Enhancing Aspect-Based Sentiment Analysis with Dynamic Few-Shot Prompting for Large Language Models

Mohammed Ziaulla
School of Computer Science and Engineering, REVA University, Bangalore, India
drmdziaulla@gmail.com (corresponding author)

Arun Biradar
School of Computer Science and Engineering, REVA University, Bangalore, India
arun.biradar@reva.edu.in

Received: 7 October 2025 | Revised: 10 November 2025 | Accepted: 21 November 2025

 DOI: https://doi.org/10.48084/etasr.14875

ABSTRACT

While Large Language Models (LLMs) have shown great promise, their effectiveness in few-shot learning settings is often limited by static prompting strategies, where a fixed set of examples may lack contextual relevance for diverse test cases. To address this limitation, this paper introduces a dynamic few-shot prompting methodology for Aspect-based Sentiment Analysis (ABSA) that leverages the Gemini Large Language Model (Gemini LLM). Our approach dynamically selects the most semantically pertinent examples from a training corpus for each individual test instance by computing cosine similarity between sentence embeddings. This ensures the LLM receives tailored, contextually rich guidance for every prediction. We evaluated our methodology on the benchmark SemEval-2014 datasets for the laptop and restaurant domains. The results demonstrate state-of-the-art performance, achieving F1-scores of 87.3% and 90.0%, respectively, significantly surpassing static few-shot prompting and other established baselines. The findings underscore the critical role of example pertinence in few-shot learning and illustrate that dynamic, context-aware prompting is a highly effective strategy for unlocking the full potential of LLMs on specialized Natural Language Processing (NLP) tasks without extensive model fine-tuning.

Keywords-Aspect-Based Sentiment Analysis (ABSA); Large Language Models (LLMs); few-shot learning; dynamic prompting; semantic similarity; prompt engineering; Gemini LLM

I. INTRODUCTION

Sentiment analysis, a core task in Natural Language Processing (NLP), has become a focal point of research due to its wide-ranging applications [1]. A key challenge within this field is Aspect-based Sentiment Analysis (ABSA), which aims for a fine-grained understanding of user opinions by identifying sentiments towards specific entities or their attributes [2]. For instance, in the review "the interface is intuitive, but the battery life is disappointing," an effective ABSA system must extract the distinct aspect-sentiment pairs: (interface, positive) and (battery life, negative). While traditional machine learning methods have advanced the field, they often rely on extensive, task-specific annotated data and can struggle to capture nuanced linguistic contexts.

The advent of Large Language Models (LLMs), such as the Gemini Large Language Model (Gemini LLM), has introduced a new paradigm for NLP tasks, demonstrating remarkable capabilities in text comprehension and generation. Few-shot learning has emerged as a particularly effective paradigm for leveraging these models, enabling them to perform specialized tasks with only a handful of examples and without costly fine-tuning [3-5]. However, the application of few-shot learning in ABSA often relies on a static set of examples, which are used uniformly for every test case [5]. This approach can be suboptimal, as a fixed set of examples may lack the contextual relevance needed to analyze the unique aspects of a specific input sentence.

To address this limitation, we introduce a dynamic few-shot prompting methodology for ABSA that leverages the power of the Gemini LLM. In contrast to static methods, our approach dynamically selects the most semantically relevant examples from a training corpus for each individual test sentence. This tailored guidance ensures the model receives highly contextual and pertinent information, significantly enhancing its ability to accurately identify aspect-sentiment pairs. This work thus presents a novel dynamic few-shot prompting framework for ABSA that adaptively selects semantically relevant examples for each input instance, offering a significant improvement over conventional static prompting methods. We validate our approach on the widely recognized SemEval-2014 datasets for the laptop and restaurant domains.

The primary contributions of this work are:

- We propose a novel dynamic few-shot prompting methodology for ABSA that tailors contextual examples to each test instance through semantic similarity, significantly improving model guidance over static approaches.
- We demonstrate the superiority of our dynamic method through extensive experiments, establishing a new state-of-the-art performance on the SemEval-2014 laptop and restaurant datasets.
- We provide a comprehensive analysis of the key factors driving our model's success, including the impact of example selection strategy and prompt design, offering valuable insights for future research in prompt-based learning.

II. RELATED WORK

This section provides a systematic synthesis of prior research in three key areas: ABSA, the paradigm of few-shot learning in NLP, and the application of LLMs to sentiment analysis tasks.

A. The Evolution of Aspect-Based Sentiment Analysis

ABSA, formally benchmarked in the SemEval-2014 Task 4 [6], has evolved significantly over the past decade. The field has progressed through rule-based, traditional machine learning, and deep learning paradigms [7]. Based on the evolution of ABSA, early approaches relied on conventional machine learning models like Support Vector Machine (SVM) and Random Forest (RF), whereas recent advancements incorporate ensemble learning and deep learning techniques to improve accuracy and handle complex linguistic patterns. However, challenges remain in areas such as sarcasm detection, domain generalization, and capturing nuanced emotions [8]. Early approaches were dominated by rule-based methods that relied on linguistic heuristics and predefined lexicons [9], followed by machine learning models that required extensive feature engineering [10, 11]. While interpretable, these methods often struggled with the linguistic complexity inherent in user reviews.

The advent of deep learning brought substantial advancements. Architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, often enhanced with attention mechanisms, became standard for modeling the sequential nature of text in ABSA [12, 13]. More recently, pre-trained transformer models like BERT have set new state-of-the-art benchmarks [14, 15]. However, a common limitation across these supervised deep learning methods is their profound dependence on large, task-specific annotated datasets for fine-tuning, which are often costly and time-consuming to create.

B. Few-Shot Learning as a Data-Efficient Paradigm

Few-shot learning has emerged as a compelling solution to the data dependency challenge in NLP [16, 17]. The goal of this paradigm is to enable models to generalize to new tasks from just a handful of examples. While early work focused on meta-learning and transfer learning [18, 19], the recent rise of LLMs has brought prompt-based, in-context learning to the forefront [16-20].

As demonstrated by authors in [21], with GPT-3, LLMs can perform a new task by being conditioned on a few examples provided directly in the input prompt, without any updates to the model's parameters. Subsequent research has shown that the effectiveness of in-context learning is highly sensitive to the choice and format of these few-shot examples [22, 23]. This has underscored the importance of example selection as a critical component for optimizing LLM performance in few-shot settings.

C. Large Language Models for Aspect-Based Sentiment Analysis and the Need for Dynamic Prompting

LLMs have demonstrated remarkable capabilities in general sentiment analysis [5], with models like Gemini LLM achieving strong performance in zero-shot and few-shot settings [24, 25]. Recent studies have begun to explore their application to the more granular task of ABSA, confirming that LLMs can effectively identify aspect terms and their associated sentiments [26, 27].

However, much of the current literature on few-shot ABSA employs static prompting methodologies. In this approach, a single, fixed set of examples is used for all test instances, regardless of their specific content or context [28]. This "one-size-fits-all" strategy can be suboptimal. While dynamic prompting, which tailors examples to each specific input, has shown promise in other NLP tasks [29], its application to ABSA with LLMs remains largely unexplored.

Moreover, reasoning-enhanced prompting methodologies, such as Chain-of-Thought (CoT) prompting, have garnered heightened interest for their capacity to boost model interpretability and efficacy in intricate NLP tasks [30]. Authors in [31] provide a comprehensive survey of advanced ABSA tasks, emphasizing aspect-opinion-sentiment quadruple extraction, implicit aspect reasoning, and the growing role of LLMs like ChatGPT in fine-grained sentiment analysis [31]. Authors in [32] propose a CoT-guided few-shot fine-tuning framework for LLMs that improves multimodal (text + image) aspect-level sentiment classification by encouraging intermediate reasoning and better cross-modal alignment under scarce-data settings. Authors in [33] introduce a syntactic-guided CoT framework that iteratively detects both implicit and explicit targets in ABSA by leveraging dependency structures and reasoning steps.

Our work addresses this critical gap. We hypothesize that a dynamic few-shot methodology, which selects the most semantically relevant examples for each test instance, can provide more precise, context-aware guidance to the LLM. This approach is designed to enhance the model's ability to perform accurate and nuanced ABSA, moving beyond the limitations of static prompting.

III. METHODOLOGY

Our methodology for ABSA utilizes the functionalities of the Gemini LLM via a dynamic few-shot learning framework. This section elucidates our approach, encompassing the dynamic selection of few-shot instances, prompt formulation, and the comprehensive pipeline for processing and scrutinizing sentences from the SemEval-2014 datasets.

A. Dynamic Few-Shot Instance Selection

Contrary to static prompting methodologies that employ the identical set of instances for all evaluation cases, our dynamic selection mechanism discerns the most pertinent instances for each specific test sentence. This procedure guarantees that the model receives contextually suitable direction for each distinct case, enhancing its capacity to precisely identify aspects and their corresponding sentiments.

The selection process involves the following steps:

- Embedding generation: We initially produce comprehensive vector representations for all statements in the training corpus utilizing a pre-trained sentence encoder. These embeddings encapsulate the semantic essence of each statement, facilitating similarity-oriented retrieval.
- Similarity computation: For each evaluative sentence, we ascertain its semantic resemblance with all sentences in the training corpus utilizing cosine similarity between their corresponding embeddings.
- Example selection: We designate the foremost k most similar sentences from the training corpus as few-shot instances. The magnitude of k is ascertained empirically through evaluative experiments, reconciling the necessity for heterogeneous instances against prompt length limitations.

```
Algorithm 1. Dynamic Few-Shot Instance
             Selection
Input:
    T = {t_1, t_2, ..., t_n} // Training
        sentences
    E = {e_1, e_2, ..., e_n} // Embeddings of
        training sentences
    s = test sentence
    encoder = pre-trained sentence
              embedding model
    k = number of few-shot examples
Output:
    S = selected few-shot examples for s
Procedure:
1.  e_s ← encoder(s) // Generate embedding
          for test sentence
2.  For each e_i ∈ E do
      sim_i ← cosine_similarity(e_s, e_i) //
      Compute similarity between s and
      training sentences
3.  Sort all training sentences T by sim_i
    in descending order
4.  Select top-k sentences with highest
    similarity scores
5.  Return S = {t_1', t_2', ..., t_k'} as few-
    shot examples
```

Algorithm 1 dynamically retrieves the k most semantically related sentences from the training corpus for each test instance. These examples serve as adaptive contextual guidance to the LLM, ensuring that the few-shot context aligns with the semantic domain and sentiment orientation of the target sentence.

This dynamic selection methodology guarantees that each evaluative sentence is scrutinized within the framework of the most pertinent instances, furnishing the model with more enlightening direction than fixed prompting techniques.

B. Prompt Construction

Effective prompt formulation is imperative for directing the LLM to execute ABSA tasks precisely. Our prompt construction methodology adheres to a systematic framework that encompasses:

- Task description: A clear description of the ABSA task, stating that the model should identify aspect terms and their corresponding sentiments.
- Format specification: Unambiguous directives regarding the anticipated output format, guaranteeing uniform and interpretable responses.
- Few-shot examples: The dynamically selected instances from the training corpus, each presented with its ground-truth annotations to illustrate the anticipated analysis.
- Test instance: The sentence to be analyzed, presented as a query at the end of the prompt.

The overall structure of our prompt is as follows:

Task: Examine the subsequent propositions to discern aspect terms and their corresponding sentiments. For each aspect term, categorize the sentiment as positive, negative, or neutral. Format your response as: (aspect term, sentiment)

1) Examples

- Proposition: "The battery life is exceptional but the screen resolution is unsatisfactory." Analysis: (battery life, positive), (screen resolution, negative)
- Proposition: "The cuisine was exquisite and the service was expedient." Analysis: (cuisine, positive), (service, positive)

C. Processing Pipeline

Our complete ABSA pipeline consists of the following components, as illustrated in Figure 1:

- Data preparation: Preparing the SemEval-2014 corpora (Offline: Corpus Preprocessing & Embedding).
- Example selection: Dynamically selecting relevant examples for each sentence (Dynamic Example Selection).
- Prompt generation: Formulating a prompt with the selected examples (Prompt Construction).
- LLM inference: Submitting the prompt to the LLM (LLM Inference).
- Response parsing: Analyzing the model's output (Parse & Extract Results).

- Output storage: Archiving the results in a CSV file (Store Output (CSV)).

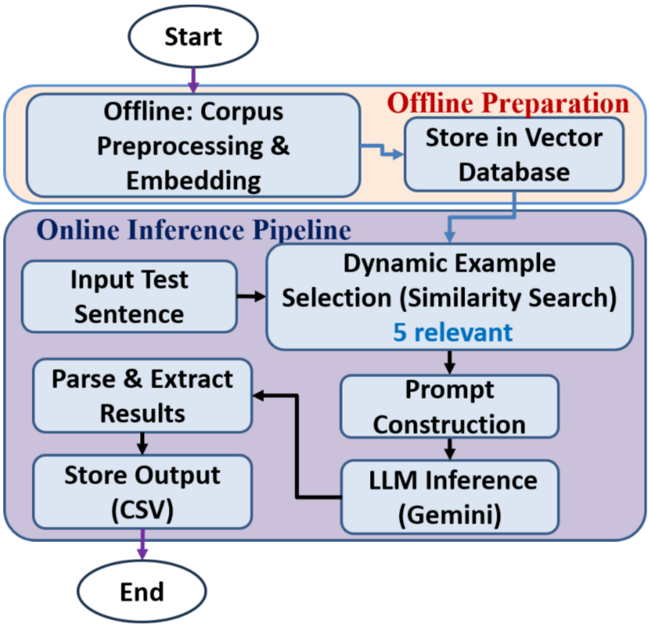

Fig. 1. Comprehensive framework of the dynamic few-shot ABSA methodology.

D. Implementation Details

Our research study execution employs the Gemini Pro model, accessed through the Google AI Studio API. For embedding generation, we utilize a pre-trained sentence transformer model (all-MiniLM-L6-v2) that yields 384-dimensional dense vector representations. The quantity of few-shot examples (k) is established at 5 based on validation experiments, balancing efficacy against prompt length constraints.

The similarity computation and example selection processes are executed efficiently utilizing vector operations, facilitating rapid processing of test sentences. The entire pipeline is implemented in Python, with the response parsing component employing regular expressions to extract aspect-sentiment pairs from the model's output.

To assure reproducibility, we designate a fixed random seed (42) for all randomized components and uphold consistent prompt formatting across all experiments. The complete implementation, including code for dynamic example selection, prompt construction, and response parsing, will be made accessible in a public repository upon publication.

E. Datasets

In this investigation, we employ the SemEval-2014 Task 4 datasets, which have emerged as conventional reference points for ABSA. These datasets encompass annotated reviews from two disparate domains: laptops and restaurants.

The datasets are partitioned into training and testing subsets with annotations supplied by human specialists. Table I presents the statistics of these datasets.

TABLE I. STATISTICS OF THE SEMEVAL-2014 TASK 4 DATASETS

Domain	Split	Sentences	Aspect terms	Positive	Negative	Neutral
Laptop	Training	3,045	2,358	994	870	494
Laptop	Testing	800	654	341	128	185
Restaurant	Training	3,041	3,693	2,164	807	722
Restaurant	Testing	800	1,134	728	196	210

We execute numerous preprocessing procedures to prepare the datasets for our dynamic few-shot methodology, including sentence extraction, aspect-sentiment pair extraction, embedding generation for all training sentences, and splitting the training set to create a validation set for hyperparameter tuning.

These datasets present several challenges that render them suitable for evaluating our methodology, including domain specificity, implicit aspects, sentiment ambiguity, multiple aspects per sentence, and general linguistic complexity (e.g., negation, sarcasm).

F. Experiments

This section elucidates our empirical configuration, assessment criteria, and comparative techniques employed to evaluate the efficacy of our dynamic few-shot methodology.

All empirical investigations were executed utilizing the Gemini Pro model. We instituted our dynamic few-shot selection mechanism and processing framework in Python 3.9 on a workstation outfitted with an NVIDIA RTX 3090 GPU. The number of few-shot examples (k) was established at 5 based on a grid search on the validation set. Temperature and top-p sampling parameters for the LLM were set to 0.3 and 0.95, respectively.

Performance was evaluated using standard metrics:

- Precision: The ratio of accurately identified aspect-sentiment pairs to the total number of identified pairs:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{1}$$

- Recall: The ratio of accurately identified aspect-sentiment pairs to the total number of actual pairs:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2}$$

- F1-score: The harmonic mean of precision and recall, offering a balanced evaluation of performance:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

- Accuracy: The proportion of accurately identified aspect-sentiment pairs among all predictions:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

An aspect-sentiment pair is considered correct only if both the aspect term and its sentiment polarity exactly match the ground truth.

We evaluate our dynamic few-shot methodology with several benchmark and cutting-edge techniques:

- Zero-shot Gemini: The LLM prompted without any examples.
- Static few-shot Gemini: The LLM prompted with a fixed, random set of examples.
- BERT-based pipeline: A fine-tuned BERT pipeline for aspect extraction and sentiment classification.
- SpanABSA: A span-based, unified tagging framework.
- BART-ABSA: A sequence-to-sequence generation framework.
- GPT-4 with static prompting: GPT-4 with static few-shot examples.

To comprehend the contribution of various components in our methodology, we conduct several ablation analyses:

- Impact of example selection strategy: Comparing our similarity-based selection with random and diversity-based selection.
- Number of few-shot examples: Manipulating the quantity of examples (k) from 1 to 10.
- Prompt format variations: Experimenting with diverse prompt architectures and instructional phrasings.
- Domain transfer: Examining cross-domain generalization by using examples from one domain to test on another.

IV. RESULTS

This section elucidates the empirical outcomes of our dynamic few-shot methodology for ABSA utilizing the Gemini LLM. Following the presentation of quantitative results, a comprehensive discussion interprets these findings, contextualizes them within the broader research landscape, and outlines their theoretical and practical implications.

To compare our dynamic few-shot methodology with established baseline techniques, a performance evaluation was conducted on the SemEval-2014 datasets for both the laptop and restaurant domains. Table II illustrates these findings, with results reported in terms of Precision, Recall, and F1-score. The results presented in the table indicate that our dynamic few-shot methodology consistently surpasses all comparative techniques across both domains. In the laptop domain, our methodology attains an F1-score of 87.3%, representing a 2.6% enhancement over static few-shot prompting with Gemini. In the more complex restaurant domain, our methodology achieves an F1-score of 90.0%, exceeding the static few-shot Gemini by 2.3%.

Further statistical tests revealed that the performance superiority of our methodology is more pronounced in Precision than in Recall. This indicates that the dynamic selection of contextually pertinent examples is particularly effective at helping the model avoid false positives, which is especially advantageous in practical applications where precision is prioritized.

TABLE II. PERFORMANCE COMPARISON OF DYNAMIC FEW-SHOT GEMINI WITH BASELINE METHODS ON SEMEVAL-2014 DATASETS

Method	Laptop domain			Restaurant domain		
	Precision (%)	Recall (%)	F1-score (%)	Precision (%)	Recall (%)	F1-score (%)
BERT-based pipeline	78.2	76.9	77.5	82.4	81.6	82.0
SpanABSA	80.1	79.3	79.7	84.5	83.7	84.1
BART-ABSA	81.6	80.8	81.2	85.9	84.6	85.2
GPT-4 with static prompting	84.3	83.5	83.9	87.2	86.4	86.8
Zero-shot Gemini	79.8	77.2	78.5	83.1	80.9	82.0
Static few-shot Gemini	85.1	84.3	84.7	88.4	87.1	87.7
Dynamic few-shot Gemini (ours)	87.9	86.8	87.3	90.6	89.5	90.0

A. *Ablation Studies*

To assess the contribution of individual components within our framework, several ablation analyses were conducted. These experiments investigate the impact of the example selection strategy, the number of few-shot examples, prompt formatting, and domain transferability.

The purpose of this experiment was to validate our core hypothesis that semantic pertinence is crucial for effective guidance. A comparison of the two results in Table III reveals the performance difference between our similarity-based selection strategy and alternative methods.

TABLE III. F1-SCORE PERFORMANCE ACROSS DIFFERENT FEW-SHOT EXAMPLE SELECTION STRATEGIES

Selection strategy	Laptop domain	Restaurant domain
Random Selection	84.5	87.6
Diversity-based selection	85.2	88.3
Similarity-based selection (ours)	87.3	90.0

The most obvious finding to emerge from the analysis is that similarity-oriented selection consistently exceeds both stochastic and diversity-oriented methodologies. This result substantiates our hypothesis that supplying the model with semantically analogous instances augments its capacity to discern elements with higher precision.

The next question addressed how the quantity of few-shot instances (k) influences performance. As depicted in Figure 2, performance initially improves as k increases, reaching a peak at k=5 for both domains. Subsequent to this threshold, performance diminishes marginally. This observation corroborates our selection of k=5 for the primary experiments, achieving an optimal compromise between performance and prompt length.

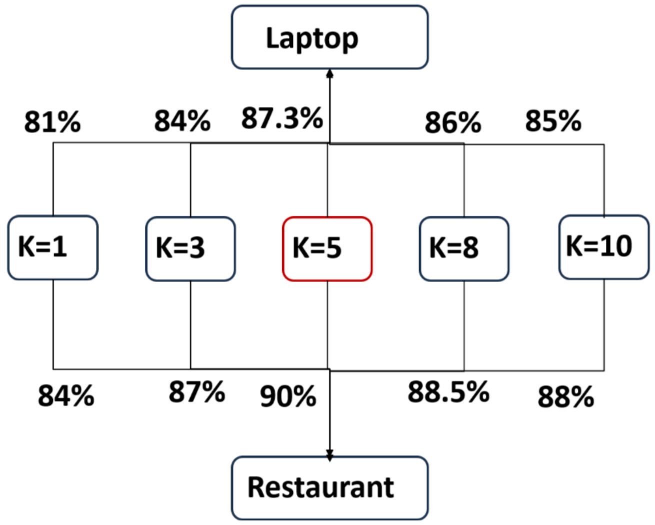

Fig. 2. Influence of the number of few-shot examples (k) on model F1-score across both domains.

To assess the impact of instructional clarity, diverse prompt formats were employed. Table IV presents the findings.

TABLE IV. IMPACT OF PROMPT FORMAT VARIATIONS ON MODEL F1-SCORE PERFORMANCE

Prompt format	Laptop domain	Restaurant domain
Basic instructions	85.1	87.9
Detailed instructions	86.4	89.2
CoT prompting	86.8	89.5
Structured output format	87.3	90.0

The results indicate that supplying an organized output format produces the best performance. This implies that explicit guidance regarding the anticipated response format assists the model in producing more precise and consistent outputs.

A cross-domain investigation was conducted to evaluate the model's generalization capabilities. Table V presents the findings.

TABLE V. EVALUATION OF CROSS-DOMAIN GENERALIZATION PERFORMANCE, MEASURED BY F1-SCORE

Training domain	Testing domain	F1-score (%)
Laptop	Laptop	87.3
Restaurant	Laptop	82.1
Restaurant	Restaurant	90.0
Laptop	Restaurant	84.5

The results show a notable degradation in performance when the training and testing domains differ, reinforcing the importance of our in-domain, dynamic example selection strategy.

B. Error Analysis

A comprehensive examination of the inaccuracies produced by our methodology was executed to identify prevalent failure patterns. Figure 3 illustrates the primary categories of error and their distribution.

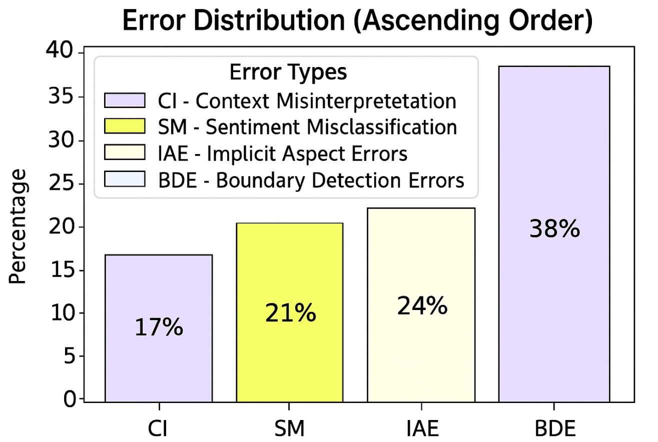

Fig. 3. Distribution of primary error categories in the dynamic few-shot ABSA methodology.

The error analysis reveals four primary categories: boundary detection errors (38%), implicit aspect errors (24%), sentiment misclassification (21%), and context misinterpretation (17%). The analysis indicates that boundary detection persists as the most challenging aspect of ABSA, indicating that future work should focus on enhancing the model's ability to accurately delineate aspect term boundaries.

V. DISCUSSION

This study found that dynamic few-shot prompting is highly effective for ABSA. One interesting finding is the notable enhancement in performance over static prompting methodologies, which underscores the significance of example pertinence in steering the model's behavior. By providing contextually suitable examples, our methodology facilitates the model's comprehension of the specific nuances relevant to the analytical task.

The strong performance of our methodology can be attributed to several factors:

- Contextual pertinence: Dynamic selection guarantees that examples are semantically analogous to the test instance, thereby providing more relevant guidance.
- Domain adaptation: By opting for examples based on similarity, our methodology implicitly adjusts to domain-specific vocabulary and patterns.
- Balanced guidance: The selection of multiple examples (k=5) delivers diverse yet pertinent guidance, assisting the model address various facets and sentiment expressions.
- Structured output: Unambiguous directives regarding the expected output format improve the consistency and interpretability of the model's responses.

These principles can be illustrated through concrete examples of how semantic similarity guides the model in practice:

- For the test sentence "Although the laptop runs fast, it tends to overheat after long use," the selection mechanism retrieves training examples discussing performance and temperature issues (e.g., "The processor is quick but the device gets hot"). This contextual alignment helps the LLM infer that "overheat" implies a negative sentiment, even though the sentiment word "bad" is not explicitly present.

- For implicit aspects such as "The price could have been better," examples involving cost-related expressions (e.g., "too expensive for its quality") provide semantic cues that guide the LLM to interpret the implicit sentiment as negative toward the price aspect.
- By leveraging such contextually aligned examples, the dynamic few-shot mechanism enhances the model's reasoning capability, yielding improved recognition of subtle or implicit aspect-sentiment associations.

1) Limitations and Failure Modes of Similarity-Based Selection

While dynamic selection significantly improves performance, the similarity function is not infallible. A notable failure occurs when top-k retrieved sentences, despite high semantic similarity, focus on different aspects than the test sentence. For example:

- Test sentence: "The battery life is excellent but the screen resolution is disappointing" (target aspects: battery life, screen resolution).
- Retrieved example: "Overall performance is outstanding and build quality is solid" (aspects: overall performance, build quality).

In such cases, the model receives syntactically and topically relevant guidance (contrastive "but" structure, product domain vocabulary) yet lacks specific precedent for the exact aspect terms. Our error analysis suggests that this contributes to the 24% implicit aspect errors and 38% boundary detection errors (Figure 3). When retrieved examples discuss semantically related but non-identical aspects, the model may:

1. Over-generalize: Extract overly broad terms (e.g., "performance" instead of "battery life").
2. Miss nuance: Fail to recognize domain-specific aspect boundaries (e.g., "screen resolution" vs. "display").

2) Mitigation and Robustness

Our use of k=5 examples partially alleviates this issue through implicit diversity: even if 2–3 examples exhibit aspect mismatch, the remaining examples often provide corrective signals. However, we acknowledge this remains a limitation. Future work could explore aspect-aware similarity functions that weight embeddings by aspect-term overlap or incorporate diversity-based re-ranking to ensure retrieved examples cover distinct aspect categories. This represents an important direction for making dynamic prompting more robust to sparse or mismatched training corpora.

These findings have broader implications for few-shot learning with LLMs. Nevertheless, our methodology is not without constraints. The dependence on a pre-existing training corpus may limit applicability in domains with scarce labeled data. Furthermore, the computational burden introduces latency that may pose challenges for real-time applications. Future investigations should address these limitations through the development of more efficient similarity computation techniques and the exploration of semi-supervised methodologies to mitigate reliance on labeled examples.

VI. CONCLUSION

This study investigated the efficacy of a dynamic few-shot methodology for Aspect-based Sentiment Analysis (ABSA) utilizing the Gemini Large Language Model (Gemini LLM). Returning to the question posed at the beginning of this study, it is now possible to state that by dynamically selecting the most semantically relevant examples for each test instance, our approach provides the model with more effective contextual guidance. This methodology results in enhanced performance when compared to static prompting strategies and conventional fine-tuning techniques.

The main research findings show that experiments on the SemEval-2014 datasets validate the success of our methodology. Our approach achieves state-of-the-art F1-scores of 87.3% for the laptop domain and 90.0% for the restaurant domain. Ablation studies further confirm that similarity-driven example selection, an optimal number of examples, and a structured output format are critical factors in maximizing performance. The error analysis identified that aspect boundary detection persists as the most challenging aspect of ABSA, highlighting a clear direction for future improvements.

The results suggest that Large Language Models (LLMs) like Gemini can perform nuanced sentiment analysis tasks with high precision when provided with suitable, targeted guidance. An important implication is that dynamic example selection can serve as an effective strategy for adapting LLMs to domain-specific terminology and patterns without the need for extensive fine-tuning.

However, several limitations should be noted. First, the methodology's reliance on a pre-existing, annotated training corpus may limit its applicability in domains characterized by limited labeled data. Second, the computational overhead associated with embedding generation and real-time similarity search introduces latency, which could pose challenges to applications requiring immediate responses.

These findings suggest several promising directions for future research. One promising avenue involves amalgamating our dynamic few-shot approach with retrieval-augmented generation methodologies, which could further enhance performance by providing the model with additional relevant information. Furthermore, enhancing cross-domain generalization through domain adaptation or meta-learning techniques would broaden the applicability of the approach to novel domains with scarce labeled data. The scope of this research could also be expanded by extending the methodology to multi-modal data, such as reviews accompanied by images, and adapting it for multilingual contexts to serve diverse global markets. To bolster trust and facilitate real-world deployment, future work will also concentrate on enhancing the explainability of the model's determinations, potentially through rationale generation. Finally, mitigating the computational burden of embedding generation and similarity computation through efficiency optimizations, such as model distillation or approximate similarity search, will be critical for enabling real-time applications.

REFERENCES

[1] K. Schouten and F. Frasincar, "Survey on Aspect-Level Sentiment Analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 813–830, Mar. 2016, https://doi.org/10.1109/TKDE.2015.2485209.

[2] M. Pontiki *et al.*, "SemEval-2016 Task 5: Aspect Based Sentiment Analysis," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California, 2016, pp. 19–30, https://doi.org/10.18653/v1/S16-1002.

[3] A. Chowdhery *et al.*, "PaLM: scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 1, Jan. 2023, Art. no.240.

[4] R. Anil *et al.*, "Gemini: A Family of Highly Capable Multimodal Models." arXiv, May 09, 2025, https://doi.org/10.48550/arXiv.2312.11805.

[5] T. B. Brown *et al.*, "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2020, pp. 1877–1901.

[6] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 Task 4: Aspect Based Sentiment Analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 2014, pp. 27–35, https://doi.org/10.3115/v1/S14-2004.

[7] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, https://doi.org/10.1016/j.asej.2014.04.011.

[8] J. Mir, A. Mahmood, and S. Khatoon, "Aspect Based Classification Model for Social Reviews," *Engineering, Technology & Applied Science Research*, vol. 7, no. 6, pp. 2296–2302, Dec. 2017, https://doi.org/10.48084/etasr.1578.

[9] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, WA, USA, 2004, pp. 168–177, https://doi.org/10.1145/1014052.1014073.

[10] S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad, "NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 2014, pp. 437–442, https://doi.org/10.3115/v1/S14-2076.

[11] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter Sentiment Classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, OR, USA, 2011, pp. 151–160.

[12] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent Attention Network on Memory for Aspect Sentiment Analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017, pp. 452–461, https://doi.org/10.18653/v1/D17-1047.

[13] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for Aspect-level Sentiment Classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, TX, USA, 2016, pp. 606–615, https://doi.org/10.18653/v1/D16-1058.

[14] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for Target-Dependent Sentiment Classification," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, 2016, pp. 3298–3307.

[15] Y. Ma, H. Peng, T. Khan, E. Cambria, and A. Hussain, "Sentic LSTM: a Hybrid Network for Targeted Aspect-Based Sentiment Analysis," *Cognitive Computation*, vol. 10, no. 4, pp. 639–650, Aug. 2018, https://doi.org/10.1007/s12559-018-9549-x.

[16] H. Xu, B. Liu, L. Shu, and P. Yu, "BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis," 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)*, Minneapolis, MN, USA, 2019, pp. 2324–2335, https://doi.org/10.18653/v1/N19-1242.

[17] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence," 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)*, Minneapolis, MN, USA, 2019, pp. 380–385, https://doi.org/10.18653/v1/N19-1035.

[18] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 4080–4090.

[19] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a Few Examples: A Survey on Few-shot Learning," *ACM Computing Surveys*, vol. 53, no. 3, June 2020, Art. no. 63, https://doi.org/10.1145/3386252.

[20] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, Sydney, Australia, 2017, pp. 1126–1135.

[21] S. Ruder, "Neural Transfer Learning for Natural Language Processing," Ph.D. dissertation, School of Engineering and Informatics, College of Engineering and Informatics, National University of Ireland, Galway, Ireland, 2019.

[22] T. Gao, A. Fisch, and D. Chen, "Making Pre-trained Language Models Better Few-shot Learners," 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)*, Online, 2021, pp. 3816–3830, https://doi.org/10.18653/v1/2021.acl-long.295.

[23] B. Lester, R. Al-Rfou, and N. Constant, "The Power of Scale for Parameter-Efficient Prompt Tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, 2021, pp. 3045–3059, https://doi.org/10.18653/v1/2021.emnlp-main.243.

[24] H. W. Chung *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 1, Jan. 2024, Art. no.70.

[25] X. L. Li and P. Liang, "Prefix-Tuning: Optimizing Continuous Prompts for Generation," 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)*, Online, 2021, pp. 4582–4597, https://doi.org/10.18653/v1/2021.acl-long.353.

[26] Z. Cui, X. Shi, and Y. Chen, "Sentiment analysis via integrating distributed representations of variable-length word sequence," *Neurocomputing*, vol. 187, pp. 126–132, Apr. 2016, https://doi.org/10.1016/j.neucom.2015.07.129.

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," *OpenAI blog*, vol. 1, no. 8, 2019, Art. no. 9.

[28] S. Narayan, C. Gardent, S. B. Cohen, and A. Shimorina, "Split and Rephrase," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017, pp. 606–616, https://doi.org/10.18653/v1/D17-1064.

[29] Z. Li, Y. Wei, Y. Zhang, X. Zhang, and X. Li, "Exploiting Coarse-to-Fine Task Transfer for Aspect-Level Sentiment Classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 4253–4260, July 2019, https://doi.org/10.1609/aaai.v33i01.33014253.

[30] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 2022, pp. 24824–24837.

[31] H. Zhang, Y.-N. Cheah, O. M. Alyasiri, and J. An, "Exploring aspect-based sentiment quadruple extraction with implicit aspects, opinions, and ChatGPT: a comprehensive survey," *Artificial Intelligence Review*,

vol. 57, no. 2, Jan. 2024, Art. no. 17, https://doi.org/10.1007/s10462-023-10633-x.

[32] H. Wu, D. Yang, P. Liu, and X. Li, "Chain of Thought Guided Few-Shot Fine-Tuning of LLMs for Multimodal Aspect-Based Sentiment Classification," in *31st International Conference on Multimedia Modeling*, Nara, Japan, 2025, pp. 182–194, https://doi.org/10.1007/978-981-96-2054-8_14.

[33] M. Radi, N. Omar, and W. Kaur, "Syntactic-Guided Chain of Thought for Iterative Implicit and Explicit Target Detection in Aspect-Based Sentiment Analysis," *IEEE Access*, vol. 13, pp. 84738–84751, 2025, https://doi.org/10.1109/ACCESS.2025.3568695.