured opinion–aspect pairs but also unstructured, sentiment-rich content as input. Similarly, pseudo-summary selection proves to be an effective weak supervision approach; replacing it with random reviews significantly degrades the model's ability to learn input–output alignment.

Furthermore, although our ROUGE-1 and ROUGE-2 scores are slightly lower than those of the strongest baseline, this can be attributed to our model's more flexible phrasing and higher lexical variation. ROUGE-1 and ROUGE-2 are based on exact n-gram overlap, which may favor templated or generic wording, while our summaries tend to use more expressive and semantically equivalent alternatives. This trade-off suggests that our model prioritizes meaningful diversity and linguistic naturalness over surface-level overlap—a desirable property in practical opinion summarization systems.

Chapter 5

Conclusion

The final chapter summarizes the thesis contributions while outlining prospective directions for upcoming research initiatives. The proposed framework's primary components and experimental outcomes receive their summary in section 5.1. Section 5.2 explores several promising directions to enhance the proposed framework.

5.1 Summary

This thesis proposed a weakly supervised framework for opinion summarization that did not rely on human-written reference summaries. To construct effective training data, we introduced a mix-structured input formulation that combined both structured opinion–aspect (OA) pairs and unstructured implicit sentences (ISs), enriched with sentiment information. To automatically construct a ground-truth summary for a set of reviews, a sentiment-aware pseudo-summary selection strategy was developed by balancing polarity and content coverage.

The summarization model adopted a dual-encoder architecture, in which OA and IS components were processed separately and fused via attention during decoding. Through extensive experiments on the Yelp and Amazon datasets, our method achieved strong performance in both the quality of the generated summary and lexical diversity, outperforming multiple baselines in ROUGE-L and self-BLEU metrics. The effectiveness of the components, including OA pairs, ISs, sentiment filtering, and pseudo-summary sampling strategy, was empirically validated through ablation studies.

Overall, this work demonstrated that structured and sentiment-aware weak supervision could serve as an effective substitute for manual annotations in opinion summarization, paving the way for scalable and domain-adaptable

opinion summarization systems.

5.2 Future Work

One potential opportunity lies in scaling up the volume of synthetic training data. Our existing implementation generates 43k and 11k mix-structured samples for Yelp and Amazon, but these numbers fall short of the 100k and 90k samples employed by previous benchmarks. Exploring ways to generate a larger volume of high-quality data could further enhance model performance and generalization.

Another promising direction is to improve the quality and reliability of the input structures, particularly the OA extraction component. The semantic and syntactic errors in LLM-based OA extraction discussed in section 4.9 create noise that affects training signals. Future work could explore the use of confidence scoring, post-hoc filtering, or alignment with external knowledge sources to improve the preciseness of this step.

A modeling approach could benefit from exploring advanced architectures that include fully LLM-based summarization together with retrieval-augmented generation under few-shot or instruction-tuned paradigms. The method should be tested for its generality by implementing domain adaptation together with multilingual extensions because our current tests focus only on product and service reviews in English.

Finally, the proposed framework enables practical applications, including e-commerce platforms, along with review aggregation systems and tools that summarize opinions for personal use. The system becomes more adaptable and user-focused when it incorporates user feedback into the learning process and produces summaries with controllable sentiment polarity.

Bibliography

[1] Mohammed Elsaid Moussa, Ensaf Hussein Mohamed, and Mohamed Hassan Haggag. A survey on opinion summarization techniques for social media. *Future Computing and Informatics Journal*, 3(1):82–109, 2018.

[2] Nur Hayatin, Suraya Alias, and Lai Po Hung. Trends and challenges in sentiment summarization: a systematic review of aspect extraction techniques. *Knowledge and Information Systems*, 66(7):3671–3717, 2024.

[3] Asim Ullah Jan, Mohammad Abid Khan, and Neelam Mukhtar. Opinion mining and summarization: A comprehensive review. *Journal of Information Communication Technologies and Robotic Applications*, pages 76–96, 2020.

[4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[5] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[6] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.

[7] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. *Mining text data*, pages 415–463, 2012.

[8] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134, 2021.

[9] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.

[10] Zhao Yan-Yan, Qin Bing, and Liu Ting. Integrating intra-and inter-document evidences for improving sentence sentiment classification. *Acta Automatica Sinica*, 36(10):1417–1425, 2010.

[11] Alejandro Moreo, Manuel Romero, JL Castro, and Jose Manuel Zurita. Lexicon-based comments-oriented news sentiment analyzer system. *Expert Systems with Applications*, 39(10):9166–9180, 2012.

[12] Peter D Turney and Michael L Littman. Measuring praise and criticism: Inference of semantic orientation from association. *acm Transactions on Information Systems (tois)*, 21(4):315–346, 2003.

[13] Qing Cao, Wenjing Duan, and Qiwei Gan. Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems*, 50(2):511–521, 2011.

[14] Seongik Park and Yanggon Kim. Building thesaurus lexicon using dictionary-based approach for sentiment classification. In *2016 IEEE 14th international conference on software engineering research, management and applications (SERA)*, pages 39–44. IEEE, 2016.

[15] I Chetviorkin and N Loukachevitch. Extraction of russian sentiment lexicon for product meta-domain in proceedings of coling 2012. *Mumbai, India*, pages 593–610, 2012.

[16] Jyoti Prakash Singh, Seda Irani, Nripendra P Rana, Yogesh K Dwivedi, Sunil Saumya, and Pradeep Kumar Roy. Predicting the "helpfulness" of online consumer reviews. *Journal of Business Research*, 70:346–355, 2017.

[17] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, et al. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204. Valletta, 2010.

[18] Teng-Kai Fan and Chia-Hui Chang. Blogger-centric contextual advertising. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1803–1806, 2009.

[19] Nan Hu, Indranil Bose, Noi Sian Koh, and Ling Liu. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision support systems*, 52(3):674–684, 2012.

[20] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[21] Wei Zhao, Ziyu Guan, Long Chen, Xiaofei He, Deng Cai, Beidou Wang, and Quan Wang. Weakly-supervised deep embedding for product review sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 30(1):185–197, 2017.

[22] Mohammad Ehsan Basiri, Shahla Nemati, Moloud Abdar, Erik Cambria, and U Rajendra Acharya. Abcdm: An attention-based bidirectional cnn-rnn deep model for sentiment analysis. *Future Generation Computer Systems*, 115:279–294, 2021.

[23] Alper Kursat Uysal and Yi Lu Murphey. Sentiment classification: Feature selection based approaches versus deep learning. In *2017 IEEE International Conference on Computer and Information Technology (CIT)*, pages 23–30. IEEE, 2017.

[24] Zhigang Yuan, Sixing Wu, Fangzhao Wu, Junxin Liu, and Yongfeng Huang. Domain attention model for multi-domain sentiment classification. *Knowledge-Based Systems*, 155:1–10, 2018.

[25] LDCS Subhashini, Yuefeng Li, Jinglan Zhang, Ajantha S Atukorale, and Yutong Wu. Mining and classifying customer reviews: a survey. *Artificial Intelligence Review*, pages 1–47, 2021.

[26] Yahia Baashar, Hitham Alhussian, Ahmed Patel, Gamal Alkawsi, Ahmed Ibrahim Alzahrani, Osama Alfarraj, and Gasim Hayder. Customer relationship management systems (crms) in the healthcare environment: A systematic literature review. *Computer Standards & Interfaces*, 71:103442, 2020.

[27] Nikolas Ruffer, Johannes Knitza, and Martin Krusche. # covid4rheum: an analytical twitter study in the time of the covid-19 pandemic. *Rheumatology International*, 40(12):2031–2037, 2020.

[28] Franco Valencia, Alfonso Gómez-Espinosa, and Benjamín Valdés-Aguirre. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *entropy*, 21(6):589, 2019.

[29] Kudakwashe Zvarevashe and Oludayo O Olugbara. A framework for sentiment analysis with opinion mining of hotel reviews. In *2018 Conference on information communications technology and society (ICTAS)*, pages 1–4. IEEE, 2018.

[30] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679, 2021.

[31] Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, and De Rosal Ignatius Moses Setiadi. Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1029–1046, 2022.

[32] Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66, 2017.

[33] Yogesh Sankarasubramaniam, Krishnan Ramanathan, and Subhankar Ghosh. Text summarization using wikipedia. *Information Processing & Management*, 50(3):443–461, 2014.

[34] Aashka Sahni and Sushila Palwe. Topic modeling on online news extraction. In *Intelligent Computing and Information and Communication: Proceedings of 2nd International Conference, ICICC 2017*, pages 611–622. Springer, 2018.

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

[36] Elena Baralis, Luca Cagliero, Naeem Mahoto, and Alessandro Fiori. Graphsum: Discovering correlations among multiple terms for graph-based summarization. *Information Sciences*, 249:96–109, 2013.

[37] Madhurima Dutta, Ajit Kumar Das, Chirantana Mallick, Apurba Sarkar, and Asit K Das. A graph based approach on extractive summarization. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2*, pages 179–187. Springer, 2019.

[38] Lamees Mahmoud Al Qassem, Di Wang, Zaid Al Mahmoud, Hassan Barada, Ahmad Al-Rubaie, and Nawaf I Almoosa. Automatic arabic

summarization: a survey of methodologies and systems. *Procedia Computer Science*, 117:10–18, 2017.

[39] Rasim M Alguliyev, Ramiz M Aliguliyev, Nijat R Isazade, Asad Abdi, and Norisma Idris. Cosum: Text summarization based on clustering and optimization. *Expert Systems*, 36(1):e12340, 2019.

[40] Jianpeng Cheng and Mirella Lapata. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*, 2016.

[41] Yogesh Kumar Meena and Dinesh Gopalani. Evolutionary algorithms for extractive automatic text summarization. *Procedia Computer Science*, 48:244–249, 2015.

[42] N Vijay Kumar and M Janga Reddy. Factual instance tweet summarization and opinion analysis of sport competition. In *Soft Computing and Signal Processing: Proceedings of ICSCSP 2018, Volume 2*, pages 153–162. Springer, 2019.

[43] Atif Khan, Naomie Salim, and Haleem Farman. Clustered genetic semantic graph approach for multi-document abstractive summarization. In *2016 International Conference on Intelligent Systems Engineering (ICISE)*, pages 63–70. IEEE, 2016.

[44] Huong Thanh Le and Tien Manh Le. An approach to abstractive text summarization. In *2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR)*, pages 371–376. IEEE, 2013.

[45] Litton J Kurisinkel, Yue Zhang, and Vasudeva Varma. Abstractive multi-document summarization by partial tree extraction, recombination and linearization. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 812–821, 2017.

[46] Pierre-Etienne Genest and Guy Lapalme. Fully abstractive approach to guided summarization. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 354–358, 2012.

[47] Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 45–53, 2014.

[48] M Jishma Mohan, C Sunitha, Amal Ganesh, and A Jaya. A study on ontology based abstractive summarization. *Procedia Computer Science*, 87:32–37, 2016.

[49] NS Ranjitha and Jagadish S Kallimani. Abstractive multi-document summarization. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1690–1694. IEEE, 2017.

[50] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

[51] Wanwan Miao, Guiping Zhang, Yu Bai, and Dongfeng Cai. Improving accuracy of key information acquisition for social media text summarization. In *2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS)*, pages 408–415. IEEE, 2019.

[52] Renxian Zhang, Wenjie Li, Dehong Gao, and You Ouyang. Automatic twitter topic summarization with speech acts. *IEEE transactions on audio, speech, and language processing*, 21(3):649–658, 2012.

[53] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. 2010.

[54] Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond Ng, and Bita Nejat. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1602–1613, 2014.

[55] Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. Multi-document abstractive summarization using ilp based multi-sentence compression. *arXiv preprint arXiv:1609.07034*, 2016.

[56] Elena Lloret, Ester Boldrini, Tatiana Vodolazova, Patricio Martínez-Barco, Rafael Muñoz, and Manuel Palomar. A novel concept-level approach for ultra-concise opinion summarization. *Expert Systems with Applications*, 42(20):7148–7156, 2015.

[57] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

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

[59] Giuseppe Di Fabbrizio, Amanda Stent, and Robert Gaizauskas. A hybrid approach to multi-document summarization of opinions in reviews. In *Proceedings of the 8th international natural language generation conference (inlg)*, pages 54–63, 2014.

[60] Seyed-Ali Bahrainian and Andreas Dengel. Sentiment analysis and summarization of twitter data. In *2013 IEEE 16th International Conference on Computational Science and Engineering*, pages 227–234. IEEE, 2013.

[61] Jung-Yeon Yang, Jaeseok Myung, and Sang-goo Lee. The method for a summarization of product reviews using the user's opinion. In *2009 International Conference on Information, Process, and Knowledge Management*, pages 84–89. IEEE, 2009.

[62] Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Sujian Li, and Houfeng Wang. Entity-centric topic-oriented opinion summarization in twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 379–387, 2012.

[63] Yizhu Liu, Qi Jia, and Kenny Zhu. Opinion summarization by weak-supervision from mix-structured data. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3086–3096, 2022.

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

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

[66] Sri Raghava Muddu, Rupasai Rangaraju, Tejpalsingh Siledar, Swaroop Nath, Pushpak Bhattacharyya, Swaprava Nath, Suman Banerjee, Amey Patil, Muthusamy Chelliah, Sudhanshu Shekhar Singh, et al. Distilling

opinions at scale: Incremental opinion summarization using xl-opsumm. *arXiv preprint arXiv:2406.10886*, 2024.

[67] Mir Tafseer Nayeem and Davood Rafiei. Lfosum: Summarizing long-form opinions with large language models. *arXiv preprint arXiv:2410.13037*, 2024.

[68] Jian Wang, Yuqing Sun, Yanjie Liang, Xin Li, and Bin Gong. Iteratively calibrating prompts for unsupervised diverse opinion summarization. In *ECAI 2024*, pages 3939–3946. IOS Press, 2024.

[69] Pratik Deelip Korkankar, Alvyn Abranches, Pradnya Bhagat, and Jyoti Pawar. Aspect-based summaries from online product reviews: A comparative study using various llms. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 562–568, 2024.

[70] Tejpalsingh Siledar, Swaroop Nath, Sankara Sri Raghava Ravindra Muddu, Rupasai Rangaraju, Swaprava Nath, Pushpak Bhattacharyya, Suman Banerjee, Amey Patil, Sudhanshu Shekhar Singh, Muthusamy Chelliah, et al. One prompt to rule them all: Llms for opinion summary evaluation. *arXiv preprint arXiv:2402.11683*, 2024.

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[72] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019.

[73] Eric Chu and Peter Liu. Meansum: a neural model for unsupervised multi-document abstractive summarization. In *International conference on machine learning*, pages 1223–1232. PMLR, 2019.

[74] Arthur Bražinskas, Mirella Lapata, and Ivan Titov. Unsupervised opinion summarization as copycat-review generation. *arXiv preprint arXiv:1911.02247*, 2019.

[75] Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. Opiniondigest: A simple framework for opinion summarization. *arXiv preprint arXiv:2005.01901*, 2020.

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

[77] Arthur Bražinskas, Mirella Lapata, and Ivan Titov. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, 2020.

[78] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[79] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100, 2018.

Appendix A

Human Evaluation of Opinion-Aspects Pairs Extraction

This appendix provides 50 representative samples of each dataset used for the human evaluation of OA pair extraction. Each entry includes the original review, the OA pairs automatically extracted by the proposed system, and the corresponding human-annotated OA pairs. These examples support the error analysis and performance discussion presented in Section 4.7.

Table A.1: Human Evaluation of OA Extraction Model in Yelp dataset

ID	Review	
	OA by Model	**OA by Human**
1	A lil cozy bar/restaurant on Frenchman Street. Small plates, good music and service.	
	(Frenchman Street, good); (small plates, good); (service, good)	(bar, cozy); (restaurant, cozy); (plates, small); (music, good); (service, good)
2	Would it be legal in my state for a man to marry a business? Surely, if corporations are considered people, that would have to mean they can be married... isn't that what love is all about? Imagine the catering at the wedding.	
	(business, legal); (corporations, people); (catering, wedding)	Null
3	Awesome food, friendly staff, byob, cute cozy restaurant. I always get the duck curry (spicy), pad se eww, papaya salad and naam salad.	
	(restaurant, cozy); (food, awesome); (staff, friendly); (duck curry, spicy); (pad se eww, spicy); (papaya salad, delicious); (nam salad, delicious)	(food, awesome); (staff, friendly); (restaurant, cute); (restaurant, cozy)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
4	I love Jay's!! It's affordable, they deliver, and it's really good, creative sushi! Plus, the eggplant lady thing is amazing! Yes, it's true, they are not fluent in english but they know enough to take your order, and they are super friendly. I even like this place better than some of the big name sushi places! I highly recommend!!	
	(affordable, love); (deliver, good); (creative, love); (friendly, super); (know, enough); (name, better)	(sushi, good); (sushi, creative); (the eggplant lady thing, amazing); (english, not fluent); (they, friendly)
5	Dr Mack is awesome. Great staff and quality over quantity.. I came to him after another eye surgeon (around the corner) totally messed up my eye from eye lid lifts., i was in bad shape. I had these masses in my lower lids that needed removal., he got me in very fast and helped me calm the damage down., a few days later he removed the masses. I am way better and hopefully when my eyes calm down he will fix my left eyelid that the other dr **itco left uneven. Thanks Dr Mack	
	(masses, awesome); (lower lids, great); (other dr, bad); (left eyelid, uneven)	(dr Mack, awesome); (staff, great); (shape, bad); (way, better)
6	My husband and I stayed at Hotel Mazarin on our honeymoon. We really found it to be a wonderful place. My husband is a paraplegic and is in a manual wheelchair. The staff and facilities were sufficiently accommodating to those needs, as well as our generlw needs such as directions, restaurant suggestions, etc. The daily (FREE) breakfast is better than most- everything from fruit to grits- and the coffee is actually good. The location is great for wandering anywhere in the French Quarter, as there is little that is even a full mile away. Streetcars and bus tours are easily accessible from here as well. Even with being centrally located, it's very quiet. Highly recommend!	
	(Hotel Mazarin, wonderful); (staff, accommodating); (facilities, accommodating); (breakfast, better); (location, great); (streetcars, accessible); (bus tours, accessible); (quiet, quiet)	(place, wonderful); (facilities, accommodating); (free breakfast; better); (coffee, good); (location, great); (streetcars, accessible); (bus tours, accessible); (location, quiet)
7	Been coming here for a while, parking can be a pain but it's usually worth it. Weapon of choice is the Ethiopian slow pour and poached eggs, can't really go wrong there. Scones are another highlight. I give it 4 stars because of how expensive/crowded it can get, food quality has gone downhill as well since they have had to expand. Worth checking out for the coffee and atmosphere.	
	(parking, worth it); (food, good); (atmosphere, good); (pricing, expensive); (quality, downhill)	(parking, pain); (food quality; gone downhill)
8	The best Thai we have ever eaten. We are foodies and travelers and by far were blown away by the flavor and freshness of the ingredients. Great service and atmosphere too. The place is a must.	
	(food, blown away); (ingredients, fresh); (atmosphere, great); (place, must)	(ingredients, flavor); (ingredients, freshness); (service, great); (atmosphere, great); (place, must)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
9	This was so disappointing. I love Chick fil A. My son even works at the Avon location, which is fabulous. This is not the same. I tried to order my usual, that I get several times a week in Avon, and they received to accommodate. I only wanted the bread from the mini rolls. Go to Avon, very close and such a wonderful place to eat.	
	(mini rolls, fabulous); (Avon, wonderful); (bread, disappointing)	(location, fabulous); (Avon, close); (place, wonderful)
10	The new management are doing a decent job there. Greens are good and withstood the hurricane very well. Love the final four golf holes with the water of Tampa Bay as a factor and a great view. The clubhouse was nice as well. People that work here were nice. I would play here again!	
	(greens, decent); (clubhouse, nice); (water, great); (people, nice)	(job, decent); (greens, good); (view, great); (clubhouse, nice) (people, nice)
11	This place is insanely good. It's very small but that didn't stop it from being amazing. The waitress was great and the cook knows what he's doing. The tesla burger is very good. It's a much try.	
	(small, insanely good); (waitress, great); (tesla burger, very good); (burger, much try)	(place, insanely good); (place, small); (waitress, great); (tesla burger, very good);
12	I love this place! Had four different cocktails and an order of the reviled eggs over a couple of visits this week. The cocktails were uniformly excellent and the eggs were quite good as well. I also enjoyed interacting with the bartenders, who not only knew their craft, but were very good without being pretentious. Great spot!	
	(cocktails, excellent); (eggs, quite good); (bartenders, good)	(cocktails, excellent); (eggs, quite good); (bartenders, good)
13	Food is okay, for mass-produced, commissary-prepared ingredients that are cooked and/or assembled on site, as it is in most large, corporate establishments like this one. Nothing is particularly good, and nothing is particularly bad. I imagine the decision where to go vs Chipotle, Qdoba, Moe's, and Blue Coast largely comes down to price and/or convenience. For me, these four in the area are pretty much interchangeable otherwise, besides little things like one's chips are slightly better, etc. This location does have a Coca-Cola Freestyle machine, FWIW. Qdoba has recently started to offer all add-ons for no additional charge – even guacamole. It will be interesting to see if and how this new wrinkle may change the level of business – I've never seen this location in any danger of being even half full.	
	(food, okay); (location, interchangeable); (chips, slightly); (business, in danger)	(food, okay); (area, interchangeable); (chips, slightly better)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
14	Op suggestie van yelp- hier geweest op onze eerste avond in NOLA tijdens onze roadtrip (dag3). It's the best place. Wachttijden kunnen oplopen, maar waar Nederland vaak faliekant in faalt, gaat hier geweldig: er wordt door t personeel vlijmscherp in de gaten gehouden wie er aan de beurt is, zelfs voor de krukken aan de bar. Bijna gingen er mensen die later waren gearriveerd dan wij er met onze lang verwachte krukken vandoor, maar de barman regelde dit op sublieme wijze. Het moraal van dit verhaal- je raakt hier niet geïrriteerd omdat andere mensen voor schnitzelen. Ondertussen word je getrakteerd op de lekkerste live muziek en hebben ze er heerlijke cocktails en prima wijnen. Eten was ondergeschikt deze avond, maar wat we hadden was bijzonder goed. Dikke aanrader- wellicht reserveren als je niet van eindeloos wachten houdt.	
	(bar, vlijmscherp); (personnel, geweldig); (krukken, lang verwachte); (mensen, schnitzelen); (muziek, lekker); (cocktails, prima); (wijnen, bijzonder); (eten, ondergeschikt)	no English
15	I come here every time when I am in university city. Food is good specially the chicken tika. Some of there other dishes or a bit on the greasy side. But overall is a good restaurant to take your family to for a nice dinner.	
	(chicken tika, good); (other dishes, a bit on the greasy side)	(food, good); (other dishes, greasy side); (restaurant, good); (dinner, nice)
16	The atmosphere is really cool and I love their large deck where you can sit with friends and have beers. However, the employees are typically rude and rush you through the process and the food is bland and boring. I have tried this location several times since it is close to my office, and it is consistently "ok." Definitely not somewhere I would want to take out-of-town guests or business associates.	
	(atmosphere, cool); (employees, rude); (food, bland); (location, ok)	(atmosphere, cool); (deck, large); (employees, rude); (food, bland); (food, boring)
17	They have very fresh raw oysters , which fits perfectly with beer. Fried food good as well.	
	(raw oysters, fresh); (fried food, good)	(raw oysters, fresh); (fried food, good)
18	This is my family's go to when we eat out for Chinese. The place is not big but it's decently modern. The quantity and the quality of their food are always good. Some items on the menu might be a little pricy, but most of the items off their menu are worth the money. One only improvement I can see is maybe they should hire a couple more staff during rush hours at dinner, food can come out a little late sometimes and the staff looks overwhelmed all the time, we've had the server forgetting a dish we ordered before because they were so busy. Nevertheless, this is a great spot and by far the best Chinese food restaurant in the area.	
	(food, good); (staff, overwhelmed); (menu, pricy); (dinner, late); (area, best)	(place, not big); (place, modern); (food, good); (some items, little pricy); (money, worth); (food, late); (staff, overwhelmed); (spot, great)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
19	Fantastic appetizers, average service, average main course. As a couple we had calamari and steak carpaccio for appetizer. Some of the best cooked calamari I've ever had. Had langoustine with pasta and gnocchi. Pasta didn't seem hand made. Gnocchi was but was heavy. Large portions. Decor ok. Looked like the back office was in the restaurant which was a little off-putting. Will likely go back just, if just for the calamari!	
	(calamari, fantastic); (langoustine, heavy); (pasta, average); (gnocchi, heavy); (decor, ok)	(appetizers, fantastic); (service, average); (main course, average); (calamari, best cooked); (gnocchi, heavy); (portions, large); (decor, ok); (office, off-putting)
20	Food was quite expensive for portion sizes. Was charged 18.00 to receive two chicken legs in sauce. My husbands food was wrong took over 25 minutes to correct it. Service was so so, it was super loud, we were packed in and it was freezing every time the door opened. The owner did come and talk to us after we paid our bill knowing we were very displeased. He was nice but instead of making it right and reducing the bill he offered if we came back again he'd take care of us. No sir. That's not how it's done. The food didn't even taste good enough to return not alone all the other issues. Too bad because I love this type of food.	
	(chicken legs, expensive); (service, loud); (food, not good); (owner, nice); (bill, wrong)	(food, expensive); (service, so so); (owner, nice)
21	If your a light to medium sleeper don't come here.. A nice place to stay, rooms were comfortable. Several things are right, however, one major drawback is the noise. The hotel is set up with exterior rooms and interior (atrium) rooms. We stayed in an interior room, every noise that was made in the atrium echoed and came in our room. Little noise dampening was on the wall, door and window that separates the room from the atrium. I was woke up at 6 am by people walking by our room having a conversation, additionally this was added to by children screaming in the atrium, I don't know if they were in the pool, it's in the atrium...	
	(room, comfortable); (noise, woke); (wall, little); (door, little); (window, little); (atrium, echoed); (conversation, 6 am); (children, screaming)	(place, nice); (rooms, comfortable); (several things, right); (drawback, noise); (atrium, echoed)
22	Megabus is mega-bullshit. They've got us packed like sardines into this tin can on wheels, and there's nowhere to store my stuff. I had my bags on a luggage rack and they made me take them off so that I had to shove them under my feet. No overhead compartments. And the wifi blocks streaming sites so I can't even listen to NPR. No room, cramped conditions. Miserable	
	(packed, cramped); (stuff, nowhere); (luggage rack, had to shove); (wifi, blocks); (room, miserable)	(megabus, mega-bullshit); (conditions, cramped)
23	Best coffee place on state. Honey hazelnut latte is amazing!!!	
	(Honey hazelnut latte, amazing); (coffee place, best)	(Honey hazelnut latte, amazing); (coffee place, best)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
24	A unique gem. The outside design just attracts me. It is on the corner of this huge building. On the inside, it has two floors with lots of seating and a bar in the middle of the first floor. They have a patio, too. The only thing I do not like about it is the motion sensor light in the restroom. I am glad I tried their food. I ordered the blackened salmon sliders appetizer. I enjoyed it because I asked to hold all the toppings except the spinach and to add pickles. The salmon looks like about 2 oz. each for the 3 sliders. I choose my side to be fries over tator tots. Delicious, tasty, satisfying, healthy, just to name a few. At 10 pm., it becomes an adult only late night restaurant. The clientele is small but seems upscale and from Carmel. They have lots of specials; join The Mob.	
	(corner, attracts); (restroom, does not like); (salmon, enjoy); (toppings, ask to hold); (side, delicious); (late night, becomes); (specials, join); (patio, has)	(gem, unique); (building, huge); (food, delicious); (food, tasty); (food, satisfying); (food, healthy); (lientele, small); (clientele, upscale)
25	I used to dine at Albasha weekly in Baton Rouge. When I found out they were also in Metairie, I made my fiancée abandon our old middle eastern favorites. Albasha is all we eat now! It's worth the drive. Typically do take out. Love the big mezze platter of everything . Or the combo chicken schwarmA and gyros is my favorite ! We love Albasha!	
	(Platter, worth); (Mezze, big); (Chicken, favorite); (Albasha, love)	(drive, worth); (combo chicken, favorite); (mezze platter, big); (Albasha, love)
26	Came to see my son's band Last Frontier. The staff and food here is top notch! Had the porkbelly and fish nachos. Awesome. Great selection of whisky. Great sound system for the band.	
	(porkbelly, awesome); (food, top notch); (whisky, great); (staff, top notch); (band, great)	(staff, top notch); (food, top notch); (selection, great); (sound system, great)
27	Easily the WORST Mexican food I have ever eaten. I asked for cilantro on my taco and they said the didn't have any but I could get some salsa at the salsa bar that had some in it. (What kind of Mexican restaurant doesn't have cilantro!?!?) Both salsa options they had out had absolutely no flavor whatsoever. The beef in my taco tasted as if it hadn't been seasoned with anything at all. Just cooked. I will never go here again. It was just awful.	
	(taco, awful); (salsa, flavorless); (beef, seasoned); (restaurant, worst)	(mexican food, the worst); (salsa options, no flavor)
28	Great little dive restaurant with fantabulous seafood. Nothing fancy on the decor. Waitresses are very attentive. Food is hot, fresh, and some of the best oysters in the NOLA area. They also have a fish market next door that has fresh fish and other Cajun specialty foods. The restaurant is a great "hole in the wall" type restaurant with no frills. Oh and they have a full bar with drink specials! This place is a great place if you want to go low key for an amazing meal!	
	(restaurant, great); (food, fantabulous); (waitresses, attentive); (oysters, best); (fish, fresh); (Cajun, specialty); (restaurant, hole in the wall); (bar, full)	(restaurant, great); (seafood, fantabulous); (waitresses, attentive); (oysters, best); (food, hot); (food, fresh); (fish, fresh); (bar, full); (place, great); (meal, amazing)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
29	Lily and her team really surpass all other nail salons with their warm welcome, their beautiful remodeled and updated salon with the water cooler right on the floor for their customers, friendly and skilled staff and have an endless supply of beautiful nail services and colors. You can get everything from chrome to the dip to acrylic to gel to a normal manicure.	
	(salon, warm); (staff, friendly); (services, beautiful); (colors, endless); (manicure, normal)	(welcome, warm); (salon, beautiful remodeled); (salon, updated); (staff, friendly); (staff, skilled);(services, endless); (color, endless)
30	Great service, ambiance, and food! This is my favorite restaurant to visit when I'm in NOLA. The truffle butter served on top of the filet is such a delight!	
	(ambiance, great); (filet, delight); (truffle, delight)	(service, great); (ambiance, great); (food, great); (restaurant, favorite); (truffle butter, delight)
31	Town And Country veterinary hospital is one of the best vets on the face the planet, in my opinion. Both doctors are professional, highly qualified, friendly, approachable, and never rush you out the door to get to the next patient. They have endless patience with questions, and get down on the floor with your pet before starting any sort of examination which really helps relax nervous pups.	
	(Town And Country, best); (doctors, professional); (doctors, highly qualified); (doctors, friendly); (doctors, approachable); (doctors, never rush); (pets, relax); (examination, get down)	(vets, best); (doctors, professional); (doctors, highly qualified); (doctors, friendly); (doctors, approachable); (doctors, never rush); (patience, endless)
32	Great selection, good service, and can't complain that I have gone before and gotten free pint glasses for buying a case of beer.	
	(selection, good); (service, great); (pint glasses, free); (buying, good)	(selection, great); (good, service); (pint glasses, free)
33	This place is beautiful. Most Kimpton properties I've stayed at have been. They always have unique interior design and at this hotel, the design was noticeable. From the carpets to the chairs and colors in the room. Really well done. I guess that's one of my things whenever I stay somewhere. It's always the little stuff that you notice that makes you say "that's nice"... and this hotel had a lot of those little things. From the exterior.. entrance..the lobby... elevators.. and of course, when you walk in the room. (And of course the bed) If you're not disappointed in any way at any of those points.. you've got yourself a great hotel. It's also larger than other Kimpton properties that I've stayed at, (in Washington D.C.) which was nice. It's right in the heart of the city, and a great place to stay. Highly recommend it.	
	(carpets, well done); (chairs, nice); (room, great); (entrance, nice); (elevators, nice); (bed, great); (lobby, nice); (hotel, great)	(place, beautiful); (interior design, unique); (design, noticeable); (hotel, great); (place, great)
34	Really good ice cream!! We went through the drive-thru, think that was a mistake. We waited about 35 minutes with 6 cars ahead of us. Looks like people walking up were getting served quicker. But it was excellent, everyone enjoyed! It was worth the wait. The girl who took our order was very pleasant. Definitely will be back!	
	(ice cream, good); (wait, worth); (girl, pleasant); (order, pleasant)	(ice cream, good); (wait, worth); (girl, pleasant)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
35	So I have to start this with saying that I never write reviews. I never just take the time, but I always read reviews. I've been going to the rock gym for years with my daughter and she loves it but yesterday they were truly phenomenal! I had a group of 10 kids coming in last minute (as in two hours before they were going to arrive) and the gym not only accommodated them completely but gave them an awesome experience! I have no complaints and kids seem to all have a blast doing the high climbs, the bouldering and the slack line. But really it was the customer service that was above and beyond.	
	(rock gym, phenomenal); (kids, had a blast); (customer service, above and beyond)	(gym, accommodated); (experience, awesome); (customer service, above and beyond)
36	Truly amazing meal. Favorite part by far was the hamachi curry. Cocktails were excellent. Also ordered the smoked tuna tartare, conch croquettes, broiled shrimp and goat curry- the most tender meat I've had in a long time! Our server Claire was very knowledgeable and fun. Highly recommend for a special night out!	
	(hamachi curry, excellent); (smoked tuna tartare, tender); (conch croquettes, tender); (broiled shrimp, tender); (goat curry, tender)	(meal, amazing); (cocktails, excellent); (meat, tender); (server, knowledgeable); (server, fun)
37	My new favorite brunch spot! Great place to eat, friendly staff and the food is delicious. They amount of items on the menu I would like to try is really incredible.	
	(food, delicious); (staff, friendly); (menu, incredible)	(brunch spot, new); (brunch spot, favorite); (place, great); (staff, friendly); (food, delicious)
38	I have been going to Amys Flowers for years because my Mother and Sister live there and have used them on many ,many occasions through the years for birthdays, anniversaries , Valentines day, Hospital situations and they have always delivered the most beautiful arrangements, But I just spent 250$ on a what I wanted to be an extra Gorgeous arrangement for my Mother and from the photos I was a bit disappointed. only 4 roses with some exotic flowers, but not the Grandiose I was hoping for, would of spent more to get the effect I wanted but thought $250 would of done it but just fell short! Bummer!	
	(arrangements, beautiful); (roses, disappointed); (photos, disappointed); (grandiose, hoped); (effect, hoped)	(arrangements, beautiful); (extra gorgeous arrangement, disappointed); (flowers, exotic)
39	Finally, a place in New Hope with unique, high quality food. I expected everything to be spicy as is my experience with Cuban food previously. Rich flavor characterized our meals best. I can't wait to go back and try something else.	
	(food, unique); (food, high quality); (meals, rich flavor)	(food, unique); (food, high quality); (flavor; rich)
40	We stopped in last night for dinner and really enjoyed this place. The food was great as was the service. We did find their prices were a little high given the quantity of food you get, but we are looking forward to trying their lunch buffet which is a much more attractively priced option.	
	(food, great); (service, good); (prices, high); (lunch buffet, attractively priced)	(food, great); (service, great); (prices, little high); (lunch buffet, attractively priced)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
41	I came here in a group and they were very nice about accommodating all of us. I loved their basil ginger lemonade, which had just the right amount of sugar. We shared a bunch of appetizers and two pizzas and everything was rather tasty. I enjoyed the meatballs too! Nothing was particularly memorable but it was a pleasant time and the patio looks like a lovely place to chill on a spring/summer night.	
	(basil ginger lemonade, loved); (appetizers, rather tasty); (meatballs, enjoyed); (patio, lovely)	(amount of sugar, right); (bunch of appetizers, tasty); (pizzas, tasty); (meatballs, enjoyed); (time, pleasant); (patio, lovely)
42	Good food, fast to go orders. The red beans and rice was great and ton of food! And the crawfish pasta was delicious.	
	(red beans, great); (crawfish pasta, delicious)	(food, good); (orders, fast); (red beans, great); (rice, great); (crawfish pasta, delicious)
43	Haha what!? Can't I Yelp this. Yes 4 star I'm a fan. I was walking along Bourbon St and there were these girls up in the front inviting me in. At the time I didn't fully understand the check-in feature on the Yelp app so I checked in for the humor of letting my friends know where I was and how my road trip was doing but of course it was a secret check-in and no one saw it. More of a 3 star place ain't nothing wrong with that. It's close to all the bars just in case you Yelpers that's doing the YSB need any UYE ideas.	
	(bar, 3); (girls, 3); (app, 3); (road trip, 3); (friend, 3); (UYE, 3); (YSB, 3)	(check-in, secret)
44	I love this place! At first, I was a little worried because I'm not a fan of spicy food and I was told that Sichuan cuisine differs a little from the Chinese cuisine most people might be accustomed to but I loved it. We had a waiter who was Chinese who gave us a lot of good information and was able to alter some dishes for us so for those who do not like their dishes too spicy or can't handle a lot of heat, do not worry, because it seems as though they can alter the spiciness to most of their dishes [...] We were also recommended the Double Cooked Pork Belly, which is something I don't usually like because it's kind of fatty, but it turned out to be great as well. They also have a cool little happy hour with some cheaper drinks so it's cool to get there a little before dinner, like we did and enjoy some cheaper drinks before dinner. I wish they had some light beer available since I'm a girl and I'd like to have some lower calorie/lighter options for beer, but it's okay since they have a good wine selection too so I just chose to go with wine. The food is relatively affordable, it's not Chinatown cheap, but it's not expensive either. Overall, it was a very delicious meal in a fun, comfortable environment. Will definitely go back.	
	(restaurant, delicious); (waiter, good); (food, affordable); (environment, fun); (Dan Dan Noodles, less spicy); (Double Cooked Pork Belly, great); (happy hour, cheaper); (wine, good); (light beer, not available)	(food, spicy); (Double Cooked Pork Belly, fatty); (Double Cooked Pork Belly, great); (happy hour, cool); (drinks, cheaper); (wine, good); (food, affordable); (meal, delicious); (environment, fun); (environment, comfortable)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
45	Used MegaBus for the first time on May2 & 4, 2014. The only thing it has going for it is the price. The driver got lost trying to find the CalTrain station in Sacramento and drove for over one hour on the surface streets of Sacramento, wandering and looking for the station in residential areas. Passengers tried to help her with directions from smart phones, but she wans't interested in any assistance. By the time we pulled into the station over an hour late, we could smell that something was overheating. First the driver claimed it was the system to open the luggage doors, as these were not working and could not be opened to offload and reload passengers luggage. She finally got that to work, and told us we would have to wait for a hot tire to cool off. [...] And the replacement was not a megabus, no wi-fi, etc. The portapotty called a bathroom was filthy, burn your shoes if you have to use it. The trip back was slightly better, also very late and had a stop to explore a smell coming from the bus. After my experience, I will pay to fly Southwest into Oakland and take BART into the city. I have no confidence that these bus drivers could handle a real emergency while driving and did not feel safe.	
	(bus, unsafe); (driver, unhelpful); (tire, overheating); (luggage, difficult); (off-ramp, unsafe); (shopping center, unsuitable); (bathroom, filthy)	(driver, lost); (tire, overheating); (driver, unglued); (portapotty, filthy); (bathroom, filthy)
46	Was in Nashville this past weekend with some friends. Decided to try this place as it was close to where we were staying. The place is very quaint on the inside. Dark lighting, the type of spot that could suffice a first date, a work event or a family get together. My friend and I showed up around 3pm and strangely there was a ten minute wait (Which is weird, because it's that awkward time in between lunch and dinner, but whatever.) Our server is super friendly and attentive. I order the pan seared tuna with a side of the sweet potato fries with jalapeno ketchup. Our food comes. I take a bite of my tuna and it appears raw, almost entirely uncooked. I try again and can't bare it. The waiter has no problem swapping my order for the salmon which is very kind of him. My friend appears to enjoy the burger. The fries are fine– not seven dollar side worthy fine, but decent enough. The jalapeno ketchup tastes like regular ketchup. The salmon doesn't particularly taste like fish (almost like an overly marinated chicken, which wouldn't be a bad thing if it wasn't brought under the guise of fish) but I run through it because I'm starved. [...] Nice, attentive staff and sweet vibe. Food was awful.	
	(food, awful); (server, friendly); (place, quaint); (lighting, dark); (ketchup, tastes like regular ketchup); (salmon, overly marinated); (meal, below average); (cost, expensive)	(place, quaint); (lighting, dark); (server, friendly); (server, attentive); (tuna, raw); (tuna, uncooked); (fries, fine); (chicken, overly marinated); (vibe, sweet); (food, awful)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
47	I don't always order pizza, but when I do, I order it completely bombed at 3 in the morning. [...] There have been restaurants in this space coming and going for years and I believe (and hope) that Nannie Franco's will fill the necessary void for another amazing neighborhood pizza place. After spending a couple hours doing the normal Friday night South Philly arm curls I reached home with an outstanding urge for some food ill probably regret in the morning. With limited options I dacked my brain. [...] The lady behind the phone also informed me that we get a free Stromboli with our order, score! [...] The driver was friendly enough and the pizza came hot. The crust was crispy and the cheese stuck just long enough for my mouth flap to engulf the entirety of a slice. My friends barley came up for air between the free Stromboli and the pizza. I throughly enjoyed the red sauce and how crispy the chunks of sausage were. The pizza and Stromboli contained enough pig to satisfy my craving for Babe, if only until breakfast. I tried the left overs in the morning and they were just as good as the night before. [..]	
	(pizza, bombed); (Stromboli, crispy); (sausage, crispy); (chunks, crispy); (cheese, stuck); (left overs, good); (cheese, stayed)	(neighborhood pizza place, amazing); (urge, outstanding); (options, limited); (stromboli, free); (driver, friendly); (pizza, hot); (crust, crispy); (red sauce, enjoyed); (sausage, crispy); (left overs, good)
48	My friend and I decided to have lunch at Mr. B's on Memorial Day. The atmosphere is moderately upscale. Fortunately for us, there was no long line or crowd, so we got seated immediately. For the most part the servers were nice, however our waitress could have improved her customer service etiquette. [...] Second, the waitress was unfamiliar with the menu. She did not seem to know whether or not the shrimp in the various dishes was deveined. My third issue was with waiting 40 minutes for the crab cake appetizer. One crab cake is $18. The crab cake was actually good and savory, but it was disappointing waiting 40 minutes for one crab cake ... Our main course arrived about 15 minute after the crab cake. I ordered the shrimp and angel hair pasta. The buttery taste overwhelmed the dish. Since the dish lacked acidity, I requested lemons from the waitress. The taste of the pasta was not a big deal ... Nothing a little lemon juice couldn't fix. My fourth issue was with the waitress bringing the small cut lemons from the bar, instead of lemon wedges ... The lemon cuts were so small that I was unable to squeeze lemon juice from the lemons ... Overall, I wanted to enjoy Mr. B's more than I did... I have no plans of returning.	
	(table, moderately upscale); (waitress, nice); (bread, crispy); (crumbs, considerable); (menu, unfamiliar); (appetizer, disappointing); (crab cake, good); (pasta, buttery); (lemons, small)	(atmosphere, upscale); (servers, nice); (tastes, good); (waitress, unfamiliar); (crab cake, good); (crab cake, savory); (waiting, disappointing); (buttery taste, overwhelmed); (lemon cuts, small)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
49	It was OK but I won't be back. They treat their white customers waaayyy better than their Black customers, and keep in mind the owner and all the staff are Black. I walked in behind two white guys and the lady serving the food explained to them with great gusto the food options of the day. Another lady passed by and joked with them. It's my turn and I'm excitedly expecting the same treatment. The lady serving the food looks at me with a bored look on her face and mumbles "what do you want." [...] In general it was a very awkward experience because I felt like even the white guys were looking and noticing the very different treatment I was getting. As a Black woman, it was not a welcoming experience at all. It was so unwelcoming that I won't be back. Quite embarrassing. With regards to the food, I was not impressed at all. It was all very OK. The flavors were OK. The ice cream was too sweet and not creamy (I got the praline ice cream). I prefer for my food to have more texture. Everything was sort of mushy–bread pudding, grits, mac & cheese, and okra gumbo. Didn't feel like I was getting too many vitamins. Plus, it was hard to believe it was all vegan. Would help if they listed the ingredients in their dishes.	
	(food, OK); (staff, friendly); (service, poor); (treatment, unfair); (ingredients, unlisted); (texture, mushy); (vitamins, lacking)	(look, bored); (experience, awkward); (treatment, different); (experience, unwelcoming); (food, ok); (flavors, ok); (ice cream, too sweet); (ice cream, not creamy)
50	Went here for dinner w/ the fam earlier this month. tuesday night dinner at 8:30p - found awesome street parking (meter, yuck) - but it was right outside the restaurant. got seated right away, awesome. our service was great. now the food review: appetizer: mac & cheese - must do it! there were 3 of us, and it was the perfect amount to share. not too much where u fill up. it was really delicious, bread crumb topping too. perfect crunch. mom had the fried chicken & was very happy with it. sister had the pot roast meal - also really delicious. i had to get the grilled cheese & tomato soup almost out of default. i always wanna get that to compare w/ other restaurants to see who has the best. definitely didnt care for their grilled cheese. the tomato soup was pretty good though. gave my leftovers to my little brother. made me realize i should not go to comfort foods restaurants anymore and get back to eating healthy!	
	(parking, awesome); (food, delicious); (bread crumbs, perfect crunch); (grilled cheese, not good); (tomato soup, pretty good)	(parking, awesome); (seated, right away); (service, great); (mac&cheese, delicious); (amount, perfect); (crunch, perfect); (tomato soup, good); (foods, comfort)

Table A.2: Human Evaluation of OA Extraction Model in Amazon dataset

ID	Review	
	OA by Model	**OA by Human**
1	I tested this by magnetizing a screwdriver to the point where it would pick up a paper clip. Then I demagnetized it using this device. I found I had to hold it very close, actually touching the screwdriver for it to be most effective. When it is working, it will buzz & vibrate slightly. I then used the device to demagnetize the sprocket on a Copenhagen Wheel, and it worked well.	
	(screwdriver, effective); (device, effective); (sprocket, well)	(touching the screwdriver, most effective)
2	This was the prettiest pair of shoes for a little girl. I loved that it was not a tie shoe. Simple velcro. It fit perfect for my 5 year old. Im very skeptical about online purchasing, but I am happy with this purchase.	
	(shoes, pretty); (fit, happy); (online, skeptical)	(pair of shoes, prettiest); (girl, little); (velcro, simple); (online purchasing, skeptical); (purchase, happy)
3	This was a surprise for my wife. She loves it. It is as advertised and of a very good quality. Just the right size for a casual evening out on the town.	
	(size, good); (quality, good)	(quality, very good); (size, right)
4	I love the color, exactly the same green as online. Soft material, and very flattering. I did buy the large since I'm between medium and large. It fits perfectly. Love it!	
	(color, love); (material, soft); (size, fits); (large, perfectly)	(material, soft); (fits, perfectly)
5	Has lots of pockets to keep everything organized. Looks as pictured. DOES NOT look nice for very long. Wear on the corners happened pretty fast and there are threads coming off of the straps where the plastic edges have cracked and worn. Not satisfied with durability as a daily use bag for work.	
	(pockets, organized); (looks, does not look nice); (corners, worn); (straps, cracked); (durability, not satisfied)	(everthing, organized); (corners, wear); (threads, coming off); (plastic edges, cracked); (plastic edges, worn); (durability, not satisfied)
6	I've worn it for over a week with no issues, and I have very sensitive ears. Would definitely purchase again! Love that it comes in multiple colors and sizes. Clicker close is sturdy and easy to use. Great hoop and a good price!	
	(hoop, good); (clicker, sturdy); (colors, love); (sizes, love)	(ears, sensitive); (colors, multiple); (sizes, multiple); (clicker close, sturdy); (clicker close, easy to use); (hoop, great); (price, good)
7	I bought this Alex and Ani on sale for my Mom for mother's day. She loved it! I especially like the unique style of this bracelet compared to the standard charm style Alex and Ani bracelets.	
	(bracelet, loved); (style, unique)	(style, unique)
8	I absolutely adore this tee!! It's sooo soft and comfortable, I'm immediately buying more. The color is vibrant and of course I love cats too. The box it came in is also a work of art. Completely satisfied!! Must buy.	
	(tee, adore); (box, work of art); (color, vibrant); (cats, love)	(tee, adore); (tee, soft); (tee, comfortable); (color, vibrant)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
9	This is pretty much the highest quality pure mink oil you can buy. It doesn't contain the fillers and silicone that the cheaper brands use. It also does a great job at making leather more water resistant and conditioned.¡br /¿¡br /¿Just be aware that it will darken your leather about 2 shades.	
	(leather, high quality); (fillers, cheaper); (water resistant, conditioned); (darken, 2 shades)	(quality, highest); (leather, more water resistant); (leather, darken)
10	This watch is absolutely gorgeous!!! Everything about it is breathtaking, the black wrist band is a gorgeous steel, it doesnt stain my wrist by the end of the day and the picture of tinkerbell and stones are magnificent! have used mine daily for months and not even one has fallen off!!¡br /¿i would recommend this to every Tinkerbell lover!!	
	(watch, gorgeous); (black wrist band, breathtaking); (stones, magnificent); (Tinkerbell, lovely)	(watch, gorgeous); (black wrist band, gorgeous steel); (picture of tinkerbell, magnificent); (stones, magnificent)
11	I purchased the black multicolor. I wanted to love these shoes. I have two other pairs of Go Walks that are very comfortable. The color and style is great. However, the sole seemed to be warped or defective in one shoe. It was as if the sole had an uneven thickness in one area. Returned the shoes.	
	(shoe, loved); (sole, uneven); (color, great); (style, great)	(two other pairs of Go Walks, comfortable); (color, great); (style, great); (sole, warped); (sole, defective); (sole, uneven thickness)
12	I bought this to for vacation but it was just a little too tight and rode up when I lifted my arms. I didn't have time to reorder in a bigger size before my trip, but it was cute enough that I would have tried 1 size up if I had time for it to get here before I left.	
	(bathing suit, too tight); (arms, rode up); (size, bigger); (time, tried)	no OA pairs
13	I order this watch thinking it was a little darker gold, however it is true to the picture, it is a light gold. I would recommend it if you like light gold. All in all I am okay with my purchase.	
	(watch, light gold); (purchase, okay)	(purchase, okay)
14	The shoes are so cute and I want to love them, however, they were too tight to wear to my event. I will try to stretch the out. I you have really narrow feet these will work for you.	
	(shoes, cute); (tight, too tight); (feet, narrow)	(shoes, cute); (shoes, too tight); (feet, narrow)
15	I really adore seeing my 15 month old toddle around with this adorable backpack on. However, I don't think the quality is incredible, particularly with the zipper on the front pocket. The little side pouch pockets are not very functional either. Cute, but not durable.	
	(backpack, adorable); (zipper, not incredible); (side pouch, not very functional); (cute, not durable)	(backpack, adorable); (quality, not incredible); (little side pouch pockets, not very functional)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
16	I got these for working in my garage. I was expecting them to take a couple weeks to break in, but they were already very comfortable after about 2 days. The leather is high quality and the waxed laces are very strong. I sized down a half size when ordering them, some reviews say to go down a whole size, but there is no way they would fit if I had gone down that much.	
	(shoes, comfortable); (leather, high quality); (laces, strong)	(leather, high quality); (waxed laces, strong)
17	Boots are great except the loop on the back about 6 inches above the heel: I imagine only the Chinese have small enough little fingers to pull the boot on with loop, I cannot even get my smallest finger in the loop, but of course I am American made! thanks Lare	
	(Boots, great); (loop, small); (Chinese, small); (American, made); (finger, small)	(boots, great); (loop, small)
18	I am a fit older male who has always had a problem keeping my pants up. This was due to the belt holes not being perfectly placed and/or belt being to stretchable and, of course, my lack of a belly. This belt is perfect. You can adjust it perfectly and with little effort re-adjust it as needed. The price is awesome as well!	
	(belt, perfect); (adjust, awesome); (price, awesome)	(male, fit older); (belt, perfect); (price, awesome)
19	The watch is nice for the price, however it was described as a child's watch and it isn't.¡br /¿I had purchased it for my 10 year old grandson and it is way too big.. I probably will be returning it..	
	(watch, nice); (watch, described); (watch, big); (grandson, too); (watch, probably)	(price, nice); (watch, too big)
20	Very cute Jeans I naught for my daughter. Not super thin material. They fit a little big but I don't mind failure she can wear for a while before out growing.	
	(Jeans, cute); (material, thin); (fit, big); (fit, a little big)	(jeans, cute); (material, not super thin)
21	Stitching was a little rough and chafed around the back of the band near the hooks. I just had to stitch some softer fabric around thethose parts. In all other respects it was comparable to other bras of similar design.	
	(back, rough); (fabric, softer); (other, comparable)	(stitching, rough); (back of the band, chafed); (bras, comparable)
22	The shoes fit to tight but the cushion inside feels nice. After a few hours of wearing my feet were hurting because they were to tight even though I ordered the shoe in the same size as my other shoes.	
	(shoes, tight); (cushion, nice); (feet, hurting); (size, same)	(shoes, tight); (cushion, nice); (feet, hurting)
23	These plastic inserts fit perfectly in my wallet. Thin enough not to add bulk to my wallet when filled with picture and cards but still strong enough to hold all said stuff.	
	(wallet, thin); (insert, fit)	(plastic, fit)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
24	I bought two sizes up expecting that it will shrink. The problem is that this is a fitted cut so when it shrinks it's no so much the length that becomes an issue (if you buy at least a size up), but the part where the zipper goes from the leg to the body - it becomes very tight in that spot. Otherwise a great, soft, and breathable pijama for the baby	
	(pijama, great); (zipper, tight)	(zipper, tight); (pijama, great); (pijama, soft); (pijama, breathable)
25	this is the first time I purchased a bathing suit online and luckily it worked out very well. the fit is perfect, love the color and is very comfortable. it's just thick enough to hold you in and be comfortable. I am happy with my purchase.	
	(bathing suit, comfortable); (fit, perfect); (color, love)	(fit, perfect); (bathing suit, comfortable); (purchase, happy)
26	These were the weirdest fitting socks I've ever bought! Even though I have a size 15 foot, the knit heel portion went well past my natural heel. More like they were made for a size 20 foot. Bought these not knowing that around $2.00 per pair is a pretty common price for any brick and mortar store. Don't seem to any better quality either: medium weight with wimpy elastic. Really regret buying 10 pair...	
	(Socks, weird); (Knit heel, past); (Size, 20); (Quality, wimpy); (Regret, buying)	(socks, weirdest); (socks, fitting); (price, common); (quality; not better); (weight, medium); (elastic, wimpy)
27	In this case, I feel like you get what you pay for.¡br /¿¡br /¿It's not an awful piece, but it also isn't something you would probably find in a nice jewelry store. It's nice enough and quite honestly a little bit smaller than I had expected. But my mom liked it enough and we haven't had any issues with it yet so I guess that's something.	
	(necklace, nice); (small, nice); (mom, liked)	(piece, not awful); (jewelry store, nice)
28	A little small but the packaging and everything else was SO CUTE. got it for a gift for a friend for her birthday who is obsessed with butterflies and she's loved it!! Not sure if I'd buy it for anyone else but still so cute!!	
	(packaging, cute); (butterflies, loved); (friend, obsessed); (gift, loved)	(packaging, cute); (butterflies, obsessed)
29	Love love love these sneakers. I don't believe I could be happier with them. The colors are perfect and they fit just as well. I have different pairs of DCs but these are probably my favorite. Nice and slender, jazz up any outfit	
	(shoes, happy); (colors, perfect); (fit, slender); (outfit, jazz up)	(colors, perfect); (sneakers, nice); (sneakers, slender)
30	I love these glasses. I've just had cataract surgery and need to control the temporary light blindness. No matter the day, cloudy or sunny, one of these pairs of glasses is perfect for the job.	
	(glasses, love); (cloudy, perfect); (sunny, perfect)	(glasses, perfect)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
31	I would differently buy from seller again, I never take it off especially using my smartphone as much as I do, I wear it on the left side were I had wrist surgery three months ago, suffering really bad in pain, horrible scars decided to wear as cover up, so I thought my swelling has gone down, and the pain has honestly gone, when I take it off for period of time, aches return, not sure if the bracelet was suppose to do this, but it did, absolutely gorgeous stones.	
	(bracelet, gorgeous); (stones, horrible); (wrist, bad); (pain, honestly); (swelling, gone); (aches, return)	(scars, horrible); (swelling, gone down); (stones, gorgeous)
32	Bought these for my gf for her birthday and she said they're actually very comfortable and the heals are very cushiony. She actually wants me to buy her a few more pairs lol.	
	(shoes, comfortable); (heals, cushiony)	(heals, cushiony)
33	This hat is really very stylish and pretty...even for those of us who are hat-challenged. The brim is wide enough to protect my whole face from the sun...just what this fair-skinned lady needs! I highly recommend!	
	(hat, stylish); (brim, wide); (sun, protect); (lady, highly recommend)	(hat, stylish); (hat, pretty); (brim, wide); (lady, fair-skinned)
34	The belt broke with the buckle pulling off the belt. Really had just normal wear on it so I'm not sure why it happened. It was after the return period so I had no recourse but to throw it away. Disappointed.	
	(belt, disappointed); (buckle, disappointed); (wear, normal); (return, no)	(belt, broke); (wear, normal)
35	Works great. Have done a few hikes in these, 3+ miles each, over mostly packed snow. My buddies have some knock off brands of these and they fall off after an hour. These have stayed on every time.	
	(shoes, great); (snow, mostly packed); (time, every time)	(works, great)
36	I'm so happy with these sneakers! I ordered a half-size bigger because I knew I would be wearing thick socks. You could easily wear them without socks because the material is very breathable. They were a perfect fit! I slipped them on and went straight to the gym. They did not slip off my feet when I walked. I love that they don't have strings. The insoles are so soft! Great pair of sneakers. So comfortable!	
	(shoes, comfortable); (material, breathable); (fit, perfect); (insoles, soft)	(sneakers, happy); (material, breathable); (material, breathable); (insoles, soft); (sneakers, great); (sneakers, comfortable)
37	NOT AS PICTURED AND ADVERTISED. RETURNING THIS. ORDERED DESIGN IN NUMERAL/STICK (2nd photo) FORM AS ADVERTISED, BUT RECEIVED STONES (1st photo). IT LOOKS BEAUTIFUL BUT I DOUBT THIS IS AUTHENTIC AS IT FEELS SO LIGHT, CHEAP AND FADED. MY FOSSIL WATCHES BEFORE WERE QUITE HEAVY TO THE WRIST. NOT SURE IF THEY MADE MODIFICATION ON THEIR WATCH WEIGHT. I HAVE NOT OWNED FOSSIL WATCH FOR A WHILE. DEFINITELY RETURNING THIS	
	(stones, light); (watches, heavy); (modification, cheap)	(stones, beautiful); (stones, so light); (stones, cheap); (stones, faded); (watches, quite heavy)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
38	As you can see above in the photo this sports bra fits a bit too tight around and quit a bit of side boob sticks out. I am usually a 36 D, so I ordered a 36 D and it's very uncomfortable around my torso.As a sports bra I will order another one. I feel it is so comfortable and when I do CrossFit and do a lot of jumping my boobs do not move at all. I am impressed, I just wished it did not hurt around the torso.	
	(torso, uncomfortable); (boobs, comfortable)	(sports bra, too tight); (torso, uncomfortable); (sports bra, comfortable)
39	If you're looking for a modestly priced, reliable pair of sunglasses, these are a great choice. The frames are surprisingly sturdy, and they fit well. Taking off a star because the latch on the case they came in fell apart after one day, but have had no problems with the glasses so far	
	(frames, sturdy); (case, fell apart); (glasses, no problems so far)	(sunglasses, modestly priced); (sunglasses, reliable); (frames, sturdy); (frames, fit well)
40	I am always worried about the fit of clothes bought online but I took a chance with this product and it did not disappoint. The size was spot on and it was a perfect fit.The quality and color was as expected. I previously had purchased a pair and liked them so much that I ordered another pair.	
	(clothes, worried); (size, perfect); (quality, expected); (product, liked)	(size, spot on); (fit, perfect); (quality, as expected); (color, as expected)
41	I got them in the mail. I literally took them out of the package and the heart necklace broke. The chain pretty much crumbled into pieces. I was not to thrilled. If I could give it half a star I would.	
	(heart necklace, broke); (chain, crumbled); (package, not thrilled)	(heart necklace, broke); (chain, crumbled)
42	Significantly lighter than my old pair of harpoons. They flex alot more leading me to believe that they aren't manufactured as well as they used to be. We shall see. They do seem to have the same great crystal clear lenses though.	
	(lenses, great); (harpoons, manufactured); (flex, well); (believe, great)	(harpoons, lighter); (lenses, great)
43	These boots are comfortable, practical and stylish. Other than changing out the insoles and ordering up a size as recommended by a fellow reviewer this boot is just right for wet/mucky days and nights. I mostly use these to walk/run the dog. Easy clean - so when I come in I sometimes do not want to take them off. Of course, if worn too long without sox there is a tendency to get your feet stuck in for a little longer ;-) vacuums suck :-(	
	(boot, comfortable); (insoles, recommended); (foot, stuck); (vacuum, suck)	(boots, comfortable); (boots, practical); (boots, stylish); (clean, easy); (feet, stuck)
44	I want one in several colors. The fabric has plenty of give and is a nice thick weave. Very flattering, I did wear a lightweight smoother undergarment which helped a lot to have the dress look like a perfect fit. Very happy with purchase.	
	(fabric, nice); (give, flattering); (weave, nice); (fit, happy); (undergarment, flattering)	(fabric, nice); (weave, thick); (weave, nice); (fit, perfect); (purchase, happy)

(Continued on next page)

ID	Review	
	OA by Model	**OA by Human**
45	These are comfy. I am 180# 54́and wear a L/12. The large fit me well, but may be too tight in the thigh for some. If you have very thick thighs, order up. The bottom of the leg is a little wide, but I/m fine with that.	
	(leg, comfy); (thigh, tight); (bottom, wide)	(thigh, too tight); (leg, wide)
46	I ordered these masks as costumes for a community theater Shakespeare production. They're very well made, beautiful and sturdy. They've worked very well for our needs. They do tie with ribbons, however, and some of the actors expressed a preference to have elastics instead, both for comfort and security. We kept the ribbons though, and we haven't had any fall off even though the actors are doing very active choreographed stage fighting. A very good purchase!!	
	(masks, well made); (ribbons, tied); (actors, comfortable); (stage fighting, active)	(masks, well made); (masks, beautiful); (masks, sturdy); (purchase, good)
47	Carhartt can kiss my ass. They cheaped out. Material now shrinks, way thinner and sizes are inconsistent. They were made in Haiti, then Mexico and now Guatemala. Less for more is the motto of American companies.	
	(Material, thinner); (sizes, inconsistent); (Haiti, cheaped out); (Mexico, cheaped out); (Guatemala, cheaped out)	(material, shrinks); (way, thinner); (sizes, inconsistent)
48	Ive had a pair of these for 2 years now that have been to hell and back. I just ordered a new pair and one of them is missing the logo. It says that they are made in china but i cant tell where my old pair is made.	
	(pair of shoes, to hell and back); (logo, missing); (China, cant tell)	(logo, missing)
49	These boots are cute in the picture, but in reality they look like the cheap shoes my parents used to buy me when I was growing up. I really wanted to keep them, but can't get past the plastic look of the materials. Sorry.	
	(shoes, cheap); (materials, plastic)	(boots, cute); (materials, plastic look)
50	Just nice to order a product that is TRUE to size. Thongs can be tricky to order online- I just wanted a simple, lined, cotton, COMFORTABLE thong I can be comfortable in during a long day of traveling, walking, working and not worry about thin straps, rolling etc. also a great value with a perfect color mix with neutrals as well (the navy/wine/black combo) looking forward to reporting back- so far so good!	
	(size, true); (thongs, comfortable); (color, great); (value, perfect); (combo, looking forward)	(size, true); (value, great); (color, perfect)

This thesis was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and VNU University of Engineering and Technology.