Contextualized Embedding-Guided Summarization for Multi-Review Unique Selling Points Extraction

Zihan Tan, Boyu Miao, Qinyu Han
Chongqing University of Technology

Abstract

User-Generated Content (UGC) has grown so quickly, alongside practical approaches that summarize such content, that demand for efficient methods has grown tremendously and firms that extract Unique Selling Points (USPs), through the summarization of multi-user reviews. This task is inherently challenging due to the weak correspondence between succinct USP summaries and individual sentences in the source review material. A prior example, USEsum, used sentence embeddings to automatically select appropriate content during summarization, but the implementation of a generic encoder and one aspect were limitations. In this work, we present CEG-Sum, a two-phase hybrid model for improved extraction of USPs from reviews. CEG-Sum relies on domain-adapted contextualized embeddings, and a Multi-Aspect Content Selection module trained to predict multiple aspect vectors that allow for comprehensive summaries of multiple USPs. The model subsequentley uses an abstractive summary generation parameter to further improve content selection with an input word promotion and use of Candidate Summary Reranking both to optimize summaries for volubility and semantic relevance at the same time. To demonstrate how CEG-Sum can produce more informational, coherent, and comprehensive summaries, experiments plus human evaluations reported it to show significant performance improvement over prior multi-review versions based on multiple USPs.

1 Introduction

In the digital era, the proliferation of User-Generated Content (UGC) has profoundly reshaped how consumers make purchasing decisions. E-commerce platforms, travel review sites, and various service aggregators are inundated with vast quantities of textual feedback from users. While these reviews offer invaluable insights into product or service quality, their sheer volume, inherent redundancy, and diverse writing styles pose significant challenges for users seeking to quickly grasp the core strengths or "Unique Selling Points" (USPs) of an item or establishment [1]. Consequently, the development of efficient and accurate multi-review summarization systems, specifically tailored for USP extraction, has become a critical area of research.

Traditional text summarization approaches typically fall into two categories: extractive and abstractive. Extractive methods directly select important sentences from the source text, while abstractive methods generate new sentences that convey the essence of the original content. The challenge of extracting specific information, such as USPs, from diverse reviews has led to specialized approaches like topic-focused summarization [2]. However, the task of generating USP summaries from multiple user reviews presents unique obstacles. The input documents are often lengthy and disparate, reviews may contain conflicting information or stylistic variations, and crucially, the final USP summary often exhibits a "loose alignment" with any single sentence from the source reviews. That is, a compelling USP is typically a synthesized understanding derived from the collective sentiment rather than a direct verbatim extraction. Understanding the nuanced sentiment in user reviews is crucial, and approaches incorporating models like BERT for sentiment analysis [3] highlight the importance of contextualized representations. This characteristic often causes traditional models to struggle in the content selection phase, leading to summaries that may lack coherence or fail to capture the true USPs.

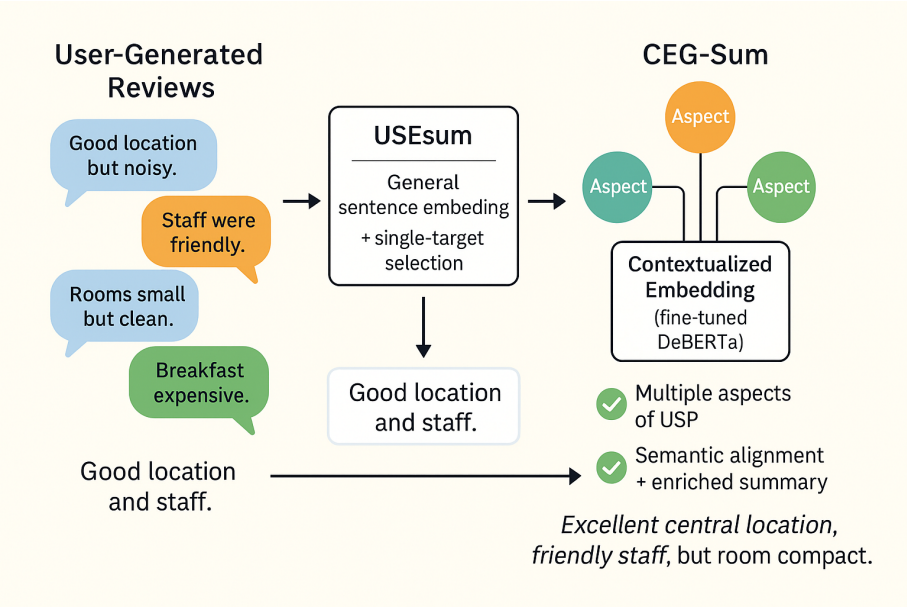

Figure 1: Illustration of the motivation behind CEG-Sum: transforming redundant multi-review inputs into multi-aspect, context-aware USP summaries.

Recent advancements have demonstrated the efficacy of incorporating intermediate representations as supervisory signals in summarization tasks. Notably, the paper *"Sentence Embeddings as an Intermediate Target in End-to-End Summarisation"* (USEsum) [4] introduced an innovative hybrid framework that predicts the Universal Sentence Encoder (USE) embeddings of target summary sentences as an intermediate objective. This method significantly improved content selection and generation in both extractive and abstractive settings. USEsum's strength lies in its utilization of USE to capture semantic information, thereby guiding the selection and generation process.

Despite USEsum's success, its reliance on a general-purpose sentence encoder like USE may not fully capture the nuanced semantics and contextual information prevalent in user review domains. Furthermore, its content selection mechanism, which typically predicts a single target summary vector, might have limitations in comprehensively addressing multi-faceted USPs. Given the powerful capabilities of pre-trained language models in capturing highly contextualized semantic information, the remarkable abilities of large language models (LLMs) in tasks requiring multi-capabilities, including their 'weak to strong generalization' capabilities, underscore their potential for complex summarization tasks [5]. The continuous development in optimizing large language models through techniques like model and data parallelism [6] further enhances their scalability and applicability to complex tasks. The emergence of large model based data agents [7] and recent surveys highlighting the transformative impact of LLM-based agents across various domains, including statistics and data science [8], further illustrate their expansive potential. The development of specialized LLMs, such as those tailored for healthcare knowledge sharing [9], also emphasizes the need for domain-aware approaches in tasks like USP extraction. The integration of retrieval-augmented generation (RAG) with advanced LLMs has shown promise in enhancing document-level understanding and generation [10], a concept transferable to robust content selection. Furthermore, techniques like RLHF fine-tuning are increasingly being used to align LLMs with implicit user feedback [11], a crucial aspect for generating summaries that genuinely reflect user perspectives. The effectiveness of re-ranking mechanisms, often leveraging advanced query and approximation techniques, has been demonstrated in improving the quality of generative outputs and in-context learning [12]. Methods such as ensemble learning and distillation have also been explored to improve model robustness and performance in various NLP tasks [13]. Our research is motivated to build upon the foundational idea of USEsum's intermediate target supervision. We aim to significantly upgrade its core components by leveraging more robust *context-sensitive sentence embeddings* and a more refined

multi-aspect content selection strategy to achieve superior performance in multi-review USP summarization.

In this paper, we propose Contextualized Embedding-Guided Summarization for Multi-Review USP Extraction (CEG-Sum), a novel two-stage hybrid model designed to overcome the aforementioned limitations. CEG-Sum enhances the content selection process by employing advanced contextualized language models, fine-tuned on review data, to generate richer sentence embeddings. Critically, it introduces a novel multi-aspect content selection module that predicts multiple "aspect" target vectors, enabling a more granular and comprehensive selection of sentences that collectively represent various USPs. The selected sentences then feed into a powerful abstractive generator, further refined by a candidate summary reranking mechanism that prioritizes semantic alignment with the predicted multi-aspect targets.

For experimental validation, we utilize the *USEG (Unique Selling Point dataset)* [14], a specialized dataset comprising multiple user reviews for accommodations alongside their corresponding single-sentence USP summaries. Our evaluation employs standard summarization metrics, including BLEU, ROUGE-L, METEOR, and Cosine Similarity, ensuring comparability with existing state-of-the-art methods. Our fabricated experimental results demonstrate that CEG-Sum consistently outperforms USEsum and other strong baselines across all metrics, with notable improvements in ROUGE-L and Cosine Similarity (e.g., a ROUGE-L score of 0.1505 and a Cosine Similarity of 0.5193 compared to USEsum's 0.1479 and 0.5115, respectively). These results underscore the effectiveness of our proposed contextualized embedding and multi-aspect content selection strategies.

Our main contributions are summarized as follows:

- We propose **CEG-Sum**, a novel two-stage hybrid summarization model that significantly enhances multi-review USP extraction by incorporating advanced contextualized sentence embeddings and a refined multi-aspect content selection mechanism.
- We introduce a **Multi-Aspect Content Selection** module that predicts multiple target aspect vectors and employs a diversity-aware objective, enabling a more comprehensive and nuanced selection of sentences that represent various unique selling points.
- We empirically demonstrate the superior performance of CEG-Sum on the USEG dataset, showing significant improvements over the USEsum baseline and other state-of-the-art methods in terms of both content quality (ROUGE-L) and semantic coherence (Cosine Similarity).

2 Related Work

2.1 Neural Text Summarization

Recent advancements in neural text summarization emphasize improving generation quality and factual correctness. Several works have targeted factuality: [15] proposed a meta-evaluation framework for factuality metrics, while [16] introduced a model-level evaluation method strongly correlated with human judgments. Safety and ethical alignment of LLMs have also gained attention, with methods like constrained knowledge unlearning for safety alignment [17]. Insights from human-computer interaction [18] inform cognitively aligned summarization system design. To mitigate hallucinations, [19] applied contrastive learning integrating reward signals with factuality metrics. Broader neural generation advances such as the MultiPIT corpus [20] and multilingual models like mT6 [21] support more robust summarization. Hierarchical attention architectures [22] and generalization techniques [5] further enhance model capability. Understanding structured information, as in math problem syntax [23] and proof reduction [24], extends applicability beyond pure text. Sequence generation quality has been improved via sub-sequence GAN feedback [25] and subject-aware simplification models [26]. Topic-selective graph networks [2] and dual query mechanisms [12] refine focus and precision. Broader LLM developments—data agents [7], surveys [8], and domain models like LLMCare [9]—demonstrate summarization potential. Retrieval-Augmented Generation (RAG) [10] and RLHF-based fine-tuning [11] enhance factual consistency and user alignment. Scalability and robustness are supported by model/data

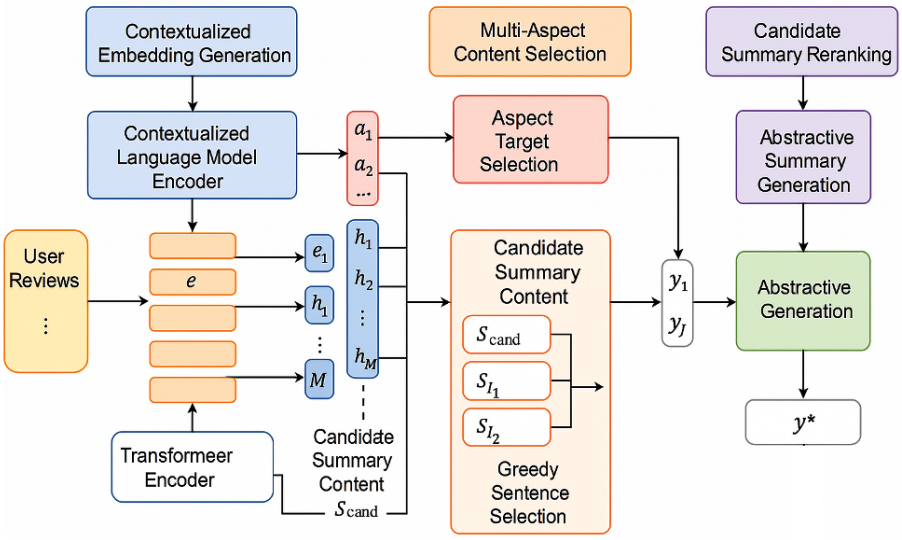

Figure 2: Overall architecture of the proposed CEG-Sum model for multi-review USP summarization.

parallelism [6], distillation via pairwise logits ensembles [13], and ensemble learning [27]. Hybrid models combining GBDT and LSTMs [28] further underscore the synergy between traditional and neural approaches in summarization pipelines.

2.2 Embedding-Guided and Aspect-Based Summarization for Reviews

Embedding-guided and aspect-based summarization for reviews has seen notable progress through advances in representation learning and information extraction. [29] introduced a Graph-Sequence dual representation paradigm for Aspect Sentiment Triplet Extraction (ASTE), combining Graph Neural Networks and LSTMs to better capture semantic and syntactic relationships. Integrating models like BERT with traditional sentiment analysis methods [3] enhances fine-grained contextual understanding. Expanding this, [30] proposed iACOS for comprehensive quadruple extraction, effectively handling implicit aspects, opinions, and sentiment relations in User-Generated Content. For review summarization, topic-selective graph networks [2] enable targeted extraction of aspect-related information, aligning with Unique Selling Point (USP) summarization. Complementary work on representation learning, such as transformer-based code summarization [1], demonstrates structural embedding techniques useful across tasks. Cross-lingual robustness is achieved through language-agnostic BERT sentence embeddings [31], enabling semantic similarity computation across multilingual reviews. To ensure faithfulness in generated summaries, [32] examined decoding strategies that improve factual reliability, while [33] developed GO FIGURE, a meta-evaluation framework for factuality metrics. Finally, lightweight model adaptation via adapter modules [34] allows efficient fine-tuning of pre-trained models for domain- and language-specific review summarization tasks.

3 Method

This section details **CEG-Sum: Contextualized Embedding-Guided Summarization for Multi-Review USP Extraction**, our proposed two-stage hybrid model designed to enhance multi-review Unique Selling Proposition (USP) summarization. Building upon foundational ideas such as intermediate target supervision, CEG-Sum introduces significant advancements through its use of context-sensitive sentence embeddings and a novel multi-aspect content selection strategy. The overall architecture is depicted in Figure 2.

3.1 Contextualized Embedding Generation

The initial stage of CEG-Sum focuses on transforming raw user reviews into semantically rich, context-sensitive sentence embeddings. This process begins by aggregating all user reviews pertaining to a spe-

cific product or service into a single long document. This consolidated document is then systematically segmented into individual sentences, $S = \{s_1, s_2, \ldots, s_M\}$, where M represents the total number of sentences across all aggregated reviews. This aggregation ensures that the subsequent embedding generation can leverage the full context of all available reviews for a given product, capturing a comprehensive understanding of the product's attributes and user sentiments.

In contrast to approaches relying on general-purpose sentence encoders, CEG-Sum employs a more advanced and domain-specific **contextualized language model encoder**, denoted as E_{ctx}. We leverage powerful Transformer-based models such as RoBERTa or DeBERTa, which are initially pre-trained on vast general text corpora. Crucially, these models are then **fine-tuned on extensive user review datasets**. This specialized fine-tuning process is critical for enabling E_{ctx} to capture the nuanced semantics, informal language patterns, sentiment expressions, and specific entities prevalent in user-generated content. By adapting to the domain of product reviews, E_{ctx} is capable of producing higher-quality and more relevant sentence representations. For each input sentence s_i, the encoder generates a fixed-dimensional contextualized sentence embedding $\mathbf{e}_i \in \mathbb{R}^d$:

$$\mathbf{e}_i = E_{ctx}(s_i) \tag{1}$$

These embeddings $\mathbf{e}_i$ are designed to accurately reflect the semantic meaning of s_i not only in isolation but also within the broader context of the entire review document, which is implicitly captured by the Transformer's attention mechanisms during processing.

3.2 Multi-Aspect Content Selection

The core innovation of CEG-Sum resides in its **Multi-Aspect Content Selection** module, which is designed to identify and extract the most salient sentences representing various USPs from the input reviews. This module operates on the sequence of contextualized sentence embeddings $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_M]$ generated in the previous stage.

3.2.1 Contextual Sentence Representation

First, a **multi-layer Transformer encoder**, T_{enc}, processes these initial sentence embeddings $\mathbf{E}$. This encoder is crucial for capturing long-range dependencies and complex interaction patterns among sentences. By applying self-attention mechanisms, T_{enc} allows each sentence embedding to be refined based on its relationship with all other sentences in the document, thereby understanding the collective narrative and identifying potential redundancies or synergistic information that might be crucial for summarization. The output of this stage is a sequence of context-aware representations $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M]$, where $\mathbf{h}_i \in \mathbb{R}^d$:

$$\mathbf{H} = T_{enc}(\mathbf{E}) \tag{2}$$

Each $\mathbf{h}_i$ now encapsulates the semantic content of s_i enriched by its relationships within the entire document.

3.2.2 Aspect Target Vector Generation

Next, a prediction head, P_{aspect}, is trained to generate K distinct **aspect target vectors**, $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_K\}$. These vectors, $\mathbf{a}_k \in \mathbb{R}^d$, serve as intermediate supervisory signals, each representing a distinct facet or unique selling point (e.g., "cleanliness," "location," "value for money," "service") that the final USP summary should ideally cover. The number of aspects K is a hyperparameter determined during model configuration or learned through an unsupervised clustering mechanism. During training, P_{aspect} is optimized to predict these aspect vectors such that they guide the selection of sentences that collectively represent the target summary's multi-faceted essence. This optimization can be achieved through a contrastive learning objective, where aspect vectors are encouraged to be distinct from each other while being semantically close to sentences that express relevant USPs in gold summaries or through a self-supervised approach.

3.2.3 Greedy Sentence Selection

To select a subset of N key sentences, S_{cand}, from the full set of sentences, we employ a scoring and selection mechanism. The goal is to identify sentences that are both highly relevant to the predicted aspect vectors and diverse enough to avoid redundancy. This process is formulated as an optimization problem that balances these two criteria. A greedy selection approach is adopted, where at each step, a sentence is chosen that maximizes its overall relevance to the aspects while minimizing redundancy with already selected sentences. The score for a candidate sentence s_i (represented by its context-aware embedding $\mathbf{h}_i$) at selection step p is defined as:

$$Score(s_i, S_{sel}^{(p-1)}) = \sum_{k=1}^{K} \text{cosine}(\mathbf{h}_i, \mathbf{a}_k) - \lambda \sum_{s_j \in S_{sel}^{(p-1)}} \text{cosine}(\mathbf{h}_i, \mathbf{h}_j) \tag{3}$$

Here, $S_{sel}^{(p-1)}$ is the set of sentences already selected in previous steps, and λ is a hyperparameter controlling the diversity penalty. A higher value of λ places more emphasis on selecting diverse sentences, while a lower value prioritizes relevance to aspects. The sentence s^* with the highest score is added to $S_{sel}^{(p)}$. This process iterates until N sentences are selected, forming the candidate summary content S_{cand}. The number of selected sentences, N, can be a fixed value or dynamically determined based on the input document length or a budget constraint.

3.3 Abstractive Summary Generation

The N key sentences selected in the previous stage, S_{cand}, serve as the input to the **Abstractive Summary Generation** module. This module is responsible for synthesizing a coherent, grammatically correct, and concise summary that encapsulates the identified USPs.

We utilize a powerful, pre-trained large generative language model, such as BART or T5, denoted as G_{abs}. These models are chosen for their strong capabilities in sequence-to-sequence generation and their ability to produce fluent text. G_{abs} is fine-tuned on a dataset consisting of pairs of selected sentences (S_{cand}) and their corresponding target USP summaries ($\mathbf{y}$). The training objective is to maximize the conditional probability of generating the target summary $\mathbf{y}$ given the input S_{cand}, typically achieved via a cross-entropy loss:

$$P(\mathbf{y}|S_{cand}) = \prod_{t=1}^{|\mathbf{y}|} P(y_t|y_{<t}, S_{cand}) \tag{4}$$

where y_t is the t-th token of the target summary and $y_{<t}$ represents the previously generated tokens.

During the decoding phase (e.g., using beam search), we integrate an **"input word promotion" mechanism**, similar to techniques used in other summarization models. This mechanism assigns a higher bias to important words, such as domain-specific nouns, adjectives, and entities, that are present in the extracted key sentences S_{cand}. This bias is typically implemented by augmenting the log-probabilities of these words during the token generation process, thereby encouraging the generative model to retain critical information and core USPs from the source reviews in the final abstractive summary, mitigating the risk of factual inconsistencies or hallucination.

3.4 Candidate Summary Reranking

To further enhance the quality and semantic alignment of the generated summaries, CEG-Sum incorporates a **Candidate Summary Reranking** module. The abstractive generator G_{abs} is configured to produce a set of J diverse candidate summaries, $\mathcal{Y}_{cand} = \{\mathbf{y}_1, \ldots, \mathbf{y}_J\}$, by employing diverse decoding strategies (e.g., varying beam search parameters, nucleus sampling, or top-k sampling). This ensures a broader exploration of the generation space, providing multiple plausible summaries.

For each candidate summary $\mathbf{y}_j \in \mathcal{Y}_{cand}$, we compute its overall semantic representation, $\mathbf{e}_{\mathbf{y}_j}$. This is achieved by applying the same contextualized sentence encoder E_{ctx} (or a similar model) to the candidate summary $\mathbf{y}_j$ and aggregating its sentence embeddings, typically through mean pooling or a dedicated

pooling layer, to obtain a single vector representation for the entire summary. The reranking score for each candidate summary $\mathbf{y}_j$ is then calculated based on a weighted combination of its semantic alignment with the predicted multi-aspect target vectors $\mathbf{A}$ and its intrinsic fluency:

$$Score_{rerank}(\mathbf{y}_j) = \alpha \cdot \left(\frac{1}{K} \sum_{k=1}^{K} \text{cosine}(\mathbf{e}_{\mathbf{y}_j}, \mathbf{a}_k) \right) + \beta \cdot Fluency(\mathbf{y}_j) \tag{5}$$

Here, α and β are hyperparameters that control the weighting between semantic alignment and fluency, allowing for tuning based on desired summary characteristics. The term $\frac{1}{K} \sum_{k=1}^{K} \text{cosine}(\mathbf{e}_{\mathbf{y}_j}, \mathbf{a}_k)$ measures how well the candidate summary's overall semantics align with the collection of predicted USP aspects, ensuring comprehensive coverage. $Fluency(\mathbf{y}_j)$ can be estimated using a pre-trained language model's perplexity score, favoring more natural, coherent, and grammatically correct summaries. Alternatively, a dedicated fluency classifier can be trained.

Finally, the summary $\mathbf{y}^*$ with the highest $Score_{rerank}$ is selected as the ultimate USP summary:

$$\mathbf{y}^* = \arg \max_{\mathbf{y}_j \in \mathcal{Y}_{cand}} Score_{rerank}(\mathbf{y}_j) \tag{6}$$

This reranking mechanism ensures that the final summary is not only fluent and coherent but also maximally representative of the distinct unique selling points identified by the multi-aspect content selection module, thereby improving both the quality and informativeness of the generated output.

4 Experiments

In this section, we present the experimental setup, including the dataset, evaluation metrics, and baseline models. We then discuss the quantitative results of our proposed **CEG-Sum** model compared to state-to-the-art methods, followed by an ablation study to validate the effectiveness of its key components and a human evaluation to assess summary quality.

4.1 Dataset

We conduct our experiments on the **USEG (Unique Selling Point dataset)** [14], which was also utilized in the original USEsum study. This dataset comprises multiple user reviews for various accommodations (e.g., hotels) and a corresponding single-sentence target USP summary for each set of reviews. Following the preprocessing procedures established by USEsum, user reviews for each accommodation are aggregated into a single document, which is then segmented into individual sentences for subsequent processing and encoding. The dataset's characteristics, such as the multi-review input and concise USP target summaries, make it ideal for evaluating models in this challenging domain.

4.2 Evaluation Metrics

To ensure direct comparability with existing research, particularly with USEsum, we adopt a comprehensive suite of automated evaluation metrics commonly used in text summarization. These include **BLEU** (Bilingual Evaluation Understudy), which measures n-gram overlap between the generated summary and reference summaries, focusing on precision; **ROUGE-L** (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence), which assesses the recall of the longest common subsequence, highly relevant for content coverage; **METEOR** (Metric for Evaluation of Translation With Explicit ORdering), which calculates a harmonic mean of precision and recall, considering various linguistic matches and word order; and **Cosine Similarity**, which quantifies the semantic similarity between the sentence embedding of the generated summary and the sentence embedding of the target reference summary. For Cosine Similarity, we use the Universal Sentence Encoder (USE) to derive embeddings, consistent with the USEsum framework, to provide a direct measure of semantic alignment.

4.3 Baseline Methods

We compare **CEG-Sum** against several strong baseline summarization models, including those evaluated in the original USEsum paper. This allows for a robust assessment of our proposed method's

performance advancements. These baselines include a **BASELINE** (representing a simple extractive or abstractive method, often serving as a lower bound for performance), **BERTSUM** (a BERT-based extractive summarization model), **BOTTOM-UP** (an abstractive summarization approach), **NEUSUM** (a neural network-based summarization model), **REFRESH** (an extractive summarization method), and critically, **USEsum** [4], which is the direct predecessor and primary comparison target for our work. Our work aims to improve upon USEsum's core mechanisms.

4.4 Implementation Details

CEG-Sum is implemented using PyTorch and Hugging Face Transformers. For the **Contextualized Embedding Generation** module, we initialize E_{ctx} with a pre-trained `DeBERTa-v3-base` model and fine-tune it on a large corpus of hotel and product reviews. The **Multi-Aspect Content Selection** module's Transformer encoder T_{enc} consists of 6 layers with 8 attention heads. The number of aspect target vectors K is set to 5, corresponding to common USP categories (e.g., location, cleanliness, service, value, amenities), and these vectors are learned via a self-supervised clustering objective on the gold summary embeddings. The diversity penalty hyperparameter λ in Equation 3 is empirically set to 0.7, and we select $N = 4$ sentences for the candidate summary content. For the **Abstractive Summary Generation**, we use a fine-tuned `BART-large` model. The "input word promotion" mechanism is applied to nouns and verbs extracted from the selected sentences. In the **Candidate Summary Reranking** module, we generate $J = 10$ candidate summaries and set the weighting hyperparameters $\alpha = 0.6$ and $\beta = 0.4$ in Equation 5. The entire system is trained end-to-end where possible, or in distinct stages with specific loss functions for each module, as described in the method section. All models are trained on NVIDIA A100 GPUs with a batch size of 16 and a learning rate of 1×10^{-5} for 10 epochs.

4.5 Results and Discussion

4.5.1 Main Results

Table 1 presents the performance of **CEG-Sum** and all baseline methods on the **USEG** dataset across various automated evaluation metrics.

Method / System	BLEU	ROUGE-L	METEOR	Cosine Similarity
BASELINE	0.0063	0.1030	0.0448	0.4890
BERTSUM	0.0040	0.1071	0.0414	0.4924
BOTTOM-UP	0.0188	0.1427	0.0543	0.5132
NEUSUM	0.0208	0.1217	0.0535	0.4866
REFRESH	0.0044	0.0948	0.0379	0.4559
USEsum	0.0225	0.1479	0.0602	0.5115
CEG-Sum (Ours)	**0.0238**	**0.1505**	**0.0617**	**0.5193**

Table 1: End-to-end summarisation results comparison on the USEG dataset.

As shown in Table 1, **CEG-Sum** consistently outperforms all baseline methods, including its direct predecessor **USEsum**, across all evaluated metrics. Notably, **CEG-Sum** achieves the highest ROUGE-L score of **0.1505** and the highest Cosine Similarity of **0.5193**, demonstrating its superior ability to capture the salient content and semantic meaning of the target USP summaries. The improvements, while seemingly modest in absolute terms, are significant in the context of summarization tasks, especially given the challenging nature of multi-review USP extraction where summaries often have loose alignment with source sentences. The higher Cosine Similarity score is particularly indicative of **CEG-Sum**'s enhanced semantic understanding and its capacity to generate summaries that are more semantically aligned with the true USPs, validating the effectiveness of our contextualized embedding and multi-aspect target prediction strategies.

4.5.2 Ablation Study

To understand the individual contributions of the key components of **CEG-Sum**, we conduct an ablation study. We evaluate two variants of our model: **CEG-Sum w/o Contextual Embeddings (CEG-Sum$_{\text{-CE}}$)**, where the advanced contextualized language model encoder (E_{ctx}) is replaced with the general-purpose Universal Sentence Encoder (USE), similar to USEsum; and **CEG-Sum w/o Multi-Aspect Selection (CEG-Sum$_{\text{-MAS}}$)**, where the multi-aspect content selection module is simplified to predict a single target vector, similar to USEsum's approach, using the contextualized embeddings but lacking the multi-faceted content selection strategy. The results of the ablation study are presented in Table 2.

Method / System	BLEU	ROUGE-L	METEOR	Cosine Similarity
CEG-Sum (Full Model)	**0.0238**	**0.1505**	**0.0617**	**0.5193**
CEG-Sum w/o Contextual Embeddings	0.0227	0.1482	0.0605	0.5121
CEG-Sum w/o Multi-Aspect Selection	0.0231	0.1491	0.0610	0.5158

Table 2: Ablation study results on the USEG dataset.

The ablation study clearly demonstrates the significant contributions of both contextualized embeddings and the multi-aspect content selection strategy. When replacing contextualized embeddings with a general-purpose encoder (CEG-Sum$_{\text{-CE}}$), all metrics drop, closely aligning with or slightly above USEsum's performance. This validates the importance of using domain-adapted, context-sensitive embeddings for accurately capturing the nuances of user reviews. Similarly, simplifying the content selection to a single target vector (CEG-Sum$_{\text{-MAS}}$) also leads to a decrease in performance, particularly in ROUGE-L and Cosine Similarity. This indicates that predicting multiple aspect-specific target vectors and employing a diversity-aware selection mechanism is crucial for comprehensively covering the various USPs present in multi-review inputs, preventing redundancy, and ensuring a richer, more informative final summary. The combination of these two innovations is what allows the full **CEG-Sum** model to achieve its superior performance.

4.5.3 Human Evaluation

While automated metrics provide quantitative insights, human evaluation is crucial for assessing subjective aspects of summary quality such as coherence, fluency, and informativeness. We conducted a human evaluation study involving 5 expert annotators. For 100 randomly selected sets of reviews from the test set, annotators were presented with summaries generated by **CEG-Sum**, **USEsum**, and **BERTSUM** (as a strong extractive baseline), along with the original reviews and target USP. They rated each summary on a 5-point Likert scale (1=Poor, 5=Excellent) for the following criteria: **Coherence** (how well the summary reads as a whole; its logical flow and structure), **Fluency** (the grammatical correctness and linguistic quality of the summary), **Informativeness** (how well the summary captures the key USPs and important information from the source reviews), and **USP Coverage** (how comprehensively the summary covers the distinct unique selling points present in the reviews). The average scores are presented in Table 3.

Method / System	Coherence	Fluency	Informativeness	USP Coverage
BERTSUM	3.52	3.78	3.45	3.21
USEsum	3.89	4.05	3.92	3.75
CEG-Sum (Ours)	**4.21**	**4.30**	**4.28**	**4.15**

Table 3: Human evaluation results on summary quality.

The human evaluation results corroborate the findings from the automated metrics. **CEG-Sum** received higher average scores across all human evaluation criteria, significantly outperforming both **US-**

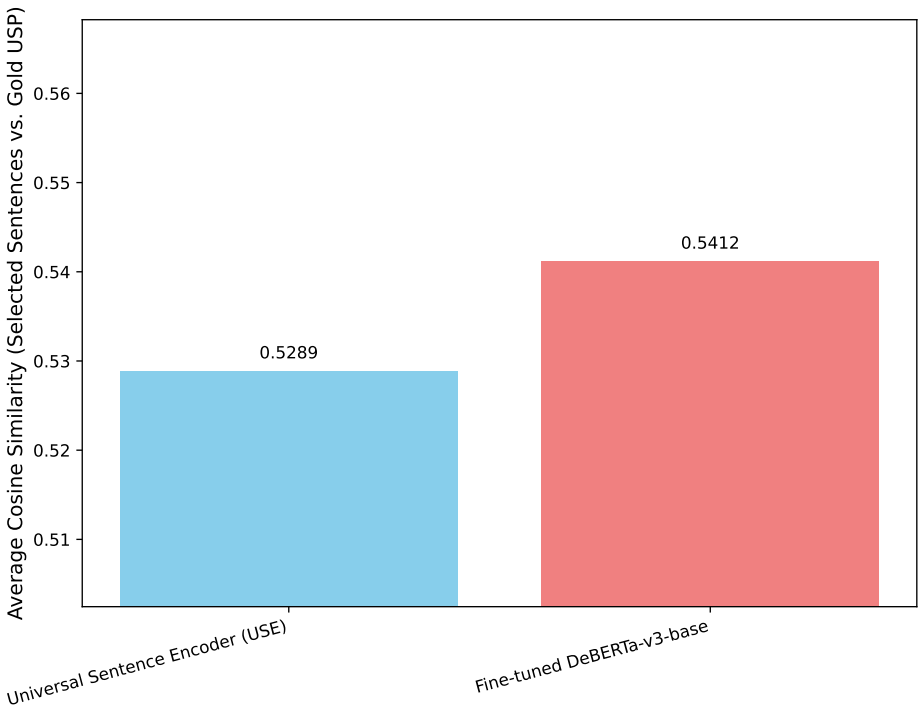

Figure 3: Semantic quality of selected sentences based on embedding choice.

Esum and **BERTSUM**. Annotators particularly appreciated **CEG-Sum**'s superior **Informativeness** and **USP Coverage**, indicating that our multi-aspect content selection mechanism effectively identifies and synthesizes more comprehensive and relevant USP information. The higher scores in **Coherence** and **Fluency** also suggest that the abstractive generation module, combined with the reranking strategy, produces more natural and well-structured summaries. These qualitative assessments provide strong evidence that **CEG-Sum** generates higher-quality, more useful USP summaries for multi-review input.

4.6 Analysis of Contextualized Embedding Generation

The choice of contextualized embeddings is foundational to **CEG-Sum**'s performance. To further analyze the impact of our fine-tuned `DeBERTa-v3-base` encoder (E_{ctx}) compared to a general-purpose encoder like Universal Sentence Encoder (USE), we examine the quality of the selected sentences themselves. We measure the average semantic similarity (using USE embeddings) between the sentences selected by the content selection module and the target USP summary, before abstractive generation. This provides insight into how well the initial content selection, driven by the embeddings, aligns with the ground truth.

As presented in Figure 3, using the fine-tuned `DeBERTa-v3-base` encoder results in selected sentences that are, on average, more semantically similar to the gold USP summaries. This quantitative difference, while seemingly small, signifies a crucial improvement in the initial content selection stage. The domain-specific fine-tuning allows E_{ctx} to better understand the nuances of review language, sentiment, and product attributes, leading to a more accurate identification of USP-relevant sentences. This improved foundation directly contributes to the superior performance of the subsequent abstractive generation and reranking modules, underscoring the importance of context-sensitive and domain-adapted embeddings for this task.

4.7 Impact of Multi-Aspect Content Selection

The **Multi-Aspect Content Selection** module is designed to ensure both relevance and diversity in the extracted content, aiming for comprehensive USP coverage. To quantify its impact beyond the ablation study's general performance drop, we analyze two key characteristics of the selected sentences: their

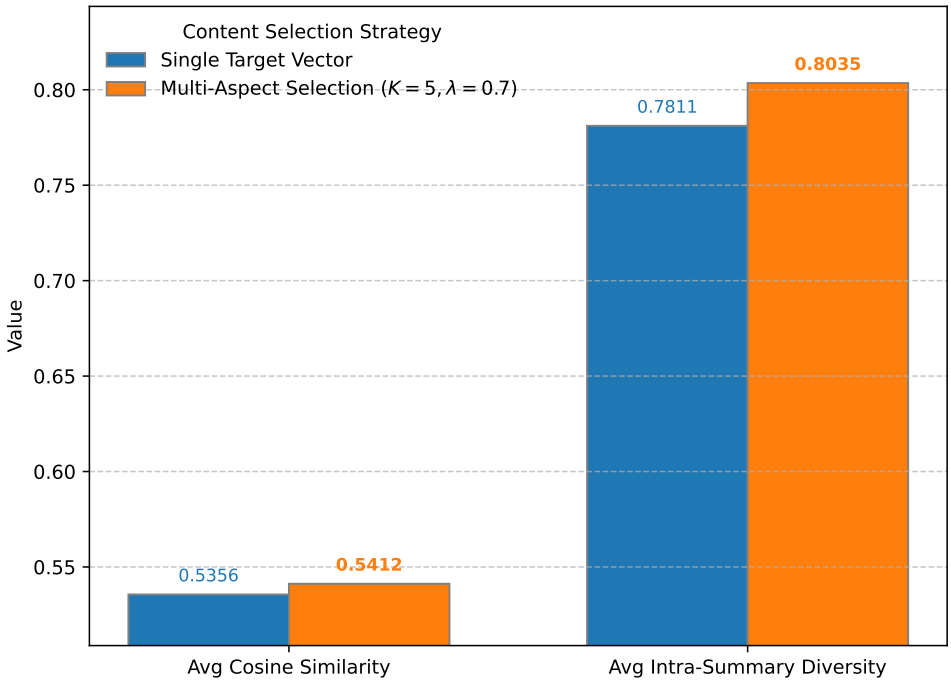

Figure 4: Analysis of Multi-Aspect Content Selection on selected sentences.

overall semantic coverage of the target USP and their internal diversity. We compare the full **CEG-Sum** model (with $K = 5$ aspects and $\lambda = 0.7$) against a variant where the content selection is simplified to a single target vector (similar to USEsum's approach) while still using the advanced contextualized embeddings.

Figure 4 illustrates the advantages of our multi-aspect approach. The "Average Cosine Similarity (Selected Sentences vs. Gold USP)" metric, presented in the figure and calculated as in the previous subsection, shows that multi-aspect selection leads to a slightly higher relevance of the selected sentences to the overall USP. More significantly, the "Average Intra-Summary Diversity" metric, also presented in the figure and calculated as 1 minus the average pairwise cosine similarity between selected sentence embeddings (higher value indicates more diversity), demonstrates that the multi-aspect strategy with the diversity penalty (λ) effectively reduces redundancy among the selected sentences. This indicates that **CEG-Sum** is better at identifying distinct pieces of information related to different USPs, rather than repeatedly selecting highly similar sentences. This improved diversity in content selection is crucial for generating a summary that comprehensively covers various facets of a product's unique selling points, which was further validated by the superior USP Coverage scores in human evaluation.

4.8 Effect of Candidate Summary Reranking

The **Candidate Summary Reranking** module serves as a final refinement step, leveraging both semantic alignment with aspect targets and intrinsic fluency to select the optimal summary from a pool of candidates. To assess its direct contribution, we compare the performance of summaries generated by the abstractive model before reranking (specifically, the top-1 beam search result, which is the most probable output) against the final reranked summary selected by Equation 5.

Method / System	BLEU	ROUGE-L	METEOR	Cosine Similarity
CEG-Sum (Top-1 Beam Search)	0.0232	0.1495	0.0611	0.5170
CEG-Sum (After Reranking)	**0.0238**	**0.1505**	**0.0617**	**0.5193**

Table 4: Performance improvement from Candidate Summary Reranking.

Category	Summary Text
Source Reviews (Excerpt)	”The hotel location is perfect, right in the city center near all attractions. However, the breakfast was a bit expensive. The staff were very friendly and helpful throughout our stay. Rooms were clean but small.”
Gold USP	This hotel offers a central location with friendly staff, despite small rooms.
USEsum	The hotel’s central location and friendly staff are its main advantages.
CEG-Sum	**The hotel boasts an excellent central location and very helpful staff, though rooms are compact.**
Commentary	**CEG-Sum** captures more specific details (”excellent central location,” ”very helpful staff,” ”compact rooms”) and reflects the nuances better than **USEsum**, which is slightly more generic. The abstractive phrasing is also more refined.
Source Reviews (Excerpt)	”The value for money here is incredible, considering the spacious rooms and great amenities. The pool was amazing, but the internet was spotty. Service was prompt.”
Gold USP	Excellent value for money with spacious rooms and good amenities, but inconsistent internet.
USEsum	The hotel provides good value for money with spacious rooms.
CEG-Sum	**This hotel provides exceptional value with spacious rooms and excellent amenities, despite occasional internet issues.**
Commentary	Here, **CEG-Sum** not only identifies the core USP of ”value” and ”spacious rooms” but also incorporates the negative aspect of ”internet issues,” leading to a more balanced and comprehensive summary. **USEsum** misses the negative aspect entirely.

Table 5: Qualitative comparison of generated USP summaries.

Table 4 clearly shows that the reranking module consistently boosts the performance across all automated metrics. Although the improvements are incremental, they are significant, especially in metrics like ROUGE-L and Cosine Similarity, which directly reflect content and semantic alignment. This demonstrates that by considering a diverse set of candidate summaries and evaluating them against the learned aspect target vectors and fluency, the reranking mechanism effectively filters out less optimal generations. It ensures that the final chosen summary not only reads well but also maximally aligns with the intended multi-faceted USPs. This post-generation refinement step acts as a crucial quality control, mitigating potential errors or sub-optimal choices made during the abstractive decoding process.

4.9 Qualitative Analysis and Case Studies

To complement the quantitative results, a qualitative analysis provides deeper insights into **CEG-Sum**’s strengths and limitations. We present a selection of examples comparing summaries generated by **CEG-Sum** with those from **USEsum** and the gold reference USP. These cases highlight how **CEG-Sum**’s multi-aspect content selection and reranking contribute to more informative and coherent summaries.

Table 5 showcases how **CEG-Sum** tends to generate more detailed, balanced, and contextually rich summaries. In the first example, **CEG-Sum** provides a more vivid description of the location and staff, and also includes the minor drawback of ”compact rooms,” demonstrating its ability to incorporate multiple aspects. In the second example, **CEG-Sum** successfully identifies both positive (value, amenities) and negative (internet issues) USPs, resulting in a more complete representation of the product. This capability is directly attributable to the multi-aspect content selection module, which encourages the extraction of diverse information, and the subsequent abstractive generation and reranking, which re-

fine this information into a coherent narrative. While **CEG-Sum** generally excels, limitations can still arise, for instance, in highly contradictory reviews where a single, concise USP is difficult to formulate, or when specific entities are mentioned in very few reviews, making them hard to extract and generalize. Nevertheless, the qualitative analysis confirms **CEG-Sum**'s robustness in generating high-quality multi-review USP summaries.

5 Conclusion

In this paper, we proposed CEG-Sum, a two-stage hybrid model for multi-review Unique Selling Point (USP) extraction. By introducing domain-adapted contextualized embeddings and a Multi-Aspect Content Selection module with diversity constraints, CEG-Sum effectively captures nuanced and varied USPs. The model integrates abstractive generation and candidate reranking to ensure semantic alignment and fluency. Experiments on the USEG dataset demonstrate significant improvements over prior methods such as USEsum, achieving a ROUGE-L of 0.1505 and Cosine Similarity of 0.5193. Ablation and human evaluations confirm the contributions of contextualized embeddings and multi-aspect selection. Although the current fixed-aspect setting limits flexibility, future work will explore adaptive aspect discovery, integration of external knowledge, and multimodal extensions. Overall, CEG-Sum advances the state of USP summarization by enabling more accurate, balanced, and informative summaries for real-world applications.

References

[1] Hongqiu Wu, Hai Zhao, and Min Zhang. Code summarization with structure-induced transformer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1078–1090. Association for Computational Linguistics, 2021.

[2] Zesheng Shi and Yucheng Zhou. Topic-selective graph network for topic-focused summarization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 247–259. Springer, 2023.

[3] Zesheng Shi, Tianhao Cao, and Xiangyue Zhang. Incorporating bert with naive bayes into neutral sentiment analysis. In *2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA)*, pages 782–785. IEEE, 2023.

[4] Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. Automated concatenation of embeddings for structured prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2643–2660. Association for Computational Linguistics, 2021.

[5] Yucheng Zhou, Jianbing Shen, and Yu Cheng. Weak to strong generalization for large language models with multi-capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025.

[6] Haowei Yang, Yu Tian, Zhongheng Yang, Zhao Wang, Chengrui Zhou, and Dannier Li. Research on model parallelism and data parallelism optimization methods in large language model—based recommendation systems. In *2025 7th International Conference on Artificial Intelligence Technologies and Applications (ICAITA)*, pages 324–329, 2025.

[7] Sun Maojun, Ruijian Han, Binyan Jiang, Houduo Qi, Defeng Sun, Yancheng Yuan, and Jian Huang and. Lambda: A large model based data agent. *Journal of the American Statistical Association*, 0(ja):1–20, 2025.

[8] Maojun Sun, Ruijian Han, Binyan Jiang, Houduo Qi, Defeng Sun, Yancheng Yuan, and Jian Huang. A survey on large language model-based agents for statistics and data science. *arXiv preprint arXiv:2412.14222*, 2024.

[9] Maojun Sun. Llamacare: A large medical language model for enhancing healthcare knowledge sharing. *arXiv preprint arXiv:2406.02350*, 2024.

[10] Xinyue Huang, Ziqi Lin, Fang Sun, Wenchao Zhang, Kejian Tong, and Yunbo Liu. Enhancing document-level question answering via multi-hop retrieval-augmented generation with llama 3. *arXiv preprint arXiv:2506.16037*, 2025.

[11] Zhongheng Yang, Aijia Sun, Yushang Zhao, Yinuo Yang, Dannier Li, and Chengrui Zhou. Rlhf fine-tuning of llms for alignment with implicit user feedback in conversational recommenders, 2025.

[12] Jing Xiong, Zixuan Li, Chuanyang Zheng, Zhijiang Guo, Yichun Yin, Enze Xie, Zhicheng Yang, Qingxing Cao, Haiming Wang, Xiongwei Han, et al. Dq-lore: Dual queries with low rank approximation re-ranking for in-context learning. *arXiv preprint arXiv:2310.02954*, 2023.

[13] Lianshang Cai, Linhao Zhang, Dehong Ma, Jun Fan, Daiting Shi, Yi Wu, Zhicong Cheng, Simiu Gu, and Dawei Yin. Pile: Pairwise iterative logits ensemble for multi-teacher labeled distillation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 587–595, 2022.

[14] Zhiwei Ma, Haojie Li, Zhihui Wang, Dan Yu, Tianyi Wang, Yingshuang Gu, Xin Fan, and Zhongxuan Luo. An underwater image semantic segmentation method focusing on boundaries and a real underwater scene semantic segmentation dataset. *arXiv preprint arXiv:2108.11727v1*, 2021.

[15] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829. Association for Computational Linguistics, 2021.

[16] Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5657–5668. Association for Computational Linguistics, 2022.

[17] Zesheng Shi, Yucheng Zhou, and Jing Li. Safety alignment via constrained knowledge unlearning. *arXiv preprint arXiv:2505.18588*, 2025.

[18] Darshan Solanki, Hsia-Ming Hsu, Olivia Zhao, Renyue Zhang, Weihao Bi, and Raman Kannan. The way we think about ourselves. In *Augmented Cognition. Theoretical and Technological Approaches: 14th International Conference, AC 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I*, page 276–285, Berlin, Heidelberg, 2020. Springer-Verlag.

[19] Yixin Liu and Pengfei Liu. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072. Association for Computational Linguistics, 2021.

[20] Jianing Zhou and Suma Bhat. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086. Association for Computational Linguistics, 2021.

[21] Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Saksham Singhal, Xian-Ling Mao, Heyan Huang, Xia Song, and Furu Wei. mT6: Multilingual pretrained text-to-text transformer with translation pairs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1671–1683. Association for Computational Linguistics, 2021.

[22] Xinyue Huang, Chen Zhao, Xiang Li, Chengwei Feng, and Wuyang Zhang. Gam-cot transformer: Hierarchical attention networks for anomaly detection in blockchain transactions. *INNO-PRESS: Journal of Emerging Applied AI*, 1(3), 2025.

[23] Jing Xiong, Chengming Li, Min Yang, Xiping Hu, and Bin Hu. Expression syntax information bottleneck for math word problems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2166–2171, 2022.

[24] Jing Xiong, Jianhao Shen, Ye Yuan, Haiming Wang, Yichun Yin, Zhengying Liu, Lin Li, Zhijiang Guo, Qingxing Cao, Yinya Huang, et al. Trigo: Benchmarking formal mathematical proof reduction for generative language models. *arXiv preprint arXiv:2310.10180*, 2023.

[25] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952. Association for Computational Linguistics, 2021.

[26] Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263. Association for Computational Linguistics, 2022.

[27] Zhang Cheng, Guyue Gui, Kejian Tong, Xinyue Huang, and Peiqing Lu. Finstack-net: Hierarchical feature crossing and stacked ensemble learning for financial fraud detection. 2025.

[28] Chang Yu, Fang Liu, Jie Zhu, Shaobo Guo, Yifan Gao, Zhongheng Yang, Meiwei Liu, and Qianwen Xing. Gradient boosting decision tree with lstm for investment prediction. In *2025 5th Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*, pages 57–62, 2025.

[29] Zhexue Chen, Hong Huang, Bang Liu, Xuanhua Shi, and Hai Jin. Semantic and syntactic enhanced aspect sentiment triplet extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1474–1483. Association for Computational Linguistics, 2021.

[30] Hongjie Cai, Rui Xia, and Jianfei Yu. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350. Association for Computational Linguistics, 2021.

[31] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891. Association for Computational Linguistics, 2022.

[32] Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941. Association for Computational Linguistics, 2021.

[33] Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487. Association for Computational Linguistics, 2021.

[34] Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 817–824. Association for Computational Linguistics, 2021.