From Context to Aspects: LLM Based Implicit Aspect Extraction with Paraphrased Input and Knowledge Graph Support

Lujain Alawwad * and Mohamed El Bachir Menai

Posted Date: 19 May 2026

doi: 10.20944/preprints202605.1240.v1

Keywords: implicit aspect extraction; transformers architecture; transfer learning; text paraphrasing; knowledge graphs

 doi:10.20944/preprints202605.1240.v1

Article

From Context to Aspects: LLM Based Implicit Aspect Extraction with Paraphrased Input and Knowledge Graph Support

Lujain Alawwad [1,2,*] and Mohamed El Bachir Menai [1]

[1] Department of Computer Science, King Saud University, Riyadh 11451, Saudi Arabia
[2] Department of Computing and Informatics, Saudi Electronic University, Riyadh 13323, Saudi Arabia
* Correspondence: L.alawwad@seu.edu.sa

Abstract

While aspect-based sentiment analysis (ABSA) has gained significant progress in the identification of explicit opinion targets, the more challenging case, implicit aspects, has not been sufficiently studied. Implicit aspect extraction is particularly challenging as it relies on contextual and semantic cues and requires systems to infer what reviewers mean rather than just say. In this paper, we propose a four-component hybrid solution for explicit and implicit aspect extraction that formulates aspect extraction as a controlled text generation task. The solution combines: (i) a fine-tuned decoder-only large language model as a generative baseline, (ii) an iterative residual generation strategy that recovers multiple aspects through successive regeneration passes, (iii) paraphrase-based input transformation to broaden the contextual signal, and (iv) domain-specific knowledge graphs activated by linguistic signals to infer implicit aspects. The novelty is not in the individual components themselves, but in the principled orchestration of these components and the gating logic for when each stage is activated. Extensive experiments are conducted on eight benchmark ABSA datasets in both English and Arabic including SemEval 2014, 2015, 2016, ACOS and M-ABSA for English and SemEval 2016, HAAD, and M-ABSA for Arabic. The proposed solution consistently outperforms strong baseline methods and recent state-of-the-art models on English datasets with F1-scores of 0.8533, 0.713, 0.7859, 0.793 and 0.664 respectively, and F1-scores of 0.7336, 0.4765 and 0.7601 on Arabic datasets respectively. These results demonstrate the effectiveness of generative modeling, iterative generation, paraphrasing and structured knowledge for aspect extraction, and the potential of the proposed approach for implicit aspect identification in particular for morphologically rich and low-resource languages such as Arabic.

Keywords: implicit aspect extraction; transformers architecture; transfer learning; text paraphrasing; knowledge graphs

1. Introduction

When a reviewer writes "the laptop feels like carrying a brick," they are not describing the laptop's material composition — they are complaining about its *weight*, an aspect never mentioned by name. Humans decode such expressions effortlessly through pragmatic inference and world knowledge; automated systems, for the most part, do not. This gap sits at the heart of opinion mining.

The volume of opinion-bearing text produced online — reviews, ratings, and commentary on products, services, and public issues — represents an extraordinarily rich source of behavioral insight for researchers and practitioners alike. Sentiment analysis (SA) was developed to exploit this resource systematically, with the goal of modeling the subjective orientation expressed in natural language text. Document- and sentence-level SA methods, however, assign a single polarity to an entire text, obscuring cases where a reviewer simultaneously praises one feature and criticizes another. Aspect-based sentiment analysis (ABSA) addresses this limitation by decomposing opinion into its constituent tar-

 doi:10.20944/preprints202605.1240.v1

gets: discrete product or service features — such as battery life, performance, food quality, or service — each carrying its own sentiment value.

ABSA subsumes two interdependent sub-tasks: aspect extraction, which identifies the features being evaluated, and aspect sentiment classification, which assigns a polarity label to each identified feature. Within aspect extraction, the distinction between explicit and implicit aspects is of particular consequence. Explicit aspects appear verbatim in the text, as in "The *screen* is bright." Implicit aspects, by contrast, must be inferred from contextual and pragmatic cues: "The laptop is heavy" implicates *weight* without naming it. Models lack the world knowledge and pragmatic reasoning this inference requires, making implicit aspect extraction a persistently difficult problem — and one that is too pervasive in authentic reviews to ignore.

The research community has formalized interest in aspect extraction through several benchmark shared tasks, most notably SemEval-2014 Task 4 [1], SemEval-2015 Task 12 [2], and SemEval-2016 Task 5 [3], which established widely adopted evaluation protocols. Complementing these, a number of datasets supporting the aspect term extraction (ATE) sub-task have been introduced, including ACOS [4] and the recent multilingual benchmark M-ABSA [5]. Methodological approaches have evolved in parallel: early work relied on statistical and feature-based machine learning [1–5], before the field shifted toward deep learning [6–10]. More recently, large language models (LLMs) have demonstrated substantial gains across NLP tasks and are increasingly being evaluated for ABSA [11–13].

Several challenges, however, remain unresolved. First, a single sentence may implicate multiple aspects simultaneously: *"The place was loud and overpriced"* conveys opinions about both *ambience* and *price* without naming either. Second, the same opinion word may index different aspects depending on domain: *"fast"* implies *service* in a restaurant context but *performance* in a laptop review. Third, annotated corpora that label implicit aspects explicitly remain scarce — particularly for Arabic — constraining both the availability of supervision and the scope of evaluation. These challenges motivate the research questions addressed in this work.

The contributions of this work are as follows:

- We propose an *iterative residual generation* strategy for generative aspect extraction, in which the model performs successive generation passes over a progressively masked input, enabling recovery of aspects missed in single-pass decoding. A paraphrase-based input strategy further broadens the contextual signal available at each pass.
- We introduce *lonely adjective detection* as a linguistically motivated gating signal that selectively activates knowledge-graph inference when opinion-bearing adjectives lack an explicit syntactic noun head — a surface configuration strongly associated with implicit aspect expression.
- We extend the full pipeline to Arabic, incorporating morphology-aware adaptations at each stage to accommodate the language's rich inflectional structure.
- We provide empirical evaluation across eight English and Arabic benchmarks, with ablation studies isolating the contribution of each pipeline component.

2. Background and Preliminaries

2.1. Text Paraphrasing

Text paraphrasing — the reformulation of a source expression into alternate wording while preserving its meaning — is a well-established technique in natural language processing, with applications spanning data augmentation, machine translation, and question answering [14]. In the context of aspect extraction, paraphrasing serves a distinct purpose: by introducing syntactic variation while

[1] https://alt.qcri.org/semeval2014/task4/
[2] https://alt.qcri.org/semeval2015/task12/
[3] https://alt.qcri.org/semeval2016/task5/
[4] https://github.com/NUSTM/ACOS
[5] https://github.com/swaggy66/M-ABSA

 doi:10.20944/preprints202605.1240.v1

holding semantics constant, paraphrased inputs can surface implicit aspects that remain latent under a single syntactic realization.

2.2. Knowledge Graphs

A knowledge graph (KG) represents facts as typed nodes and relations, formalized as triples (h, r, t), enabling models to reason over entities and their neighborhoods rather than isolated tokens. For implicit aspect extraction, a KG provides grounded opinion-to-aspect mappings that purely data-driven models cannot reliably infer from distributional statistics alone.

General-purpose resources fall short for this task. Topic models such as LDA [15] produce soft probabilistic assignments rather than precise mappings; in our preliminary experiments, their signals consistently degraded extraction performance. WordNet [16] and its Arabic counterpart [17] lack domain-specific opinion-to-aspect coverage. ConceptNet [18] encodes broader commonsense relations and has been applied to sentiment analysis [19,20], but its knowledge is domain-general, its Arabic subgraph is sparse, and its undirected relations lack the domain hierarchy needed to resolve ambiguous opinion words across domains (e.g., *slow* mapping to SERVICE#GENERAL in restaurants versus LAPTOP#PERFORMANCE in laptop reviews). Domain-specific KGs have demonstrated more consistent ABSA gains [20–23]; the KG constructed in this paper follows that direction, encoding directed, domain-specific opinion-to-aspect mappings derived from training data.

2.3. Arabic NLP and Its Challenges

Arabic presents several NLP challenges due to its extensive inflectional morphology, whereby one word may have many surface forms based on gender, number, and syntactic position. Three challenges are of direct relevance to this work.

Orthographic inconsistency. Writers inconsistently interchange letters such as أ، إ with ا, ي with ى, and ة with ه, causing the same term to appear in multiple surface forms [24]. Left unaddressed, this variability causes identical aspect terms to be treated as distinct tokens during model inference, paraphrase comparison, and knowledge graph lookup.

Diacritics. Short vowels alter meaning fundamentally (e.g., كَتَبَ *wrote* vs. كُتُب *books*) yet are routinely omitted in informal text, introducing systematic ambiguity that complicates morphological analysis and aspect matching.

Dialectal variation and resource poverty. Arabic social media largely follows dialectal rather than Modern Standard Arabic (MSA) conventions, for which syntactic analyzers and domain lexicons remain scarce. No large, publicly annotated Arabic dataset for implicit aspects exists, constraining both model training and evaluation.

These challenges directly motivate the Arabic-specific design decisions described in Section 4.2.

2.4. Research Questions

This work addresses the following research questions:

1. How can a generative model be adapted to recover multiple implicit aspects from a single sentence, where each opinion expression must be mapped to its corresponding aspect?
2. How can domain-ambiguous opinion words be resolved to their intended aspect targets across different review domains?
3. To what extent does instruction-tuned generative modeling improve implicit aspect extraction performance relative to prior approaches?
4. Does paraphrase-based input diversification yield measurable gains in implicit aspect recall, and under what conditions?

3. Related Work

In 2014, the research community devoted substantial effort to characterizing the ABSA task, leading to the release of benchmark datasets and shared evaluation settings. This review focuses on Arabic

 doi:10.20944/preprints202605.1240.v1

ABSA, and emphasizes deep learning, hybrid, and large language model (LLM)–based approaches, which currently dominate aspect extraction research. Non-Arabic studies are limited to recent work that targets implicit aspect extraction. The reviewed studies are grouped into deep learning methods and hybrid approaches that integrate neural models with structured or symbolic components.

3.1. Deep Learning Approaches

The adoption of deep learning marked a significant shift in aspect extraction by enabling models to capture richer contextual representations. In Arabic, early neural architectures applied CNNs [25] and hierarchical BiLSTMs [26], followed by models combining CNNs with LSTMs [27] or GRU-based architectures with bilingual embeddings [6]. More recent work incorporated transformer-based encoders, including multi-task extensions such as MTL-AraBERT for joint aspect extraction and sentiment classification [28].

Transformer architectures have since become central to Arabic ABSA. AraBERT has been integrated with stacked LSTM–GRU layers and attention mechanisms [7], CRF-based tagging pipelines [29], and sequence-to-sequence models generating aspect–sentiment pairs [30]. Domain-adaptive fine-tuning has also proven effective, as demonstrated by CAMeLBERT for noisy and dialectal text [8]. Comparative evaluations consistently show transformer-based models outperforming recurrent neural architectures.

Beyond Arabic, deep learning has driven substantial advances in aspect extraction. BERT-based embeddings have been combined with CNN, BiLSTM, RCNN, and attention mechanisms [31], ensemble BiLSTMs [32], topic-aware CNNs [9,10], and graph convolutional networks [33]. Further refinements include attention-augmented BiLSTMs [34,35], BiLSTM–CRF architectures [36], and domain-adaptive BiLSTMs with multi-head attention [37]. Additional contributions include span-level joint extraction models [38], constituency-aware architectures [39].

3.2. Deep Learning Approaches for Implicit Aspect Extraction

While most deep learning research focuses on explicit aspects, a smaller body of work directly targets implicit aspect extraction. Soni and Rambola [40] combined recurrent neural networks with semantic similarity measures derived from WordNet and spaCy to map opinion expressions to latent aspects. More recent studies have explored LLMs, with [11] evaluating GPT-3.5 and GPT-4 for aspect category detection and demonstrating that prompting and fine-tuning strategies can capture implicit signals. Similarly, [41] employed BERT in a zero-shot setting to address multiple ABSA subtasks, including implicit categories. Data augmentation strategies, such as Word2Vec-based replacements in [42], aimed to enrich training signals, although their impact remained limited.

3.3. Hybrid Approaches

Hybrid approaches integrate neural models with symbolic, statistical, or structural components to enhance robustness. In Arabic, early hybrid work integrated stacked LSTMs with logistic regression for aspect classification [43]. Ontology-driven and neural hybrid designs have also been explored, including BiLSTM–CRF and BiGRU–CRF pipelines combined with fastText or AraBERT embeddings [44–48]. These models leverage recurrent encoders with structured decoders to better capture sequence dependencies. More recent expansions include pseudo-labeling with ChatGPT prior to transformer fine-tuning [49] and pipelines combining BERTopic clustering, CAMeLBERT sentiment analysis, and AraBART summarization [12].

Beyond Arabic, hybrid approaches have similarly evolved. Some combine syntactic rules with BERT embeddings and neural classifiers [50], while others extend encoder–decoder architectures with contrastive learning for structured quadruple extraction [51].

3.4. Hybrid Approaches for Implicit Aspect Extraction

A limited but growing body of hybrid research directly addresses implicit aspect extraction. Ahmed et al. [52] combined topic modeling with BERT to learn implicit aspect-specific representa-

 doi:10.20944/preprints202605.1240.v1

tions, while Feng et al. [53] and Lazhar [54] integrated CNNs and association rules to link opinion expressions to hidden aspects. More recent work has shifted toward transformer-based hybrids. Li et al. [55] introduced generative T5-based frameworks augmented with graph neural networks and prompt fusion, surpassing earlier ACOS baselines. Other studies proposed multi-stage pipelines combining BERT embeddings, semantic similarity, and classical classifiers [56], frequency–syntax–CRF frameworks addressing both explicit and implicit aspects [57], and representation-driven hybrids such as WoSe [58] employed dual-level encoders and CRF decoding to better capture implicit aspects, while prompting and synthetic augmentation strategies using large language models were explored in [59].

3.5. Critical Assessment and Research Gap

Three gaps emerge from this review, first, Arabic ABSA research overwhelmingly targets explicit aspects. The few studies engaging with implicit aspects in Arabic [11,41,42] address category detection or label augmentation rather than proposing a dedicated implicit-to-aspect inference mechanism. Second, even with transformer and LLM-based models, current solutions typically rely on one-shot inference and model-internal reasoning, which limits their ability to recover multiple or implicit aspects. Third, while KG-augmented inference has been explored in English [55], no prior work constructs an Arabic-specific knowledge graph for implicit aspect extraction, leaving this pathway entirely unexplored for Arabic. These gaps motivate the proposed solution.

4. Methodology

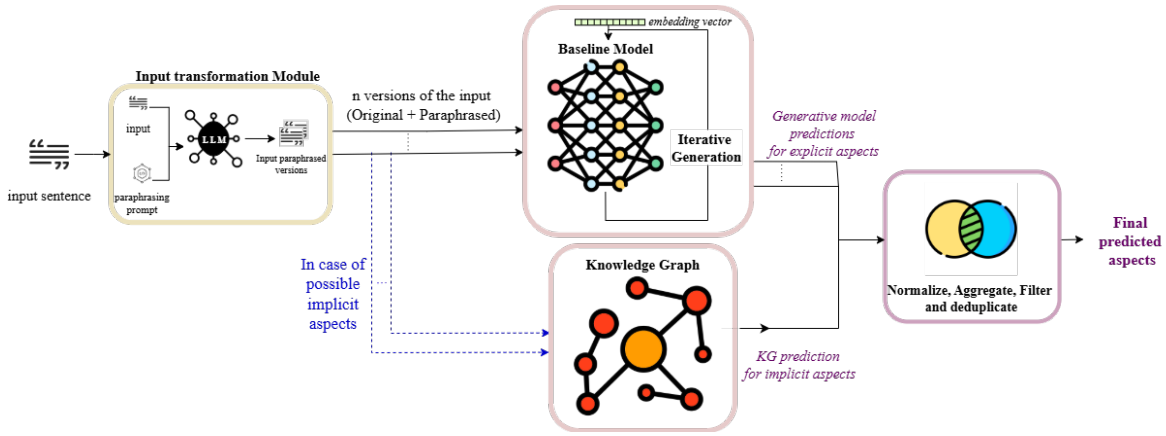


Figure 1. Model architecture.

This section presents the proposed four-component pipeline for explicit and implicit aspect extraction. The pipeline comprises: a fine-tuned decoder-only baseline model (BL) that provides the generative foundation; an iterative residual generation component (IG) that progressively recovers multiple aspects from a single sentence; a paraphrase-based input variation module (PARA) that enforces cross-wording agreement and reveals latent aspects; and a domain-specific knowledge graph (KG) that provides structured opinion-to-aspect mappings for implicit aspect inference. The overall architecture is illustrated in Figure 1.

This section is organized as follows: each component is described in detail, followed by the end-to-end inference procedure that formalizes their interaction, and concluding with the Arabic-specific adaptations and preprocessing steps required to extend the pipeline to Arabic.

4.1. Proposed Solution Components

4.1.1. Baseline Model: Fine-Tuned Generative Model

A decoder-only LLM is fine-tuned for aspect extraction using instruction-based supervised fine-tuning (SFT). Each training instance is formatted as an (**instruction**, **sentence**, **aspect**) triple using the following prompt template:

 doi:10.20944/preprints202605.1240.v1


```
Below is an instruction that describes a task, paired with an input sentence. Write a
response that appropriately completes the request.

### Instruction: {instruction}

### Sentence: {sentence}

### Aspect: {aspect}
```

The instruction field is fixed across all instances:

Act as an Natural Language Processing expert. Extract aspect term(s) explicitly mentioned in the sentence. Only return terms that appear verbatim; do not infer or invent related words. If no aspect is found, return "none".

Aspect extraction is framed as text generation: the model produces explicit aspect terms present in the input, or returns `none`. This fine-tuned model serves as the baseline (**BL**) component. At inference, the same template is used with the *Aspect* field left empty for the model to predict.

4.1.2. Iterative Residual Generation

Iterative residual generation decomposes multi-aspect extraction into successive generation rounds. After each round, the extracted term is masked from the input, redirecting the model's attention to the remaining content. Each round targets the residual left by previous rounds, with extraction terminating when the model produces a null output.

This design resolves a structural limitation of beam search, which biases toward the most salient aspect regardless of beam width, since all beams are conditioned on the same unmodified input [60]. Masking extracted terms between rounds removes this attractor at each step, producing a monotonic increase in recall that beam search cannot achieve.

4.1.3. Input Transformation via Paraphrasing

Several paraphrases are generated for every sentence, and each one is processed through the same iterative generation process. The predictions are then combined and filtered such that retained aspect terms must either explicitly appear in the original sentence or satisfy a set of validation constraints, including a minimum semantic similarity to the original sentence and sufficient agreement across paraphrases. These consistency checks allow the fine-tuned model to recover latent aspects while suppressing generative drift.

Paraphrases are generated using the GPT-4.1-mini model [61]. For each input sentence, ten paraphrases are requested using the following prompt:

"Act as an expert linguist in English. Give 10 different paraphrases to the following text in English, trying to explicitly state every noun described in the sentence. Only include the paraphrased sentence with no additional text before or after."

4.1.4. Knowledge Graph for Implicit Aspect Identification

While the preceding components target explicit aspect extraction, many opinionated sentences convey sentiment without naming an aspect term. In such cases, the fine-tuned generative model may correctly return NULL, indicating the absence of lexically grounded aspects. A NULL output, however, may correspond either to an objective sentence or to an implicit-bearing sentence where the aspect is expressed indirectly through evaluative clues; distinguishing between these cases requires reasoning beyond direct extraction.

Accordingly, the KG component is activated conditionally: when the generative component returns NULL, the KG component scans the input sentence for implicit aspect clues. It searches for two categories of cue: *lonely adjectives*, sentiment adjectives that syntactically modify no explicit noun in the sentence; and *verb-phrase clues*, sentiment verbs that signal an evaluative stance without naming an aspect. When at least one such cue is detected, the component maps it to a domain-consistent aspect concept. When no cue is detected, the component produces no implicit prediction for that sentence.

 doi:10.20944/preprints202605.1240.v1

Knowledge Graph Structure

The KG represents domain knowledge using four vertex types: *domains*, *entities*, *aspects*, and *clues*. Clue vertices correspond to opinion-bearing expressions detected in text, while aspect vertices represent semantic aspect concepts, grouped under higher-level entity vertices and ultimately associated with domain vertices. The graph encodes three primary relationship types: *entity is_relevant_to domain*, *aspect belongs_to entity*, and *clue describes_aspect*. This layered structure enables multi-hop reasoning from surface-level clues to domain-specific aspects.

The full formal definition of the KG, relation composition, and algorithmic details are provided in Appendix A.1.

Knowledge Graph Construction

Both English and Arabic KGs share the same hierarchical structure, constructed by extracting opinion clues from review texts and aligning them to aspects, entities, and domains via training set labels. For English, lexical coverage was expanded with synonyms, antonyms, and related expressions across adjectives, verbs, nouns, and interjections that function as implicit opinion signals (e.g., *clean*, *spotless*, *filthy* → Cleanliness). For Arabic, clue nodes group multiple lexical realizations under unified concepts to accommodate nominalized and derivational forms (e.g., نظافة عدم قذارة، وسخ، قذر، → Cleanliness), and evaluative constructions are captured directly (e.g., السعر يستحق لا → Prices). Figure 2 shows a KG subgraph and Table 1 reports the scale of both KGs.

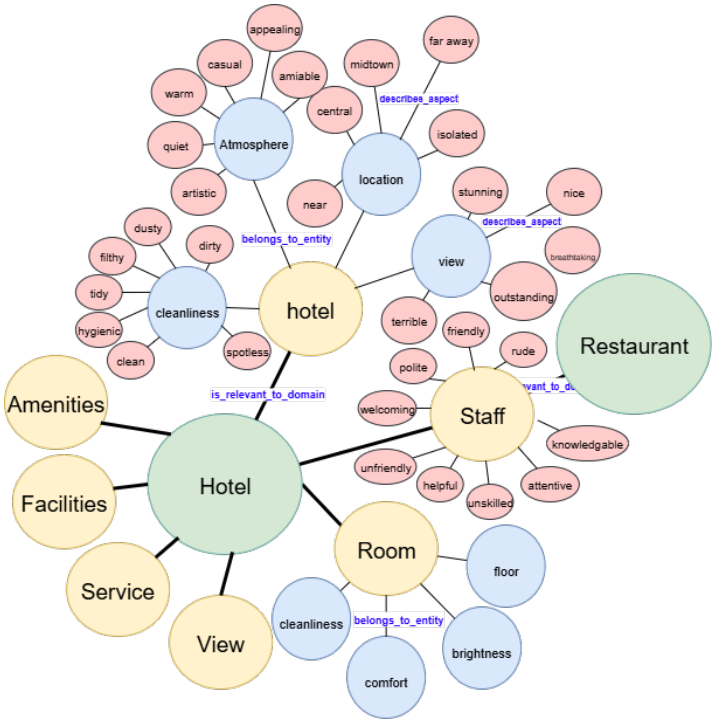


Figure 2. A subgraph of the English KG. Domain vertices: green; entity vertices: yellow; aspect vertices: blue; clue vertices: pink.

Table 1. Knowledge graph statistics.

	Coverage				Edges			
	Domains	**Entities**	**Aspects**	**Clues**	***describes_aspect***	***belongs_to_entity***	***is_relevant_to_domain***	**Total**
English KG	7	85	136	1,179	2,630	301	71	3,435
Arabic KG	8	89	3,258	3,255	5,026 *	5,026 *	5,026 *	5,026

* In the Arabic KG, clue-to-aspect, aspect-to-entity, and entity-to-domain mappings are encoded within a single unified edge structure rather than as separate edge types, yielding identical counts across all three relation columns.

 doi:10.20944/preprints202605.1240.v1

Algorithm 1: End-to-End Inference for Explicit and Implicit Aspect Extraction

Input : s: input sentence; M: fine-tuned generative LLM; G: domain knowledge graph; R: maximum IG rounds; N: number of paraphrases; τ: similarity threshold; *min_consensus*: minimum prediction agreement across paraphrases
Output : A: final set of extracted aspects

$A \leftarrow \emptyset$; // Final aspect set
$C \leftarrow \emptyset$; // Intermediate aspect predictions
// Step 1: Generate paraphrases
$P \leftarrow \{s\}$;
for $i \leftarrow 1$ **to** N **do**
  $p_i \leftarrow LLM(\text{ParaphrasePrompt}(s))$;
  $P \leftarrow P \cup \{p_i\}$;
// Step 2: Iterative generation over original and paraphrases
foreach $x \in P$ **do**
  $x' \leftarrow x$;
  for $r \leftarrow 1$ **to** R **do**
    // Prompt the fine-tuned model to generate a prediction
    $a \leftarrow M(\text{GeneratePrediction}(x'))$;
    if $a = none$ **then**
      break;
    $C \leftarrow C \cup \{a\}$;
    // Remove the extracted terms from the sentence
    $x' \leftarrow \text{Remove}(x', a)$;
// Step 3: Paraphrase agreement and semantic filtering
$C' \leftarrow \emptyset$;
foreach $a \in C$ **do**
  $sim \leftarrow \text{SemanticSimilarity}(a, s)$;
  $votes \leftarrow \text{consensus}(a, P)$; // number of paraphrases giving this prediction
  if $sim \geq \tau \wedge votes \geq min\_consensus$ **then**
    $C' \leftarrow C' \cup \{a\}$;
// Step 4: Knowledge-guided implicit aspect inference
$L \leftarrow \text{IdentifyClues}(s)$; // search for cues in s
if $L \neq \emptyset$ **then**
  $I \leftarrow$ Algorithm 2(s, P, G);
  $C' \leftarrow C' \cup I$;
// Step 5: Deduplication and normalization
$A \leftarrow \text{NormalizeAndDeduplicate}(C')$;
return A

4.1.5. End-to-End Inference Procedure

Algorithm 1 presents the end-to-end inference procedure of the proposed solution. Given an input sentence, the solution first generates paraphrases to introduce syntactic variation while preserving meaning, then applies iterative residual generation jointly to the original sentence and its paraphrases, allowing the fine-tuned model to recover multiple explicit aspects missed under single-pass decoding. The resulting candidate pool is filtered through semantic similarity and paraphrase consensus constraints to discard spurious predictions. In parallel, the KG component scans the sentence for opinion-bearing clues with no syntactic noun head, operating as an independent implicit aspect identification path that maps each detected cue to a domain-consistent aspect. Finally, all candidates are normalized, deduplicated, and aggregated into a unified output set.

4.2. Arabic Adaptation

The Arabic adaptation addresses the challenges identified in Section 2.3 through four preprocessing steps applied uniformly across all components.

1. **Clitic segmentation:** Aspect clues frequently appear inside affixed tokens (e.g., وبأسعار). CAMeL Tools [62] separates functional prefixes from content-bearing stems (و+ب+ أسعار → أسعار → lemma سعر), ensuring aspect cues are not obscured by affixation.
2. **Orthographic normalization:** Segmented tokens undergo hamzah unification (آ أ إ → ا), diacritic and tatweel removal, Unicode normalization, and feminine ending (ة) removal from adjectives, ensuring consistent token identity across all pipeline components.

 doi:10.20944/preprints202605.1240.v1

3. **Lexical clustering:** Semantically related surface forms sharing a common referent (e.g., أسعار، سعر، مرتفع غالي،) are unified to a single aspect target (e.g., PRICES), reducing sparsity in knowledge graph lookups and evaluation matching.
4. **Morphological disambiguation:** CAMeL-derived lemmas are matched against knowledge graph clue nodes, ensuring that inflected variants across gender and number (نظيف، نظيفة، نظيفًا) map to the same canonical form, increasing clue coverage without manual enumeration of inflected forms.

5. Experimental Study

5.1. Experimental Setup

5.1.1. Implementation Details

A decoder-only large language model, Llama-3-8B [63], was fine-tuned on task-specific datasets using a parameter-efficient fine-tuning (PEFT) method, specifically LoRA [64]. Under LoRA, the pre-trained model parameters are kept frozen, while low-rank adaptation layers are introduced and optimized during the training.

Fine-tuning was performed on an NVIDIA A100 SXM4 GPU with 40 GB VRAM, accessed via Google Colaboratory with the High-RAM runtime enabled. The High-RAM runtime is required at LoRA rank $r = 128$.

Hyperparameter configurations shown in Table 2 followed the configuration of Neveditsin et al. [59], who fine-tuned the same LLM backbone under an identical parameter-efficient fine-tuning setup and reported competitive results, providing a well-validated starting point for our experimental conditions.

Table 2. Fine-tuning hyperparameters.

Component	Setting
LoRA rank (`r`)	128
LoRA alpha	32
LoRA dropout	0.1
Batch size (per device)	8
Gradient accumulation	1
Warmup steps	2
Max training steps	Dataset-dependent *
Learning rate	1×10^{-4}
Optimizer	`paged_adamw_32bit`
Eval strategy	`steps`, every 200 steps
Save strategy	`steps`, every 200 steps
Model selection	`load_best_model_at_end=True`, `metric_for_best_model=eval_loss`

* Corresponds to roughly 4–6 epochs depending on dataset size.

 doi:10.20944/preprints202605.1240.v1

Table 3. Software environment and NLP tools.

Component	Version / Details
Runtime	
Operating system	Ubuntu 22.04 LTS
Python	3.12.13 (Colab default)
PyTorch	2.10.0 (CUDA 12.1 build)
CUDA toolkit	12.1 (Colab-managed)
NVIDIA driver	535.x (Colab-managed)
Fine-tuning	
Transformers	4.40.0 (HuggingFace)
PEFT	0.10.0 (HuggingFace)
TRL	0.8.6 (HuggingFace)
BitsAndBytes	0.43.1 (4-bit quantization)
Unsloth	2024.x (LoRA optimization)
Accelerate	0.29.3 (HuggingFace)
Paraphrasing	
OpenAI API	Chat Completions (GPT-4.1-mini)
NLP Processing	
spaCy	3.7.x (`en_core_web_sm`)
Stanza	1.11.0 (`tokenize, pos, lemma, depparse`)
CAMeL-Tools	1.5.2 (MLE disambiguator)
SentenceTransformers	`paraphrase-multilingual-MiniLM-L12-v2`

Table 3 lists the full software environment and NLP tools used across all experiments.

The implementation of the proposed solution, including training, inference, and evaluation scripts, is publicly available at https://github.com/LujainAlawwad/ImplicitAspectExtraction

5.1.2. Datasets

We evaluate the proposed solution on both English and Arabic ABSA benchmarks. For English, we use the SemEval 2014, 2015, and 2016 shared-task datasets, ACOS, and the English portion of the Multilingual ABSA dataset (M-ABSA). The Arabic evaluation covers the SemEval 2016 Arabic Hotel reviews, the Human Annotated Arabic Dataset (HAAD) [6] of book reviews, and the Arabic portion of M-ABSA; all datasets are publicly available. Tables 4 and 5 summarize review-level statistics for all splits, distinguishing explicit, implicit, mixed, and objective sentences.

Table 4. English ABSA datasets used in our experiments. All counts are at the review level. #Mix.: sentences containing both explicit and implicit aspects.

Dataset	Domain	Split	#Reviews	#Expl.	#Impl.	#Mix.	#Obj.
SE14	Rest.	Train	3041	2021	1020	0	0
	Rest.	Test	800	606	194	0	0
	Lapt.	Train	3045	1488 *	0	0	1557
	Lapt.	Test	800	422	378	0	0
SE15	Rest.	Train	1315	818	302	0	195
	Rest.	Test	684	401	180	0	103
SE16	Rest.	Train	2000	1210	498	0	292
	Rest.	Test	586	419	167	0	0
ACOS	Rest.	Train	1701	1225	476	0	0
	Rest.	Test	583	412	171	0	0
	Lapt.	Train	2485	1710	775	0	0
	Lapt.	Test	816	645	171	0	0
$\text{M-ABSA}_{\text{en}}$	Multi	Train	8796	5043	1864	323	1566
	Multi	Test	3794	2564	986	171	73

* 1488 subjective sentences; official labels do not annotate aspect categories.

[6] https://github.com/msmadi/HAAD

 doi:10.20944/preprints202605.1240.v1

Table 5. Arabic ABSA datasets. Statistics as in Table 4.

Dataset	Domain	Split	#Reviews	#Expl.	#Impl.	#Mix.	#Obj.
SE16	Hotel	Train	4776	4445	331	0	0
	Hotel	Test	1225	1029	196	0	0
HAAD	Books	Train	1198	1187	11	0	0
	Books	Test	299	299	0	0	0
M-ABSA$_{ar}$	Multi	Train	8792	5005	1893	328	1566
	Multi	Test	3793	2504	1039	177	73

5.1.3. Evaluation Metrics

Explicit and implicit aspects require separate evaluation protocols and cannot be collapsed into a single overall score: explicit sentences carry a surface aspect term that the model must identify, whereas implicit sentences carry only an abstract `Entity#Aspect` category with no surface mention to extract. We therefore report two metrics throughout: Metric A for explicit aspect extraction and Metric B for implicit aspect identification via the KG module.

We evaluate system performance using micro-precision (P), micro-recall (R), and micro-F_1 score (F_1), which are the standard metrics for assessing aspect-based sentiment analysis performance. These metrics are computed globally by aggregating true positives (TP), false positives (FP), and false negatives (FN) over all predictions:

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN}, \qquad F_1 = \frac{2PR}{P + R} \tag{1}$$

Metric A: Explicit Aspect Extraction.

For explicit aspect-bearing sentences, predictions are compared against the surface-level `aspect` field of the gold annotations using micro-$P/R/F_1$. A predicted aspect term is counted as a true positive if it matches a gold aspect term, with a one-to-one matching constraint ensuring each prediction matches at most one gold term and vice versa.

Metric B: Implicit Aspect Identification.

The gold annotations of implicit sentences carry `NULL` in the aspect field but provide a complete `Entity#Aspect` pair in the category field (e.g., `RESTAURANT#PRICES`). Accordingly, KG module predictions are evaluated against the aspect component of the category field.We report *partial matching*: a prediction is counted as correct if the predicted aspect matches the gold aspect after normalization, accepting token-subset overlaps in addition to exact matches to accommodate lexical variation in compound expressions (e.g., *swimming pool* and *pool*). A strict bipartite matching constraint ensures each prediction is aligned to at most one gold aspect and vice versa.

5.2. Experiment Results

This section presents the experimental results across all datasets For English and Arabic languages. We begin with an ablation study analyzing the contribution of each system component, followed by performance evaluation, sensitivity analysis, Error analysis and comparative results against published SOTA.

5.2.1. Proposed Solution Evaluation on English Datasets

1. Ablation Study

 doi:10.20944/preprints202605.1240.v1

Table 6. Explicit aspect extraction performance (micro-precision, -recall, -F_1) on English datasets. ZS: zero-shot; BL: fine-tuned baseline; BL+IG: baseline with iterative generation; BL+IG+Para: baseline with iterative generation and paraphrase-based input variation.

	SE14			SE15			SE16			ACOS			M-ABSA$_{en}$		
Method	***P***	***R***	F_1	***P***	***R***	F_1	***P***	***R***	F_1	***P***	***R***	F_1	***P***	***R***	F_1
ZS	0.3425	0.2322	0.2767	0.2663	0.2037	0.2308	0.3020	0.2160	0.2523	0.3323	0.3323	0.3323	0.2425	0.1701	0.2000
BL	0.8894	0.6030	0.7187	0.7298	0.6110	0.6651	0.8618	0.6491	0.7405	0.8270	0.8270	0.8270	0.7065	0.5990	0.6483
BL+IG	0.8906	0.7864	0.8353	0.7367	0.6776	0.7059	0.8249	0.7995	0.8120	0.7389	0.8541	0.7923	0.6609	0.6353	0.6478
BL+IG+Para	0.8957	0.8148	**0.8533**	0.7402	0.6878	**0.7130**	0.7839	0.7879	**0.7859**	0.7104	0.8969	**0.7930**	0.6878	0.6415	**0.6639**

Table 6 reports micro-F_1 scores across five English benchmarks: SemEval-2014, SemEval-2015, SemEval-2016, ACOS, and M-ABSA. For ACOS, results are aggregated over both the Restaurant and Laptop subsets.

Across all datasets, fine-tuning the foundation model substantially improves performance over the zero-shot (ZS) setting, confirming the necessity of task adaptation. The strongest absolute performance is observed on SemEval-2014 ($F_1 = 0.8533$), reflecting the relatively constrained and well-structured annotation scheme of this benchmark. Iterative generation (+IG) consistently improves recall over the baseline across all datasets, with the most pronounced gain on SE14 (F_1 : $0.7187 \rightarrow 0.8353$). The paraphrase-based component (+Para) achieves the best overall F_1 on all five benchmarks. Its contribution is largest on SE14 and ACOS, while remaining more conservative on SE15, SE16, and M-ABSA$_{en}$. On M-ABSA$_{en}$, the multi-domain nature of the benchmark — spanning restaurant, hotel, laptop, phone, food, sight, and Coursera reviews — presents a broader vocabulary challenge; the system nevertheless achieves $F_1 = 0.6639$, with the largest per-domain gains observed in the hotel ($F_1 = 0.7695$) and laptop ($F_1 = 0.7692$) domains.

2. Performance Evaluation

(a) Explicit vs. Implicit Performance

Table 7. Explicit vs. implicit pipeline performance (micro-F_1). For implicit rows, correct abstention (NULL prediction vs. NULL gold) is counted as TP. BL: Baseline; +Para: BL+IG+Para.

Dataset	**Type**	**BL**	**+IG**	**+Para**	**Δ vs BL**
SE14	Explicit	0.6193	0.7916	**0.8245**	+0.2052
	Non-Explicit	0.9381	0.9231	**0.8906**	−0.0475
SE15	Explicit	0.5274	0.6009	**0.6224**	+0.0950
	Non-Explicit	0.9429	0.9347	**0.9309**	−0.0120
SE16	Explicit	0.6932	0.7963	**0.7739**	+0.0807
	Non-Explicit	0.8862	0.8683	**0.8299**	−0.0563
ACOS-Rest.	Explicit	0.8398	0.7722	**0.7743**	−0.0655
	Non-Explicit	0.8555	0.8300	**0.8415**	−0.0140
ACOS-Lap.	Explicit	0.8127	0.8037	**0.7937**	−0.0190
	Non-Explicit	0.8246	0.8580	**0.8290**	+0.0044
M-ABSA$_{en}$	Explicit	0.5718	0.5824	**0.5993**	+0.0275
	Non-Explicit	0.9392	0.9245	**0.9332**	−0.0060

Table 7 separates performance across explicit and non-Explicit subsets. Non-explicit subset include implicit, mixed and objective rows. For implicit rows, the gold label is NULL; a correct NULL prediction counts as a true positive.

On explicit rows, the pipeline improves consistently from BL to +Para across SemEval datasets, with the largest gain on SE14 ($\Delta F_1 = +0.205$) driven by iterative generation recovering multiple aspects. On ACOS, the baseline achieves the highest precision, while +IG and +Para expand recall at the cost of false positives, consistent with ACOS's denser multi-aspect annotation. On M-ABSA$_{en}$ explicit rows, all components contribute incre-

 doi:10.20944/preprints202605.1240.v1

mentally (F_1 : 0.572 → 0.582 → 0.599), though overall performance is lower, reflecting the challenge of generalising across seven heterogeneous domains.
On implicit rows, all system components achieve high abstention scores ($F_1 = 0.83$–0.94), confirming reliable NULL prediction across all datasets and leaving implicit aspect evaluation to Metric B.

(b) Implicit Aspect Identification via Knowledge Graph
The KG module is evaluated against the `category` gold annotation (ENTITY#ASPECT format). SemEval-2014 is excluded because its `category` field is empty in the laptop domain, or uses a generic label (*food*, *service*, *anecdotes/miscellaneous*, *ambience*, *price*) in the restaurant domain.

Table 8. KG implicit aspect identification performance. Any-TP@row: proportion of implicit rows with ≥1 correct prediction.

Dataset	P	R	F_1	Any-TP@row
SE15	0.7119	0.3077	0.4297	0.3306
SE16	0.7021	0.3708	0.4853	0.3952
ACOS-Rest.	0.6842	0.3757	0.4851	0.3757
ACOS-Lap.	0.4479	0.2515	0.3221	0.2515
M-ABSA$_{en}$	0.5420	0.2821	0.3710	0.3095

Table 9. KG inference step distribution across implicit rows (%).

Dataset	`STOPPED`	`INFERRED`	`NO_MATCH`
SE15	54.7	40.4	4.9
SE16	49.1	44.9	6.0
ACOS-Rest.	49.7	45.0	5.2
ACOS-Lap.	54.4	40.4	5.3
M-ABSA$_{en}$	48.7	41.8	9.5

The KG achieves high precision on restaurant and hotel domains ($P = 0.68$–0.71), confirming reliable predictions when inference fires. Recall is moderate ($R = 0.31$–0.37), reflecting the coverage ceiling of lonely-adjective and verb-clue detection: roughly half of implicit rows are suppressed due to no detectable syntactic or lexical cue (Table 9). On ACOS-Laptop, performance drops substantially ($F_1 = 0.32$) as technical aspects such as *operation_performance* and *portability* require domain-specific clues underrepresented in the KG lexicon. On M-ABSA$_{en}$, the KG achieves $F_1 = 0.37$ overall, with strongest coverage on restaurant ($F_1 = 0.50$) and hotel ($F_1 = 0.41$) subsets, and weaker coverage on sight ($F_1 = 0.12$) and Coursera ($F_1 = 0.31$), where implicit sentiment patterns are inadequately represented.

(c) Domain-level Performance Restaurant benefits most from iterative generation and paraphrasing across SemEval benchmarks (up to +0.144 on SE14-Restaurant). SE14-Laptop shows an inversion where +Para outperforms +IG (0.875 > 0.813), suggesting paraphrase-based variation is particularly effective for technical vocabulary. SE15-Hotel shows the smallest gain ($\Delta F_1 = +0.021$), consistent with lower KG recall ($R = 0.183$), indicating more restrained language with fewer detectable implicit cues. On ACOS, both domains show slight F_1 decreases from BL to +Para due to precision–recall trade-offs: the baseline predicts conservatively, while +IG and +Para expand coverage at the cost of false positives. Among M-ABSA$_{en}$ domains, Hotel and Laptop achieve the highest F_1 (0.770 and 0.769); Sight is the most challenging ($F_1 = 0.481$), as tourist attraction reviews use idiomatic and metaphorical language poorly represented in training data, yielding low KG coverage ($R = 0.073$ on implicit Sight rows).

 doi:10.20944/preprints202605.1240.v1

Table 10. Per-domain explicit extraction performance (micro-F_1). BL: Baseline; +Para: BL+IG+Para.

Dataset	Domain	BL	+IG	+Para	Δ vs BL
SE14	Restaurant	0.6917	0.8528	**0.8357**	+0.1440
SE14	Laptop	0.7500	0.8132	**0.8754**	+0.1254
SE15	Restaurant	0.6826	0.7260	**0.7421**	+0.0595
SE15	Hotel	0.6257	0.6599	**0.6466**	+0.0209
SE16	Restaurant	0.7405	0.8120	**0.7859**	+0.0454
ACOS	Restaurant	0.8444	0.7876	**0.7919**	-0.0525
ACOS	Laptop	0.8148	0.8128	**0.7993**	-0.0155
M-ABSA$_{en}$	Hotel	0.7402	0.7451	**0.7695**	+0.0293
M-ABSA$_{en}$	Laptop	0.7532	0.7418	**0.7692**	+0.0160
M-ABSA$_{en}$	Restaurant	0.7403	0.7412	**0.7483**	+0.0080
M-ABSA$_{en}$	Food	0.6451	0.6427	**0.6728**	+0.0277
M-ABSA$_{en}$	Phone	0.5668	0.5889	**0.5852**	+0.0184
M-ABSA$_{en}$	Coursera	0.6836	0.6724	**0.6877**	+0.0041
M-ABSA$_{en}$	Sight	0.4640	0.4556	**0.4813**	+0.0173

Table 11. Boundary identification errors.

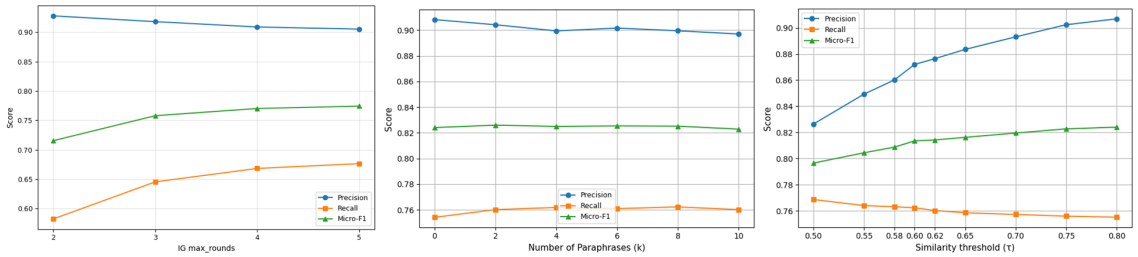
Issue	Gold	Predicted
Subspan	product quality	quality
Subspan	Apple Help	Help
Superspan	shipping	two day shipping
Superspan	start up	quick start up

3. Error Analysis

- Boundary Mismatch:
  Generative models produce aspect spans that are either subspans or superspans of the gold annotation. Table 11 shows representative cases. Boundary errors account for 8–37% of English false positives (highest on SemEval-2014: 37%, $n = 34$) These mismatches represent an annotation convention disagreement rather than a genuine inference failure.
- Spurious Hallucination. Spurious predictions sharing no token overlap with any gold term dominate English false positives (84–93% across all English datasets). They are most prevalent in sentences with domain-specific terminology or multi-aspect structures absent from fine-tuning data, suggesting targeted augmentation or constrained decoding as remedies.
- Taxonomy Mismatch:
  The most frequent KG error across all restaurant and hotel datasets is a fine-grained taxonomy mismatch between the KG's internal aspect vocabulary and the gold annotation scheme. The KG predicts *taste* for sentences expressing food quality, while the gold annotation uses *quality*. For example, the sentence *"Mmmmmmmmm so delicious"* is predicted as `food#taste` but annotated as `food#quality`.
- KG's Coverage issue: No-Cue Detection (STOPPED rows):
  Approximately 49–55% of implicit rows across all datasets are suppressed (`STOPPED_no_lonely_adj`) because no lonely adjective or verb clue is detected. Examination of these sentences reveals three recurring patterns. First, implicit sentiment expressed through noun phrases alone without any adjective, such as *"Two thumbs up!"* (gold: `food#quality`). Second, figurative expressions that carry no surface-level sentiment marker the KG can detect, such as textit"You'll be there for every anniversary…" (gold: `restaurant#miscellaneous`). Third, sentences where the implicit aspect is carried entirely by a verb with no accompanying adjective, such as *"We were charged full price"* (gold: `service#general`).

 doi:10.20944/preprints202605.1240.v1

4. Sensitivity Analysis
   (a) **Iterative Generation rounds** Figure 3a: Increasing rounds from 2 to 3 yields the largest F_1 gain (+9.4% recall, −2.3% precision); beyond 3 rounds improvements diminish. Maximum rounds set to 3 for English.
   (b) **Paraphrase-Aided Parameters** Figures 3b,c:



(**a**) (**b**) (**c**)

Figure 3. Sensitivity analysis: (**a**) iterative generation rounds; (**b**) number of paraphrases; (**c**) semantic similarity threshold τ.

   i. **Number of paraphrases**: Two paraphrases give the best F_1; beyond two, marginal recall gains are offset by precision loss.
   ii. **Similarity threshold** τ: Precision improves smoothly with τ; recall stays nearly constant; F_1 saturates at $\tau \geq 0.70$.
   iii. **Candidate retention, consensus, explicitness**: All three have negligible impact once semantic filtering is active. Full results in Appendix B.1.1.

5. Statistical Significance Tests
   Paired bootstrap resampling [65] ($n = 10{,}000$ iterations) and McNemar's test [66] were applied to assess whether observed F_1 differences between pipeline systems reflect genuine improvements rather than sampling variation. Three pairwise comparisons were evaluated: BL vs. BL+IG, BL vs. BL+IG+PARA, and BL+IG vs. BL+IG+PARA. Each comparison was tested across three evaluation views: all rows, aspect-extraction only (rows with at least one non-null gold aspect), and explicit rows only. The KG component was evaluated separately, comparing BL against KG on implicit rows only, using the ASPECT part of the `entity#aspect` category label as gold. Full numerical results are reported in Appendix B.2.
   BL→BL+IG and BL→BL+IG+PARA are bootstrap-significant ($p<0.001$) across all views on SE14, SE15, and SE16, confirming that both IG and the full pipeline deliver genuine gains over the baseline. The marginal step BL+IG→BL+IG+PARA is additionally significant on SE14 ($p<0.001$, explicit rows), making it the only dataset where every pairwise comparison is bootstrap-confirmed; on SE15 it does not reach significance, indicating partial redundancy between IG and PARA on that dataset. On ACOS, BL is the best system and all additions degrade aggregate performance; bootstrap p-values near 1.0 confirm the degradation is not due to chance, while a significant McNemar result indicates the components shift which sentences are correctly predicted without improving the overall score. On M-ABSA$_{en}$, BL+IG alone does not improve significantly over BL by bootstrap, whereas BL→BL+IG+PARA and BL+IG→BL+IG+PARA are both significant ($p \leq 0.003$), suggesting that the paraphrase component drives the improvement and IG is insufficient on its own. The KG component is bootstrap-significant on all four eligible English datasets ($\Delta F_1 = +0.28$ to $+0.37$, $p<0.001$; Table A8).

5.2.2. Proposed Solution Evaluation on Arabic Datasets

1. Ablation Study

 doi:10.20944/preprints202605.1240.v1

Table 12. Explicit aspect extraction performance on Arabic datasets. ZS: Zero-Shot; BL: fine-tuned baseline; BL+IG: baseline with iterative generation; BL+IG+Para: baseline with iterative generation and paraphrase-based input variation.

	M-ABSA$_{ar}$			SE16			HAAD		
Method	***P***	***R***	***F*$_1$**	***P***	***R***	***F*$_1$**	***P***	***R***	***F*$_1$**
ZS	0.0482	0.0476	0.0479	0.0702	0.0363	0.0478	0.1818	0.0952	0.125
BL	0.7542	0.5907	0.6625	0.7950	0.7339	0.7632	0.4708	0.4378	0.4537
BL+IG	0.8094	0.7264	0.7656	0.6383	0.8249	0.7197	0.3712	0.6462	0.4716
BL+IG+Para	0.7768	0.7441	**0.7601**	0.6630	0.8210	**0.7336**	0.4098	0.5692	**0.4765**

Table 12 reports micro-F_1 on explicit rows across the three Arabic datasets. On M-ABSA$_{ar}$, iterative generation delivers the largest gain, driven by improved recall as the iterative component recovers multiple aspect mentions missed in single-pass decoding. The paraphrase component moderates precision, with the best result at $F_1 = 0.760$. Notably, +IG alone outperforms the full pipeline, indicating that Arabic paraphrase variants introduce more spurious candidates than in English — a consequence of greater morphological ambiguity reducing paraphrase precision. On HAAD, the pattern is consistent: +IG improves recall most substantially ($R : 0.438 \rightarrow 0.646$), while +Para recovers precision ($P : 0.371 \rightarrow 0.410$), yielding the best $F_1 = 0.477$. On SE16, the baseline achieves the highest precision ($P = 0.795$, $F_1 = 0.763$). Iterative generation expands recall substantially ($R : 0.734 \rightarrow 0.825$) at a notable precision cost, and +Para partially recovers precision while maintaining high recall ($F_1 = 0.734$). The SE16 hotel reviews average 2.05 gold aspects per explicit row, explaining why iterative generation is particularly effective at recovering additional aspects despite increased false positives.

2. Performance Evaluation

(a) Explicit vs. Implicit Performance

Table 13. Explicit vs. implicit pipeline performance on Arabic datasets (micro-F_1, BL+IG+Para). Implicit rows have NONE gold aspect; correct abstention is counted as TP. Mixed rows (M-ABSA$_{ar}$ only) contain one explicit and one implicit aspect per sentence.

Dataset	**Type**	**BL**	**+IG**	**+Para**	**Δ vs BL**
M-ABSA$_{ar}$	Explicit (1,356)	0.7542	0.8094	**0.7768**	+0.0226
	Implicit (718)	0.9392	0.9245	**0.9332**	−0.0060
	Mixed (104)	0.7586	0.8130	**0.6667**	−0.0919
SE16	Explicit (1,130)	0.7950	0.6383	**0.6630**	−0.1320
	Implicit (96)	0.8862	0.8683	**0.8299**	−0.0563
HAAD	Explicit (299)	0.4708	0.3712	**0.4098**	−0.0610

Table 13 shows per-type performance on M-ABSA$_{ar}$ under BL+IG+Para. The pipeline achieves $F_1 = 0.760$ on explicit rows and $F_1 = 0.671$ on mixed rows. The implicit abstention score ($F_1 = 0.933$) reflects the system's reliable identification of sentences with no surface aspect, consistent with the English results ($F_1 = 0.83$–0.94). Mixed-row performance is lower than explicit-only rows because mixed sentences require the system to simultaneously extract a surface aspect and correctly output NULL for the implicit component — the latter is addressed separately by the KG module.

(b) Implicit Aspect Identification via Knowledge Graph

The Arabic KG is evaluated against ENTITY#ASPECT category annotations independently of explicit extraction. Two strategies are compared: Strategy 1 uses the English KG with multilingual embeddings; Strategy 2 uses direct lexical lookup on the Arabic KG. HAAD is excluded as its test split contains no implicit aspects.

Table 14 shows Strategy 2 outperforms Strategy 1 by $\Delta F_1 = 0.211$ on M-ABSA$_{ar}$ and $\Delta F_1 = 0.193$ on SE16, gaining on both precision and recall. The improvement stems

 doi:10.20944/preprints202605.1240.v1

from strict domain filtering reducing cross-domain noise, and the Arabic clue lexicon with verb-phrase fallback recovering more implicit rows than cross-lingual embedding similarity alone.

Table 14. Strategy 1 (cross-lingual English KG) vs. Strategy 2 (Arabic domain-specific KG) for implicit aspect identification.

Dataset	Strategy	P	R	F_1
M-ABSA$_{ar}$	S1	0.4841	0.0951	0.1590
	S2	0.6060	0.2541	**0.3580**
SE16	S1	0.4545	0.1429	0.2174
	S2	0.6981	0.3524	**0.4684**

Table 15. Arabic KG inference step distribution across implicit rows (%).

Dataset	`STOPPED`	`INFERRED`	`NO_MATCH`
M-ABSA$_{ar}$	52.1	45.1	2.8
SE16	42.7	53.1	4.2

The Arabic KG achieves high precision when it fires ($P = 0.61$–0.70), comparable to the English KG ($P = 0.68$–0.71). Recall is lower ($R = 0.25$–0.35) due to high STOPPED rates (Table 15): 52.1% of M-ABSA$_{ar}$ and 42.7% of SE16 implicit rows produce no prediction, reflecting Arabic morphological complexity and the absence of detectable adjective or sentiment verb cues. The verb-phrase fallback contributes substantially: on M-ABSA$_{ar}$, 23.7% of implicit rows are recovered via sentiment verbs such as أوصي and أعجبني; on SE16, the `INFERRED` share reaches 53.1%, reflecting the prevalence of recommendation expressions in hotel reviews.

Table 16. Per-domain KG implicit identification performance on M-ABSA$_{ar}$ implicit rows (Strategy 2).

Domain	P	R	F_1
Hotel	0.7778	0.3333	**0.4667**
Restaurant	0.6338	0.2830	0.3913
Food	0.5333	0.2149	0.3064
Laptop	0.4792	0.1983	0.2805

Table 17. Evaluation on M-ABSA$_{ar}$ mixed sentences (104 rows). Metric A: full system explicit extraction; Metric B: KG implicit identification (Strategy 2, partial matching).

Metric	Method	P	R	F_1
A (Explicit)	BL+IG+Para+KG	0.6736	0.6831	0.6783
B (Implicit)	S2 KG	0.6071	0.2629	0.3669

(c) Domain-level Performance (M-ABSA$_{ar}$)

The Hotel domain achieves the highest $F_1 = 0.467$ with $P = 0.778$, reflecting strong alignment between the hotel-focused SemEval-2016 training edges in the Arabic KG and the hotel domain reviews in M-ABSA$_{ar}$. Restaurant and Food domains perform moderately ($F_1 = 0.391$ and 0.306 respectively), while Laptop is the weakest ($F_1 = 0.281$), consistent with the English results: laptop implicit aspects require highly specific technical vocabulary — *operation_performance*, *design_features*, *portability* — that is underrepresented in the Arabic KG's adjective lexicon, which was primarily enriched from hotel and restaurant review corpora.

 doi:10.20944/preprints202605.1240.v1

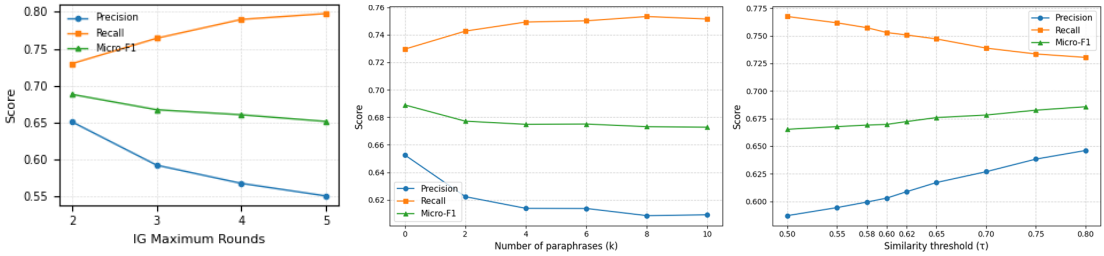
(d) KG performance on Mixed Sentence
Mixed sentences contain both an explicit and an implicit aspect in the same sentence, requiring the solution to handle both tasks simultaneously (Table 17). Explicit extraction achieves $F_1 = 0.678$, a modest decline from $F_1 = 0.777$ on purely explicit rows, suggesting that competing opinion signals in mixed sentences introduce some additional difficulty for aspect boundary identification. Implicit identification achieves $F_1 = 0.367$, closely matching the $F_1 = 0.370$ obtained on purely implicit rows, confirming that the lonely adjective detection mechanism operates independently of whether an explicit aspect is also present. Mixed sentences are therefore a moderately harder case for explicit extraction but not for implicit identification.

3. Error Analysis

- Boundary Mismatch. Boundary errors account for 2–8% of Arabic false positives, where Arabic morphological structure causes the model to either include genitive complements (superspan) or omit clitic-bearing head nouns (subspan).
- Arabic Annotation Inconsistency. The Arabic datasets exhibit uneven granularity: fine-grained aspects (e.g. Prices, Cleanliness) coexist with coarse General annotations that collapse multiple implicit aspects. Table 18 illustrates cases where the solution correctly identifies specific aspects but is penalised against the coarser gold label. Generic opinion adjectives with no specific aspect noun constitute the largest false positive category on M-ABSA$_{ar}$ (47%, $n = 277$), SemEval-2016 Arabic (24%, $n = 302$), and HAAD (22%, $n = 108$). Nevertheless, General-label noise remains non-trivial because the boundary between general sentiment and fine-grained aspects is inconsistently drawn in the annotations themselves.
- KG Coverage: No-Cue Detection (STOPPED rows)
STOPPED rates of 52.1% (M-ABSA$_{ar}$) and 42.7% (SE16) arise from three structural categories: (i) purely factual or nominal sentences with no sentiment marker (e.g., حزم ثلاث طلبت, gold: `food#general`), where the aspect is inferred from purchase behaviour rather than linguistic cues; (ii) exclamatory expressions (e.g., !أووووه) conveying sentiment through pragmatic cues alone; and (iii) comparative constructions (e.g., أمي طبخ بعد شيء أفضل هو المكان هذا), where the comparative adjective modifies a noun phrase and is correctly blocked by the lonely-adjective gate. All three require discourse-level or pragmatic reasoning beyond lexical KG inference.
Table 19 summarizes errors across all English and Arabic datasets. English FPs are dominated by spurious hallucination (84–93%), with boundary errors secondary in the laptop domain. Arabic FPs are more varied: General-label noise dominates on M-ABSA$_{ar}$, clitic duplication on HAAD, and spurious hallucination on SE16-Arabic. False negatives are consistent across datasets: 76–85% were never predicted by any stage; 14–24% were correctly identified by the baseline but dropped by later voting stages

4. Sensitivity Analysis:

 doi:10.20944/preprints202605.1240.v1



(**a**) (**b**) (**c**)

Figure 4. Sensitivity analysis of the proposed model. (**a**) Effect of iterative generation rounds. (**b**) Effect of number of paraphrases. (**c**) Effect of semantic similarity threshold.

(a) **Iterative Generation Component Parameter:** Figure 4a: Unlike English, increasing rounds consistently degrades micro-F_1 in Arabic. Additional rounds surface noisy candidates, causing a sharp precision drop that outweighs recall gains.

(b) **Paraphrase-Aided Component Parameters:**

i. **Number of paraphrases:** Figure 4b: Adding paraphrases increases recall but causes a monotonic precision and micro-F_1 decline. Unlike English, the optimal Arabic configuration uses no or very few paraphrases, as paraphrasing amplifies lexical noise rather than providing semantic reinforcement.

ii. **Similarity threshold:** Figure 4c: Raising the threshold improves precision while recall decreases gradually, yielding a monotonic micro-F_1 gain. Conservative semantic filtering is therefore essential for suppressing paraphrase-induced noise in Arabic.

(c) On M-ABSA$_{ar}$, BL→BL+IG and BL→BL+IG+PARA are highly significant across all views ($p<0.001$), confirming that both components contribute meaningful gains; however, BL+IG→BL+IG+PARA is not significant by either test, indicating PARA provides no additional benefit once IG is applied. On SE16$_{ar}$, both BL→BL+IG and BL→BL+IG+PARA are associated with statistically significant *degradation* (McNemar $p<0.001$, bootstrap $p\approx1.0$), reflecting a precision collapse under generative augmentation in morphologically rich Arabic; BL remains the best system. The only positive comparison on SE16$_{ar}$, BL+IG→BL+IG+PARA ($\Delta F_1 = +0.019$, $p<0.001$), shows that PARA partially recovers from the damage introduced by IG, but not sufficiently to surpass BL. On HAAD, no comparison reaches bootstrap significance ($p = 0.05$–0.10), attributable to the small evaluation set ($n = 299$); a significant McNemar result nonetheless confirms that the components alter error patterns. Despite the inconsistent pipeline behavior across Arabic datasets, the KG component is bootstrap-significant on both datasets ($\Delta F_1 = +0.37$ and $+0.41$, $p<0.001$; Table A10), establishing it as the primary and most reliable mechanism for implicit aspect identification in Arabic.

Table 18. Annotation inconsistency in the Arabic SemEval 2016 dataset: the pipeline identifies specific aspects that are collapsed into General in the gold annotation.

Example	Prediction	Gold aspect
مزعجة ونوادي سيء **استقبال**	COMFORT, SERVICE	GENERAL
والإستجمام الراحة مسافر كل مطلب بالوعود الوفاء وعدم التعامل في المستوى دون	COMFORT, CLEANLINESS	GENERAL
للنوم جدًا **مزعج**	COMFORT	GENERAL

 doi:10.20944/preprints202605.1240.v1

Table 19. False positive and false negative breakdown across all datasets (explicit rows; full test sets). FP categories: Spur. = spurious hallucination; Bound. = boundary error; Gen. = General-label noise; Clit. = clitic duplication (Arabic only). FN categories: Never = never predicted; Lost = predicted by baseline but filtered later.

							FP breakdown (%)				FN breakdown (%)	
Dataset	***n***	**TP**	**FP**	**FN**	***P***	***R***	**Spur.**	**Bound.**	**Gen.**	**Clit.**	**Never**	**Lost**
SE14 (EN)	1028	1389	91	391	0.939	0.780	63	37	0	—	85	14
SE15 (EN)	605	474	116	324	0.803	0.594	93	7	0	—	85	14
SE16 (EN)	419	478	131	133	0.785	0.782	89	11	0	—	79	19
ACOS Rest. (EN)	412	378	200	34	0.654	0.917	93	7	0	—	76	24
ACOS Lap. (EN)	801	727	267	74	0.731	0.908	84	16	0	—	77	14
MABSA (EN)	2576	2307	951	1654	0.708	0.582	91	8	1	—	78	15
MABSA (AR)	1356	1280	594	439	0.683	0.745	31	8	47	15	78	18
SE16 (AR)	1130	1912	1248	411	0.605	0.823	45	2	24	29	77	18
HAAD (AR)	299	326	487	245	0.401	0.571	36	2	22	40	79	16

5.3. Comparative Analysis

The comparative analysis validates the core design objective of the proposed system: robust aspect identification encompassing both explicit and implicit expressions across multiple domains and benchmarks.

1. English datasets Comparison results

Table 20. Performance comparison with state-of-the-art methods on English benchmark datasets. Target: Explicit = explicit aspects only, Both = explicit and implicit aspects. —: not reported by the study.

Study	**Dataset**	**Domain**	**Target**	***P***	***R***	F_1
SemEval-2014						
[39]	SE-14	Lap.	Explicit	—	—	0.8635
[67]	SE-14	Rest.	Explicit	0.8481	0.8619	0.8693
[67]	SE-14	Lap.	Explicit	—	—	0.8299
[68]	SE-14	Rest.	Explicit	0.7987	0.6603	0.7230
[59]	SE-14	Lap.	Both 0.91	0.72	0.81	
[59]	SE-14	Rest.	Both	0.88	0.77	0.82
[10]	SE-14	Combined	Explicit	—	—	0.8200
[9]	SE-14	Rest.	Explicit	—	—	0.7596
[9]	SE-14	Lap.	Explicit	—	—	0.5174
Ours	SE-14	Rest.	Both	**0.8909**	**0.7869**	**0.8357**
Ours	SE-14	Lap.	Both	**0.9014**	**0.8508**	**0.8754**
SemEval-2015						
[67]	SE-15	Rest.	Explicit	—	—	0.7551
[58]	SE-15	Rest.	Both	—	—	0.7286
Ours	SE-15	Rest.	Both	**0.7402**	**0.6878**	**0.7130**
SemEval-2016						
[39]	SE-16	Rest.	Explicit	—	—	0.8163
[57]	SE-16	Rest.	Both	0.912	0.8953	0.91
[58]	SE-16	Rest.	Both	—	—	0.8113
Ours	SE-16	Rest.	Both	**0.7839**	**0.7879**	**0.7859**
ACOS						
[39]	ACOS	Rest.	Explicit	0.8054	0.8208	0.8163
[39]	ACOS	Lap.	Explicit	0.8635	0.8635	0.8635
Ours	ACOS	Rest.	Both	**0.7085**	**0.8974**	**0.7919**
Ours	ACOS	Lap.	Both	**0.7193**	**0.8992**	**0.7993**

Table 20 validates the proposed solution across both explicit and implicit aspect identification. On SemEval-2014, the solution outperforms [9] and [68] on Restaurant, and surpassing [67] and [39] on Laptop. The closest competitor targeting both aspect types, [59], both slightly below our results. On SemEval-2015, the solution falls short of [67] (explicit only) and [58] , reflecting the additional challenge of targeting implicit aspects in a smaller dataset. On SemEval-2016, [57] achieves the highest F_1; the gap is partly attributable to differences in evaluation scope and

 doi:10.20944/preprints202605.1240.v1

annotation coverage across studies. On ACOS, the solution substantially exceeds [39] in recall (explicit only), demonstrating the effectiveness of iterative and paraphrase-based expansion for increasing aspect coverage. The lower precision relative to explicit-only systems reflects the inherent trade-off of targeting both aspect types.

2. Arabic datasets Comparison results

Table 21. Performance comparison with state-of-the-art methods on Arabic benchmark datasets. Target: Explicit = explicit aspects only; Both = explicit and implicit aspects. Subtask: ATE = Aspect Term Extraction; ACD = Aspect Category Detection. −: not reported.

Study	Dataset	Target	Subtask	P	R	F_1
SemEval-2016 Arabic						
[28]	SE-16	Explicit	ATE	0.8096	0.7970	0.8032
[29]	SE-16	Explicit	ACD	–	–	0.7100
[42]	SE-16	Explicit	ACD	–	–	0.6610
Ours	SE-16	**Both**	**ATE**	**0.6630**	**0.8210**	**0.7336**
HAAD						
[11]	HAAD	Explicit	ACD	–	–	0.3200
[13]	HAAD	Explicit	ACD	–	–	0.4830
[41]	HAAD	Explicit	ATE	–	–	0.3447
[42]	HAAD	Explicit	ATE	–	–	0.661
Ours	HAAD	**Both**	**ATE**	**0.4098**	**0.5692**	**0.4765**

Table 21 compares the proposed solution against SOTA Arabic methods. On SE16, explicit-only systems report higher F_1 — [28] but are trained and evaluated on explicit aspects only, with no requirement to handle implicit sentences. Our system achieves higher recall, confirming that iterative generation effectively surfaces multiple aspects per sentence, at the cost of additional false positives. Against category-detection systems [29,42], the system is competitive while additionally identifying implicit aspects. On HAAD, the system exceeding [11]despite the challenging KG coverage of Books domain. Strong recall ($R = 0.569$) reflects iterative generation's robustness under domain shift; lower precision ($P = 0.410$) indicates paraphrase-induced noise outside the training domain.

5.4. Limitations

The error analysis (Sections 3 and 3) reveals four system-level limitations that transcend individual datasets.

1. Pragmatic implicit aspects: The lonely adjective gate and verb-clue fallback address surface-level implicit signals but cannot handle sentiment expressed through purchase behaviour, exclamatory fragments, or figurative language. Resolving these cases requires discourse-level or pragmatic reasoning beyond the scope of lexical KG inference.
2. Boundary identification: Generative decoding produces subspans and superspans of gold aspect terms, particularly in morphologically complex Arabic constructions. Constrained decoding or span-extraction post-processing could reduce this failure mode.
3. Arabic annotation noise. The inconsistent granularity of Arabic gold annotations — where both fine-grained and coarse General labels appear for structurally similar sentences — introduces irreducible noise that system design alone cannot resolve. Standardising Arabic ABSA annotation guidelines is a prerequisite for progress on this front.
4. Domain vocabulary coverage. The KG's clue lexicon was built from hotel, restaurant, and food corpora. Laptop, Coursera, and sight domains are underrepresented, leading to systematic false negatives on technical and non-restaurant implicit aspects. Expanding the KG through domain-adaptive construction from larger Arabic and English corpora is the most direct mitigation.

 doi:10.20944/preprints202605.1240.v1

6. Conclusion

This paper presented a multi-component solution for implicit aspect extraction combining a fine-tuned generative backbone, iterative residual generation, paraphrase-based input transformation, and a domain-specific knowledge graph governed by linguistically motivated gating. The central finding is that no single component is sufficient: the KG component provides gains precisely where the generative backbone returns NULL — sentences with isolated opinion adjectives and no syntactic noun head — while iterative generation recovers multi-aspect sentences that single-pass decoding misses. Ablation results confirm that this complementarity holds across both language families.

Experiments across eight datasets demonstrate strong performance, outperforming strong baselines and recent state-of-the-art methods on six of eight datasets. On Arabic, primary implicit aspect evidence comes from SemEval-2016 Hotels and M-ABSA$_{ar}$; HAAD is evaluated for explicit extraction only, as its annotation scheme does not mark implicit aspects. The English–Arabic performance gap reflects greater annotation consistency, stronger boundary alignment, and lower morphological complexity in English.

A consistent finding across both languages is that strict span-based evaluation penalizes semantically correct predictions that differ lexically or boundary-wise from gold annotations, systematically underestimating generative model performance. Semantic equivalence metrics are a prerequisite for fairly benchmarking generative ABSA systems, and their development constitutes a natural next step for the field.

Two directions follow directly from these results. First, extending the Arabic KG to improve morphological coverage across dialects and derivational forms, increasing robustness beyond the MSA-dominant training corpora used here. Second, investigating cross-lingual transfer to bootstrap implicit aspect extraction in lower-resource Arabic domains where annotated data remains scarce.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ABSA	Aspect-Based Sentiment Analysis
ATE	Aspect Term Extraction
BL	Baseline component
IG	Iterative Generation component
PARA	Paraphrase-based input variation component
KG	Knowledge Graph
LLM	Large Language Model
LoRA	Low-Rank Adaptation
MSA	Modern Standard Arabic
NLP	Natural Language Processing
SA	Sentiment Analysis

Appendix A. Knowledge Graph Notation

Appendix A.1. Formal Notation

- **Graph Definition.** Let $G = (V, E)$ be a labeled multigraph with four vertex types:

$$V = V_d \cup V_e \cup V_a \cup V_c,$$

 doi:10.20944/preprints202605.1240.v1

representing *domains*, *entities*, *aspects*, and *clues* respectively, connected by three typed relations:

$$R_{\mathrm{c}\to\mathrm{a}} \subseteq V_c \times V_a \text{ (clue describes aspect)},$$
$$R_{\mathrm{a}\to\mathrm{e}} \subseteq V_a \times V_e \text{ (aspect belongs to entity)},$$
$$R_{\mathrm{e}\to\mathrm{d}} \subseteq V_e \times V_d \text{ (entity relevant to domain)}.$$

- **Candidate Retrieval.** Given a matched clue node $c \in V_c$, aspect candidates are retrieved as:

$$C = N_G^+(c; R_{\mathrm{c}\to\mathrm{a}}) = \{\, u \in V_a \mid (c,u) \in R_{\mathrm{c}\to\mathrm{a}} \,\}.$$

- **Domain-Consistent Filtering.** Domain consistency is enforced via relation composition:

$$R_{\mathrm{a}\to\mathrm{d}} = R_{\mathrm{a}\to\mathrm{e}} \circ R_{\mathrm{e}\to\mathrm{d}},$$

yielding the domain-filtered candidate set for inferred domain $d \in V_d$:

$$C_d = \{\, a \in C \mid (a,d) \in R_{\mathrm{a}\to\mathrm{d}} \,\}.$$

- **Scoring and Selection.** Each $a \in C_d$ receives a relevance score combining clue–aspect and sentence–aspect similarity:

$$\mathrm{score}(a) = \mathrm{Sim}(\mathrm{Embed}(c), \mathrm{Embed}(a)) + \mathrm{Sim}(\mathrm{Embed}(s), \mathrm{Embed}(a)), \tag{A1}$$

where $\mathrm{Embed}(\cdot)$ denotes SBERT embeddings and $\mathrm{Sim}(\cdot,\cdot)$ cosine similarity. The implicit aspect set is then:

$$I = \operatorname*{arg\,topk}_{a \in C_d,\ \mathrm{score}(a) \geq \tau_{\mathrm{asp}}} \mathrm{score}(a).$$

Appendix A.2. Algorithm

Algorithm 2 operationalizes the formal definitions above. Lonely adjectives are detected and mapped to clue nodes V_c; the sentence domain is inferred by selecting the most semantically similar domain vertex. For each clue, candidates are retrieved via $R_{\mathrm{c}\to\mathrm{a}}$, filtered for domain consistency through $R_{\mathrm{a}\to\mathrm{d}}$, and scored via Eq. A1; if no candidate survives filtering, a *general* fallback is assigned. Top-ranked aspects above τ_{asp} are normalized, deduplicated, and returned as the implicit aspect set I.

Algorithm 2: KG-Based Implicit Aspect Extraction

Input : s: input sentence; P: paraphrased sentences; $G = (V, E)$: labeled digraph; $D \subseteq V$: domain vertices
Output : $I \subseteq V_a$: implicit aspect set
Variables : $L \subseteq V_c$: matched clue vertices; $C, C_d \subseteq V_a$: candidate and filtered aspects; $d \in D$: inferred domain

$L \leftarrow \mathrm{IdentifyClues}(\{s\} \cup P) \cap V_c$; // Detect lonely adjectives and map to clue nodes
$d \leftarrow \arg\max_{d' \in D} \mathrm{Sim}(\phi(s), \phi(d'))$; // Infer sentence domain
$I \leftarrow \emptyset$;
foreach $c \in L$ **do**
  $C \leftarrow \{\, a \in V_a \mid (c,a) \in R_{\mathrm{c}\to\mathrm{a}} \,\}$; // Retrieve aspect candidates
  $C_d \leftarrow \{\, a \in C \mid \exists e : (a,e) \in R_{\mathrm{a}\to\mathrm{e}} \wedge (e,d) \in R_{\mathrm{e}\to\mathrm{d}} \,\}$; // Domain-consistent filtering
  if $C_d = \emptyset$ **then** $C_d \leftarrow \{general\}$;
  foreach $a \in C_d$ **do**
    $\mathrm{score}(a) \leftarrow \mathrm{Sim}(c,a) + \mathrm{Sim}(s,a)$; // Eq. A1
  $I \leftarrow I \cup \operatorname{arg\,topk}_{a \in C_d,\ \mathrm{score}(a) \geq \tau_{\mathrm{asp}}} \mathrm{score}(a)$;
$I \leftarrow \mathrm{FilterAndNormalize}(I)$;
return I

 doi:10.20944/preprints202605.1240.v1

Appendix B. Sensitivity and Statistical Significance Tests

Appendix B.1. Sensitivity Analysis

All sensitivity experiments use the same gold-annotated test splits as the main evaluation, varying only inference-time parameters with no retraining. Each parameter is adjusted independently while all others are fixed at their defaults. Performance is measured using micro-P, R, and F_1 as the main evaluation.

Appendix B.1.1. Sensitivity Analysis on English Datasets

Tables A1–A3 report the numerical results. The English results are largely stable across hyperparameters; semantic similarity thresholding and iterative generation depth show the clearest monotonic effects, while remaining parameters introduce only marginal variation.

Table A1. Paraphrases sensitivity.

Setting	P	R	F_1
Original (k=0)	0.9082	0.7542	0.8241
k=2	0.9042	0.7602	0.8260
k=4	0.8994	0.7619	0.8250
k=6	0.9016	0.7610	0.8254
k=8	0.8995	0.7623	0.8252
k=10	0.8970	0.7602	0.8229

Table A2. Similarity threshold sensitivity.

Thr.	P	R	F_1
0.50	0.8264	0.7686	0.7965
0.55	0.8493	0.7640	0.8044
0.58	0.8601	0.7631	0.8087
0.60	0.8720	0.7623	0.8135
0.62	0.8764	0.7602	0.8142
0.65	0.8835	0.7585	0.8162
0.70	0.8931	0.7572	0.8195
0.75	0.9024	0.7559	0.8227
0.80	0.9069	0.7551	0.8240

Table A3. IG rounds sensitivity.

Rounds	P	R	F_1
2	0.9278	0.5822	0.7154
3	0.9180	0.6453	0.7579
4	0.9089	0.6682	0.7702
5	0.9053	0.6763	0.7742

Appendix B.1.2. Sensitivity Analysis on Arabic Dataset

Tables A4–A6 summarize the numerical sensitivity results for Arabic dataset.In contrast to English, Arabic sensitivity trends exhibit stronger trade-offs between precision and recall. Paraphrase quantity and iterative generation depth significantly increase recall but introduce substantial noise.

 doi:10.20944/preprints202605.1240.v1

Table A4. Paraphrases sensitivity.

Setting	P	R	F_1
k=0	0.6526	0.7296	0.6890
k=2	0.6223	0.7427	0.6772
k=4	0.6138	0.7493	0.6749
k=6	0.6137	0.7502	0.6751
k=8	0.6085	0.7533	0.6732
k=10	0.6091	0.7515	0.6728

Table A5. Similarity threshold sensitivity.

Thr.	P	R	F_1
0.50	0.5871	0.7674	0.6653
0.55	0.5944	0.7617	0.6677
0.58	0.5994	0.7573	0.6691
0.60	0.6030	0.7529	0.6697
0.62	0.6088	0.7507	0.6723
0.65	0.6171	0.7471	0.6759
0.70	0.6269	0.7388	0.6782
0.75	0.6383	0.7335	0.6826
0.80	0.6461	0.7304	0.6857

Table A6. IG rounds sensitivity.

Rounds	P	R	F_1
2	0.6512	0.7300	0.6884
3	0.5926	0.7643	0.6676
4	0.5680	0.7898	0.6608
5	0.5510	0.7977	0.6518

Appendix B.2. Statistical Significance Tests

Table A7 reports significance results for English datasets, and Table A8 reports the KG component significance on implicit rows. SemEval2014 is excluded from KG evaluation owing to absent category annotations.

Table A9 reports significance results for Arabic datasets, and Table A10 reports the KG component on implicit rows. HAAD is excluded from KG evaluation owing to absent category annotations.

 doi:10.20944/preprints202605.1240.v1

Table A7. Statistical significance of pipeline components on English datasets. Bootstrap p (Boot.) and McNemar p (McN.) reported for three evaluation views: all rows (All), aspect-extraction only (Asp.), explicit only (Exp.). F_1 scores and ΔF_1 shown for the explicit-only view. $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$, — not significant. Negative ΔF_1 indicates degradation relative to system A.

		All		Asp.		Exp.		Exp. F_1		
Dataset	**Comparison**	**Boot.**	**McN.**	**Boot.**	**McN.**	**Boot.**	**McN.**	**F_1(A)**	**F_1(B)**	**ΔF_1**
SE14 (n=1028)	BL → BL+IG	***	***	***	***	***	***	0.6202	0.7921	+0.1719
	BL → BL+IG+PARA	***	***	***	***	***	***	0.6202	0.8228	+0.2026
	BL+IG → BL+IG+PARA	*	—	***	—	***	—	0.7921	0.8228	+0.0306
SE15 (n=605)	BL → BL+IG	***	*	***	**	***	**	0.5256	0.6005	+0.0749
	BL → BL+IG+PARA	***	*	***	**	***	**	0.5256	0.6207	+0.0950
	BL+IG → BL+IG+PARA	—	—	—	—	—	—	0.6005	0.6207	+0.0202
SE16 (n=419)	BL → BL+IG	***	*	***	**	***	**	0.6822	0.7880	+0.1058
	BL → BL+IG+PARA	***	—	***	—	***	—	0.6822	0.7628	+0.0807
	BL+IG → BL+IG+PARA	—	**	—	**	—	**	0.7880	0.7628	−0.0251
ACOS (n=1213)	BL → BL+IG	—	***	—	***	—	***	0.8175	0.7870	−0.0305
	BL → BL+IG+PARA	—	***	—	***	—	***	0.8175	0.7824	−0.0352
	BL+IG → BL+IG+PARA	—	***	—	***	—	***	0.7870	0.7824	−0.0046
M-ABSA$_{en}$ (n=2576)	BL → BL+IG	—	**	—	*	—	*	0.5719	0.5824	+0.0106
	BL → BL+IG+PARA	**	—	***	—	***	—	0.5719	0.5993	+0.0274
	BL+IG → BL+IG+PARA	***	**	**	**	**	**	0.5824	0.5993	+0.0169

Table A8. Statistical significance of the KG component on English implicit rows. Gold = ASPECT part of `entity#aspect`. SE14 excluded (no category annotations). $^{***}p<0.001$.

Dataset	**N_{impl}**	**F_1(BL)**	**F_1(KG)**	**ΔF_1**	**Boot.**	**McN.**
SE15$_{en}$	245	0.0000	0.3218	+0.3218	***	***
SE16$_{en}$	167	0.0000	0.3739	+0.3739	***	***
ACOS$_{en}$	344	0.0058	0.2885	+0.2827	***	***
M-ABSA$_{en}$	913	0.0010	0.3016	+0.3005	***	***

Table A9. Statistical significance of pipeline components on Arabic datasets. $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$, — not significant. Negative ΔF_1 indicates degradation relative to system A.

		All		**Asp.**		**Exp.**		**Exp. F_1**		
Dataset	**Comparison**	**Boot.**	**McN.**	**Boot.**	**McN.**	**Boot.**	**McN.**	**F_1(A)**	**F_1(B)**	**ΔF_1**
M-ABSA$_{ar}$ (n=1356)	BL → BL+IG	***	***	***	***	***	***	0.6559	0.7744	+0.1185
	BL → BL+IG+PARA	***	***	***	***	***	**	0.6559	0.7642	+0.1083
	BL+IG → BL+IG+PARA	—	—	—	—	—	—	0.7744	0.7642	−0.0103
SE16$_{ar}$ (n=1130)	BL → BL+IG	—	***	—	***	—	***	0.7676	0.7156	−0.0521
	BL → BL+IG+PARA	—	***	—	***	—	***	0.7676	0.7343	−0.0333
	BL+IG → BL+IG+PARA	***	—	***	—	***	—	0.7156	0.7343	+0.0187
HAAD (n=299)	BL → BL+IG	—	***	—	***	—	***	0.4465	0.4720	+0.0254
	BL → BL+IG+PARA	—	***	—	***	—	***	0.4465	0.4743	+0.0278
	BL+IG → BL+IG+PARA	—	—	—	—	—	—	0.4720	0.4743	+0.0024

Table A10. Statistical significance of the KG component on Arabic implicit rows. Gold = ASPECT part of `entity#aspect`. HAAD excluded (no category annotations). $^{***}p<0.001$.

Dataset	**N_{impl}**	**F_1(BL)**	**F_1(KG)**	**ΔF_1**	**Boot.**	**McN.**
SE16$_{ar}$	96	0.0000	0.4099	+0.4099	***	***
M-ABSA$_{ar}$	718	0.0000	0.3725	+0.3725	***	***

References

1. Karagoz, P.; Kama, B.; Ozturk, M.; Toroslu, I.H.; Canturk, D. A framework for aspect based sentiment analysis on Turkish informal texts. *Journal of Intelligent Information Systems* **2019**, *53*, 431–451. https://doi.org/10.1007/S10844-019-00565-W.

 doi:10.20944/preprints202605.1240.v1

2. Nandhini, M.D.S.; Pradeep, G. A Hybrid Co-occurrence and Ranking-based Approach for Detection of Implicit Aspects in Aspect-Based Sentiment Analysis. *SN Computer Science* **2020**, *1*. https://doi.org/10.1007/S42979-020-00138-7.
3. Al-Ayyoub, M.; Gigieh, A.; Al-Qwaqenah, A.; Al-Kabi, M.; Talafha, B.; Alsmadi, I. Aspect-Based Sentiment Analysis of Arabic Laptop Reviews. *The International Arab Journal of Information Technology* **2017**.
4. Alassaf, M.; Qamar, A.M. Aspect-Based Sentiment Analysis of Arabic Tweets in the Education Sector Using a Hybrid Feature Selection Method. In Proceedings of the 2020 14th International Conference on Innovations in Information Technology (IIT), 2020, pp. 178–185. https://doi.org/10.1109/IIT50501.2020.9299026.
5. Sana, T.; Boujelben, I.; Jamoussi, S.; Benayed, Y. A hybrid method for Arabic aspect-based sentiment analysis. *International Journal of Hybrid Intelligent Systems* **2020**, *16*, 1–12. https://doi.org/10.3233/HIS-200285.
6. Al-Smadi, M.; Hammad, M.M.; Al-Zboon, S.A.; Al-Tawalbeh, S.; Cambria, E. Gated Recurrent Unit with Multilingual Universal Sentence Encoder for Arabic Aspect-Based Sentiment Analysis. *Knowledge-Based Systems* **2021**, p. 107540. https://doi.org/10.1016/j.knosys.2021.107540.
7. Alghamdi, A.; Taileb, M.; Almani, N. Stacked LSTM-GRU with Double Attention model for Arabic Aspect-Based Sentiment Analysis. In Proceedings of the 2025 2nd International Conference on Advanced Innovations in Smart Cities (ICAISC). IEEE, May 2025. https://doi.org/10.1109/ICAISC64594.2025.10959531.
8. Aljomah, F.; Aldhafeeri, L.; Alfadel, M.; Alshahrani, S.; Abbas, Q.; Alhumoud, S. Enhancing Arabic Sentiment Analysis with Pre-Trained CAMeLBERT: A Case Study on Noisy Texts. *Computers, Materials & Continua* **2025**, *84*, 5318–5335. https://doi.org/10.32604/cmc.2025.062478.
9. Kalra, S. The Aspect Extraction using Topic-aware dynamic Convolutional Neural Network. In Proceedings of the 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON). IEEE, 2023, pp. 1–7. https://doi.org/10.1109/SMARTGENCON60755.2023.10442593.
10. Zhang, Z.; Rao, Y.; Lai, H.; Wang, J.; Yin, J. TADC: A Topic-Aware Dynamic Convolutional Neural Network for Aspect Extraction. *IEEE Transactions on Neural Networks and Learning Systems* **2023**, *34*, 3912–3924. https://doi.org/10.1109/TNNLS.2021.3119026.
11. Youssef, L.; Elhoussaine, Z.; Soufiane, N.; Noureddine, M. Enhancing Arabic aspect category detection using large language models (LLMs). *Results in Engineering* **2025**, *26*, 105049. https://doi.org/10.1016/j.rineng.2025.105049.
12. Alotaibi, A.; Nadeem, F. An Unsupervised Integrated Framework for Arabic Aspect-Based Sentiment Analysis and Abstractive Text Summarization of Traffic Services Using Transformer Models. *Smart Cities* **2025**, *8*, 62. https://doi.org/10.3390/smartcities8020062.
13. AlNasser, S.; AlMuhaideb, S. Listening to Patients: Advanced Arabic Aspect-Based Sentiment Analysis Using Transformer Models Towards Better Healthcare. *Big Data and Cognitive Computing* **2024**, *8*, 156. https://doi.org/10.3390/bdcc8110156.
14. Gadag, A.; Sagar, B.M. A review on different methods of paraphrasing. In Proceedings of the 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 2016, pp. 188–191. https://doi.org/10.1109/ICEECCOT.2016.7955212.
15. Titov, I.; McDonald, R. Modeling Online Reviews with Multi-grain Topic Models. In Proceedings of the Proceedings of the 17th International Conference on World Wide Web, Beijing, China, 2008; pp. 111–120. https://doi.org/10.1145/1367497.1367513.
16. Miller, G.A. WordNet: a lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. https://doi.org/10.1145/219717.219748.
17. Elkateb, S.; Black, W.; Rodríguez, H.; Alkhalifa, M.; Vossen, P.; Pease, A.; Fellbaum, C. Building a WordNet for Arabic. In Proceedings of the Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, 2006.
18. Speer, R.; Chin, J.; Havasi, C. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In Proceedings of the Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017, pp. 4444–4451.
19. Liao, J.; Wang, M.; Chen, X.; Wang, S.; Zhang, K. Dynamic Commonsense Knowledge Fused Method for Chinese Implicit Sentiment Analysis. *Information Processing & Management* **2022**, *59*, 102934. https://doi.org/10.1016/j.ipm.2022.102934.
20. Xu, M.; Wang, D.; Feng, S.; Yang, Z.; Zhang, Y. KC-ISA: An Implicit Sentiment Analysis Model Combining Knowledge Enhancement and Context Features. In Proceedings of the Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 6906–6915.
21. Liang, B.; Su, H.; Gui, L.; Cambria, E.; Xu, R. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge-Based Systems* **2022**, *235*, 107643.

 doi:10.20944/preprints202605.1240.v1

22. Teo, A.; Wang, Z.; et al. Knowledge Graph enhanced Aspect-Based Sentiment Analysis Incorporating External Knowledge. In Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW), 2023, pp. 791–798.
23. Dubey, G.; Dubey, A.K.; Kaur, K.; Raj, G.; Kumar, P. Adaptive Contextual Memory Graph Transformer with Domain-Adaptive Knowledge Graph for Aspect-Based Sentiment Analysis. *Expert Systems with Applications* **2025**, *278*, 127300. https://doi.org/10.1016/j.eswa.2025.127300.
24. Shaalan, K.; Siddiqui, S.; Alkhatib, M.; Monem, A.A., Challenges in Arabic Natural Language Processing. In *Computational Linguistics, Speech and Image Processing for Arabic Language*; Series on Language Processing, Pattern Recognition, and Intelligent Systems, 2018; chapter Chapter 3, pp. 59–83. https://doi.org/10.1142/9789813229396_0003.
25. Ruder, S.; Ghaffari, P.; Breslin, J.G. INSIGHT-1 at SemEval-2016 Task 5: Deep Learning for Multilingual Aspect-based Sentiment Analysis. In Proceedings of the Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016); Bethard, S.; Carpuat, M.; Cer, D.; Jurgens, D.; Nakov, P.; Zesch, T., Eds., San Diego, California, June 2016; pp. 330–336. https://doi.org/10.18653/v1/S16-1053.
26. Ruder, S.; Ghaffari, P.; Breslin, J.G. A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis. In Proceedings of the Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; Su, J.; Duh, K.; Carreras, X., Eds., Austin, Texas, November 2016; pp. 999–1005. https://doi.org/10.18653/v1/D16-1103.
27. Al-Dabet, S.; Tedmori, S.; AL-Smadi, M. Enhancing Arabic aspect-based sentiment analysis using deep learning models. *Computer Speech & Language* **2021**, *69*, 101224. https://doi.org/https://doi.org/10.1016/j.csl.2021.101224.
28. Fadel, A.; Saleh, M.; Salama, R.; Abulnaja, O. MTL-AraBERT: An Enhanced Multi-Task Learning Model for Arabic Aspect-Based Sentiment Analysis. *Computers* **2024**, *13*, 98. https://doi.org/10.3390/computers13040098.
29. Youssef, L.; ELhoussaine, Z. Arabic Aspect Category Detection using traditional neural networks and AR-BERT. In Proceedings of the 2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM). IEEE, 2024. https://doi.org/10.1109/WINCOM62286.2024.10657409.
30. Chouikhi, H.; Jarray, F.; Alsuhaibani, M. A Sequence-to-Sequence Neural Network for Joint Aspect Term Extraction and Aspect Term Sentiment Classification Tasks. In Proceedings of the Proceedings of the 15th International Conference on Agents and Artificial Intelligence, INSTICC, mar 2023, pp. 117–123. https://doi.org/10.5220/0011620500003393.
31. Wang, L.; Yao, C.; Li, X.; Yu, X. BERT-based implicit aspect extraction. In Proceedings of the Proceedings of 2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology, ICCASIT 2021, 2021, pp. 758–761. https://doi.org/10.1109/ICCASIT53235.2021.9633578.
32. Busst, M.M.A.; Anbananthen, K.S.M.; Kannan, S.; Krishnan, J.; Subbiah, S. Ensemble BiLSTM A Novel Approach for Aspect Extraction From Online Text. *IEEE Access* **2024**, *12*, 3528–3539.
33. Xavier, A.; Author, C. You Only Read Once Constituency-Oriented Relational Graph Convolutional Network for Multi-Aspect Multi-Sentiment Classification. In Proceedings of the IEEE Access, 2024. Note: Placeholder entry based on review document.
34. Chauhan, G.S.; Nahta, R.; Meena, Y.K. Improved Deep Learning Model for Aspect Words Extraction. In Proceedings of the 2023 IEEE World Conference on Applied Intelligence and Computing (AIC), 2023, pp. 286–292. https://doi.org/10.1109/AIC57670.2023.10263853.
35. Chauhan, G.S.; Nahta, R.; Meena, Y.K. A Review Level Aspect Extraction in Sentiment Analysis. In Proceedings of the 2024 2nd International Conference on Computer, Communication and Control (IC4). IEEE, 2024, pp. 1–7. https://doi.org/10.1109/IC457434.2024.10486257.
36. Jayanthi, S.; Arumugam, S.S. Refining Sentiment Analysis: Explicit Aspect Extraction with Diverse Datasets and Advanced Models. In Proceedings of the 2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI). IEEE, 2024, pp. 182–187. https://doi.org/10.1109/ICOICI62503.2024.10696470.
37. Wang, T. Domain Adaptive English Aspect Word Extraction Method Based On Bidirectional Long And Short-term Memory Network And Multi-head Attention Mechanism. *Journal of Applied Science and Engineering* **2025**, *28*, 2661–2669. https://doi.org/10.6180/jase.202512_28(12).0013.
38. Huang, Z.; Li, L.; Li, Z.; Zhang, H.; Yan, J. Bidirectional interaction and inference strategy joint approach for span-level aspect sentiment triplet extraction. *International Journal of Machine Learning and Cybernetics* **2025**. https://doi.org/10.1007/s13042-025-02675-0.

 doi:10.20944/preprints202605.1240.v1

39. Quan, X.; Min, Z.; Li, K.; Yang, Y. Compound Aspect Extraction by Augmentation and Constituency Lattice. *IEEE Transactions on Affective Computing* **2023**, *14*, 2323–2335. https://doi.org/10.1109/TAFFC.2022.3161683.
40. Soni, P.K.; Rambola, R.K. Deep Learning, WordNet, and spaCy based Hybrid Method for Detection of Implicit Aspects for Sentiment Analysis. In Proceedings of the 2021 International Conference on Intelligent Technologies, CONIT 2021, June 2021. https://doi.org/10.1109/CONIT51480.2021.9498372.
41. Hammad, M.; AbuEnnab, N.; Al-Refai, M. An aspect-based sentiment analysis model for Arabic game reviews based on hybrid transformers models. *Neural Computing and Applications* **2025**, *37*, 10309–10331. https://doi.org/10.1007/s00521-025-11032-9.
42. Almasre, M.A. Enhance the Aspect Category Detection in Arabic Language using AraBERT and Text Augmentation. In Proceedings of the 2022 Fifth National Conference of Saudi Computers Colleges (NCCC). IEEE, December 2022, pp. 7–10. https://doi.org/10.1109/NCCC57165.2022.10067648.
43. Tamchyna, A.; Veselovská, K. UFAL at SemEval-2016 Task 5 - Recurrent Neural Networks for Sentence Classification. In Proceedings of the Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California, june 2016; pp. 367–371. https://doi.org/10.18653/v1/S16-1059.
44. Al-Smadi, M.; Talafha, B.; Al-Ayyoub, M.; Jararweh, Y. Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. *International Journal of Machine Learning and Cybernetics* **2019**, *10*, 2163–2175. https://doi.org/10.1007/s13042-018-0799-4.
45. Abdelgwad, M.M.; Soliman, T.H.A.; Taloba, A.I.; Farghaly, M.F. Arabic aspect based sentiment analysis using bidirectional GRU based models. *Journal of King Saud University - Computer and Information Sciences* **2021**. https://doi.org/10.1016/j.jksuci.2021.08.030.
46. Fadel, A.S.; Saleh, M.E.; Abulnaja, O.A. Arabic Aspect Extraction Based on Stacked Contextualized Embedding With Deep Learning. *IEEE Access* **2022**, *10*, 30526–30535. https://doi.org/10.1109/ACCESS.2022.3159252.
47. Bensoltane, R.; Zaki, T. Towards Arabic aspect-based sentiment analysis: a transfer learning-based approach. *Social Network Analysis and Mining* **2022**, *12*, 1–16. https://doi.org/10.1007/S13278-021-00794-4/METRICS.
48. Hamam, H.; Rehman, A.U.; Othman, M.T.B.; Chouikhi, H.; Alsuhaibani, M.; Jarray, F. BERT-Based Joint Model for Aspect Term Extraction and Aspect Polarity Detection in Arabic Text. *Electronics* **2023**, *12*, 515. https://doi.org/10.3390/ELECTRONICS12030515.
49. Alamoudi, E.; Solaiman, E. EHSAN: Leveraging ChatGPT in a Hybrid Framework for Arabic Aspect-Based Sentiment Analysis in Healthcare. *Proceedings of the International Conference on Intelligent Data Engineering and Automated Systems (IDEAS) 2025* **2025**. Preprint, https://doi.org/10.5281/zenodo.15418860.
50. Thi, T.L.; Tran, T.K.; Phan, T.T. Deep Learning Using Context Vectors to Identify Implicit Aspects. *IEEE Access* **2023**. https://doi.org/10.1109/ACCESS.2023.3268243.
51. Xiong, H.; et al. BART-based contrastive and retrospective network for aspect-category-opinion-sentiment quadruple extraction. *International Journal of Machine Learning and Cybernetics* **2023**, pp. 1–13. https://doi.org/10.1007/S13042-023-01831-8/METRICS.
52. Ahmed, M.; et al. BERT-ASC: Implicit Aspect Representation Learning through Auxiliary-Sentence Construction for Sentiment Analysis. *arXiv preprint arXiv:2203.11702* **2022**. https://doi.org/10.48550/arxiv.2203.11702.
53. Feng, J.; Cai, S.; Ma, X. Enhanced sentiment labeling and implicit aspect identification by integration of deep convolution neural network and sequential algorithm. *Cluster Computing* **2019**, *22*, 5839–5857. https://doi.org/10.1007/S10586-017-1626-5/METRICS.
54. Lazhar, F. Implicit feature identification for opinion mining. *International Journal of Business Information Systems* **2019**, *30*, 13–30. https://doi.org/10.1504/IJBIS.2019.097042.
55. Li, X.; Wang, X.; Yao, C.; Li, Y. Graph-enhanced implicit aspect-level sentiment analysis based on multi-prompt fusion. *Scientific Reports* **2025**, *15*, 1–19. https://doi.org/10.1038/s41598-025-02609-4.
56. Author1, F.; Author2, S. Attention-Based Sentence Extraction for Aspect-Based Sentiment Analysis with Implicit Aspect Cases in Hotel Reviews Using Machine Learning Algorithm, Semantic Similarity, and BERT. *International Journal of Computers, Communications & Control* **2023**, *18*, 450–468.
57. Kabir, M.M.; Othman, Z.A.; Yaakub, M.R. A Hybrid Frequency Based, Syntax, and Conditional Random Field Method for Implicit and Explicit Aspect Extraction. *IEEE Access* **2024**, *12*, 72361–72373. https://doi.org/10.1109/ACCESS.2024.3403479.

 doi:10.20944/preprints202605.1240.v1

58. Agathangelou, P.; Katakis, I.; Kasnesis, P. Word- and Sentence-Level Representations for Implicit Aspect Extraction. *IEEE Transactions on Computational Social Systems* **2024**, *11*, 5935–5948. https://doi.org/10.1109/TCSS.2024.3391833.
59. Neveditsin, N.; Lingras, P.; Mago, V.K. From Annotation to Adaptation: Metrics, Synthetic Data, and Aspect Extraction for Aspect-Based Sentiment Analysis with Large Language Models. In Proceedings of the Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop). Association for Computational Linguistics, 2025, p. 142–161. https://doi.org/10.18653/v1/2025.naacl-srw.14.
60. Wiseman, S.; Rush, A.M. Sequence-to-Sequence Learning as Beam-Search Optimization. In Proceedings of the Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, 2016; pp. 1296–1306. https://doi.org/10.18653/v1/D16-1137.
61. OpenAI. GPT-4.1-mini: A compact instruction-following model, 2025.
62. Obeid, O.; Zalmout, N.; Khalifa, S.; Taji, D.; Oudah, M.; Alhafni, B.; Inoue, G.; Eryani, F.; Erdmann, A.; Habash, N. CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing. In Proceedings of the Proceedings of the Twelfth Language Resources and Evaluation Conference; Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; et al., Eds., Marseille, France, May 2020; pp. 7022–7032.
63. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The Llama 3 Herd of Models, 2024, [arXiv:cs.AI/2407.21783].
64. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models, 2021, [arXiv:cs.CL/2106.09685].
65. Berg-Kirkpatrick, T.; Burkett, D.; Klein, D. An Empirical Investigation of Statistical Significance in NLP. In Proceedings of the Proceedings of EMNLP, 2012, pp. 995–1005.
66. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157.
67. Chen, M.; Mao, Y.; Hua, Q.; Wu, J. Aspect Extraction from E-Commerce Reviews Based on Word–Word Relationship Classification. In Proceedings of the 2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT), 2023, pp. 226–231. https://doi.org/10.1109/ACAIT60137.2023.10528431.
68. Haq, B.; Daudpota, S.; Ali, I.; Kastrati, Z.; Noor, W. A Semi-Supervised Approach for Aspect Category Detection and Aspect Term Extraction from Opinionated Text. *Computers, Materials, & Continua* **2023**, *77*, 115–137.

Short Biography of Authors

Mohamed El Bachir Menai is a Professor of Computer Science at King Saud University. He previously was an Associate Professor of Computer Science at King Saud University and at the University Center of Tebessa (Algeria). He also was a Research Scientist at LIASD Laboratory (Laboratoire d'Informatique Avancée de Saint-Denis) at the University of Paris 8 (France).

Lujain A. Alawwad is a PhD researcher in Computer Science at King Saud University, College of Computer and Information Sciences. and a lecturer at Saudi Electronic University, College of Computing and Informatics. Her research interests include Artificial Intelligence and Natural Language Processing, with a primary focus on the Arabic language. She is particularly interested in leveraging deep learning and structured knowledge to advance multilingual text understanding.