Scientific Research Publishing

Journal of Computer and Communications, 2023, 11, 31-48
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Knowledge-Integrate Cross-Domain Data Generation Method for Aspect and Opinion Co-Extraction

Hao Zhang[1], Yegang Li[1], Jiachen Yang[1], Rujiang Bai[2]

[1]School of Computer Science and Technology, Shandong University of Technology, Zibo, China
[2]Institute of Information Management, Shandong University of Technology, Zibo, China
Email: ZH3025520@outlook.com

How to cite this paper: Zhang, H., Li, Y.G., Yang, J.C. and Bai, R.J. (2023) A Knowledge-Integrate Cross-Domain Data Generation Method for Aspect and Opinion Co-Extraction. *Journal of Computer and Communications*, **11**, 31-48.
https://doi.org/10.4236/jcc.2023.1112003

Received: November 28, 2023
Accepted: December 24, 2023
Published: December 27, 2023

Abstract

To address the difficulty of training high-quality models in some specific domains due to the lack of fine-grained annotation resources, we propose in this paper a knowledge-integrated cross-domain data generation method for unsupervised domain adaptation tasks. Specifically, we extract domain features, lexical and syntactic knowledge from source-domain and target-domain data, and use a masking model with an extended masking strategy and a re-masking strategy to obtain domain-specific data that remove domain-specific features. Finally, we improve the sequence generation model BART and use it to generate high-quality target domain data for the task of aspect and opinion co-extraction from the target domain. Experiments were performed on three conventional English datasets from different domains, and our method generates more accurate and diverse target domain data with the best results compared to previous methods.

Keywords

Knowledge-Integrate, Domain Adaptation, Text Generation, Aspect and Opinion Co-Extraction

1. Introduction

Aspect-level sentiment Classification of texts aimed at analyzing and understanding people's perspectives at the aspectual level has received increasing attention [1]. Aspect terms and opinion words extraction are two basic subtasks of aspect-based sentiment analysis. It aims to extract aspect terms and opinion words from reviews. For example, given the comment, "The pizza is delicious." The

aspect term is "pizza", and the opinion word is "delicious". Due to the gradual maturity of deep learning techniques, especially the great success of large-scale pre-trained models such as Bert, most supervised models have achieved excellent results in ABSA tasks. However, the high cost of annotating the data required in aspect-level text sentiment analysis tasks still exists in most domains where there exists a lack of richly annotated resources, which greatly restricts the performance of the models.

Therefore, the method of unsupervised domain adaptation that transfers knowledge from source domains with richly annotated data to target domains with no annotated data is very attractive [2]. The main challenge of unsupervised domain-adaptive tasks comes from the distributional differences between the data in the source domain and the target domains.

Traditional domain adaptation methods reduce the differences between domains through new feature representations [3] or redistribution of weights [4]. Because of the complexity of the fine-grained sentiment analysis task, they are mainly applied to coarse-grained cross-domain sentiment analysis. Only a few studies have attempted to address the fine-grained sentiment analysis task's domain self-adaptation problem.

A knowledge-integrate cross-domain data generation framework is proposed to address the issues using sufficient domain-invariant knowledge and differences between domains. It applies to the task of aspect and opinion co-extraction. Therefore, how to fully use domain-invariant knowledge and select domain-specific features are key factors in determining the generation of high-quality target domain comments and fine-grained annotations.

To address the above issues, a masking model that includes a re-masking strategy and an expand-masking strategy is proposed to generate accurate domain-independent comments which are comments with domain-specific features removed. In addition, the new cross-domain data generation model generates corresponding text snippets and fine-grained labels by integrating target domain-specific features into the context of domain-independent comments.

Our approach more adequately masks domain-specific features between domains than previous methods. It breaks the restriction on the number of reviews generated correspondingly and fully exploits domain-invariant knowledge, such as contextual knowledge, lexical knowledge, and syntactic knowledge between domains, to generate higher-quality comments of the target domain. The main contributions of this paper can be summarized as follows:

- The knowledge-integrate cross-domain data generation framework is proposed for unsupervised domain adaptation, which incorporates a masking model for domain-independent comment generation and a sequence-to-sequence data generation model for generation of reviews and annotations in the target domain.
- In the aspect and opinion co-extraction task, the framework's effectiveness is demonstrated in many experiments on three different domain datasets. The framework generates higher quality and more diverse comments on the target domains than previous methods and significantly improves the Micro-F1 values

achieving the best results compared to existing methods.

2. Related Works

2.1. Aspect and Opinion Co-Extraction

Most existing work treats the aspect term and opinion word extraction task as a sequence labeling task [5]. Early methods of extracting aspect opinion words relied heavily on feature engineering or direct extraction using opinion word dictionaries. For example, Jin *et al.* [6] proposed a vocabulary-based HMM model to extract aspect and opinion terms from comments. Liu *et al.* [7] processed the corpus through syntactic rules and then completed the extraction of aspect and opinion words through a bidirectional propagation approach. With the development of deep learning and pre-trained models, many supervised methods have achieved desirable results in most domains, Chen *et al.* [8] used CNN and Bi-GRU models to extract aspect terms and opinion words, Chen *et al.* [9] trained classifiers by introducing word interconnections into global knowledge. However, these methods rely on rich training data and thus have difficulty training robust models in certain domains with insufficient annotated data. Therefore, unsupervised domain-adaptive methods are introduced to solve the problem of insufficient data in certain domains.

2.2. Unsupervised Domain Adaptation

Several domain-adaptive methods have been used for coarse-grained text classification tasks. Ganin *et al.* [10] and Guo *et al.* [11], the basic idea of their approach is to align domain-specific features with domain-independent centric words and learn an autoencoder-based domain-invariant representation. Ganin *et al.* [12] and Li *et al.* [13] used a domain-adversarial approach for the cross-domain text classification task.

However, only a few domain-adaptive methods have been proposed for ABSA tasks. Xu *et al.* [14] post-trained Bert on a cross-domain corpus to enhance its domain adaptation. Li *et al.* [15] exploit manual syntactic rules an opinion seeds to extract aspects and opinions. Ding *et al.* [16] use artificial syntactic rules and public opinion seeds to extract aspect terms and opinion items. Wang *et al.* [17] predict the relation between any two adjacent words in the dependency tree by building structural correspondences and generate an auxiliary task. Pereg *et al.* [18] combine external syntactic information into Bert with an attentional mechanism that aids in the task, and Chen *et al.* [19], learning the domain-invariant features through bridging. Most of them rely too much on the quality of manual rule-making or fail to take full advantage of the important knowledge of the target domain.

2.3. Data Enhancement

Data augmentation is an essential solution to address the scarcity of domain datasets, especially in sentence-level sentiment analysis [20] and text categorization

[21]. For aspect/opinion extraction, Ding *et al.* [22] devised a data augmentation approach using a language model trained on linearized labeled sentences to generate large amounts of labeled data. Hsu *et al.* [23] used a masked language model, Bert, to replace unimportant words in sentences to enhance the diversity of the data. However, these studies only focused on tasks within the domain and did not address transfer to other domains.

In a recent study, Yu *et al.* [24] proposed a cross-domain data generation method based on the masking language model Bert, which replaces the source-specific aspects and comments in the labeled source domain comments with target-specific aspects and comments. Li *et al.* [25] generated feature words and corresponding labels simultaneously through the BART model to generate target domain comments with fine-grained annotations. Related studies by Yu and Li *et al.* have demonstrated the superiority of data-based augmentation adaptive methods. However, they do not consider domain attributes such as lexical knowledge syntactic knowledge, ignore some domain-invariant knowledge, and have limitations on the number of target domain comments to be generated, which limit the quality and quantity of the generated target domain data as well as the model's adaptability.

Therefore, the framework proposed in this paper can generate more flexible and accurate target domain comments through source domain data with fine-grained annotations better adapted to unsupervised domain adaptation tasks.

3. Methods

We view the aspect and opinion co-extraction task as a sequence annotation problem, where the input text with n words is denoted as a sequence of token $x=\left[x_1,x_2,\cdots,x_n\right]$, with the corresponding labels $y=\left[y_1,y_2,\cdots,y_n\right]$. The aspect and opinion co-extraction task is to predict the sequence of labels for a comment $y_i\in\left\{\text{B-ASP},\text{I-ASP},\text{B-OP},\text{I-OP},\text{O}\right\}$. For the unsupervised domain adaptation task, the labeled data can only be obtained from the source domain. Thus, the task relies on the labeled source domain comments $D_s=\left\{\left(x_i^s,y_i^s\right)\right\}_{i=1}^{N^s}$ and the unlabeled target-domain comment $D_u=\left\{x_i^u\right\}_{i=1}^{N^u}$ to predict the labelled sequences of the test data in the target domain $D_t=\left\{\left(x_i^t,y_i^t\right)\right\}_{i=1}^{N^t}$ for the labels y^t.

Our proposed framework contains four modules: knowledge extraction, feature masking, cross-domain data generation, and data processing, referred to as A, B, C, and D respectively. The framework is called CDDG-IK and its flowchart is shown in **Figure 1**.

3.1. Knowledge Extraction

In order to fully and accurately utilize the between domain-invariant features, the contextual information and the sequence labels are considered as domain-invariant features and more fully extract lexical information and syntactic distance information as domain-invariant features. An unsupervised approach is used to extract fragments of domain features as domain features in the reviews of both domains. These will be key knowledge for more fully masking the domain features and extending the generation of target domain reviews.

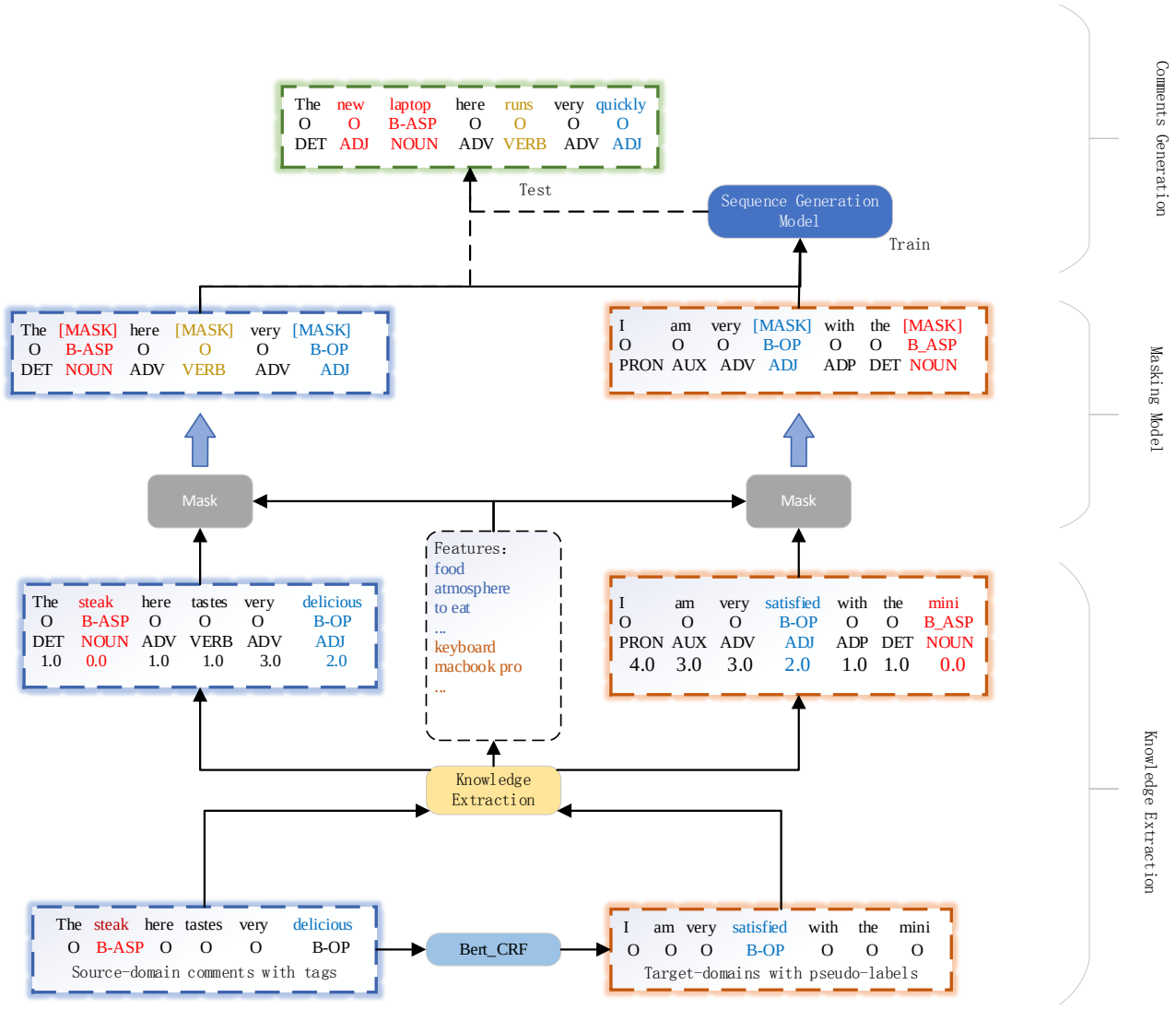

Figure 1. The overall architecture of CDDG-IK.

1) Domain feature extraction

In order to obtain features between different domains, a frequency ratio approach [26] is used to define text segments that occur more frequently in one of the domains as domain feature segments. All sentences are split in the two domains into word segments of different lengths, and then the relative frequency of the n-gram segments in the dataset is calculated with the following formula:

$$s\left(w, D_v\right) = \frac{count\left(w, D_v\right) + \lambda}{\sum_{v' \in V, v' \neq V} count\left(w, D_{v'}\right) + \lambda} \tag{1}$$

2) The Pos tagging and the syntactic relative distance

17 pos tags from the spacy library in Python to automatically annotate the pos tags of the comments in the source and target domains, recognizing each word in the comments as a noun, verb, adjective, etc., and it will become weakly supervised data in the text generation model.

In order to get better access to hidden domain features, we measure the syntactic relative distance of an aspect term from other words by the shortest distance of the corresponding node of the word in the syntactic parse tree, centered on the aspect term. If the aspect term consists of more than one word, the relative distance between the aspect term and the other words is the average distance between the constituent words and the other words. As shown in **Figure 2**, where the aspect term is sirloin steak.

SRD (sirloin, delicious) = 3.

SRD (steak, delicious) = 2.

SRD (sirloin steak, delicious) = 1.5.

3) The pseudo-label generation

First, a base classifier on the labeled data from the source domain DS is trained, which employs a pre-trained BERT model [27] to obtain the contextualized word representation and a Conditional Random Field (CRF) layer for sequence labeling. The trained classifier is used to perform fine-grained label prediction on the target domain comment D_u to obtain the pseudo-labeled target domain comment D_{tp}.

3.2. Domain-Specific Feature Mask

Generating high-quality target domain data depends heavily on the quality of the domain-independent comments, so it is crucial to mask domain-specific features as much as possible and avoid masking out domain-invariant features. Therefore, an expand-masking strategy and a re-masking strategy are proposed, where expansion masking is also a solution to address the annotated data shortage to improve the quantity and quality of domain-independent comment generation. Examples are shown in **Table 1**. The specific practices are as follows:

Table 1. The sample of masking strategy.

Source Domain Comments	The sirloin steak here tastes very delicious.
Domain-Specific Segment Mask	The [mask] [mask] here tastes very delicious.
Expand-Masking Strategy	The [mask] [mask] here tastes very delicious. The [mask] here tastes very delicious.
Re-Masking Strategy	The [mask] [mask] here [mask] very delicious. The [mask] here [mask] very delicious

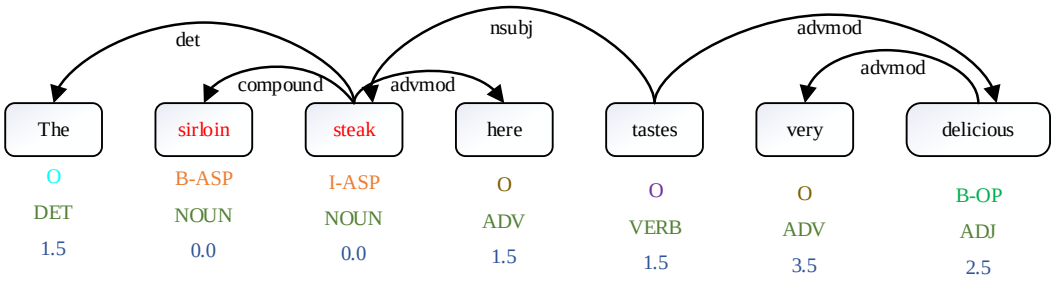

Figure 2. The sample of semantic-relative distance.

1) Feature masking

The set of domain features extracted in A is the more frequently occurring text fragments in a domain. Although these contain most domain-specific phrases, they also contain a lot of domain-invariant knowledge, such as [screen is, this mac, the place is], and words like [screen, mac, place] are domain-specific knowledge. In contrast, like [is, this, the] such deactivated words, if masked, will bring some noise to the target domain text generation process. Therefore, the forward maximum matching calculus is used to match the domain feature fragments in the set appearing in the reviews and replace the words that do not stop words in the matched fragments with special tokens [MASK]. It is worth noting that as long as one word of a domain-specific phrase is masked, the whole phrase will be masked.

2) Expand-Masking strategy

Domain-independent comments are sentences after replacing domain-specific features with special tokens [mask], and their number determines the diversity of generated target domain comments. In order to obtain diverse domain-independent comments, the expand-masking strategy is used to expand and remove [mask] tags according to a certain probability. Specifically, expanding mask segments with only one [mask] to two [mask] tokens with a 60% probability and selectively removing mask segments with multiple [mask] tokens with a 40% probability. Which follow the sequence tagging rules and lexical collocation laws to delete or expand the corresponding sequence tags and lexical tags when modifying the [mask] tokens. For example, suppose the sequence label corresponding to the mask position is an aspect term or opinion word. In that case, the expanded sequence label is I-ASP or I-OP. Regarding expanding and removing lexical tags, it is important to follow the laws of lexical collocation. For example, adjective tags or noun tags can be added between qualifiers and nouns. For removing multiple masking tags, it is still important to follow the above rule and keep the number of tags aligned with the number of words. This allows the generation of multiple domain-independent comments from a single source domain comment, eliminating the limitation of aligning the number of generated target and source domain comments.

3) Re-Masking strategy

Since feature extraction is computed from frequency ratios under different domains, it is difficult to determine low-frequency words, as well as higher-frequency words that occur in both domains but which often do not apply to the current context. This could significantly limit the quality of the target-domain generation, these words are defined as implicit domain-specific features. As in Table 1, the sentence after domain-specific feature masking: The [mask] [mask] here tastes very delicious. Where the verb "tastes" does not fit the context, the implicit domain-specific feature makes the generated target domain comments logically incorrect due to contextual inconsistency. Also, it creates a certain amount of noise in the text generation process. Therefore, the re-masking strategy

is adopt to process the text after ordinary masking, using the aspectual word syntactic distances computed above to filter out words with syntactic distances less than 4.0 and selecting implicit domain-specific features that need to be re-masked based on their lexical labels. If the corresponding lexemes are "VERB", "NOUN", "PROPN", "ADV" and "ADJ", it is masked.

3.3. Target Domain Comment Generation

The modified pre-trained sequence-to-sequence model BART [28], uses sequence labels and lexical labels as weakly supervised information to generate more accurate target domain comments and corresponding labels. The domain-independent comments of the two domains are used as the training data for the BART model, and it should be noted that the domain-independent comments used for training here do not include the expand-masking strategy, the model is shown in **Figure 3**.

1) Train the BART model

The set of domain features extracted in A is the more frequently occurring text fragments in a domain. Although these contain most domain-specific phrases, they also contain a lot of domain-invariant knowledge, such as [screen is, this mac, the place is], and words like [screen, mac, place] are domain-specific knowledge. In contrast, like [is, this, the] such deactivated words, if masked, will bring some noise to the target domain text generation process. Therefore, the forward maximum matching calculus is used to match the domain feature fragments in the set appearing in the reviews and replace the words that do not stop

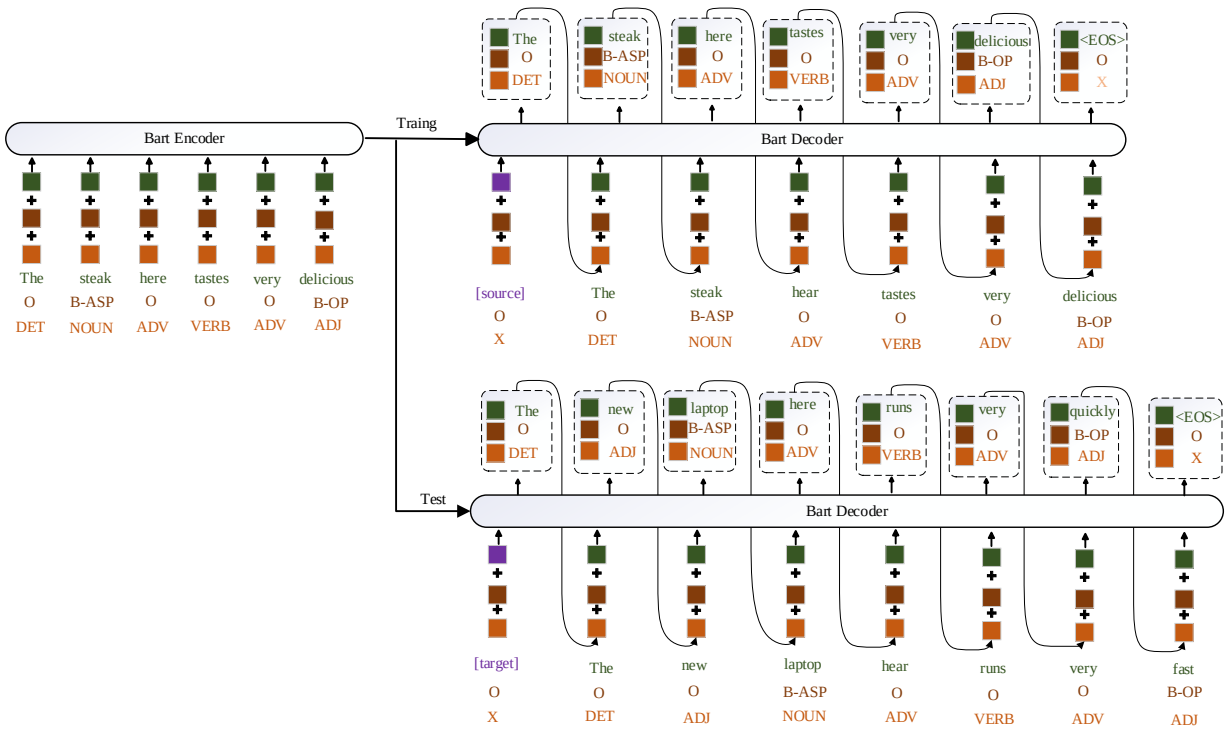

Figure 3. Sequence generation model.

words in the matched fragments with special tokens [MASK]. It is worth noting that as long as one word of a domain-specific phrase is masked, the whole phrase will be masked.

For each sample $(X, L, P) \in D_s \cup D_{TP}$, the corresponding masked domain-independent comments $(\tilde{X}, L, P)$ can be obtained as inputs to the model, where each masked sentence is $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n]$, the sequence label of each word is $L = [l_1, l_2, \cdots, l_n]$, and the lexical label is $P = [p_1, p_2, \cdots, p_n]$. In the encoder, in addition to the word embedding and positional embedding in Bart, the label embedding layer and the lexical embedding layer are added as weakly supervised data.

$$E_x = TokenEmb([\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n]) \tag{2}$$

$$E_l = TokenEmb([l_1, l_2, \cdots, l_n]) \tag{3}$$

$$E_p = TokenEmb([p_1, p_2, \cdots, p_n]) \tag{4}$$

where $E_x \in \boldsymbol{R}^{n \times d}$, $E_l \in \boldsymbol{R}^{n \times d}$, $E_p \in \boldsymbol{R}^{n \times d}$, and d is the dimension of the embedding. The output of the hidden state can be formulated as:

$$H = BartEncoder(E_x + E_l + E_p) \tag{5}$$

where $H \in \boldsymbol{R}^{n \times d'}$, d' denotes the hidden dimension. In the decoder, in order for the model to distinguish between different domain-specific features, a tuple of domain labels $([source], O, X)$ or $([target], O, X)$ is set up at the beginning of the decoder as a domain prompt. For each time step t, the decoder takes as input $(x_{<t}, l_{<t}, p_{<t})$ and the encoder output H to obtain the probability of the next word, token, and lexical with three independent linear layers:

$$P(x_t | x < t, l < t, p < t, H) = Softmax(W_x z_t + b_x) \tag{6}$$

$$P(l_t | x < t, l < t, p < t, H) = Softmax(W_l z_t + b_l) \tag{7}$$

$$P(p_t | x < t, l < t, p < t, H) = Softmax(W_p z_t + b_p) \tag{8}$$

where $W_x \in \boldsymbol{R}^{|v_x| \times d}$, $W_l \in \boldsymbol{R}^{|v_l| \times d}$, $W_p \in \boldsymbol{R}^{|v_p| \times d}$, and $|v_x|$, $|v_l|$ and $|v_p|$ refer to the dictionary size, the number of tag types5 and the number of lexical tags, respectively.17 The hidden layer vector $\boldsymbol{z}_t$ for time step t is as follows:

$$\boldsymbol{z}_t = BartDecoder(E_t) \tag{9}$$

$$E_t = TokenEmb(x_{t-1}) + TokenEmb(l_{t-1}) + TokenEmb(p_{t-1}) \tag{10}$$

For each sample, we calculate the negative log-likelihood loss for word vectors, sequence labels, and lexical labels separately:

$$Loss_x = -\sum_{t=1}^{n+1} \log\left(P(x_t | x < t, l < t, p < t, H)\right) \tag{11}$$

$$Loss_l = -\sum_{t=1}^{n+1} \log\left(P(x_t | x < t, l < t, p < t, H)\right) \tag{12}$$

$$Loss_p = -\sum_{t=1}^{n+1} \log\left(P(x_t | x < t, l < t, p < t, H)\right) \tag{13}$$

The final training loss consists of the addition of three parts:

$$Loss = Loss_x + Loss_l + Loss_p \tag{14}$$

2) Target domain comment generation

In the target domain comment generation phase, we use the masking tuple $\left(\tilde{X}_{expand+re}, L, P\right)$ obtained by the masking model to feed into the BART model encoder for each sample $\left(X, L, P\right) \in Ds$. It is worth noting that the domain-independent commenting here includes the expand-masking strategy, and $\left(\left[target\right], O, X\right)$ as the domain prompt is only provided to decode a target-domain sentence based on the auto-regressive manner and to jointly predict their sequence labels and lexical labels.

3.4. Data Processing and Task Training

As the generated text and labels will have irregularities, the generated text will be processed and filtered. First, sentences whose labels do not match the BIO pattern are deleted, and then the basic classifiers assign labels on the generated target domain data, and sentences whose assigned labels do not match the generated labels are deleted. Finally, these data are fed into the Bert-CRF model for training, and its performance is evaluated on the test set of the target domain.

4. Experimentation and Analysis

4.1. Datasets

Experiments on the publicly available datasets from three different domains are conducted, namely Restaurant(R), Laptop(L), and Device(D). R and L are two combination datasets from SemEval-2014 [29] and SemEval-2015 [30], D are collected by Hu and Liu [31] from digital devices, the statistics of which are shown in **Table 2**. We construct six cross-domain pairs (source domain → target domain) by combining datasets from different domains two by two, denoted as R → L, R → D, L → R, L → D, D → R, D → L.

4.2. Experimental Setting

- In the segmented masking method for domain feature masking, we set the length of the n-gram $w \in \left[1, 4\right]$ and set the relative frequency threshold δ to 10.0. In the expand-masking method, expand with a probability of 60% for only one [mask] position, delete with a probability of 100% for multiple [masks], and uniformly shorten it to less than three for more than four. The syntactic distance threshold for re-masking is defined as 4.0 for the re-masking method. In the sequence-to-sequence model, we set the training period to 5 and the batch size to 16, where Adam is used as the optimizer with a learning rate of 5e-5.
- A BERT-CRF classifier consisting of a Bert model and a CRF layer is used to assign the pseudo-labels task and target domains' final aspect and opinion co-extraction task. The Adam optimizer is used with different learning rates of 5e-5 and 0.01. Finally, the average Micro-F1 values of three random seeds for aspect and opinion co-extraction are used to evaluate the model.

Table 2. Statistics for the experimental dataset.

Datasets	Domains	Train	Test	Total
R	Restaurant	4381	1460	5841
L	Laptop	2884	961	3845
D	Device	2887	959	3836

4.3. Comparative Experiments

In order to demonstrate the effectiveness of our approach in cross-domain aspect and opinion co-extraction, as well as cross-domain sentiment analysis based on text generation, the comparison system is divided into two parts.

The first part is a domain adaptive based approach, models as follows:

• RNSCN: A recursive neural network for predicting syntactic structure by building structural correspondences.

• TRNN: Integration of the recursive neural network with a sequence labeling classifier by constructing dependency trees and integrating syntactic relations to model context.

• TIMN: Training a transferable interactive memory network to learn shared representations across domains by incorporating an auxiliary task and domain adversarial networks.

• SemBridge: A novel active domain adaptation method based on the CNN model, that builds semantic bridges to link source and target domains by retrieving transferable knowledge.

• SA-EXAL: A self-attention mechanism that bridges the gap across domains by coupling the Bert model and external linguistic information.

• The second part is the method based on the target domain text generation:

• CDRG: Generate target-domain reviews with fine-grained annotation by replacing specific attributes in the source domain comments with aspect and opinion words from the target domain.

• GCDDA: Co-extraction of aspects and opinions across domains is achieved by expanding the source domain data and using the Bart model to generate the target domain data and the corresponding labels.

4.4. Experimental Results and Analysis

The results of the comparisons for the two tasks, aspect extraction and opinion extraction are respectively reported in Table 3 and Table 4, and it can be observed that our model achieves optimal values on most of the cross-domain pairs. As far as the average Micro-F1 is concerned, our method achieves the best performance. Compared with the active domain-adaptive method SemBridge, the CDDG-IK model improves aspect extraction and opinion extraction performance by 7.67% and 2.67%, respectively, which proves the superiority of the cross-domain text generation method over the traditional domain-adaptive method. Compared to the latest domain-adaptive method GCDDA based on cross-domain text generation, the CDDG-IK model shows a significant improvement in all cross-domain

Table 3. Experimental results of aspect extraction.

Models	Aspect extraction-F1-scores/%						
	R → L	R → D	L → R	L → D	D → R	D → L	AVE
RNSCN	40.43	35.10	52.91	40.42	48.36	51.14	44.73
TRNN	40.15	37.33	53.78	41.19	51.17	51.66	45.99
TIMN	43.68	35.45	54.12	38.63	53.82	52.46	46.36
SemBridge	50.67	43.34	63.04	44.91	60.19	53.02	52.53
SA-EXAL	47.59	40.50	54.67	42.19	54.54	47.72	47.87
CDRG-Merge	58.23	37.96	72.88	40.62	66.79	54.26	55.12
GCDDA	66.56	44.80	62.22	45.11	68.23	57.44	57.39
CDDG-IK	**71.08**	46.26	**67.00**	**47.38**	**69.35**	**60.17**	**60.20**

Table 4. Experimental results of opinion extraction.

Models	Opinion extraction-F1-scores/%						
	R → L	R → D	L → R	L → D	D → R	D → L	AVE
RNSCN	65.85	60.17	72.51	61.51	73.75	71.18	67.44
TRNN	65.63	60.32	73.40	60.20	74.37	68.79	67.12
TIMN	68.44	59.05	73.69	62.22	76.52	69.32	67.12
SemBridge	71.51	63.46	80.48	64.15	80.21	72.63	72.08
SA-EXAL	75.79	63.33	80.05	60.19	71.57	63.98	69.15
CDRG-Merge	76.08	62.19	82.34	59.04	82.23	76.42	73.05
GCDDA	77.63	64.86	82.67	60.72	82.44	76.75	74.18
CDDG-IK	78.36	64.87	83.03	63.01	82.23	76.99	74.75

pairs, with an average F1 value improvement of 2.81% and 0.57% in aspect extraction and opinion extraction tasks, which proves that the masking method proposed in our model can more adequately mask domain-specific features and improve the diversity of the generated texts. With the incorporation of labels and lexical knowledge, target aspects or opinions can be generated more accurately and controllably at the masked locations.

4.5. Sample Analysis

In order to analyze the quality of the target domain comments generated by our model, several target domain comments generated are compared by the cross-domain pair R → L in the CDDG-IK model. The comparison examples are shown in Table 5 (where the red font corresponds to the aspect terms, the blue font represents the opinion term words, and the green font represents the implicit domain-specific features with no annotations). Through examples 1, 2, and 3 we can observe that the expand-masking strategy dramatically improves the diversity and flexibility of the generated target domain comments. Through examples 1 and 3 we can observe that due to the CDDG-IK model's re-masking strategy, some implicit features can be well masked, which makes the generated tar-

get domain comments more standardized. Comparing with example 4 we can see that adding lexical knowledge embedding in the BART model can make the generated target domain features more accurate.

4.6. Ablation Experiment

In order to validate the effectiveness of each strategy in the CDDG-IK model, we conducted ablation experiments on the re-masking approach, expand-masking approach, lexical knowledge embedding, and label embedding respectively, the results of the experiments are shown in Table 5:

- w/o-*re_mask*: Removing the re-masking strategy.
- w/o-*expand_mask*: Removing the expand-masking strategy.
- w/o-(*re_mask*+*expand_mask*): Removing the re-masking and expand-masking strategy.
- w/o-$Label_{embedding}$: Removing the label embedding.
- w/o-$POS_{embedding}$: Removing the lexical embedding.

By comparing the bar charts of the experimental results in Table 6 with those in Figure 4, we can see that the CDDG-IK model performs best in the aspect and viewpoint co-extraction task. In contrast, the model's experimental results decreased by 2.17% and 1.61% after removing the lexical knowledge embedding,

Table 5. Translated with www.DeepL.com/Translator (free version) Samples analysis.

Sequence	Models	Samples
1	Source	rao's has the best service and atmosphere in nyc.
	GCDDA	rao$^{\times}$'s has the best battery life and service in the market.
	CDDG-IK	apple$^{\surd}$'s has the best memory and operating system in mac. apple$^{\surd}$'s has the best battery and memory in mac.
2	Source	i fell in love with the egg noodles in the beef broth with shrimp dumpings and slices of bbq roast pork.
	GCDDA	i fell in love with the 13 "" macbook pro i5 .5 ghz 15 "" mac book pro and 15 "" of ram.
	CDDG-IK	i fell in love with the touchpad$^{\surd}$. i fell in love with the internet speed. i fell in love with the glass screen in the macbook.
3	Source	we all ate pasta entrees, which were great.
	GCDDA	we all ate$^{\times}$ our macbook pro, which were great.
	CDDG-IK	we all owned$^{\surd}$ mac, which were great. we all bought$^{\surd}$ mac software, which were great.
4	Source	Do not get the go go hamburgers, no matter what the reviews say.
	GCDDA	Do not get the go go hamburgers$^{\times}$, no matter what the reviews say.
	CDDG-IK	Do not get the bluetooth mouse$^{\surd}$, no matter what the reviews recommended.

Table 6. Experimental ressults of ablation study.

Sequence	Models	F1-scores, %	
		AS	OP
1	w/o-*re_mask*	59.61	74.40
2	w/o-*expand_mask*	59.37	74.22
3	w/o-(*re_mask* + *expand_mask*)	58.61	73.74
4	w/o-$Label_{embedding}$	58.35	73.27
5	w/o-$POS_{embedding}$	58.03	73.13
6	CDDG-IK	60.20	74.74

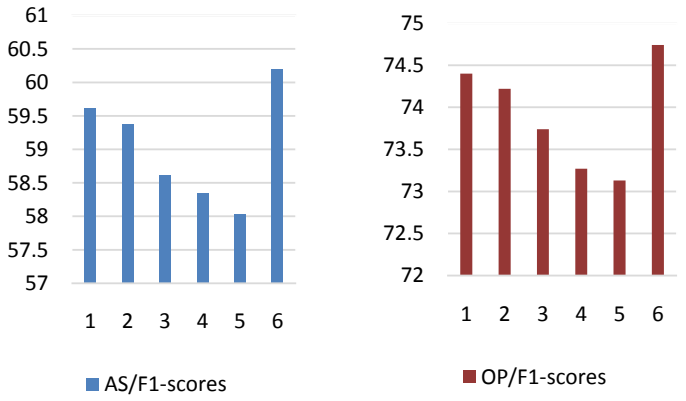

Figure 4. Bar chart comparing experimental results of ablation studies.

and 1.85% and 1.47% after removing the label embedding, respectively, which greatly affected the model's performance. This verifies that adding label embeddings and lexical knowledge embeddings as weakly supervised data to the encoder of the pre-trained Bart model enriches the linear knowledge of the model and improves the accuracy of text generation and the ability to handle more complex text generation tasks. The experimental results also show a certain degree of degradation after removing the inner and extended masking strategies, which demonstrates the effectiveness of the extended masking strategy in enhancing the data and re-masking methods for masking hidden domain-specific features and data enhancement.

5. Conclusion

In this paper, we investigate the cross-domain problem in the task of aspect and opinion co-extraction, and propose a framework for knowledge-integrated cross-domain data generation. Among them, the extended masking and re-masking strategy, a new masking strategy, can effectively augment the cross-domain generated data and greatly improve the quality of the generated text, while we improve the pre-training model, Bart, so that the target domain text and labels can be generated more accurately. Finally, the effectiveness of the CDDG-IK model is clearly verified by experiments on public datasets. Notably, our approach also

provides new methods and help for data enhancement and text generation in the domain.

Acknowledgements

This work was supported by the National Social Science Foundation Project "Research on Intelligent Intelligence Perception Driven by Multi-source Data Fusion", China (Item No. 21BTQ071).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Poria, S., Hazarika, D., Majumder, N. and Mihalcea, R. (2023) Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *IEEE Transactions on Affective Computing*, **14**, 108-132. https://ieeexplore.ieee.org/document/9260964 https://doi.org/10.1109/TAFFC.2020.3038167

[2] Li, J.J., Meng, L.C., Zhang, K., *et al.* (2021) Review of Studies on Domain Adaptation. *Computer Engineering*, **47**, 1-13. http://www.ecice06.com/CN/10.19678/j.issn.1000-3428.0060659

[3] Yu, J.F. and Jiang, J. (2017) Leveraging Auxiliary Tasks for Document-Level Cross-Domain Sentiment Classification. *Proceedings of the* 8*th International Joint Conference on Natural Language Processing*, Taipei, 1 December 2017, 654-663. https://aclanthology.org/I17-1066/

[4] Dredze, M., Kulesza, A. and Crammer, K. (2010) Multi-Domain Learning by Confidence-Weighted Parameter Combination. *Machine Learning*, **79**, 123-149. https://doi.org/10.1007/s10994-009-5148-0

[5] Jin, W., Ho, H.H. and Sriharir, K. (2009) A Novel Lexicalized HMM-Based Learning Framework for Web Opinion Mining. In: *Proceedings of the* 26*th Annual International Conference on Machine Learning*, ACM, New York, 465-472. https://doi.org/10.1145/1553374.1553435

[6] Liu, Q., Liu, B., Zhang, Y., *et al.* (2016) Improving Opinion Aspect Extraction Using Semantic Similarity and Aspect Associations. *Proceedings of the AAAI Conference on Artificial Intelligence*, **44**, 506-518. https://doi.org/10.1609/aaai.v30i1.10373

[7] Chen, X., Yang, X.-B. and Yao, Y.-H. (2021) Two-Channel Mixed Neural Network Sentiment Analysis Model Based on Character and Word Fusion. *Journal of Chinese Computer Systems*, **42**, 279-284. http://xwxt.sict.ac.cn/CN/Y2021/V42/I2/279

[8] Chen, Z. and Qian, T.Y. (2020) Enhancing Aspect Term Extraction with Soft Prototypes. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), November 2020, 2107-2117. https://aclanthology.org/2020.emnlp-main.164/ https://doi.org/10.18653/v1/2020.emnlp-main.164

[9] Ganin, Y. and Lempitsky, V.S. (2015) Unsupervised Domain Adaptation by Backpropagation. *Proceedings of the* 32*nd International Conference on Machine Learning*, *ICML* 2015, Lille, 6-11 July 2015, 1180-1189.

[10] Guo, H., Pasunuru, R. and Bansal, M. (2020) Multi-Source Domain Adaptation for

Text Classification via DistanceNet-Bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 7830-7838.
https://ojs.aaai.org/index.php/AAAI/article/view/6288
https://doi.org/10.1609/aaai.v34i05.6288

[11] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., *et al.* (2016) Domain-Adversarial Training of Neural Net-Works. *The Journal of Machine Learning Research*, **17**, 2096-2030.

[12] Li, Z., Wei, Y., Zhang, Y. and Yang, Q. (2018) Hierarchical Attention Transfer Network for Cross-Domain Sentiment Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 5852-5859.
https://doi.org/10.1609/aaai.v32i1.12055

[13] Xu, H., Liu, B., Shu, L. and Yu, P.S. (2019) BERT Post-Training for Review Reading Comprehension and Aspect-Based Sentiment Analysis. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Volume 1, 2324-2335.
https://aclanthology.org/N19-1242/

[14] Min, J., McCoy, R.T., Das, D., Pitler, E. and Linzen, T. (2020) Syntactic Data Augmentation Increases Robustness to Inference Heuristics. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, 5-10 July 2020, 2339-2352. https://doi.org/10.18653/v1/2020.acl-main.212

[15] Li, F.T., Pan, S.J., Jin, O., Yang, Q. and Zhu, X.Y. (2012) Cross-Domain Co-Extraction of Sentiment and Topic Lexicons. *Proceedings of the* 50*th Annual Meeting of the Association for Computational Linguistics*, Volume 1, 410-419.
https://aclanthology.org/P12-1043/

[16] Ding, Y., Yu, J.F. and Jiang, J. (2017) Recurrent Neural Networks with Auxiliary Labels for Cross Domain Opinion Target Extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, **31**, 3436-3442.
https://doi.org/10.1609/aaai.v31i1.11014

[17] Wang, W.Y. and Pan, S.J. (2018) Recursive Neural Structural Correspondence Network for Cross Domain Aspect and Opinion Co-Extraction. *Proceedings of the* 56*th Annual Meeting of the Association for Computational Linguistics*, Volume 1, 2171-2181. https://aclanthology.org/P18-1202/
https://doi.org/10.18653/v1/P18-1202

[18] Pereg, O., Korat, D. and Wasserblat, M. (2020) Syntactically Aware Cross-Domain Aspect and Opinion Terms Extraction. *Proceedings of the* 28*th International Conference on Computational Linguistics*, Barcelona, December 2020, 1772-1777.
https://aclanthology.org/2020.coling-main.158/
https://doi.org/10.18653/v1/2020.coling-main.158

[19] Chen, Z. and Qian, T.Y. (2021) Bridge-Based Active Domain Adaptation for Aspect Term Extraction. *Proceedings of the* 59*th Annual Meeting of the Association for Computational Linguistics and the* 11*th International Joint Conference on Natural Language Processing*, Volume 1, 317-327.
https://aclanthology.org/2021.acl-long.27/
https://doi.org/10.18653/v1/2021.acl-long.27

[20] Zhang, W., Li, X., Deng, Y., Bing, L. and Lam, W. (2023) A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 11019-11038.

[21] Chen, J.A., Wang, Z.H., Tian, R., Yang, Z.C. and Yang, D.Y. (2020) Local Additivity Based Data Augmentation for Semi-Supervised NER. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Novem-

ber 2020, 1241-1251. https://aclanthology.org/2020.emnlp-main.95/
https://doi.org/10.18653/v1/2020.emnlp-main.95

[22] Ding, B.S., Liu, L.L., Bing, L.D., Kruengkrai, C., Nguyen, T.H., Joty, S., Si, L. and Miao, C.Y. (2020) DAGA: Data Augmentation with a Generation Approach for Low-Resource Tagging Tasks. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), November 2020, 6045-6057.
https://aclanthology.org/2020.emnlp-main.488/
https://doi.org/10.18653/v1/2020.emnlp-main.488

[23] Hsu, T.-W., Chen, C.-C., Huang, H.-H. and Chen, H.-H. (2021) Semantics-Preserved Data Augmentation for Aspect-Based Sentiment Analysis. *Proceedings of the* 2021 *Conference on Empirical Methods in Natural Language Processing*, November 2020, 4417-4422. https://aclanthology.org/2021.emnlp-main.362/
https://doi.org/10.18653/v1/2021.emnlp-main.362

[24] Yu, J.F., Gong, C.G. and Xia, R. (2021) Cross Domain Review Generation for Aspect-Based Sentiment Analysis. *Findings of the Association for Computational Linguistics*: *ACL-IJCNLP* 2021, August 2021, 4767-4777.
https://aclanthology.org/2021.findings-acl.421/
https://doi.org/10.18653/v1/2021.findings-acl.421

[25] Li, J.J., Yu, J.F. and Xia, R. (2022) Generative Cross-Domain Data Augmentation for Aspect and Opinion Co-Extraction. *Proceedings of the* 2022 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Seattle, July 2022, 4219-4229.
https://aclanthology.org/2022.naacl-main.312/
https://doi.org/10.18653/v1/2022.naacl-main.312

[26] Li, J.C., Jia, R., He, H. and Liang, P. (2018) Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer. *Proceedings of the* 2018 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Volume 1, 1865-1874.
https://aclanthology.org/N18-1169/
https://doi.org/10.18653/v1/N18-1169

[27] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Volume 1, 4171-4186.
https://aclanthology.org/N19-1423/

[28] Lewis, M., Liu, Y.H., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L. (2020) BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, July 2020, 7871-7880. https://aclanthology.org/2020.acl-main.703/
https://doi.org/10.18653/v1/2020.acl-main.703

[29] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androut-sopoulos, I. and Manandhar, S. (2014) Semeval-2014 Task 4: Aspect Based Sentiment Analysis. *Proceedings of the* 8*th International Workshop on Semantic Evaluation* (*SemEval* 2014), Dublin, August 2014, 27-35. https://aclanthology.org/S14-2004/
https://doi.org/10.3115/v1/S14-2004

[30] Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S. and Androutsopoulosm, I. (2015) Semeval-2015 Task 12: Aspect Based Sentiment Analysis. *Proceedings of the* 9*th International Workshop on Semantic Evaluation* (*SemEval* 2015), Denver, June 2015, 486-495. https://aclanthology.org/S15-2082/

https://doi.org/10.18653/v1/S15-2082

[31] Hu, M.Q. and Liu, B. (2004) Mining and Summarizing Customer Reviews. *Proceedings of the* 10*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, 22-25 August 2004, 168-177. https://doi.org/10.1145/1014052.1014073