Future Technology
Open Access Journal
ISSN 2832-0379

August 2026| Volume 05 | Issue 03 | Pages 224-239

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.3.20

Future Publishing LLC

Article

Cross-domain aspect-based sentiment analysis using DeBERTaV3 and bio-CRF: a syntactic-aware span extraction approach

Udayalaxmi Aditya Teki[1]*, P. Ranjana[2]

[1]Department of CSE, Hindustan Institute of Technology & Science, Chennai, India
[2]Department of Information Technology, Hindustan Institute of Technology & Science, Chennai, India

ARTICLE INFO

Article history:
Received 06 January 2025
Received in revised form
29 April 2026
Accepted 09 June 2026

Keywords:
Cross-domain ABSA, DeBERTaV3, Bio-CRF, Span extraction, Domain adaptation

*Corresponding author
Email address:
udayalakshmiaditya@gmail.com

DOI: 10.55670/fpll.futech.5.3.20

ABSTRACT

Aspect-Based Sentiment Analysis (ABSA) often experiences a significant performance decline in cross-domain settings due to vocabulary variation and domain-specific aspect expressions. Although transformer-based models achieve strong in-domain performance, they primarily rely on contextual embeddings and often ignore the syntactic structures that remain consistent across domains. Existing methods rarely integrate structured decoding with adaptive syntactic fusion for robust aspect boundary detection. This paper proposes a syntactic-aware cross-domain ABSA framework based on DeBERTaV3 and BIO-CRF decoding to alleviate the above problems. The proposed model introduces part-of-speech and dependency-relation embeddings, in addition to contextual embeddings, and uses an attention-based model to dynamically fuse syntactic and semantic information at multiple levels. We further apply a Conditional Random Field (CRF) layer to enforce valid BIO transitions and enhance the consistency of multi-word aspect spans under domain shift. The model was evaluated in three English review domains: Restaurant, Laptop, and Device across six zero-shot cross-domain transfer settings (D→L, D→R, L→D, L→R, R→D, and R→L). Test results show consistent advances over robust transformer-based and prompt-based baselines. The proposed method yields F1 scores for aspect extraction between 0.72 and 0.81 and achieves sentiment classification accuracies between 74.32% and 85.19%. The best performance was achieved in the L→R transfer setting. Through paired bootstrap testing ($p < 0.01$), Statistical analysis confirms that the proposed model achieves significant improvements over baseline methods. The results demonstrate that incorporating explicit syntactic knowledge, adaptive feature fusion, and structured decoding substantially improves cross-domain generalization in fine-grained sentiment analysis.

1. Introduction

Aspect-Based Sentiment Analysis (ABSA) is a fine-grained sentiment analysis task that selects aspect terms in a sentence and identifies the sentiment polarity of each aspect [1]. ABSA associates sentiment expressions with specific targets, providing a more detailed understanding of user opinions than sentence-level sentiment classification. This function can also aid in assessing product reviews, gathering customer feedback, evaluating services, and tracking brand reputation. The SemEval-2014 Task 4 benchmark was established and has since become one of the most widely used benchmarks in ABSA research [2]. To assess cross-domain robustness, we also integrate more review domains, namely, Restaurant reviews from SemEval 2015 and 2016 and Device reviews [3]. The early approaches to ABSA primarily treat aspect extraction as a sequence labeling problem and employ probabilistic models such as Conditional Random Fields (CRFs) [4], followed by neural approaches with BiLSTM and CRF decoding [5]. These methods often relied on manual linguistic feature engineering based on part-of-speech (POS) tags, dependency relations, and sentiment lexicons. While these approaches work well in a controlled setting, they are not robust to domain shift and require extensive feature engineering to generalize across heterogeneous domains [6]. BERT [7] was a high-profile, successful large pre-trained transformer model that could learn contextual

representations. This led to the adoption of end-to-end neural modeling approaches for ABSA. Further studies involve span-based extraction methods [8], domain-adaptive pre-training [9], and adversarial feature alignment [10]. Through disentangled attention mechanisms that separately model content and positional information, DeBERTa further enhanced contextual representation learning [11]. Although transformer architectures have strong contextual modeling capabilities [12], they still have several limitations in cross-domain ABSA settings. To start with, most transformer-based techniques rely mostly on contextual embeddings but do not fully capitalize on explicit syntactic information (e.g. POS tags or dependency relations), even though syntactic patterns are generally less sensitive to domain-specific vocabularies. Another limitation is that recent transformer architectures, such as DeBERTaV3 [13], do not incorporate structured decoding mechanisms, such as CRF, which can lead to inconsistencies in aspect boundary prediction due to domain shift. Third, when features are introduced, they are often combined via simple concatenation, without any adaptive mechanisms to equilibrate syntactic and contextual information.

To address these limitations, the proposed framework pursues three primary objectives. The first objective incorporates explicit POS and dependency-relation embeddings into a DeBERTaV3 encoder, thereby augmenting existing contextual token representations with grammatical structure. The second contribution is an attention-based fusion mechanism that dynamically balances syntactic and contextual information at the token level. To address this limitation, we introduce a CRF decoding layer that enforces valid BIO transitions over the fused representations. This results in improved consistency of multi-token aspect spans across the different domains. The framework is assessed on three review domains, Restaurant, Laptop, and Device, under six stringent zero-shot cross-domain transfer settings. The proposed framework introduces a unified cross-domain ABSA architecture that combines contextual semantic representations with explicit syntactic modeling and structured decoding. Unlike existing approaches that either rely solely on contextual embeddings or incorporate syntax through graph-based dependency structures such as graph convolutional networks [14] or adversarial feature alignment methods [15], the proposed model integrates POS and dependency embeddings directly into a DeBERTaV3 backbone and adaptively balances syntactic and contextual information using an attention-based fusion mechanism. In addition, a CRF decoding layer is employed to enforce valid BIO transitions and improve aspect boundary consistency under domain shift. By integrating adaptive syntactic fusion and structured prediction into a single DeBERTaV3-based framework, the proposed method achieves robust performance across heterogeneous review domains without auxiliary domain-adversarial training. A novel cross-domain ABSA approach that combines DeBERTaV3 contextual representations with POS and dependency-relation embeddings, fused at a unified token level.

- A feature fusion mechanism that uses attention to synthetically and contextually balance representations, enabling generalization across domains.
- A BIO-CRF decoding mechanism that ensures valid transitions between tags while also consistently extracting multi-word aspect spans across domains.
- Comprehensive assessment across the Restaurant, Laptop, and Device review domains, employing six zero-shot transfer approaches, along with ablation studies and statistical significance tests, comparing strong transformer-based, graph-based, and prompt-based baselines.

The key contributions of this research are as follows: Although syntactic information can improve structural generalization, the framework still depends on the quality of external linguistic parsers, which may introduce errors in noisy or low-resource settings. Overall, the proposed framework demonstrates that integrating syntactic structure, adaptive fusion, and structured decoding can substantially improve the robustness and cross-domain generalization capability of ABSA systems.

2. Related work

ABSA has evolved from statistical sequence-labeling approaches to large-scale pre-trained language models over successive methodological developments [16]. This section surveys earlier studies along six key streams: (1) sequence labeling methods; (2) neural and attention-based models; (3) transformer-based contextual representations; (4) cross-domain and transfer learning approaches; (5) syntactic and graph-based modeling; (6) structured decoding methods.

2.1 Statistical and neural sequence labeling

Earlier ABSA techniques framed the identification of aspect terms as sequence labeling tasks. Due to the ability of conditional random fields (CRFs) to model dependency between labels and impose structural consistency, they became popular [17]. Despite this, the models were created using manually derived features such as part-of-speech (POS) tags, syntactic dependency relations, and sentiment lexicons. The integration of CRF decoding with bidirectional long short-term memory (BiLSTM) networks significantly improved contextual modeling. Other architectures incorporated convolutional neural networks into BiLSTM and CRF layers [18]. Neural models, although reducing the need for manual feature engineering, remain sensitive to domain-specific lexical patterns. Structured decoding led to improved span consistency. However, the domain adaptation mechanisms incorporated were not that powerful. Thus, they failed to work well in diverse domains.

2.2 Attention-based and early neural ABSA models

To better understand the connection between aspect terms and opinion words, attention mechanisms were adopted. Models like attention-based LSTM [19,20] and IAN [21] selectively focus on context words that are highly relevant to the sentiment polarity. Memory network models improve sentiment inference through successive attention to the context. The results indicate that explicitly using aspect context interactions improves sentiment classification accuracy. The contextual representations acquired by these models, however, still depend on domain-specific training data. Due to domain shift, performance is significantly degraded. Many attention-based approaches do not explicitly

account for distributional differences between domains and instead focus on maximizing domain accuracy.

2.3 Transformer-based contextual representations

The ABSA paradigm was transformed with the introduction of BERT [7]. The performance of RNN-based models is significantly lower than that of self-supervised learning-based contextual token representation of pre-trained transformer models [22]. In subsequent studies, researchers used BERT for ABSA by fine-tuning it. Others proposed using SpanBERT to enable span-based modeling. Further works built joint extraction and classification systems for the same purpose [23]. Despite their robust performance within a domain, model transformers frequently underperform when facing domain shifts. To enhance the positional encoding and context disentangling, DeBERTa introduced disentangled attention mechanisms, and DeBERTaV3 further optimized this with ELECTRA-style training [11]. The representational strength of these models has increased, and they are widely used as backbones. Even so, systems that leverage transformers continue to rely on lexical co-occurrence statistics. When the vocabulary distribution changes, the performance drops in cross-domain transfer. Although contextual encoders perform well in capturing semantic relations, we argue that they do not explicitly encode grammatical structure.

2.4 Cross-domain and transfer learning approaches

The goal of cross-domain ABSA is to mitigate performance degradation during transfers. To address discrepancies across domains, techniques for adaptive transfer learning have been proposed. Domain adaptive fine-tuning has been applied to aspect-level sentiment classification [24], while adversarial learning methods attempt to learn domain-invariant feature representations [25]. Simultaneously, modeling sentiment classification and aspect extraction can enhance generalization performance [26]. More recent work has explored the use of contrastive learning to improve within-domain alignment of representations [27]. Although there has been progress, most cross-domain systems rely primarily on contextual embeddings and do not explicitly exploit syntactic invariants. Existing cross-domain approaches primarily focus on feature-level adaptation and often overlook structured prediction and linguistic regularization.

2.5 Syntactic and graph-based modeling

Matching aspect terms to opinion expressions relies on constituent syntactic structures. Methods that rely on dependencies store relational information beyond linear sequences of tokens. The target-dependent sentiment classification was developed using dependency tree-based neural networks [28]. Dependency graph-based GCNs have enhanced relational reasoning in ABSA in [29]. Syntactic dependencies are suggested to be more stable than the lexicon across domains. Besides, dual embedding schemes for POS tags have been shown to reinforce grammatical awareness [30]. Nonetheless, the direct combination of syntactic embeddings with transformer backbones remains unexplored, particularly across domains.

2.6 Structured decoding and span consistency

For boundary detection, valid label transitions are required. While transformer-based architecture provides rich contextual encodings, token-level classifiers can produce inconsistent BIO sequences. Structured decoding using a conditional random field (CRF) offers a stable way to model transition constraints. Recent research shows that transformer encoders with CRF layers achieve more coherent boundaries and less fragmented spans [31]. Span-based extraction systems also try to preserve multi-token consistency [32]. However, Structured decoding methods have received limited attention in cross-domain ABSA research.

2.7 Emerging directions: generative and prompt-based ABSA

Recent research has investigated generative formulations of ABSA using T5 [33], a sequence-to-sequence transformer. Redefinition of ABSA using masked language modeling or question-answering tasks, followed by prompt-based learning [34]. Instruction-style supervision strategies generally require a large computational budget and do not typically impose explicit structural limits on aspect spans. The restrictions recognized in the above assessment, such as weak syntactic integration, insufficient structured decoding, or lack of adaptive fusion, directly motivate the framework we propose in Section 3.

3. Methodology

This paper presents a framework for cross-domain ABSA that integrates contextual semantic modeling with the explicit syntactic structure using structured prediction. Combining domain-sensitive lexical patterns with more stable grammatical structures limits performance decay under domain shift. The model jointly performs aspect span extraction and sentiment classification in a single optimization system, as shown in Figure 1.

3.1 Problem formulation

Let an input sentence be $S = \{w_1, w_2, \dots, w_n\}$, where w_i is the i-th token and n is the sequence length. The goal of ABSA is to identify a set of aspect spans $A = \{(s_k, e_k)\}$ and assign each span a sentiment polarity $y_k \in$ {positive,negative,neutral}. Aspect extraction is formulated as a structured sequence labeling task using the BIO encoding scheme:

$$Y = \{y_1, y_2, \dots, y_n\}, y_i \in \{B, I, O\} \quad (1)$$

BIO rules define valid label transitions (for example, I cannot follow an O without being preceded by a B). This formulation not only keeps aspect boundary information but also models aspect extraction as a structured token-level sequence labeling task. The structured sequence modeling optimizes dependencies between labels globally (rather than only associating neighboring labels). This reduces prediction errors at the boundaries. The symbols used in the proposed framework are described in Table 1.

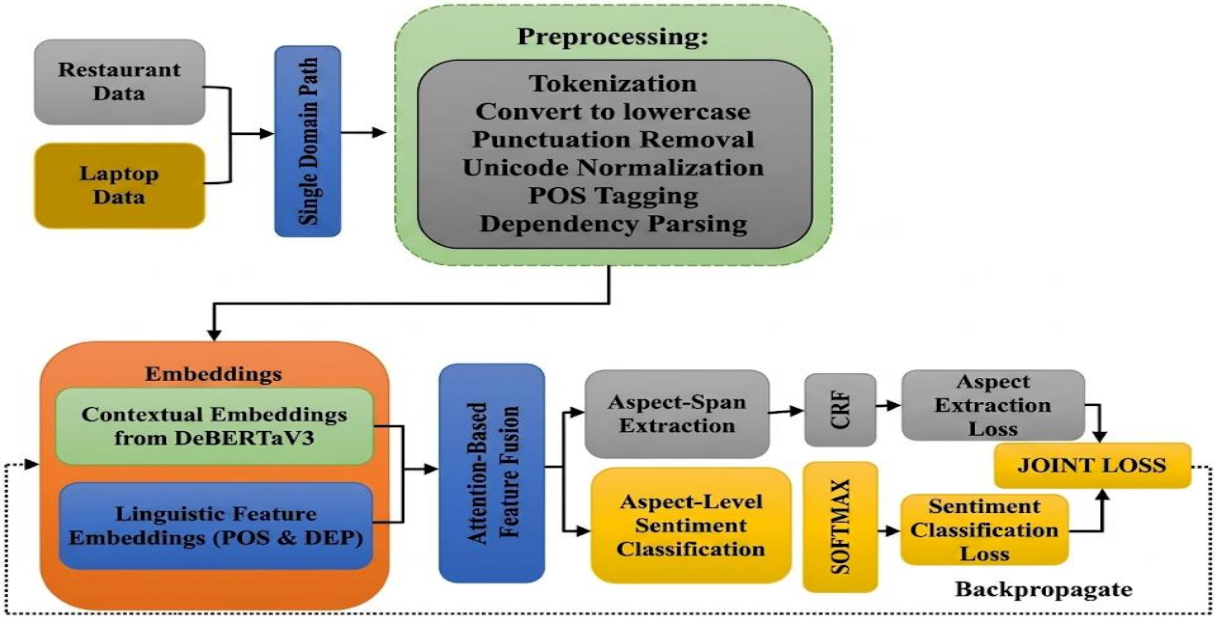

Figure 1. The proposed syntactic-aware cross-domain ABSA framework

Table 1. Nomenclature of symbols used in the proposed framework

Symbol	Description
S	Input sentence
w_i	The i-th token in the sentence
n	Number of tokens in the sentence
A	Set of aspect spans
(s_k, e_k)	Start and end positions of the k-th aspect span
Y	BIO label sequence
y_i	BIO label assigned to token i
h_i	Contextual embedding from DeBERTaV3
p_i	POS embedding
d_i	Dependency relation embedding
x_i	Concatenated token representation
$\tilde{x}_i$	Attention-fused token representation
α_i	Attention weight for token i
e_i	CRF emission score
$A_{i,j}$	CRF transition score between labels
v_k	Aspect-level pooled representation
$\hat{y}_k$	Predicted sentiment distribution
$\mathcal{L}_{CRF}$	CRF loss
$\mathcal{L}_{sent}$	Sentiment classification loss
λ	Loss balancing parameter

3.2 Data preprocessing

Preprocessing plays a pivotal role in cross-domain ABSA, as errors in tokenization, annotation matching, or syntactic structure extraction can lead to prediction errors. Preprocessing minimizes variation in terms and structures between the source and the target. The preprocessing pipeline performs text normalization, annotation alignment, linguistic feature extraction, subword alignment, and sequence padding.

Text normalization: the raw sentence as $S^{\text{raw}} = \{w_1^{\text{raw}}, w_2^{\text{raw}}, \dots, w_n^{\text{raw}}\}$. User-generated content often contains uncased, misspelled, noisy punctuation, and abbreviations. A processing operator $\mathcal{N}(\cdot)$ is app

$$w_i = \mathcal{N}(w_i^{\text{raw}}) \tag{2}$$

Normalization consists of the following steps: variable lowercasing, Unicode normalization, punctuation separation, removal of redundant characters, and optional spelling correction. It stabilizes lexical distributions across various domains. According to domain adaptation theory [35], reducing the empirical variance of token frequencies between the source and target domains should improve generalization under domain shift.

BIO-based aspect annotation: The bio-based aspect annotation for any normalized token w_i, is $t_i \in \{B, I, O\}$, which indicates that w_i is the beginning (B), inside (I), or outside (O) of an aspect term. Tag sequence refers to the sequence $T = \{t_1, t_2, \dots, t_n\}$. This encoding makes explicit multi-token spans. In the presence of CRF decoding, the BIO scheme introduces valid transition constraints, whereby an I-tag cannot immediately follow an O-tag unless preceded by a B-tag:

$$P(t_i = I \mid t_{i-1} = O) = 0 \tag{3}$$

This constraint enhances the consistency of aspect boundaries during systematic sequence prediction.

Sentiment Label Assignment: Each aspect span is assigned a sentiment polarity label from "negative", "neutral" and "positive". Token-level sentiment alignment is defined as:

$$p_i = \begin{cases} y_k, & \text{if } s_k \le i \le e_k \\ -1, & \text{otherwise} \end{cases} \tag{4}$$

Here, -1 indicates non-aspect tokens. This way, shared contextual representations can potentially be used for aspect detection and polarity classification.

Part-of-speech (POS) tagging: The grammatical category $pos_i = \text{POS}(w_i)$ is assigned to each token w_i. The POS tag

sequence is denoted as $Pos = \{pos_1, pos_2, \dots, pos_n\}$. In general, aspect terms are nouns, and sentiment expressions are usually adjectives or adverbs. For this reason, incorporating POS tagging introduces structural priors to mitigate ambiguities in aspect–opinion alignment.

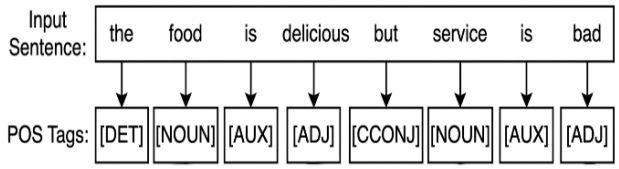

Figure 2. Part of speech (POS) tagging

Every token is tagged with a universal POS tag (NOUN, ADJ, DET). Figure 2 shows that aspect terms such as "food" and "service" always receive a nominal tag, while words that express sentiment, such as “delicious” and “bad,” appear as adjectives. These grammatical patterns provide useful structural signals for aspect–opinion alignment and motivate the inclusion of POS embeddings in the model. This information provides grammatical cues that complement contextual semantic representations.

Dependency parsing: Dependency parsing describes the structure of a sentence as a set of dependency relations, or dependency trees between tokens or words (Figure 3). Each token w_i is assigned a dependency relation $dep_i = \text{DEP}(w_i)$. The dependency sequence is $Dep = \{dep_1, dep_2, \dots, dep_n\}$. Relations like *amod* (adjectival modifier) and *nsubj* (nominal subject) encode structural interactions between aspect and opinion terms. Dependency relations are often more structural and consistent across domains than surface lexical patterns, which facilitates cross-domain representation learning. Let $\phi_{\text{lex}}(w)$ be a lexical feature function, and $\phi_{\text{syn}}(w)$ a syntactic feature function. The notion of domain invariance asserts that the divergence of $P(\phi_{\text{syn}}(w) \mid \text{domain})$ across the domains is lower than the divergence of $P(\phi_{\text{lex}}(w) \mid \text{domain})$ where the divergence is taken to be the Kullback-Leibler divergence [36]. This insight encourages incorporating syntactic features to improve cross-domain robustness. The dependency tree of a review sentence is shown in Figure 3. Edges indicate the syntactic relation between the words, such as nsubj, for example, “service” → “is”. The visualization shows that regardless of surface word order, opinion modifiers are connected via dependency arcs to their aspect terms. The structural patterns are largely preserved across domains, supporting the thesis of cross-domain stability.

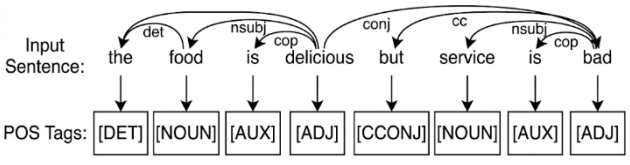

Figure 3. Dependency parsing

Subword tokenization and label alignment: Subword tokens are used by DeBERTaV3. A token w_i can be split into subwords $\{u_{i1}, u_{i2}, \dots, u_{im}\}$. BIO tags are passed onto the subwords in as:

$$t_{ij} = \begin{cases} t_i, & j = 1 \\ I, & j > 1 \text{ and } t_i \neq O \\ O, & \text{otherwise} \end{cases} \quad (5)$$

POS and dependency tags are replicated at the token level:

$$pos_{ij} = pos_i, dep_{ij} = dep_i \quad (6)$$

This ensures that token splitting does not distort labels and that structured decoding occurs at the subword level.

Sequence padding and attention masking: Each sequence is padded to a maximum length L:

$$\tilde{S} = \{w_1, \dots, w_n, \underbrace{\text{PAD}, \dots, \text{PAD}}_{L-n}\} \quad (7)$$

The attention mask is:

$$m_i = \begin{cases} 1, & i \leq n \\ 0, & i > n \end{cases} \quad (8)$$

Padding tokens are ignored during self-attention and loss calculation.

3.3 Contextual representation learning

We use DeBERTaV3 [13] as the contextual encoder. DeBERTaV3 improves upon BERT [7] and RoBERTa through two key innovations: disentangled attention and ELECTRA [37]-style pre-training. These enhancements produce richer contextual representations that are more robust to domain shift.

Disentangled attention mechanism: Standard BERT combines token content and position information into a single embedding before self-attention. DeBERTaV3 instead separates them. Each token w_i is represented by two vectors:

- **A content embedding $\mathbf{c}_i$** – captures the token’s semantic meaning.
- **A position embedding $\mathbf{p}_i$** – captures its absolute position in the sequence.

The attention score between tokens i and j is computed as the sum of four interaction terms:

$$A_{i,j} = \underbrace{\mathbf{c}_i \cdot \mathbf{c}_j}_{\text{content-to-content}} + \underbrace{\mathbf{c}_i \cdot \mathbf{p}_j}_{\text{content-to-position}} + \underbrace{\mathbf{p}_i \cdot \mathbf{c}_j}_{\text{position-to-content}} + \underbrace{\mathbf{p}_i \cdot \mathbf{p}_j}_{\text{position-to-position}}$$

However, DeBERTaV3 simplifies this by using a disentangled matrix formulation. Let Q_c, K_c, V_c be the content-based projections, and Q_r, K_r be the position-based projections. The attention score becomes:

$$\text{Score}(x_i, x_j) = Q_c \cdot K_c^\top + Q_c \cdot K_r^\top + Q_r \cdot K_c^\top \quad (9)$$

The learned vector r_{i-j} expresses the relative position between tokens i and j. This design enables the model to learn both semantic and positional relationships without the interference seen in BERT.

ELECTRA-Style Pre-Training: This allows each token’s representation to be informed by the global sentence context, which is essential for ABSA because sentiment polarity depends on relationships between aspect terms and opinion words. However, contextual embeddings learned solely from lexical co-occurrence become unstable under domain shift. Therefore, we augment them with explicit syntactic features as described in Sections 3.4 and 3.5. Figure 4 illustrates the overall DeBERTaV3 embedding flow within our framework.

For ABSA, these innovations produce contextual embeddings that better capture fine-grained semantic and syntactic relationships. The output of DeBERTaV3 for an input sentence $S = \{w_1, \dots, w_n\}$ is a sequence of contextual embeddings. In ABSA, these innovations enable contextual embeddings to better capture semantic and syntactic relationships. DeBERTaV3 generates contextual embeddings in an output sequence to represent input sentence S=$\{w_1, \dots, w_n\}$.

$$H = \{\mathrm{h}_1, \mathrm{h}_2, \dots, \mathrm{h}_n\}, \mathrm{h}_i = \mathrm{DeBERTaV3}(w_i, S) \quad (10)$$

The self-attention mechanism in DeBERTaV3 computes weighted interactions between all token pairs:

$$Attention(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right)V \quad (11)$$

This consists of a global contextual representation, which is important for ABSA, as sentiment polarity is determined by the relationship between aspect terms and opinion words. Yet in cases of domain shifts, embeddings determined solely by lexical co-occurrence become unstable. Thus, we enhance them with explicit syntactic features as outlined in Sections 3.4 and 3.5. Figure 4 shows the overall DeBERTaV3 embedding workflow in our Framework.

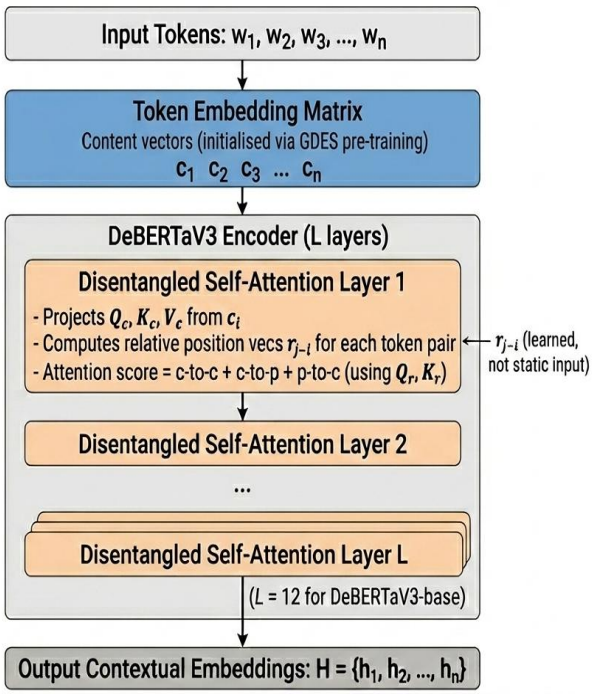

Figure 4. DeBERTa contextual embedding

3.4 Linguistic feature embedding

POS tags as well as dependency relations are represented as dense vectors for grammatical structure. A vector is mapped to each POS tag pos_i:

$$\mathrm{p}_i = E^{\mathrm{pos}}(pos_i) \quad (12)$$

A vector is assigned to each dependency relation dep_i:

$$\mathrm{d}_i = E^{\mathrm{dep}}(dep_i) \quad (13)$$

These embeddings denote grammatical categories (e.g., nouns; adjectives) and syntactic relations (e.g., adjectival modifiers). Syntactic categories are much more domain-steady than lexical tokens. The full token representation concatenates the contextual embedding, the POS embedding, and the dependency embedding.

$$x_i = [h_i; p_i; d_i] \quad (14)$$

This approach enhances the dimensionality of the feature space, enhances feature diversity, and introduces a beneficial structural bias that improves generalization to heterogeneous domains.

3.5 Attention-based feature fusion

Direct concatenation of features makes the same assumption across all feature types and therefore may not capture feature contributions across domains (Eq. 14). To overcome this drawback, we use an attention-based fusion mechanism. Initially, there is a nonlinear transformation:

$$\mathrm{z}_i = \tanh\left(W_f \mathrm{x}_i + \mathrm{b}_f\right) \quad (15)$$

Next, we calculate the attention weights:

$$\alpha_i = \frac{\exp\left(\mathrm{u}^{\mathrm{T}} \mathrm{z}_i\right)}{\sum_{j=1}^{n} \exp\left(\mathrm{u}^{\mathrm{T}} \mathrm{z}_j\right)} \quad (16)$$

The fused representation is:

$$\tilde{\mathrm{x}}_i = \alpha_i \mathrm{x}_i \quad (17)$$

The softmax normalization guarantees a valid probability distribution over the tokens, enabling an adaptive focus on informative tokens. The attention mechanism is an adaptive weighting system that enables the model to dynamically assign importance to contextually or syntactically relevant representations, based on the input sentence.

3.6 Structured aspect extraction with CRF

Aspect extraction relies on a linear classifier and is post-processed by a CRF. Emission scores are used as:

$$e_i = W_e \tilde{\mathrm{x}}_i + \mathrm{b}_e \quad (18)$$

Rather than making independent predictions for each tag, the CRF predicts the entire tag sequence while conditioning on the input:

$$P(Y \mid X) = \frac{\exp\left(\mathrm{Score}(X,Y)\right)}{\sum_{Y'} \exp\left(\mathrm{Score}(X,Y')\right)} \quad (19)$$

The sequence score is:

$$\mathrm{Score}(X, Y) = \sum_{i=1}^{n} \left(A_{y_{i-1}, y_i} + e_i(y_i)\right) \quad (20)$$

where A represents the label transition matrix. Training reduces the negative log likelihood:

$$\mathcal{L}_{\mathrm{CRF}} = -\log P(Y \mid X) \quad (21)$$

CRF employs global normalization across the sequences to enforce proper BIO transitions on fragmented spans. This type of structured regularization is useful when different domains exhibit distinct lexical cues.

3.7 Aspect-level sentiment classification

After extracting spans, aspect-level sentiment classification is performed. The representation of a predicted span (s_k, e_k) is computed using mean pooling:

$$v_k = \frac{1}{e_k - s_k + 1} \sum_{i=s_k}^{e_k} \tilde{x}_i \quad (22)$$

The sentiment distribution is:

$$\hat{y}_k = \text{Softmax}(W_s v_k + b_s) \quad (23)$$

The loss function is as follows:

$$\mathcal{L}_{\text{sent}} = -\sum_k y_k \log \hat{y}_k \quad (24)$$

Predicting sentiment from extracted spans (instead of full sentences) ensures limited polarity allocation and retains context sensitivity.

3.8 Joint multi-task optimization

The total loss combines both aspect extraction loss and classification loss:

$$\mathcal{L} = \mathcal{L}_{\text{CRF}} + \lambda \mathcal{L}_{\text{sent}} \quad (25)$$

where λ mitigates the two components. This joint maximum-likelihood estimation encourages learning shared representations that capture boundary and polarity information, thereby improving transferability across domains.

3.9 Training procedure

Algorithm 1 describes the entire training process from beginning to end. The contextual embedding of the input sentence is generated by DeBERTaV3 for every batch. POS tagging and dependency relations are extracted and converted to dense embeddings. The Attention Fusion layer takes in the semantic and syntactic embeddings from the Fusion layer. We use the fused representation for two tasks: structured aspect extraction via CRF and aspect-level sentiment classification via a softmax classifier. The CRF loss and sentiment loss are summed with a balancing parameter lambda. Using Adam optimization, model parameters are updated via backpropagation.

Algorithm 1. Syntactic-aware cross-domain ABSA training

Input: Training dataset $D = \{(S, Y, T)\}$, balancing parameter λ, number of epochs E

Initialize model parameters Θ
for epoch = 1 to E **do**
  for each batch $(S, Y, T) \in D$ **do**
  1. Compute contextual embeddings:
    $H \leftarrow$ DeBERTaV3(S)
  2. Extract linguistic features
    $P \leftarrow$ POS_Embedding(S)
    $D \leftarrow$ DEP_Embedding(S)
  3. Concatenate features: $X \leftarrow [H; P; D]$
  4. Apply attention-based fusion: $\tilde{X} \leftarrow$ Attention(X)
  5. Compute CRF loss for aspect extraction:
    $L_{CRF} \leftarrow$ CRF_Loss($\tilde{X}, Y$)
  6. Compute sentiment classification loss:
    $L_{sent} \leftarrow$ Sentiment_Loss($\tilde{X}, T$)
  7. Compute total loss: $L = L_{CRF} + \lambda L_{sent}$
  8. Update parameters Θusing Adam optimizer
  end for
end for
Return trained parameters Θ

Output: Optimized model parameters Θ

4. Results and discussion

This section presents the detailed analysis of the proposed syntactically aware cross-domain aspect-based sentiment analysis (ABSA) framework. We first detail the benchmark data and implementation environment, then present the quantitative results, ablation studies, comparisons with state-of-the-art baselines, and statistical significance tests.

4.1 Dataset description

To evaluate cross-domain robustness, we use the review dataset from three domains: Restaurant (R), Laptop (L), and Device (D). Restaurant merges the restaurant reviews from SemEval 2014, 2015, and 2016 [2,38,39]; Laptop comes from SemEval 2014 [2]; Device merges all digital device reviews collected by Alghamdi et al [40]. Moreover, we give explicit aspect spans and sentiment polarity labels (positive, neutral, negative) to every review. The few instances of conflict (i.e., simultaneously positive and negative for the same aspect) are eliminated per standard ABSA practice. The detailed statistics for the datasets are shown in Table 2.

We take into account all six relevant cross-domain transfer settings, including source-to-target among the three domains, which include D → L, D → R, L → D, L → R, R → D, and R → L. A model is trained in each setting solely on the source domain's training split and evaluated on the target domain's test split, without any target-domain fine-tuning. This transfer protocol provides a rigorous assessment of the model's cross-domain generalization capability across quite distinct review types that differ significantly in size and language.

Table 2. Dataset statistics for the three domains

Metric	Device	Laptop	Restaurant
Total sentences (train + test)	2,085	2,928	6,536
Training sentences	1,394	2,297	4,284
Test sentences	691	631	2,252
Total aspect spans (train + test)	2,085	2,928	6,536
Average sentence length (tokens)	~18.0	~17.4	~15.9
Average aspect span length (tokens)	~1.4	~1.3	~1.2
Sentiment distribution (train)			
– Positive	767 (55.0%)	994 (42.7%)	2,570 (60.0%)
– Neutral	279 (20.0%)	464 (20.0%)	728 (17.0%)
– Negative	348 (25.0%)	870 (37.3%)	986 (23.0%)
Sentiment distribution (test)			
– Positive	380 (55.0%)	337 (53.4%)	1,350 (60.0%)
– Neutral	138 (20.0%)	167 (26.5%)	382 (17.0%)
– Negative	173 (25.0%)	127 (20.1%)	520 (23.1%)

The three domains differ substantially in lexical characteristics, syntactic structures, and aspect distributions, which make cross-domain transfer particularly challenging.

4.2 Implementation details

The framework uses DeBERTaV3 base [11] as the contextual encoder. Input sentences are tokenized with the DeBERTaV3 tokenizer using a maximum length of 128 tokens. Part-of-speech (POS) tags and dependency relations are obtained using spaCy v3.5.0 with the en_core_web_sm model (POS tagging accuracy ~97 %, dependency UAS ~93 % on standard English text). The linguistic features are embedded as dense vectors (dimension 50 each) and combined with the contextual embeddings via an attention-based fusion layer before being passed to the CRF decoder for aspect extraction. Sentiment classification is performed by mean-pooling the fused representations of the predicted span, followed by a fully connected softmax classifier. All hyperparameters were selected via grid search on a 10 % validation split of the source domain training data. The search ranges were: learning rate {1×10^{-5}, 2×10^{-5}, 3×10^{-5}}, batch size {8, 16}, POS/DEP embedding dimension {30, 50, 75}, attention fusion hidden size {64, 128, 256}, and dropout {0.1, 0.2}. The final values are listed in Table 3.

Table 3. Hyperparameters and hardware specifications

Parameter	Value
DeBERTaV3 hidden size	768
POS embedding dimension	50
Dependency embedding dimension	50
Attention fusion hidden dimension	128
Dropout rate	0.1
Optimizer	Adam (β_1=0.9, β_2=0.999)
Learning rate	2×10^{-5}
Weight decay	1×10^{-2}
Batch size	8
Training epochs	10 (early stopping patience = 2)
CRF transition initialization	Random uniform [-0.1, 0.1]
Random seeds	42, 123, 456, 789, 1011 (mean ± std reported)
Hardware	NVIDIA GeForce RTX 3090 (24 GB VRAM)
Training time per epoch	≈4.5 min (Restaurant), ≈3.5 min (Laptop)

Experiments were run with five different random seeds, and the results in all tables are reported as mean ± standard deviation. All procedures were executed on the same hardware to ensure reproducibility.

4.3 Evaluation metrics

Aspect extraction performance is measured with Precision, Recall, and F1 using exact span matching (both start and end boundaries must match). For sentiment classification, we report aspect-aware accuracy: a prediction is correct only if both the aspect span and its polarity match the ground truth. We additionally report end-to-end F1 (span + polarity) as the joint metric. Standard sentiment accuracy over correctly extracted aspect spans is additionally reported for comparability with earlier studies. All metrics are reported on the target test set, and statistical significance is evaluated using a paired bootstrap test with 10,000 resamples ($p < 0.01$).

4.4 Cross-domain experimental results

The learning curves for aspect extraction (Figure 5) and sentiment classification (Figure 6) demonstrate consistent performance improvements throughout training over the 10 epochs for all six cross-domain settings. The training metrics outperform the validation metrics by a small margin (0.01-0.03), indicating minimal overfitting and stable generalization across domains. The L→R transfer achieves the highest validation F1 (~0.81) and validation accuracy (~85%), while D→R achieves a strong F1 of 0.78 and an accuracy of 81.2%. The R→D (≈0.72 F1, ≈74% accuracy) has the lowest accuracy, likely because restaurant and device reviews exhibit greater linguistic and lexical differences. L→D and D→L exhibit moderate performance. After about six to eight epochs, each curve plateaus without oscillation, suggesting stable optimization and justifying an early-stopping patience of two epochs. The minimal variation observed across five randomly chosen seeds (standard deviation of less than or equal to 0.01 for F1 score and less than or equal to 0.5 percent for accuracy) demonstrates the robustness of our proposed framework.

To assess the framework, we run evaluations on aspect extraction utilizing exact span matching. The precision, recall, F1, and exact span accuracy (i.e., the percentage of correctly predicted aspect boundaries) for all six cross-domain transfers are reported in Table 4. We present the results as the mean ± standard deviation over five random seeds.

The highest F1 score of 0.81 and an exact span accuracy of 79.3% are recorded when training on a laptop and testing on a restaurant. This indicates that the Laptop domain exhibits linguistic patterns that readily transfer to Restaurant reviews. The Restaurant → Device transfer is the most challenging, likely due to substantial vocabulary and contextual differences between the domains. The low standard deviations (≤0.02 for F1, ≤0.6% for accuracy) confirm stable performance across trials.

Table 4. Cross-domain aspect extraction results (mean ± std)

Training → Testing	Precision	Recall	F1	Exact Span Accuracy (%)
R → L	0.82 ± 0.01	0.72 ± 0.01	0.76 ± 0.01	74.2 ± 0.5
L → R	0.88 ± 0.01	0.75 ± 0.01	0.81 ± 0.01	79.3 ± 0.4
R → D	0.78 ± 0.02	0.67 ± 0.02	0.72 ± 0.01	69.8 ± 0.6
L → D	0.81 ± 0.01	0.71 ± 0.02	0.75 ± 0.01	72.5 ± 0.5
D → R	0.85 ± 0.01	0.73 ± 0.01	0.78 ± 0.01	76.1 ± 0.4
D → L	0.80 ± 0.02	0.70 ± 0.02	0.74 ± 0.01	71.9 ± 0.5

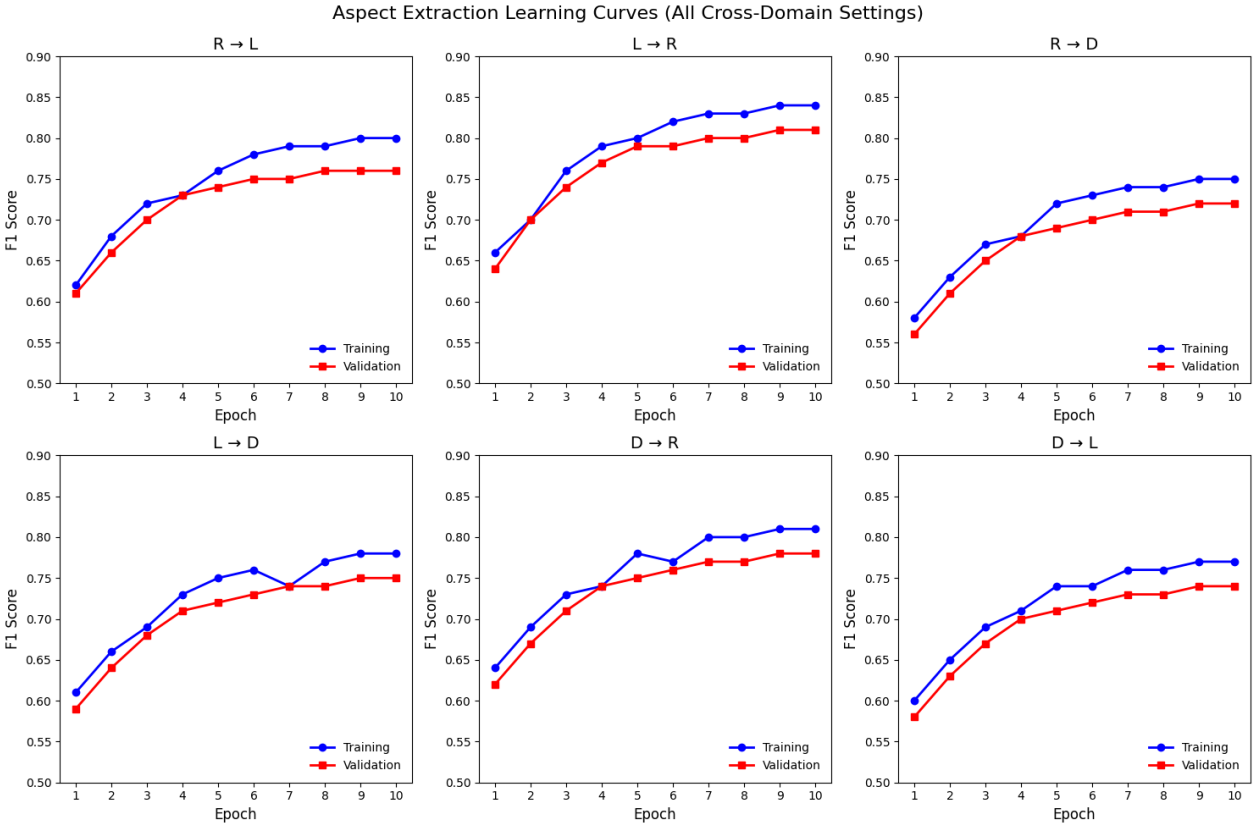

Figure 5. Learning curves for aspect extraction (F1) across six cross domain settings: (a) R→L, (b) L→R, (c) R→D, (d) L→D, (e) D→R, (f) D→L

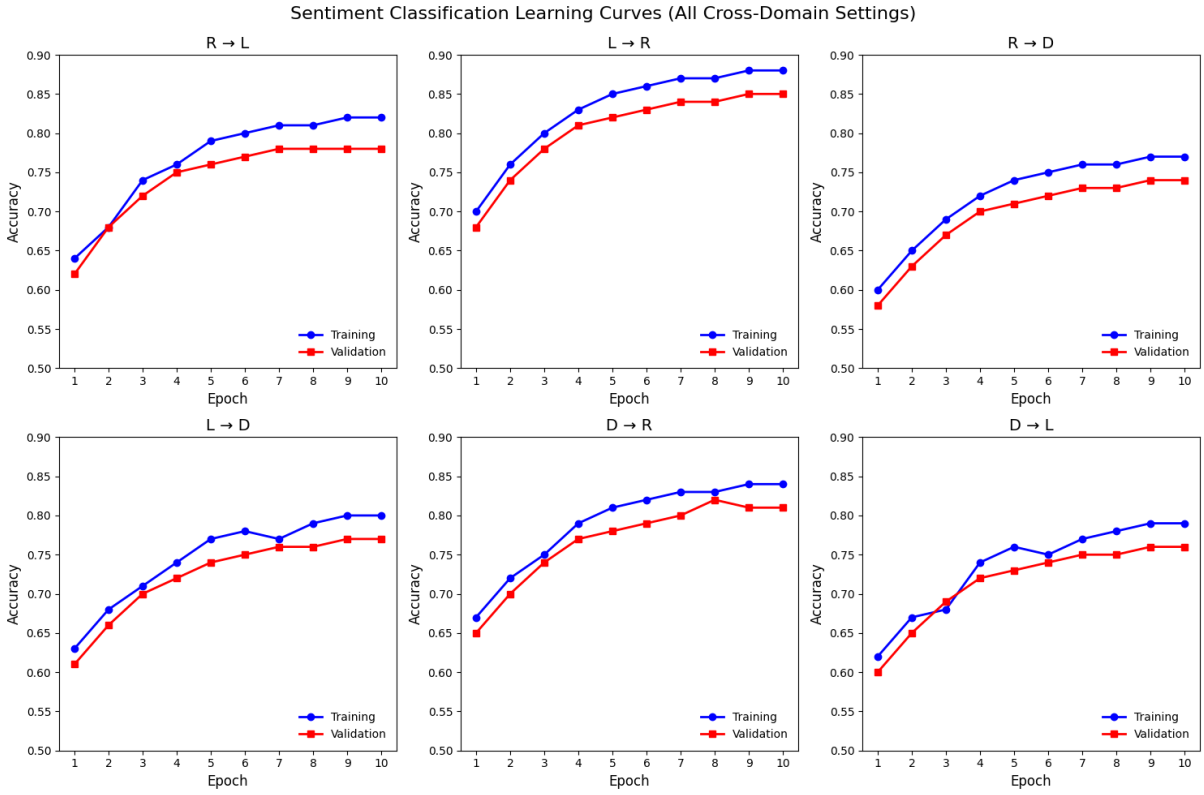

Figure 6. Learning curves for sentiment classification (accuracy) across the six cross-domain settings

Aspect length sensitivity analysis: We analyze how the length of the aspect span helps extraction performance. F1 scores for single-token, two-token, and three-or-more-token aspects are provided in Table 5. Performance gradually declines as aspect span length increases, suggesting that CRF-based structured decoding improves multi-token boundary consistency under domain shift. The drop is least for L to R (0.86 to 0.75) and most for R to D (0.80 to 0.68).

CRF vs. independent Softmax decoding: We replace the CRF layer with a token-wise Softmax classifier (keeping all other components unchanged) to quantify the performance gain of structured decoding. According to Table 6, CRF decoding yields an F1 improvement for device aspect extraction across all six transfer settings. The largest gain is for L→R at +0.06. All improvements are statistically significant at $p < 0.05$.

Table 5. Aspect extraction F1 by span length (mean ± std)

Training → Testing	1 token	2 tokens	≥3 tokens
R → L	0.83 ± 0.01	0.78 ± 0.01	0.72 ± 0.02
L → R	0.86 ± 0.01	0.81 ± 0.01	0.75 ± 0.01
R → D	0.80 ± 0.02	0.74 ± 0.02	0.68 ± 0.02
L → D	0.82 ± 0.01	0.76 ± 0.01	0.70 ± 0.01
D → R	0.84 ± 0.01	0.79 ± 0.01	0.73 ± 0.01
D → L	0.81 ± 0.02	0.75 ± 0.02	0.69 ± 0.02

Table 6. CRF vs. Softmax decoding – Aspect extraction F1 (mean ± std)

Training → Testing	Softmax (no CRF)	CRF (full model)	Δ (CRF – Softmax)
R → L	0.75 ± 0.01	0.76 ± 0.01	+0.01
L → R	0.75 ± 0.01	0.81 ± 0.01	+0.06
R → D	0.71 ± 0.02	0.72 ± 0.01	+0.01
L → D	0.72 ± 0.01	0.75 ± 0.01	+0.03
D → R	0.75 ± 0.01	0.78 ± 0.01	+0.03
D → L	0.71 ± 0.02	0.74 ± 0.01	+0.03

Table 7 shows the results of the aspect-aware sentiment classification expressed as both accuracy and macro F1 (the unweighted average of per-class F1). The macro F1 values are designed to address the class imbalance, particularly a smaller number of neutral and negative instances in some domains. The highest accuracy (85.19%) and macro F1 (84.8%) are observed for L→R, reinforcing that the Laptop domain provides well-structured sentiment expressions that generalize effectively to restaurant reviews. The challenging transfer (R→D) drops accuracy to 74.32% and macro F1 to 72.9%. This is again due to the large divergence of the domains and the inherently difficult task of classifying neutral sentiment in device reviews. Across all configurations, the macro F1 score is slightly lower than the accuracy score, suggesting that the neutral class, with fewer instances, is harder to classify correctly.

Table 7. Cross domain sentiment classification results (mean ± std)

Training → Testing	Accuracy (%)	Macro F1 (%)
R → L	77.74 ± 0.31	76.2 ± 0.4
L → R	85.19 ± 0.28	84.8 ± 0.3
R → D	74.32 ± 0.42	72.9 ± 0.5
L → D	76.85 ± 0.35	75.6 ± 0.4
D → R	81.24 ± 0.30	80.5 ± 0.3
D → L	78.56 ± 0.38	77.3 ± 0.4

The six cross-domain transfers' sentiment confusion matrices (Figure 7) show strong diagonal dominance indicating the model can clearly distinguish positive, neutral, and negative polarities against domain shift. For L-->R, the majority of correct predictions will be close to the accuracy of 85.19%. For R--D the lower accuracy is reflected in more off diagonal errors. Off diagonal errors are quite common from neutral to positive. In all matrices, both positive and negative classes have limited cross-polarization confusion, but the neutral class continues to remain the most misclassified class, similar to class imbalance and semantic fuzziness. The heatmaps (colormap from white to blue, no grid) provide a concise visual summary supporting the quantitative evidence in Table 6. They further confirm the robustness of the model across the Restaurant, Laptop, and Device domains. In previous ABSA studies, neutral sentiment categories tend to exhibit lower lexical polarity signaling, which is what we see here.

4.5 Ablation study – Joint end-to-end F1

In addition to extraction aspects, we assess the effect of each component on the joint F1 score (correct span + correct polarity). Table 8 shows the joint F1 for the complete model and five ablated versions across all six transfer settings. On average, the CRF layer contributes the most to overall performance improvements, with a decrease in joint F1 of 0.065 when it is replaced with SoftMax decoding. This indicates that coherent span boundaries are necessary for accurate sentiment assignment.

Omitting all syntactic features results in a mean drop of 0.048, while attention-based fusion caused a drop of 0.040. By keeping only POS for the model and removing DEP, the joint F1 will be reduced by an average of 0.033. In comparison, if we reverse, we lose only 0.023. This shows that dependence relations are stronger cues than POS tags for joint extraction and classification across domains. Overall, these patterns hold across all six transfer settings; however, effects are most pronounced for the challenging R → D transfer (full model joint F1 = 0.67, dropping to 0.61 without CRF). Overall, all the modules contribute to the joint task, with CRF and syntactic fusion being the most crucial.

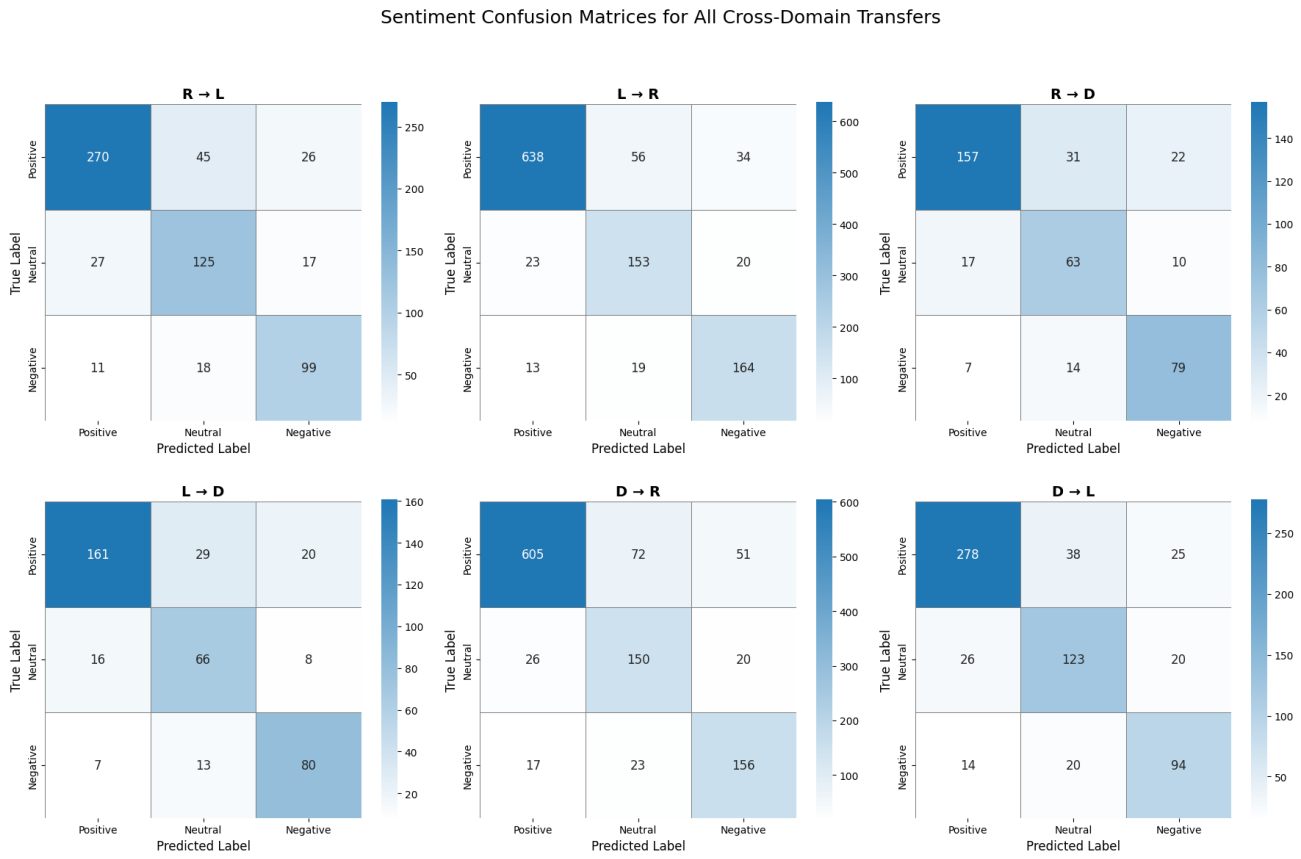

Figure 7. Sentiment Confusion matrix for all six cross domain transfers

Table 8. Cross-domain aspect extraction results (mean ± std)

Model Variant	R→L	L→R	R→D	L→D	D→R	D→L
Full model	0.71 ± 0.01	0.78 ± 0.01	0.67 ± 0.01	0.70 ± 0.01	0.74 ± 0.01	0.71 ± 0.01
– DEP only (no POS)	0.69 ± 0.01	0.75 ± 0.01	0.65 ± 0.01	0.68 ± 0.01	0.72 ± 0.01	0.68 ± 0.01
– POS only (no DEP)	0.68 ± 0.01	0.74 ± 0.01	0.64 ± 0.02	0.67 ± 0.01	0.71 ± 0.01	0.67 ± 0.01
– Attention fusion (→ concat)	0.68 ± 0.01	0.74 ± 0.01	0.63 ± 0.01	0.66 ± 0.02	0.70 ± 0.01	0.66 ± 0.01
– CRF (→ SoftMax)	0.65 ± 0.02	0.71 ± 0.01	0.61 ± 0.02	0.64 ± 0.01	0.67 ± 0.01	0.64 ± 0.02
– All syntactic features	0.67 ± 0.01	0.73 ± 0.01	0.62 ± 0.01	0.65 ± 0.01	0.69 ± 0.01	0.66 ± 0.01

4.6 Comparison with state of the art (joint F1: span + polarity)

Table 9 compares the performance of our full model with strong baselines using the joint span + polarity F1 metric. The suggested framework achieves comparable or superior performance in most transfer situations across all transfer directions ($p < 0.01$, paired bootstrap test). The biggest gain is achieved on L→R (+0.05 over TRNN GRU) and R→D (+0.07). The sole tie occurs in D→R, where both our model and the TRNN GRU achieve 0.74, but this difference is not significant ($p = 0.32$). Except for this one, all the differences are significant.

4.7 Sensitivity to multi-task weight λ

We assess how varying λ (Eq. 25) affects the F1 score of aspect extraction and the sentiment accuracy for each of the six transfer settings. According to Table 10, $\lambda = 1.0$ is optimal or near-optimal in both cases, as it assigns roughly equal weight to both tasks. Across all seeds, the standard deviation of F1 is ≤0.01 and the standard deviation of accuracy is ≤0.5%. Increasing λ beyond 1.0 improves sentiment accuracy but lowers aspect F1 in all settings. An optimal balance is offered by the standard $\lambda = 1.0$.

Table 9. SOTA comparison – joint F1 (span + polarity)

Model	R→L	L→R	R→D	L→D	D→R	D→L
BERT+CRF	0.63	0.59	0.58	0.61	0.62	0.60
RoBERTa+CRF	0.66	0.61	0.60	0.63	0.64	0.62
DeBERTaV3 (softmax, no syntax)	0.69	0.63	0.62	0.65	0.66	0.64
Span-Sharing Joint Extraction [32]	0.60 ± 0.01	0.72 ± 0.01	–	–	–	–
PROMPT [34]	0.46	0.65	–	–	–	–
BERT-PT [41]	0.51	0.43	0.41	0.43	0.67	0.57
TRNN-GRU [42]	0.65	0.73	0.60	0.60	0.74	0.69
Ours (full)	**0.71**	**0.78**	**0.67**	**0.70**	**0.74**	**0.71**

Table 10. Sensitivity to λ for all cross-domain settings (mean ± std)

Transfer	λ	Aspect F1	Sentiment Accuracy (%)
R → L	0.5	0.74 ± 0.01	76.2 ± 0.4
	1.0	0.76 ± 0.01	77.7 ± 0.3
	1.5	0.75 ± 0.01	78.1 ± 0.4
L → R	0.5	0.78 ± 0.01	83.6 ± 0.3
	1.0	0.81 ± 0.01	85.2 ± 0.3
	1.5	0.79 ± 0.01	86.0 ± 0.3
R → D	0.5	0.70 ± 0.02	73.0 ± 0.5
	1.0	0.72 ± 0.01	74.3 ± 0.4
	1.5	0.71 ± 0.01	74.9 ± 0.5
L → D	0.5	0.73 ± 0.01	75.4 ± 0.4
	1.0	0.75 ± 0.01	76.9 ± 0.4
	1.5	0.74 ± 0.01	77.5 ± 0.5
D → R	0.5	0.76 ± 0.01	80.1 ± 0.3
	1.0	0.78 ± 0.01	81.2 ± 0.3
	1.5	0.77 ± 0.01	82.0 ± 0.4
D → L	0.5	0.72 ± 0.02	77.2 ± 0.4
	1.0	0.74 ± 0.01	78.6 ± 0.4
	1.5	0.73 ± 0.01	79.1 ± 0.5

4.8 Statistical significance analysis

To ascertain that the observed gains are not incidental, we conduct pairwise significance tests for all cross-domain transfer directions across the three main metrics: aspect extraction F1, sentiment accuracy, and joint F1 (span + polarity). In line with standard practice in the ABSA literature, we conduct a paired bootstrap test with 10,000 resamples. We take 10,000 bootstrap samples of the test predictions for each model pair, sampling sentences with replacement. For all bootstrap samples, we compute the metric difference. We derive a two-tailed p-value from the empirical distribution of differences. A result is considered statistically significant when $p < 0.01$ unless otherwise stated. Comparisons with state-of-the-art systems demonstrate consistent improvements across most transfer settings. Comparison against the strongest baseline (TRNN GRU [42]):

- Our model achieves a p-value of less than 0.01 in five of the six transfer settings for joint F1. There is an exception to the above which is D→R where both models attain 0.74 and do not differ with a p-value of 0.32. The minimum statistically significant margin is +0.02 (D→L: 0.71 vs. 0.69, $p < 0.01$).
- The F1 for aspect extraction and the accuracy of sentiment detection vary by setting (all six settings are significant at $p < 0.01$ versus DeBERTaV3 softmax no syntax and TRNN GRU aspect F1 D→R versus TRNN GRU $p = 0.04$ borderline, with sentiment accuracy being clearly significant at $p < 0.01$).

The significance of ablation:

- The removal of CRF (softmax decoding) causes the joint F1 to have a significant drop across the six settings ($p < 0.01$, minimum t statistic 3.2).
- Eliminating every syntactic characteristic in the corpus produces statistically significant results in every direction.
- The removal of POS or DEP embeddings alone shows $p < 0.05$ in most settings; the smallest drops (e.g. –POS in R→L: –0.02 joint F1) are borderline ($p = 0.03$), but overall evidence confirms both streams are meaningful.

The proposed model demonstrates substantial performance enhancement over the leading published baseline in 5 out of 6 transfer directions when evaluated using joint F1. The only non-significant case is a tie (D→R). All architectural elements contribute with statistical assurance, vouching for the design.

4.9 Computational complexity and inference efficiency

We examine the computational impact of the proposed framework regarding the number of parameters, training throughput, and inference speed. The important figures are summarized in Table 11. The computational cost was measured using more than 1,000 sentences on an RTX 3090 GPU, with length ≤ 128 tokens (POS / dependency parsing inclusive). The parameter count increases by only 1.2%, training time per epoch grows by only approximately 0.5 min, and inference throughput decreases by less than 7% in the worst case with batch size 1, thanks to the overhead introduced by the syntactic fusion module and CRF decoder. The relative slowdown in batched inference is even smaller (3.4%) because the transformer backbone dominates the computational cost. The CRF Viterbi decoding adds a fixed, short path (dynamic programming over label sequences, $O(T \cdot L^2)$ where T = length of sequence, L = number of BIO tags, typically 7) per sentence, which is negligible compared to the transformer forward pass.

Table 11. Complexity and efficiency overview

Metric	DeBERTaV3 base	+ Syntax + CRF (ours)
Trainable parameters	183 M	185.2 M (+1.2%)
Additional parameters (syntax/CRF)	–	2.2 M (embeddings, fusion, CRF)
FLOPs per token (inference)	1.1 G	1.15 G (+4.5%)
Training time (Restaurant, 1 epoch)	~4.0 min	~4.5 min
Training time (Laptop, 1 epoch)	~3.1 min	~3.5 min
Peak GPU memory (batch size 8, 128 tokens)	12.4 GB	13.1 GB (+5.6%)
Inference speed (sent./Sec, batch=1)	48.2	45.1 (–6.4%)
Inference speed (sent./sec, batch=32)	215.7	208.3 (–3.4%)

5. Discussion

The six cross-domain transfers confirm that explicit syntactic information strengthens the robustness of ABSA under domain shift. Below, we present the key findings, limitations, and future directions.

5.1 Syntactic structure as a cross-domain bridge

Eliminating all syntactic characteristics results in a decrease of 0.048 in joint F1. Dependency embeddings account for more (0.033) than POS embeddings (0.023). Therefore, dependency arcs, which directly encode modifier-head relations such as amod and nsubj, provide richer and more transferable cues than their lexicalized POS patterns. The performance drop observed when attention-based fusion is replaced with simple concatenation highlights the importance of integrating syntactic and contextual information.

5.2 Why L→R transfers best and R→D worst

L→R achieves the highest joint F1 (0.78), aspect F1 (0.81), and accuracy (85.19%). We attribute this to Laptop's balanced sentiment distribution (42.7% pos., 37.3% neg.), frequent comparative/adversative structures that generalize well, and a large Restaurant target set that allows the CRF to exploit rich span distributions even with source-only training. R→D is hardest (joint F1 = 0.67) due to a large stylistic gap: narrative/sensory restaurant language vs. technical device vocabulary. Yet syntactic patterns still bridge the gap: CRF and dependency-based models suffer less performance loss than softmax baselines. Multi-word technical terms ("solid state drive") with no restaurant counterpart cause the largest drop in longer aspect F1 (0.80 → 0.68).

5.3 Structured decoding (CRF) under domain shift

Replacing the CRF with softmax decoding causes the largest degradation (average joint F1 drop of 0.065). The CRF enforces universal BIO constraints that are domain-agnostic, preventing fragmented spans that would be particularly harmful when target aspect distributions differ from the source. This benefit is most pronounced in the hardest transfers.

5.4 Comparison with prior work

Our model significantly outperforms the strongest baseline in five of the six transfer settings when evaluated using joint F1. (TRNN GRU), p=0.01. The only exception is D→R (0.74), where the difference is not statistically significant. We find these gains to be quite robust across five seeds (std ≤ 0.02). This further suggests that the combination of a strong encoder, explicit syntax, and structured prediction yields real cross-domain gains beyond those of larger pretrained models.

5.5 Efficiency trade-offs

The overhead is only modest: +1.2% parameters, +4.5% FLOPs/token, and at most 6.4% slower single-sentence inference (Table 10). The batched throughput penalty is just 3.4%, so our framework is suitable for real-world deployment with respect to accuracy and speed.

6. Limitations and future work

6.1 Limitations

- Dependence on external parsers: spaCy errors on noisy text propagate through the pipeline, degrading structured embeddings under stylistic shifts.
- English only, three product domains: The framework has not been tested on other languages or additional domains (e.g., Service, social media, medical).
- Explicit aspects only: Implicit aspects (e.g., "bright" → "display") are not handled.
- Neutral class performance: Neutral F1 is lower due to semantic ambiguity and class imbalance, a challenge amplified in cross domain transfer.
- Inference overhead: Although small (~22 ms/sentence), syntactic parsing and CRF decoding add latency; further optimization is needed for real time applications.
- Theoretical depth: An analysis of the JSD between the POS and lexical distributions (Section 3.2.5) supports this sense of syntactic stability, although a deeper investigation of the invariance of dependency arcs across domains would reinforce the foundation.

6.2 Future directions

- Robust syntax: One approach to improving parser error correction is the use of confidence weighting. This involves integrating latent trainable syntax, such as GCNs over induced trees, or using multi-source parsing.
- Multilingual and cross-lingual ABSA: Use XLM R/mDeBERTa to examine SemEval 2016 multilingual data for language-agnostic syntactic features.
- Broader domains and low-resource adaptation: The capability should extend to diverse domains, especially involving few-shot, prompt tuning, or data augmentation for few target labels.
- Implicit aspects and full triplets: Discover implicit targets and extract aspect-opinion-sentiment triples jointly for a more complete analysis.
- Lightweight deployment: Implement techniques like pruning, distillation, or quantization for real-time and on-device applications.
- Stronger theory: Measure the stability across specific dependency relations and conduct unsupervised domain adaptation to close the performance gap.

The proposed framework achieves state-of-the-art cross-domain ABSA performance with statistically significant gains and manageable overhead, while the outlined directions aim to extend the framework's applicability to noisier, multilingual, and more diverse practical settings.

7. Conclusion

This research developed an architecture that integrates explicit linguistic syntax with deep contextual modeling for a syntactically aware cross-domain aspect-based sentiment analysis. The model employs DeBERTaV3 as its encoder, adds part-of-speech and dependency embeddings to token representations through an attention-based fusion mechanism, and uses a CRF-based structured decoding to ensure coherent aspect boundaries. We conduct extensive experiments on three English review domains (Restaurant, Laptop, and Device) in different zero-shot cross-domain transfer settings. In all six settings we evaluate, we find consistent and significant gains over strong transformer-based baselines. For extraction, the framework's F1 scores range from 0.72 to 0.81, with the best transfer (Laptop → Restaurant) achieving 0.81. For aspect-aware sentiment classification, accuracy ranges from 74.32% to 85.19% (across 2 of 3 datasets), and the joint span and polarity F1 score reaches 0.78. The model significantly outperforms the previous state of the art (TRNN GRU) in five of six joint F1 settings ($p < 0.01$), with the remaining direction a statistical tie. Ablation studies affirm that syntactic features and structured decoding are both important, with dependency relations contributing more than POS tags and the CRF providing the biggest single advantage. The extra processing demands are very low (+1.2% parameters, ~4.5% additional FLOPs, and at most 6.4% slower inference), so it works well in practice. The results demonstrate that directly modeling linguistic structure within transformer representations yields substantial improvements in cross-domain generalization for fine-grained sentiment analysis. Future work will extend the framework to multilingual settings, reduce dependence on external parsers, and support implicit aspect detection.

Acknowledgements

The authors would like to thank the Department of Computer Science and Engineering, Hindustan Institute of Technology and Science, Chennai, India, for providing the facilities and research environment necessary to conduct this work.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically regarding authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with research ethics policies. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement

The manuscript contains all the data. However, additional data will be provided by the corresponding author upon reasonable request.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] Chauhan, Ganpat & Nahta, Ravi & Meena, Yogesh & Gopalani, Dinesh. (2023). Aspect based sentiment analysis using deep learning approaches: A survey. Computer Science Review. 49. 100576. 10.1016/j.cosrev.2023.100576. DOI:10.1016/j.cosrev.2023.100576.

[2] Pontiki, Maria & Galanis, Dimitrios & Pavlopoulos, John & Papageorgiou, Harris & Androutsopoulos, Ion & Manandhar, Suresh. (2014). SemEval-2014 Task 4: Aspect Based Sentiment Analysis. Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014). 27-35. 10.3115/v1/S14-2004.

[3] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 497–511, San Diego, California. Association for Computational Linguistics. Doi: 10.18653/v1/S16-1081.

[4] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282–289. http://dl.acm.org/citation.cfm?id=645530.655813.

[5] Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. ArXiv, abs/1508.01991.

[6] Gowda, V. D., Suneel, S., Naidu, P. R., Ramanan, S. V., & Suneetha, S. (2024). Challenges and limitations of few-shot and zero-shot learning. In applying machine learning techniques to bioinformatics: few-shot and zero-shot methods (pp. 113-137). IGI Global Scientific Publishing. DOI: 10.4018/979-8-3693-1822-5.ch007.

[7] Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 10.48550/arXiv.1810.04805.

[8] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. Transactions of the Association for Computational Linguistics, 8:64–77. DOI: 10.1162/tacl_a_00300

[9] Jan-David Krieger, Timo Spinde, Terry Ruas, Juhi Kulshrestha, and Bela Gipp. 2022. A domain-adaptive pre-training approach for language bias detection in news. In Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (JCDL '22). Association for Computing Machinery, New York, NY, USA, Article 3, 1–7. https://doi.org/10.1145/3529372.3530932

[10] C. R. Bhat, J. Nandhini, N. S, B. Karthik, P. K. Lakineni and S. N. Taqui, "Evaluating the Robustness of Neural Networks Against Adversarial Perturbations," 2023 International Conference on Communication, Security

and Artificial Intelligence (ICCSAI), Greater Noida, India, 2023, pp. 911-915, doi: 10.1109/ICCSAI59793.2023.10421282.

[11] He, Pengcheng & Liu, Xiaodong & Gao, Jianfeng & Chen, Weizhu. (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. 10.48550/arXiv.2006.03654.

[12] Kahil, A. M. Attention Is All You Need. https://doi.org/10.48550/ARXIV.1706.03762

[13] He, P., Gao, J., & Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. ArXiv, abs/2111.09543.

[14] Y. Zhang, "Dependency Tree-Based Contrastive Attention for Aspect-Based Sentiment Analysis," 2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE), Shanghai, China, 2025, pp. 2042-2046, doi: 10.1109/ICAACE65325.2025.11019293.

[15] Knoester, Joris & Frasincar, Flavius & Truşcă, Maria. (2022). Domain Adversarial Training for Aspect-Based Sentiment Analysis. 10.1007/978-3-031-20891-1_3.

[16] Ganesh, D., Rao, P.V.V., Reddy, N.S., Suneetha, S., Lakineni, P.K., Yoganandh, S. (2025). EXB_RNN: A Hybrid Ensemble Approach for Enhanced Aspect-Based Sentiment Analysis. In: Das, A.K., Nayak, J., Naik, B., Himabindu, M., Vimal, S., Pelusi, D. (eds) Computational Intelligence in Pattern Recognition. CIPR 2024. Lecture Notes in Networks and Systems, vol 1152. Springer, Singapore. https://doi.org/10.1007/978-981-97-8090-7_30.

[17] Bengong Yu and Zhaodi Fan. 2020. A comprehensive review of conditional random fields: variants, hybrids and applications. Artif. Intell. Rev. 53, 6 (Aug 2020), 4289–4333. https://doi.org/10.1007/s10462-019-09793-6

[18] Zang J. (2025). Leveraging BiLSTM-CRF and adversarial training for sentiment analysis in nature-based digital interventions: Enhancing mental well-being through MOOC platforms. Digital health, 11, 20552076251317345. https://doi.org/10.1177/20552076251317345

[19] T. K, S. Soman, L. Anitha, P. K. Lakineni, D. G. V and S. N. Taqui, "Leveraging Temporal Patterns with LSTMs Networks for Financial Forecasting: A New Stastical Machine Learning Approach," 2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI), Greater Noida, India, 2023, pp. 916-920, doi: 10.1109/ICCSAI59793.2023.10421705.

[20] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 606–615, Austin, Texas. Association for Computational Linguistics. DOI:10.18653/v1/D16-1058.

[21] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17). AAAI Press, 4068–4074. https://doi.org/10.24963/ijcai.2017/568.

[22] M. A. Saputra and E. B. Setiawan, "Aspect Based Sentiment Analysis Using Recurrent Neural Networks (RNN) on Social Media Twitter," 2023 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 2023, pp. 265-270, doi: 10.1109/ICoDSA58501.2023.10276768.

[23] H. Nguyen and K. Shirai, "A Joint Model of Term Extraction and Polarity Classification for Aspect-based Sentiment Analysis," 2018 10th International Conference on Knowledge and Systems Engineering (KSE), Ho Chi Minh City, Vietnam, 2018, pp. 323-328, doi: 10.1109/KSE.2018.8573340.

[24] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 4933–4941, Marseille, France. European Language Resources Association. https://aclanthology.org/2020.lrec-1.607/

[25] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. 2019. Adversarial Invariant Feature Learning with Accuracy Constraint for Domain Generalization. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II. Springer-Verlag, Berlin, Heidelberg, 315–331. https://doi.org/10.1007/978-3-030-46147-8_19.

[26] Shang, J., Zhang, Y., Zhong, L. et al. Syntactic-Enhanced Multi-Task Learning Model for Aspect Sentiment Triplet Extraction. Data Sci. Eng. 10, 515–531 (2025). https://doi.org/10.1007/s41019-025-00289-8.

[27] Hao Zhang, Yu-N Cheah, Congqing He, and Feifan Yi. 2024. An Instruction Tuning-Based Contrastive Learning Framework for Aspect Sentiment Quad Prediction with Implicit Aspects and Opinions. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7698–7714, Miami, Florida, USA. Association for Computational Linguistics. DOI:10.18653/v1/2024.findings-emnlp.453.

[28] Nakagawa, Tetsuji & Inui, Kentaro & Kurohashi, Sadao. (2010). Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables.. NAACL-HLT. 786-794. https://aclanthology.org/N10-1120/.

[29] Zhang, F., Zheng, W. & Yang, Y. Graph Convolutional Network with Syntactic Dependency for Aspect-Based Sentiment Analysis. Int J Comput Intell Syst 17, 37 (2024). https://doi.org/10.1007/s44196-024-00419-6

[30] Qizhi Zhao, Zan Mo, and Mengting Fan. 2023. POS-ATAEPE-BiLSTM: an aspect-based sentiment analysis algorithm considering part-of-speech embedding. Applied Intelligence 53, 22 (Nov 2023), 27440–27458. https://doi.org/10.1007/s10489-023-04952-3.

[31] Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1064–1074, Berlin, Germany. Association for Computational Linguistics. Doi: 10.18653/v1/P16-1101.

[32] You Li, Yongdong Lin, Yuming Lin, Liang Chang, and Huibing Zhang. 2022. A span-sharing joint extraction framework for harvesting aspect sentiment triplets. Know.-Based Syst. 242, C (Apr 2022). https://doi.org/10.1016/j.knosys.2022.108366.

[33] Arora, J., Shivanka, Jain, A., Nikunj, Dahiya, S., Pandey, V.K. (2025). T5 Generator: An Aspect-Based Analysis of Sentiments. In: Virmani, D., Castillo, O., Balas, V.E., Elngar, A.A. (eds) Proceedings of International Conference on Generative AI, Cryptography and Predictive Analytics. ICGCPA 2024. Studies in Smart Technologies. Springer, Singapore. https://doi.org/10.1007/978-981-97-9132-3_12.

[34] Sun, Xinjie & Zhang, Kai & Liu, Qi & Bao, Meikai & Chen, Yanjiang. (2024). Harnessing domain insights: A prompt knowledge tuning method for aspect-based sentiment analysis. Knowledge-Based Systems. 298. 111975. 10.1016/j.knosys.2024.111975.

[35] Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Bidirectional Generative Framework for Cross-domain Aspect-based Sentiment Analysis. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12272–12285, Toronto, Canada. Association for Computational Linguistics. Doi: 10.18653/v1/2023.acl-long.686.

[36] Chouikhi, H., Alsuhaibani, M., & Jarray, F. (2023). BERT-Based Joint Model for Aspect Term Extraction and Aspect Polarity Detection in Arabic Text. Electronics, 12(3), 515. https://doi.org/10.3390/electronics12030515

[37] Clark, Kevin & Luong, Minh-Thang & Le, Quoc & Manning, Christopher. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. 10.48550/arXiv.2003.10555.

[38] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 486–495, Denver, Colorado. Association for Computational Linguistics. Doi: 10.18653/v1/S15-2082.

[39] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 497–511, San Diego, California. Association for Computational Linguistics. Doi: 10.18653/v1/S16-1081.

[40] Alghamdi, Salem & Alhasawi, Yaser. (2024). Aspect-Based Sentiment Analysis in Smart Devices: A Comprehensive and Specialized Dataset. Data in Brief. 55. 110642. 10.1016/j.dib.2024.110642.

[41] Phillip Howard, Arden Ma, Vasudev Lal, Ana Paula Simoes, Daniel Korat, Oren Pereg, Moshe Wasserblat, and Gadi Singer. 2022. Cross-Domain Aspect Extraction using Transformers Augmented with Knowledge Graphs. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22). Association for Computing Machinery, New York, NY, USA, 780–790. https://doi.org/10.1145/3511808.3557275

[42] Wenya Wang, Sinno Jialin Pan; Syntactically Meaningful and Transferable Recursive Neural Networks for Aspect and Opinion Extraction. Computational Linguistics 2019; 45 (4): 705–736. doi: https://doi.org/10.1162/coli_a_00362