Dynamic Multi-granularity Attribution Network for Aspect-based Sentiment Analysis

Anonymous ACL submission

Abstract

Aspect-based sentiment analysis (ABSA) aims to predict the sentiment polarity of a specific aspect within a given sentence. Most existing methods predominantly leverage semantic or syntactic information based on attention scores, which are susceptible to interference caused by irrelevant contexts and often lack sentiment knowledge at a data-specific level. In this paper, we propose a novel ***D**ynamic **M**ulti-granularity **A**ttribution **N**etwork* (DMAN) from the perspective of attribution. Initially, we leverage Integrated Gradients to dynamically extract importance scores for each token, which contain underlying reasoning knowledge for sentiment analysis. Subsequently, we aggregate attribution representations from multiple semantic granularities in natural language, enhancing profound understanding of the semantics. Finally, we integrate attribution scores with syntactic information to more accurately capture the relationships between aspects and their relevant contexts during the sentence understanding process. Extensive experiments on five benchmark datasets demonstrate the effectiveness of our proposed method.

1 Introduction

Aspect-based sentiment analysis (ABSA) is a fine-grained classification task that focuses on identifying the sentiment polarity of specific aspects within a sentence (Jiang et al., 2011; Pontiki et al., 2014). For instance, given a sentence "*The street is very crowded, but the atmosphere is pleasant*", the task aims to predict sentiment polarity associated with two aspects *"street"* and *"atmosphere"*, which are negative and positive respectively.

The core challenge of ABSA is to model the connection between aspect and its contexts, especially those parts that express opinions and ideas. To this end, various studies (Tang et al., 2016; Fan et al., 2018; Chen et al., 2020; Zhang et al., 2021) concentrate on attention mechanisms to model the relationships between aspect and its context. In addition, many methods (Zhang et al., 2019; Tang et al., 2020; Li et al., 2021; Zhang et al., 2022b) leverage syntactic information derived from dependency trees to better capture the interactions between aspects and opinion expressions. Recently, methods incorporating Pre-trained Language Models (Zhang et al., 2022a; Yin and Zhong, 2024; Sun et al., 2024) have demonstrated impressive results in ABSA. Despite these significant advancements, critical challenges persist when directly applying attention mechanisms or syntactic information to this fine-grained task.

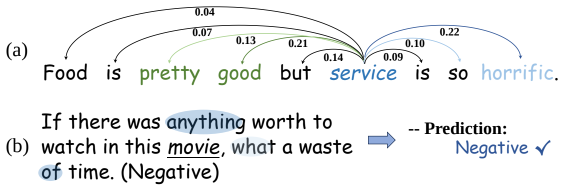


Figure 1: (a) Attention mechanism assigns high scores to words unrelated to aspect *service*. (b) We construct attention weights on irrelated words and overlook opinion words, but still yield right prediction.

Specifically, attention-based methods may inappropriately assign high attention scores to words that are irrelevant to the aspect. Li et al. (2021); Zhang et al. (2022b); Ma et al. (2023) propose that's because attention mechanisms are vulnerable to noise within sentences. As shown in Figure 1 (a) , the aspect "service" receives disproportionately high attention scores for the unrelated opinion words "pretty" and "good". Furthermore, some research that focuses on interpretability of attention mechanisms (Serrano and Smith, 2019; Jain and Wallace, 2019; Bibal et al., 2022) have indicated that attention scores do not always correlate with significance. Serrano and Smith (2019) have dicovered that removing features deemed important by attention scores leads to less prediction flip than gradient-based strategies. Besides, Jain and Wallace (2019) have observed shuffling the attention weights often does not affect the final prediction, which is consistent with our observations shown in Figure 1 (b). To sum up, while attention mechanisms have improved the performance of ABSA, they often operate as a black box, leaving their ability to accurately capture critical opinion words remains debatable. This underscores the need for methods that efficiently capture keywords for reasoning sentiment polarity. Additionally, although leveraging syntactic knowledge has shown to improve performance, it is important to recognize that not all syntactic information is equally beneficial to this fine-grained task. Syntactic information irrelevant to the aspect can be redundant and may even introduce noise rather than provide useful insights. Therefore, it is crucial to focus on extracting relevant syntactic information, emphasizing the identification of important words within sentences.

To address the aforementioned issues, we introduce attribution analysis into ABSA and propose a Dynamic Multi-granularity Attribution Network (DMAN). Attribution information reflects the importance of different tokens towards the prediction, which contain reasoning knowledge of the sentiment at data-driven level. Initially, we employ Integrated Gradients (IG) (Sundararajan et al., 2017), a well-established gradient-based attribution method, to compute the importance scores of tokens. Inspired by the observation (Brouwer et al., 2021; Zhang et al., 2022a) that the significance of essential words dynamically changes during semantic comprehension, we design multi-step attribution analysis to capture the dynamic significance of tokens during the comprehension process. More concretely, we utilize stacked self-attention blocks in conjunction with IG to calculate attribution scores for each layer, and adopt a Top-K strategy to filter out dimensions with low values, thereby reducing the impact of trivial dimensions. Subsequently, we incorporate semantic representations at both token and span levels to derive multi-granularity attribution scores, ensuring more comprehensive semantic concepts. Finally, we construct the adjacency matrices based on the dependency tree, and then use obtained attribution scores to initialize different adjacency matrices for different layers of GCNs, which facilitates the dynamic capture of key syntactic knowledge during throughout the process of sentence comprehension.

In summary, our contributions could be summarized as follows:

- To the best of our knowledge, we are the first to introduce attribution analysis into the ABSA task, which provides data-specific insights for reasoning sentiment polarity.

- We propose a novel model DMAN that leverages IG to dynamically extract attribution scores of tokens from multi-granularity perspectives. Furthermore, we integrate these scores with syntax to capture essential syntactic elements during sentence comprehension.

- Extensive experiments on five public benchmark datasets show the effectiveness and interpretability of our proposed DMAN.

2 Related Works

2.1 Aspect-based Sentiment Analysis

The goal of ABSA is to identify the sentiment polarity of specific aspect in the sentence. In recent years, various approaches have utilized attention mechanisms to investigate the semantic correlations between contexts (Tang et al., 2016; Wang et al., 2016; Ma et al., 2017; Fan et al., 2018; Tan et al., 2019; Liang et al., 2019; Pang et al., 2021; Zhang et al., 2021). For instance, Ma et al. (2017) proposed the interactive attention networks to interactively learn attentions in the contexts and targets. Fan et al. (2018) exploited a novel multi-grained attention network to capture the interaction between aspect and context. Tan et al. (2019) designed dual attention mechanisms to distinguish conflicting opinions. Zhang et al. (2021) proposed a cross-domain feature learning module with an aspect-oriented multi-head attention mechanism.

In addition, various research (Zhang et al., 2019; Huang and Carley, 2019; Sun et al., 2019; Wang et al., 2020; Tang et al., 2020; Li et al., 2021; Tian et al., 2021; Zhang et al., 2022b; Yin and Zhong, 2024) proposes different methods that leverage syntactic knowledge to model relationships between aspects and contexts. For instance, Wang et al. (2020) proposed a relational graph attention network to encode the new tree structure. Li et al. (2021) designed a dual graph convolutional network to model syntax structures and semantic correlations simultaneously. Tian et al. (2021) exploited an approach to explicitly utilize dependency types with type-aware graph convolutional networks, and Yin and Zhong (2024) proposed a double-view graph Transformer to alleviate the over-smoothing problem.

The core idea underlying these methods is to comprehend the semantics and syntax of sentences, thereby directing greater attention to significant words. Distinct from these approaches, our study pioneers the investigation of ABSA from an attribution perspective, unveiling the reasoning processes behind sentiment polarity at a data-driven level.

2.2 Attribution Analysis

The purpose of attribution analysis (Baehrens et al., 2010; Ancona et al., 2018; Brunner et al., 2020) is to assign importance scores the intermediate or input elements of a network, which matches well with the objectives of sentiment analysis. There are various types of attribution methods. Occlusion-based techniques (Zeiler and Fergus, 2014) determine the significance of each feature by occluding it and comparing the resulting output to the original. Gradient-based methods (Sundararajan et al., 2017; Ding et al., 2019; Serrano and Smith, 2019; Brunner et al., 2020; Bibal et al., 2022) use the gradient information of features to approximate their importance. Compared to occlusion-based methods, gradient-based methods are generally faster as they require only a single forward pass. Perturbation-based methods (Guan et al., 2019; De Cao et al., 2020; Ivanovs et al., 2021) add noise to features to evaluate their significance for model predictions.

Attribution analysis has not been extensively explored in aspect-based sentiment analysis. In our work, we take the initiative to investigate whether attribution analysis can enhance ABSA performance and provide more reliable interpretations.

3 Methods

In this section, we describe our proposed DMAN in detail. Specifically, we begin with the problem definition, followed by encoder module and the overall architecture of DMAN.

Problem Definition. Given a sentence-aspect pair (s, a), where $s = \{w_1, w_2, ..., w_n\}$ is a sentence with n words, and $a = \{a_1, a_2, ..., a_m\}$ is the given aspect. ABSA aims to predict sentiment polarity of aspect a in the sentence s.

Encoder. We utilize BERT as sentence encoder to extract aspect-specific context representations. We construct input as "$[CLS]$ s $[SEP]$ a $[SEP]$" to map each word into a real-value vector, getting sentence embedding $E_0 = \{e_1, e_2, ..., e_n\}$ and aspect embedding $E_a = \{e_{a_1}, e_{a_2}, ..., e_{a_m}\}$.

Overall architecture. As illustrated in Figure 2, our proposed Dynamic Multi-granularity Attribution Network comprises three primary components: (1) Multi-step Attribution Extraction, (2) Multi-granularity Attribution, and (3) Dynamic Syntax Concentration. The technical details will be elaborated on as follows.

3.1 Multi-step Attribution Extraction

Integrated Gradients. (Sundararajan et al., 2017) proposed IG for attributing the prediction of a deep network to its input or intermediate features. Formally, suppose a function F to represent a network, and let x be the input feature and x' be the baseline feature, IG considers the straight line path from x' to x and aggregate the gradients at all points along the path. The Integrated Gradients of i-th dimension is defined as $IG_i(F,\ x)$ as follows:

$$\mathrm{IG_i}(F,\ x) = (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'_i))}{\partial x_i} d\alpha. \tag{1}$$

Attribution Extraction. In this study, we design a stacked self-attention architecture to facilitate semantic comprehension and dynamically caputre attribution knowledge at each layer. Unlike traditional methods that utilize attention mechanisms for final classification, we treat the attention layers as black boxes for semantic understanding, concentrating on the gradient variations of tokens. Specifically, given sentence embedding H_0 from encoder, we process it through multiple blocks consisting of Self-Attention and Feed-Forward Networks (FFN), which can be formulated as follows:

$$E'_l = \mathrm{softmax}\left(\frac{(E_{l-1}W_l^q)(E_{l-1}W_l^k)^T}{\sqrt{d_k}}\right) E_{l-1}W_l^v, \tag{2}$$

$$E_l = \max(0,\ E'_l W_l^1 + b_l^1)W_l^2 + b_l^2, \tag{3}$$

where W_l^k, W_l^q, W_l^v, W_l^1, W_l^2 are learnable model parameters of l-th layer, $E_l \in \{e_1^l, e_2^l, ..., e_n^l\}$ is the product of l-th layer while E_{l-1} is the output from the preceding layer.

Then we map the final features from the stacked blocks into a probability distribution $P_c = [P_1, ..., P_C] \in R^C$, where c presents the sentiment polarity labels. In our approach, we denote the function $E \rightarrow P^c$ as F^c, and we conduct exhaustive attribution analysis for each dimension of input features and obtain attribution scores of i-th token, which could be denoted as $\mathrm{IG_i}$:

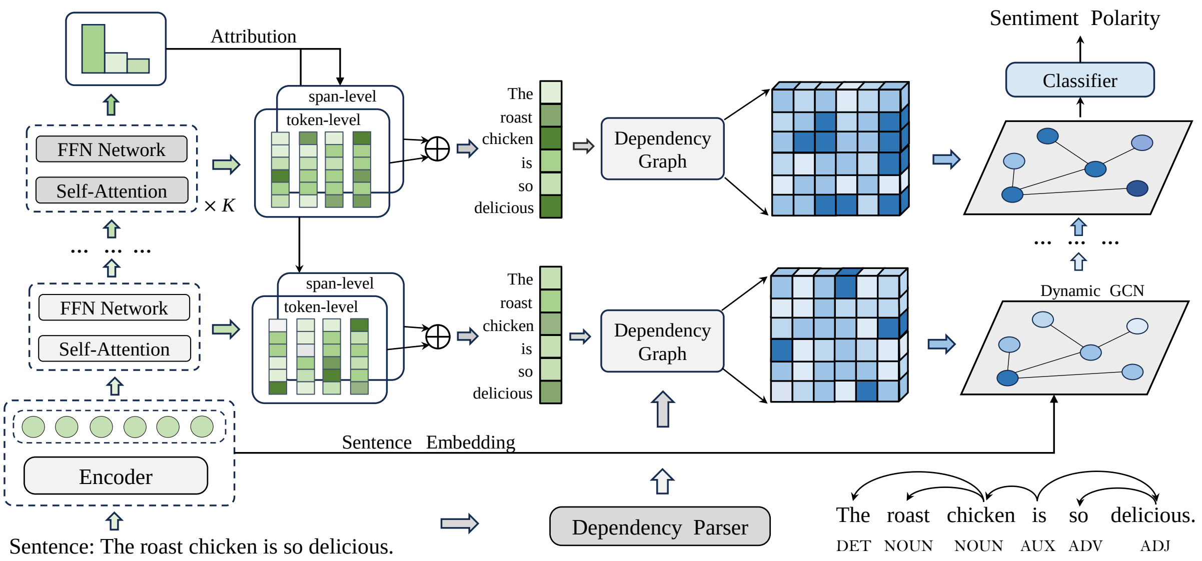


Figure 2: The overall architecture of our proposed DMAN, which consists of three modules arranged from left to right: Multi-step Attribution Extraction, Multi-granularity Attribution, and Dynamic Syntax Concentration.

$$\mathrm{IG_i}(F^c, E) = \sum_{j=1}^{m} \mathrm{IG_{ij}}(F^c, E) \tag{4}$$

$$= \sum_{j=1}^{m} (e_{ij} - e'_{ij}) \times \int_{\alpha=0}^{1} \frac{\partial F^c(e'_{ij} + \alpha \times (e_{ij} - e'_{ij}))}{\partial e_{ij}} d\alpha.$$

During the process, we employ an efficient approximation technique for estimating integral calculations, which significantly enhanced computational efficiency:

$$\mathrm{IG_{ij}}(F^c, E) \approx \sum_{t=1}^{T} < \nabla_{e_i} F^c(e'_{ij} + \Delta e_k), (e_{ij} - e'_{ij}) >$$

$$= \frac{(e_{ij} - e'_{ij})}{T} \times \sum_{t=1}^{T} \frac{\partial F^c(e'_{ij} + \frac{t}{T} \times (e_{ij} - e'_{ij}))}{\partial e_{ij}}. \tag{5}$$

In our implementation, we use zero vectors as baseline features to reflect the significance of each token. Symbols are excluded from consideration, and absolute values are used to aggregate attributions across each dimension, thereby deriving token-level attribution values. Recognizing that not all dimensions hold equal significance, selecting the crucial dimensions becomes essential. During the computational process, we observed that certain dimensions consistently maintain low values, failing to effectively differentiate between various tokens or stages. Therefore, we employ the Top-K algorithm to filter out dimensions with low attribution influence, which is denoted as:

$$\mathrm{IG'_i}(F^c,\ E) = |\ \mathrm{Top-K}(\mathrm{IG_i}(F^c,\ E))\ |. \tag{6}$$

In our study, attribution analysis is conducted on each self-attention block to thoroughly elucidate the dynamic semantic comprehension. The attribution value of the k-th layer is denoted as V_k:

$$V_K = \|_{i=1}^{n} \mathrm{IG'_i}(F^c,\ E_k), \tag{7}$$

where $\|$ represents the concatenation operation and $V_k \in \{v_1^k, v_2^k, ..., v_n^k\}$.

3.2 Multi-granularity Attribution

Most existing ABSA approaches focus on single granularity representation, overlooking the fact that texts are comprehensive representations constructed across multiple granularity levels (i.e. token, span, sentence). To the end, our method extracts attribution from both token and span granularities, providing hierarchical information that aids in a deeper understanding of the underlying motivations behind sentiment.

The first granularity is the token level. Given the vector V_k, v_i^k represents the attribution value of the i-th token, offering a fine-grained level of representation. he second granularity is the span, which may consist of consecutive words. To ensure semantic coherence, we extract phrases that convey complete meaning as spans. For instance, in the sentence "*The Mona Lisa is a famous painting housed in the Louvre Museum*", "*Mona Lisa*" and "*Louvre Museum*" are meaningful spans. We utilize spaCy[1] toolkit to construct spans $s_{span} = [s_1, s_2, ..., s_n]$, where $s_i = [w_j, ..., w_{j+q_i-1}]$ denotes i-th token

[1]We use spaCy toolkit: https://spacy.io/

belongs to a span a span starting at the j-th token and containing q_i tokens. Subsequently, for tokens belonging to a specific span, we employ mean pooling to obtain span-level attribution values:

$$\hat{v}_i^k = (\sum_j^{j+q_i-1} v_j^k) \ / \ q_i, \tag{8}$$

where $\hat{v}_i^k$ is span-granularity attribution of i-th token, thus we obtain $\hat{V}_k = \{\hat{v}_1^k, \hat{v}_2^k, ..., \hat{v}_n^k\}$. Then, we design a simple linear operation to integrate token-level and span-level attribution values:

$$\overline{V}_k = (\alpha V_k \ + \ (1-\alpha)\hat{V}_k)/\tau_k, \tag{9}$$

where $\overline{V}_K$ is integrated multi-granularity attribution score of k-th layer, α and τ_k is the coefficient hyperparameter of the k-th layer.

3.3 Dynamic Syntax Concentration

Leveraging syntactic information has significantly improved the performance of ABSA (Tang et al., 2020; Li et al., 2021; Zhang et al., 2022b). However, we propose that syntactic information within a sentence does not always hold equal importance. As semantic understanding is a dynamic process, the the critical syntactic elements also change dynamically in response to this process.

In our approach, we adjust dependency relationships based on multi-step attribution scores to achieve dynamic syntax concentration. Specifically, we construct adjacent matrix A according to the dependency tree derived from spaCy:

$$\boldsymbol{A_{ij}} = \begin{cases} 1 & \text{if link}(i,j) = True \text{ or } i = j, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where $link(i, \ j)$ represents whether i-th and j-th token have a dependency relationship. To model the dynamic changes of key syntactic information during sentence comprehension, we utilize attribution $\overline{V}_k$ to derive the dynamic adjacency matrix A^k. Then, we employ GCNs to capture syntactic knowledge, which can be formulated as:

$$A^k = \overline{V}_k \ \otimes A, \tag{11}$$

$$h_i^k = \text{ReLU}(\sum_{j=1}^{n} A_{ij}^k W^k h_j^{k-1} + b^k \ ), \tag{12}$$

where h_i^k is the i-th token representation of k-th GCN, W^k and b^k are learnable parameters. The output of the k-th layer is $H^k = \{h_0^k, h_1^k, ..., h_n^k\}$, and initial input $H_0 = E_0$. With these above calculations, we finally obtain dynamic syntax-enhanced representations for subsequent classification.

3.4 Model Training

Attribution Analysis. During the process of multi-step attribution extraction, we map the final representation into a probability distribution P, and apply the following function to extract attribution:

$$\mathcal{L}_A = -\sum_{i=1}^{M}\sum_{c=1}^{C} y_i^c \log\left(p_i^c\right), \tag{13}$$

where y_i^c is the ground truth label, C is the number of labels, M is the number of training samples.

Sentiment Classification. After obtaining dynamic syntax-enhanced representation H^k, we concatenate it with original sentence representation E_0 to get the final sentiment classification features. Then we map it to the probabilities over sentiment polarities over a softmax layer:

$$z = [H^k, \ E_0], \tag{14}$$

$$\hat{y} = softmax(W_z z + b_z), \tag{15}$$

where W_z and b_z are trainable parameters.Finally, we use cross-entropy loss as our objective function:

$$\mathcal{L} = -\sum_{i=1}^{M}\sum_{c=1}^{C} y_i^c \log\left(\hat{y}_i^c\right). \tag{16}$$

Datasets	Positive		Neutral		Negative	
	Train	Test	Train	Test	Train	Test
Lap14	994	341	464	169	870	128
Rest14	2164	728	637	196	807	196
Rest15	912	326	36	34	256	182
Rest16	1240	469	69	30	439	117
MAMS	3380	400	5042	607	2764	329

Table 1: The statistics of five benchmark datasets.

4 Experiments

4.1 Datasets

We evaluate our DMAN on five public standard datasets, including Lap14 and Rest14 from (Pontiki et al., 2014), Rest15 from (Pontiki et al., 2015), Rest16 from (Pontiki et al., 2016), and MAMs from (Jiang et al., 2019). We adopt the official data splits, which strictly keep the same as previous papers, and we use the accuracy and macro-averaged F1 value as the main evaluation metrics. Each sample in these datasets consists of a sentence, an aspect, and the sentiment polarity. The statistics of the datasets are presented in Table 1.

Models	Lap14		Rest14		Rest15		Rest16		MAMs	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BERT-SPC (Song et al., 2019)	78.99	75.03	84.46	76.98	83.40	65.28	89.54	70.47	80.11	80.34
R-GAT (Wang et al., 2020)	78.21	74.07	86.60	81.35	81.80	68.21	89.51	75.81	82.93	82.75
DGEDT (Tang et al., 2020)	79.80	75.60	86.30	80.00	84.00	71.00	91.90	79.00	-	-
DualGCN (Li et al., 2021)	81.80	78.10	87.13	81.16	84.69	72.97	89.87	77.26	83.83	83.47
T-GCN (Tian et al., 2021)	80.88	77.03	86.16	79.95	85.26	71.69	92.32	77.29	83.38	82.77
SSEGCN (Zhang et al., 2022b)	81.01	77.96	87.31	81.09	-	-	-	-	-	-
MGFN (Tang et al., 2022)	81.83	78.26	87.31	82.37	84.40	72.66	92.04	**81.57**	-	-
TF-BERT (Zhang et al., 2023)	81.80	78.46	87.09	81.15	-	-	-	-	-	-
RSC (Wang et al., 2023)	81.56	75.92	87.45	82.41	83.98	70.86	91.61	77.44	84.68	84.23
TextGT (Yin and Zhong, 2024)	81.33	78.71	87.31	82.27	-	-	-	-	-	-
Our DMAN	**82.29**	**78.91**	**87.59**	**82.47**	**86.30**	**72.97**	**92.85**	77.37	**85.55**	**85.01**

Table 2: Experiment results (%) comparison on five publicly benchmark datasets. The best scores are bolded, and the second best ones are underlined. All models are based on BERT.

4.2 Implementation Details

In the implementation, we build our framework based on *bert-based-uncased* with max length as 90. We employ the AdamW optimizer to optimize parameters. The embedding size is set to 768. The batch size is manually tested in [16, 32] and the learning rate is carefully tuned amongst [1e-5, 2e-5, 4e-5]. The dropout rate is set to 0.1. The number of Multi-step is finally set to 2 and the K value of Top-K is tested between 10 and 300. Correspondingly, the number of GCN layers is set to 2. The hyper-parameter α is set to 0.6, and τ_k is adjusted amongst [0.04, 0.07] for different layers. We conduct experiments on a single NVIDIA 4090 GPU.

4.3 Baselines

To validate the effectiveness of our approach, we compared it with advanced baseline models. To ensure a fair comparison, all selected baselines are based on the *bert-based-uncased* architecture.

BERT-SPC (Song et al., 2019) feed the contexts and aspects into the BERT model for the sentence pair classification task.

RGAT (Wang et al., 2020) generate a unified aspect-oriented dependency tree proposes a relational graph attention network to encode the tree.

DGEDT (Tang et al., 2020) propose a dependency graph dual-transformer network by considering flat representations and graph-based representations.

DualGCN (Li et al., 2021) propose a dual graph convolutional networks model that considers syntax structures and semantic correlations.

T-GCN (Tian et al., 2021) propose an approach to explicitly utilize dependency types for ABSA with type-aware graph convolutional networks.

SSEGCN (Zhang et al., 2022b) design an aspect-aware attention mechanism to enhance the node representations with GCN.

MGFN (Tang et al., 2022) leverage the richer syntax dependency relation label information and affective semantic information of words.

TF-BERT (Zhang et al., 2023) propose a novel table filling based model, which considers the consistency of multi-word opinion expressions.

RSC (Wang et al., 2023) propose two straightforward effective methods to leverage the explanation for preventing the learning of spurious correlations.

TextGT (Yin and Zhong, 2024) design a novel double-view graph Transformer on text and a new algorithm to implement edge features in graphs.

4.4 Main Results

The experiment results of different methods on five benchmark datasets are presented in Table 2. Our DMAN consistently outperforms all compared baselines on the Lap14, Rest14, Rest15, and MAMs datasets, and achieves overall better results than the baselines on the Rest16 dataset, demonstrating the effectiveness of our method. Compared to methods utilizing attention scores and dependency graphs (e.g., RGAT, DualGCN, SSEGCN), our attribution-based DMAN effectively reduces noise interference from irrelevant opinion words that could be introduced through attention scores. Compared to more methods that leverage syntactic information in different ways (e.g. T-GCN, MGFN), our DMAN still achieves better performance, validating that integrating attribution scores to dynamically capture keywords facilitates a more effective use of syntactic information. Furthermore,

Models	Lap14		Rest14		Rest15		Rest16		MAMs	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Our DMAN	**82.29**	**78.91**	**87.59**	**82.47**	**86.30**	**72.97**	**92.85**	**77.37**	**85.55**	**85.01**
w/o multi-attribution	80.88	76.37	86.34	79.95	84.63	68.84	91.87	75.74	83.83	83.04
w/o token-level	81.66	77.83	87.05	80.35	85.37	71.00	92.04	75.90	84.73	84.08
w/o span-level	81.82	78.06	87.23	81.76	85.74	71.68	92.36	76.89	85.03	84.36
w/o syntax information	81.03	77.39	86.61	81.09	85.19	70.86	91.71	75.17	84.13	83.39

Table 3: Ablation study results (%) of our DMAN on five benchmark datasets.

As MAMs is a challenging dataset that is large-scale and has multi-aspect within sentences, our method still has significant improvements. This further demonstrates DMAN's capability to effectively focus on aspect-related opinion words and capture attribution knowledge towards sentiment.

4.5 Ablation Study

To further investigate the effectiveness of each component in our model, we conducted ablation studies on the five datasets. The results are shown in Table 3. In the model without multi-granularity, the performance of DMAN suffers from a sharp degradation, with accuracy decreases of 1.41%, 1.48% and 1.72% on Lap14, Rest15 and MAMs datasets, respectively. These results demonstrate the effectiveness of our proposed multi-step attribution framework, which can accurately identify the critical words for sentiment expression and dynamically leverage the effective syntactic structures. In the model w/o syntax information, we do not initial adjacent matrix based on dependency tree. The results show that syntactic information offers crucial clues for correlations between words, effectively mitigating potential attribution errors and significantly enhancing classification precision. Moreover, we conduct experiments only using single-granularity attribution. The performance decreases demonstrate that the integration of multi-granularity representations significantly enhances the precise comprehension of semantics.

4.6 Further Analysis

Effect of Top-K. To mitigate the interference of noisy dimensions, we have employed the Top-K strategy on the attribution scores to filter out dimensions with relatively low significance. In this section, we explore the impact of varying K values. Specifically, we conducted experiments on the Rest14 and MAMs datasets, testing a range of K values from 100 to 300. The results, illustrated

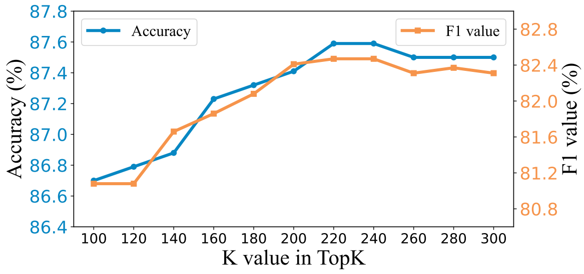


Figure 3: Accuracy (%) and macro-F1 value (%) on Rest14 dataset with different K values in Top-K strategy.

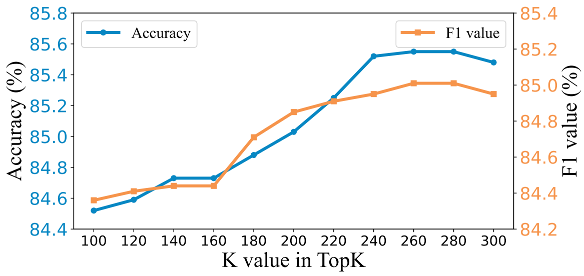


Figure 4: Accuracy (%) and macro-F1 value (%) on MAMs dataset with different K values in Top-K strategy.

in Figure 3 and Figure 4 show that accuracy and macro-F1 scores on both datasets initially improve as K increases, but then plateau or slightly decrease. We conjecture that low K values fail to adequately capture attribution knowledge, while high K values may introduce noise. Thus, selecting an appropriate K value is crucial for optimal performance.

Effect of Attribution Steps. To investigate how the number of attribution steps influences performance, we evaluated our DMAN with varying steps on the Rest14, Lap14, and MAMs datasets. Notably, to maintain compatibility with our framework, the number of GCN layers must increase correspondingly as the number of attribution steps increases. As depicted in Figure 5, our model achieves optimal performance with 2 steps, while performance significantly declines with further in-

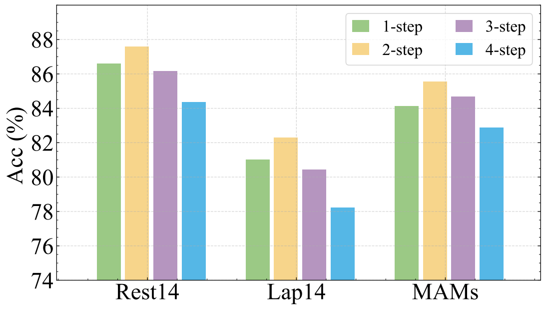
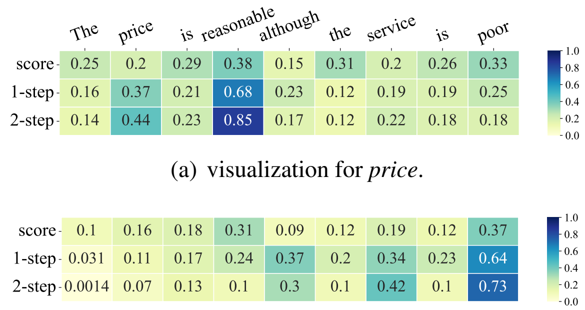


Figure 5: Accuracy (%) of DMAN on Rest14, Lap14 and MAMs datasets with different attribution steps.



(a) visualization for *price*.

(b) visualization for *service*.

Figure 6: Visualization of attention scores and multi-step attribution scores on two aspects, *price* and *service*. score denotes attention scores, 1-step and 2-step denote attribution scores of 1st and 2nd layers.

creases in the number of layers. We attribute this phenomenon to two primary factors. Firstly, when the number of GCN layers becomes excessive, node representations face the issue of over-smoothing, leading to vanishing gradients and information redundancy. Secondly, due to the relatively small size of ABSA datasets, the network is prone to overfitting as the model complexity increases, which results in a situation where gradients convey less effective attribution knowledge.

4.7 Visualization on Attribution

To demonstrate the effectiveness of attribution analysis in our approach, we selected samples with multiple aspects and visualized the attention scores and multi-step attribution scores in Figure 6 (a) and (b). Specifically, given the sentence "*The price is reasonable although the service is poor*" with two aspects, "*price*" and "*service*", attention scores are shown to be susceptible to noise within the sentence, often assigning relatively high scores to irrelevant words (e.g., "*is poor*" for "*price*"). In contrast, our proposed DMAN more accurately identifies aspect-related opinion words (e.g., "*reasonable*" for "*price*", "*poor*" for "*service*"). Furthermore, the progression of attribution scores from the first to the second step illustrates the process of semantic understanding, clearly indicating the effectiveness and interpretability of our model in dynamically capturing aspect-related contexts.

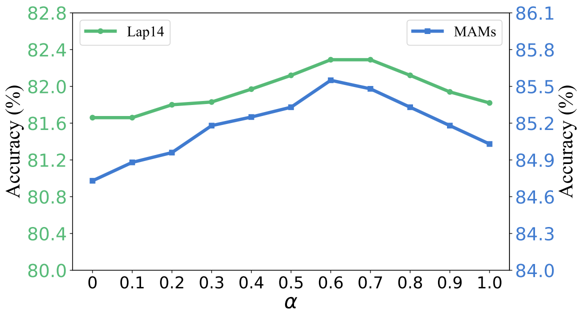


Figure 7: Accuracy (%) on Lap14 and MAMs datasets with different α values for granularity fusion.

4.8 Impact of α in Multi-granularity

In the Multi-granularity Attribution Module, we introduce α to balance token granularity and span granularity. To investigate their impact on model performance, we conducted experiments with different values of α on Lap14 and MAMs datasets. As illustrated in Figure 7, the performance improves with increasing α value and reaches a peak, and then declines. This suggests that effectively integrating multi-granularity representations can provide a more comprehensive understanding of sentence semantics. Specifically, considering that ABSA is a fine-grained classification task, we do not employ sentence-level representations.

5 Conclusion

In this paper, we propose a novel Dynamic Multi-granularity Attribution Network (DMAN) for the ABSA task, which is different from traditional models that rely on attention scores. Specifically, we first leverage Integrated Gradients to extract multi-step attribution during semantic comprehension, and Top-K strategy is adopted to filter out unimportant dimensions. We then consider multiple granularities of semantic concepts, fusing attribution representations from both token-level and span-level. Finally, we integrate these attribution values with dependency trees to dynamically capture relevant syntactic knowledge, thereby enhancing semantic understanding for sentiment classification. Extensive experiments on five public datasets demonstrate the effectiveness of our proposed DMAN.

Limitations

One of the primary limitations of our approach is that our method does not always provide accurate attributions when addressing sentences with overly complex content and structure. Actually, this is a common limitation among most ABSA methods. Additionally, Our framework comprises two components: attribution analysis and sentiment classification. The complexity of the model structure results in increased computational costs during training process.

Ethics Statement

Our work will not cause ethical issues, and the datasets we use are publicly available. Additionally, we do not involve the collection or use of any private information.

References

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:1803–1831.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland. Association for Computational Linguistics.

Harm Brouwer, Francesca Delogu, Noortje J. Venhuizen, and Matthew W. Crocker. 2021. Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, 12.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *International Conference on Learning Representations*.

Chenhua Chen, Zhiyang Teng, and Yue Zhang. 2020. Inducing target-specific latent structures for aspect sentiment classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5596–5607, Online. Association for Computational Linguistics.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3255, Online. Association for Computational Linguistics.

Shuoyang Ding, Hainan Xu, and Philipp Koehn. 2019. Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 1–12, Florence, Italy. Association for Computational Linguistics.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442, Brussels, Belgium. Association for Computational Linguistics.

Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, and Xing Xie. 2019. Towards a deep and unified understanding of deep neural models in NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2454–2463. PMLR.

Binxuan Huang and Kathleen Carley. 2019. Syntax-aware aspect level sentiment classification with graph attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5469–5477, Hong Kong, China. Association for Computational Linguistics.

Maksims Ivanovs, Roberts Kadiķis, and Kaspars Ozols. 2021. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 151–160, Portland, Oregon, USA. Association for Computational Linguistics.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6280–6285, Hong Kong, China. Association for Computational Linguistics.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6319–6329, Online. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Jinchao Zhang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2019. A novel aspect-guided deep transition model for aspect based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5569–5580, Hong Kong, China. Association for Computational Linguistics.

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, page 4068–4074. AAAI Press.

Fukun Ma, Xuming Hu, Aiwei Liu, Yawen Yang, Shuang Li, Philip S. Yu, and Lijie Wen. 2023. AMR-based network for aspect-based sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 322–337, Toronto, Canada. Association for Computational Linguistics.

Shiguan Pang, Yun Xue, Zehao Yan, Weihao Huang, and Jinhui Feng. 2021. Dynamic and multi-channel graph convolutional networks for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2627–2636, Online. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

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.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. In *International Conference on Artificial Neural Networks*.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinjie Sun, Kai Zhang, Qi Liu, Meikai Bao, and Yanjiang Chen. 2024. Harnessing domain insights: A prompt knowledge tuning method for aspect-based sentiment analysis. *Knowledge-Based Systems*, page 111975.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Xingwei Tan, Yi Cai, and Changxi Zhu. 2019. Recognizing conflict opinions in aspect-level sentiment classification with dual attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3426–3431, Hong Kong, China. Association for Computational Linguistics.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224, Austin, Texas. Association for Computational Linguistics.

Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou. 2020. Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6578–6588, Online. Association for Computational Linguistics.

Siyu Tang, Heyan Chai, Ziyi Yao, Ye Ding, Cuiyun Gao, Binxing Fang, and Qing Liao. 2022. Affective knowledge enhanced multiple-graph fusion networks for aspect-based sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5352–5362, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuanhe Tian, Guimin Chen, and Yan Song. 2021. Aspect-based sentiment analysis with type-aware graph convolutional networks and layer ensemble. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2910–2922, Online. Association for Computational Linguistics.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238, Online. Association for Computational Linguistics.

Qianlong Wang, Keyang Ding, Bin Liang, Min Yang, and Ruifeng Xu. 2023. Reducing spurious correlations in aspect-based sentiment analysis with explanation from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2930–2941, Singapore. Association for Computational Linguistics.

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.

Shuo Yin and Guoqiang Zhong. 2024. Textgt: A double-view graph transformer on text for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19404–19412.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham. Springer International Publishing.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4568–4578, Hong Kong, China. Association for Computational Linguistics.

Kai Zhang, Qi Liu, Hao Qian, Biao Xiang, Qing Cui, Jun Zhou, and Enhong Chen. 2021. Eatn: An efficient adaptive transfer network for aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35:377–389.

Kai Zhang, Kun Zhang, Mengdi Zhang, Hongke Zhao, Qi Liu, Wei Wu, and Enhong Chen. 2022a. Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3599–3610, Dublin, Ireland. Association for Computational Linguistics.

Mao Zhang, Yongxin Zhu, Zhen Liu, Zhimin Bao, Yunfei Wu, Xing Sun, and Linli Xu. 2023. Span-level aspect-based sentiment analysis via table filling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9273–9284, Toronto, Canada. Association for Computational Linguistics.

Zheng Zhang, Zili Zhou, and Yanna Wang. 2022b. SSEGCN: Syntactic and semantic enhanced graph convolutional network for aspect-based sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4916–4925, Seattle, United States. Association for Computational Linguistics.