Supervised Gradual Machine Learning for Aspect-Term Sentiment Analysis

Yanyan Wang[†‡] **Qun Chen**[*†‡] **Murtadha H.M. Ahmed**[†‡] **Zhaoqiang Chen**[†‡]
Jing Su[†‡] **Wei Pan**[†‡] **Zhanhuai Li**[†‡]

[†]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[‡]Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Ministry of Industry and Information Technology, Xi'an, China
{wangyanyan@mail., chenbenben@, murtadha@mail., chenzhaoqiang@mail., sujing@mail., panwei1002@, lizhh@}nwpu.edu.cn

Abstract

Recent work has shown that Aspect-Term Sentiment Analysis (ATSA) can be effectively performed by Gradual Machine Learning (GML). However, the performance of the current unsupervised solution is limited by inaccurate and insufficient knowledge conveyance. In this paper, we propose a supervised GML approach for ATSA, which can effectively exploit labeled training data to improve knowledge conveyance. It leverages binary polarity relations between instances, which can be either *similar* or *opposite*, to enable supervised knowledge conveyance. Besides the explicit polarity relations indicated by discourse structures, it also separately supervises a polarity classification DNN and a binary Siamese network to extract implicit polarity relations. The proposed approach fulfills knowledge conveyance by modeling detected relations as binary features in a factor graph. Our extensive experiments on real benchmark data show that it achieves the state-of-the-art performance across all the test workloads. Our work demonstrates clearly that, in collaboration with DNN for feature extraction, GML outperforms pure DNN solutions.

1 Introduction

Aspect-Term Sentiment Analysis (ATSA) is a classical fine-grained sentiment classification task (Pontiki et al., 2015, 2016). Aiming to analyze detailed opinions towards certain aspects of an entity, it has attracted extensive research interests. In ATSA, an aspect-term, also called target, has to explicitly appear in a review. For instance, consider the running example shown in Table 1, in which r_i and s_{ij} denote the review and sentence identifiers, respectively. In r_1, ATSA needs to predict the expressed sentiment polarity, *positive* or *negative*, toward the explicit targets of *space* and *food*.

The state-of-the-art solutions of ATSA have been built upon pre-trained language models, such as LCF-BERT (Zeng et al., 2019), BAT (Karimi et al., 2020a), PH-SUM (Karimi et al., 2020b), and RoBERTa+MLP (Dai et al., 2021), to name a few. It is noteworthy that the efficacy of these deep solutions depends on the independent and identically distributed (i.i.d.) assumption. However, in real scenarios, there may not be sufficient labeled training data; even if provided with sufficient training data, the distributions of training data and target data are almost certainly different to some extent.

To alleviate the limitation of the i.i.d assumption, a solution based on the non-i.i.d paradigm of Gradual Machine Learning (GML) has recently been proposed for ATSA (Wang et al., 2021). GML begins with some easy instances, which can be automatically labeled by the machine with high accuracy, and then gradually labels more challenging instances by iterative knowledge conveyance in a factor graph. Without exploiting labeled training data, the current unsupervised solution relies on sentiment lexicons and explicit polarity relations indicated by discourse structures to enable knowledge conveyance. An improved GML solution leverages unsupervised DNN to extract sentiment features beyond lexicons (Ahmed et al., 2021). It has been empirically shown that even without leveraging any labeled training data, unsupervised GML can achieve competitive performance compared with many supervised deep models. However, unsupervised sentiment

*Corresponding author.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 723–739, 2023. https://doi.org/10.1162/tacl_a_00571
Action Editor: Minlie Huang. Submission batch: 11/2022; Revision batch: 1/2023; Published 6/2023.

r_i	s_{ij}	Text
r_1	s_{11}	**Space** was limited, but the **food** made up for it.
r_2	s_{21}	The **food** is sinful.
	s_{22}	The **staff** was really friendly.

Table 1: A running example in the domain of restaurant.

features are usually incomplete and noisy. Meanwhile, even though explicit polarity relations are accurate, they are usually very sparse in real natural language corpus. Therefore, the performance of gradual learning is still limited by inaccurate and insufficient knowledge conveyance.

Therefore, there is a need to investigate how to leverage labeled training data to improve gradual learning. In this paper, we propose a supervised solution based on GML for ATSA. As pointed out by Wang et al. (2021), linguistic hints can be very helpful for polarity reasoning. For instance, as shown in Table 1, the two aspect polarities of s_{11} can be reasoned to be opposite because their opinion clauses are connected by the shift word of ''but'', while the absence of any shift word between s_{21} and s_{22} indicates their polarity similarity. Representing the most direct way of knowledge conveyance, such binary polarity relations can effectively enable gradual learning. Unfortunately, the binary relations indicated by discourse structures are usually sparse in real natural language corpora. Therefore, besides explicit polarity relations, our proposed approach also separately supervises a DNN for polarity classification and a Siamese network to extract implicit polarity relations.

A supervised DNN can usually effectively separate the instances with different polarities. As a result, two instances appearing very close in its embedding space usually have the same polarity. Therefore, we leverage a polarity classifier for the detection of polarity similarity between close neighbors in an embedding space. It can also be observed that in natural languages, there are many different types of patterns to associate opinion words with polarities. However, a polarity classifier may put the instances with the same polarity but different association patterns in far-away places in its embedding space. In comparison, metric learning can cluster the instances with the same polarity together while separating those with different polarities as far as possible (Kaya and Bilge, 2019); it can thus align different association patterns with the same polarity. Therefore, we also employ a Siamese network for metric learning, which has been shown to perform well on semantic textual similarity tasks (Reimers and Gurevych, 2019), to detect complementary polarity relations. A Siamese network can detect both *similar* and *opposite* polarity relations between two arbitrary instances, which may be far away in an embedding space.

Finally, our proposed approach fulfills knowledge conveyance by modeling polarity relations as binary features in a factor graph. In our implementation, we use the state-of-the-art DNN model for ATSA, RoBERTa+MLP (Dai et al., 2021), to capture neighborhood-based polarity similarity while adapting the Siamese network (Chopra et al., 2005), the classical model of deep metric learning (Kaya and Bilge, 2019), to extract arbitrary polarity relations. It is worth pointing out that our work is orthogonal to the research on polarity classification DNNs and Siamese networks in that the proposed approach can easily accommodate new polarity classifiers and Siamese network models.

The main contributions of this paper can be summarized as follows:

- We propose a supervised GML approach for ATSA, which can effectively exploit labeled training data to improve gradual learning;
- We present the supervised techniques to extract implicit polarity relations for ATSA, which can be easily instilled into a GML factor graph to enable supervised knowledge conveyance;
- We empirically validate the efficacy of the proposed approach on real benchmark data. Our extensive experiments have shown that it consistently achieves the state-of-the-art performance across all the test datasets.

2 Related Work

Sentiment analysis at different granularity levels, including document, sentence, and aspect levels, has been extensively studied in the literature (Ravi and Ravi, 2015). At the document (resp., sentence) level, its goal is to detect the polarity of the entire document (resp., sentence) without

regard to the mentioned aspects (Zhang et al., 2015; Johnson and Zhang, 2017; Qian et al., 2017; Reimers and Gurevych, 2019). The state-of-the-art solutions for document-level and sentence-level sentiment analysis have been built upon various DNN models (Lei et al., 2018; Long et al., 2017; Letarte et al., 2018). However, they cannot be directly applied to the finer-grained aspect-level sentiment analysis because a document or sentence may express different polarities towards different aspects. The task of aspect-level sentiment analysis has been further classified into two finer subtasks, Aspect-Term Sentiment Analysis (ATSA) and Aspect-Category Sentiment Analysis (ACSA) (Xue and Li, 2018). ATSA aims to predict the sentiment polarity associated with an explicit aspect term appearing in the text while ACSA deals with both explicit and implicit aspects. In this paper, we focus on the far more popular subtask of ATSA. But, as shown in our experimental evaluation, our proposed approach is also applicable to ACSA.

Even though early work on deep learning for ATSA employed non-attention models (Dong et al., 2014; Tang et al., 2016), more recent proposals leveraged various attention mechanisms to output aspect-specific sentiment features, such as Interactive Attention Networks (Ma et al., 2017), Recurrent Attention Network (Chen et al., 2017), Content Attention Model (Liu et al., 2018), Multi-grained Attention Network (Fan et al., 2018), Segmentation Attention Network (Wang and Lu, 2018), Attention-over-Attention Neural Networks (Huang et al., 2018), and Effective Attention Modeling (He et al., 2018) to name a few. Most recently, the focus has experienced a considerable shift towards how to leverage pre-trained language models for ATSA, e.g., BERT-SPC (Song et al., 2019), AEN-BERT (Attentional Encoder Network) (Song et al., 2019), and LCF-BERT (Local Context Focus) (Zeng et al., 2019). Since BERT is trained on Wikipedia articles and has limited ability to understand review texts, Xu proposed to first post-train BERT on both domain knowledge and task knowledge, and then fine-tune the resulting model of BERT-PT on supervised domain data (Xu et al., 2019). Since then, many models built upon BERT-PT have been proposed (Karimi et al., 2020a,b). Other variants of BERT for ATSA include Adapted BERT (BERT-ADA) (Rietzler et al., 2020), Robustly Optimized BERT (RoBERTa) (Dai et al., 2021), and BERT with Disentangled Attention (DeBERTa) (Silva and Marcacini, 2021).

Since syntax structures are helpful for aspects to find their contextual words, many syntax-enhanced models have been recently proposed for ATSA, such as Proximity-Weighted Convolution Network (PWCN) (Zhang et al., 2019), Relational Graph Attention Network (RGAT) (Bai et al., 2021), Graph Convolutional Networks (GCN) (Zhao et al., 2020), Dependency Graph Enhanced Dual-Transformer network (DGEDT) (Tang et al., 2020), Type-aware Graph Convolutional Networks (T-GCN) (Tian et al., 2021), and Knowledge-aware Gated Recurrent Memory Network with Dual Syntax Graph (KaGRMN-DSG) (Xing and Tsang, 2022). They focused on how to exploit explicit syntactic information provided by dependency-based parse trees. Other proposals investigated how to induce implicit syntactic information from pre-trained models (Dai et al., 2021).

The GML paradigm was first proposed for the task of entity resolution (Hou et al., 2022). Since then, it has also been applied to the task of ATSA (Wang et al., 2021; Ahmed et al., 2021). Without exploiting labeled training data, the performance of unsupervised GML is usually limited by inaccurate and insufficient knowledge conveyance. In this paper, we focus on how to leverage labeled examples to improve gradual learning for ATSA.

3 The GML Framework

In this section, we illustrate the GML framework by the existing unsupervised GML solution for ATSA (Wang et al., 2021). Given a corpus of reviews, R, the goal of ATSA is to predict the sentiment polarity of each aspect unit in R, $t_i = (r_j, s_k, a_l)$, where r_j denotes a review, s_k denotes a sentence in the review r_j, and a_l denotes an explicit aspect appearing in the sentence s_k. In this paper, we suppose that an aspect polarity is either positive or negative.

As shown in Figure 1, the framework consists of the following three essential steps:

3.1 Easy Instance Labeling

Gradual machine learning begins with some easy instances. Therefore, high label accuracy of easy instances is critical for GML's ultimate performance. The existing unsupervised solution for ATSA employs simple user-specified rules to

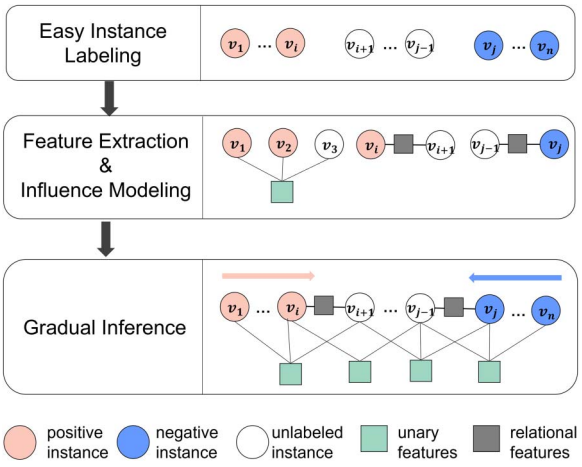

Figure 1: Unsupervised GML Solution for ATSA.

identify non-ambiguous instances as easy ones (Wang et al., 2021). Specifically, if a sentence contains some strong positive (resp., negative) sentiment words, but no negation, contrast, and hypothetical connectives, it can be reliably reasoned to be positive (resp., negative). *It is noteworthy that since this paper considers ATSA in the supervised setting, in which some labeled training data are supposed to be available, these training data with ground-truth labels can naturally serve as initial easy instances.*

3.2 Feature Extraction and Influence Modeling

Features serve as the medium to convey the knowledge obtained from labeled easy instances to unlabeled harder ones. This step extracts the common features shared by the labeled and unlabeled instances. To facilitate effective knowledge conveyance, it is desirable that a wide variety of features are extracted to capture diverse information. For each extracted feature, this step also needs to model its influence over the labels of relevant instances.

The existing unsupervised solution for ATSA presented in Wang et al. (2021) relies on sentiment lexicons and explicit polarity relations indicated by discourse structures to enable knowledge conveyance. Specifically, given a sentiment word, *positive* or *negative*, any sentence containing the word is supposed to have the same polarity as the word. Similarly, a similar (resp., opposite) polarity relation between two instances indicates that they are expected to have the same (resp., opposite) polarities. GML models word and relation features as unary and binary factors in a factor graph respectively.

3.3 Gradual Inference

This step gradually labels the instances with increasing hardness. Gradual learning is fulfilled by iterative inference on a factor graph, G, which consists of evidence variables representing labeled instances, inference variables representing unlabeled instances, and factors representing their features. The values of evidence variables once labeled remain unchanged while the values of inference variables need to be gradually inferred based on G.

Formally, suppose that a factor graph, G, consists of a set of evidence variables, $\mathbf{\Lambda}$, a set of inference variables, $\mathbf{V_I}$, and a group of factor functions of variables indicating their correlations, denoted by $\phi_{w_i}(V_i)$. In the case of ATSA, each variable in the factor graph is a boolean variable indicating the polarity of an aspect unit, the value of 1 for *positive* and 0 for *negative*. Then, the joint probability distribution over $V = \{\Lambda, V_I\}$ of G can be formulated as

$$P_{\mathbf{w}}(\Lambda, V_I) = \frac{1}{Z_{\mathbf{w}}} \prod_{i=1}^{m} \phi_{w_i}(V_i), \quad (1)$$

where V_i denotes a set of variables, w_i denotes a factor weight, m denotes the total number of factors, and $Z_{\mathbf{w}}$ denotes the normalization constant. Factor inference on G learns factor weights by minimizing the negative log marginal likelihood of evidence variables as follows:

$$\hat{\mathbf{w}} = arg\min_{\mathbf{w}} -log \sum_{V_I} P_{\mathbf{w}}(\Lambda, V_I). \quad (2)$$

In each iteration, GML generally chooses to label the inference variable with the highest degree of evidential certainty. Given an inference variable v, GML measures its evidential certainty by the inverse of entropy as follows:

$$E(v) = \frac{1}{H(v)} = \frac{1}{-\sum_{i=0,1} P_i(v) \cdot \log_2 P_i(v)}, \quad (3)$$

in which $E(v)$ and $H(v)$ denote the evidential certainty and entropy of v respectively, and $P_i(v)$ denotes the inferred probability of v having the label of 0 or 1. The iteration is repeatedly invoked until all the instances are labeled.

Algorithm 1: Scalable Gradual Inference

while *there exists any unlabeled variable in* G **do**
  $V' \leftarrow$ all the unlabeled variables in G;
  for $v \in V'$ **do**
    Measure the evidential support of v in G;
  Select top-m unlabeled variables with the most evidential support (denoted by V_m) ;
  for $v \in V_m$ **do**
    Approximately rank the entropy of v in V_m;
  Select top-k most promising variables in terms of entropy in V_m (denoted by V_k) ;
  for $v \in V_k$ **do**
    Compute the probability of v in G by factor graph inference over a subgraph of G;
  Label the variable with the minimal entropy in V_k;

To improve efficiency, GML usually implements gradual inference by a scalable approach, as sketched in Algorithm 1. It consists of three steps: measurement of evidential support, approximate ranking of entropy, and construction of inference subgraph. It first selects the top-m unlabeled variables with the most evidential support in G as the inference candidates. For each unlabeled instance, GML measures its evidential support from each feature by the degree of labeling confidence indicated by labeled observations, and then aggregates them based on the Dempster-Shafer theory.[1] It then approximates entropy estimation by an efficient algorithm on the m candidates and selects only the top-k most promising variables among them for factor graph inference. Finally, it estimates the probabilities of the finally chosen k variables by factor graph inference.

4 Supervised GML for ATSA

The overview of the proposed approach, denoted by S-GML, is shown in Figure 2. In this section, we first describe how to extract relational features, and then present their factor modeling.

[1] https://en.wikipedia.org/wiki/Dempster-Shafer_theory.

4.1 Polarity Relation Extraction

As mentioned in the Introduction, there exist some discourse relations between clauses or sentences that can provide helpful hints for polarity reasoning. Specifically, if two sentences are connected with a shift word (e.g., ‘‘but’’ and ‘‘however’’), they usually have opposite polarities. In contrast, two neighboring sentences without any shift word between them usually have similar polarities. S-GML uses the same rules as presented in Wang et al. (2021) to extract the explicit relations indicated by discourse structures. Therefore, we focus on how to extract implicit polarity relations in the rest of this subsection.

4.1.1 By Polarity Classification DNN

Since a supervised DNN can effectively separate the instances with different polarities, two instances appearing very close in its embedding space usually have the same polarity. Therefore, we supervise a DNN to automatically generate polarity-sensitive vector representations, and then exploit them for polarity similarity detection based on the nearest neighborhood.

As shown in Figure 2(b), we extract k_n-nearest neighbors of each unlabeled instance from both labeled training data and unlabeled test data, where vector distance is measured by cosine distance. To ensure that only very close instances in the embedding space are considered to be similar, we also set a high threshold (e.g., 0.05 in our implementation) to filter out unreliable pairs. Our experiments have demonstrated that the performance of supervised GML is robust w.r.t. the value of k_n provided that it is set within a reasonable range (between 5 and 9). In the implementation, we use RoBERTa+MLP (Dai et al., 2021), the-state-of-art deep model for ATSA, to learn polarity-sensitive vector representations. However, other deep models for ATSA can also be applied.

4.1.2 By Siamese Network

A polarity classifier may put the instances with the same polarity but different opinion association patterns in far-away places in its embedding space. To extract complementary polarity relations, we also employ metric learning, which can cluster the instances with the same polarity together while separating those with different polarities as far as possible, to align different association patterns with the same polarity. Metric learning can detect both similar and opposite polarity relations

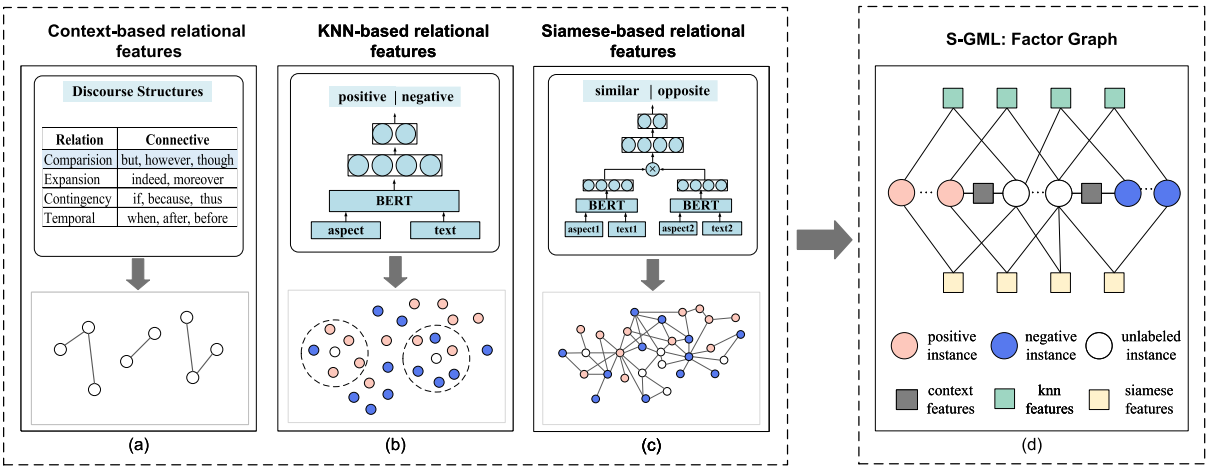

Figure 2: The overview of S-GML: 1) it extracts three types of polarity relations; 2) it models the extracted relations as binary factors to enable gradual learning.

between two arbitrary instances, which may be far away in an embedding space. In our implementation, we use the Siamese network, which has been shown to perform well on semantic textual similarity tasks (Reimers and Gurevych, 2019), to detect polarity relations between two arbitrary instances.

The structure of the Siamese network has been shown in Figure 2(c). Given two instances, $t_1 = (r_1, s_1, a_1)$ and $t_2 = (r_2, s_2, a_2)$, it first generates their vector representations by feeding the sequence of ''[CLS] + aspect + [SEP] + text + [SEP]'' into the BERT model, then computes their mutual information by multiplication, and finally uses a linear layer to predict their polarity relation, 0 for *opposite* and 1 for *similar*. The whole process can be represented by

$$v_1 = BERT(t_1), \tag{4}$$
$$v_2 = BERT(t_2), \tag{5}$$
$$p_r = softmax([v_1 \odot v_2] * W), \tag{6}$$

where d_m denotes the dimension of the BERT model, $v_1, v_2 \in \mathcal{R}^{d_m}$ denote the pooled vector representations, $W \in \mathcal{R}^{d_m \times 2}$ denotes the weights of the linear layer, $\odot$ denotes the element-wise multiplication, and $p_r \in \mathcal{R}^{1\times 2}$ denotes the output of the Siamese network, $p_r = [d, 1-d]$, where d denotes the predicted dissimilarity probability obtained from the softmax layer.

The training of a Siamese network aims to minimize the binary entropy loss defined as

$$L = -y \log \hat{y} - (1-y)\log(1-\hat{y}), \tag{7}$$

where $\hat{y}$ denotes the prediction output of two instances having the same polarity, and y denotes the ground-truth label indicating whether they are similar or opposite. Since the Siamese network is supposed to predict binary labels, 0 or 1, as usual, we set the threshold at 0.5. Certainly, its predictions are noisy, containing some false positives and false negatives. However, gradual learning does not require all the predicted relations to be correct; instead a set of noisy relations can correctly predict the label of a target instance provided that the majority of them are correct.

For the training of the Siamese network, S-GML randomly selects a fixed number of binary relations (e.g., 80 in our implementation), half of which are similar ones and the other half are opposite ones. In the prediction phase, for each unlabeled instance, S-GML randomly selects k_s from both labeled and unlabeled instances to extract its binary relations. Since polarity relation detection between two arbitrary instances is generally more challenging than polarity similarity detection between close neighbors in an embedding space, the number of relations constructed based on Siamese network per instance, denoted by k_s, is suggested to be set to be not greater than the number of its extracted nearest neighbors, namely $k_s <= k_n$. Our experiments have demonstrated that the performance of supervised GML is robust w.r.t. the values of k_n and k_s provided that they are set to be within a reasonable range (between 3 and 9). It is noteworthy that the total number of relations extracted by the Siamese network can be represented by $O(m \times k_s)$, in which m denotes the number of

unlabeled instances in a target workload. Due to the limited value of k_s, relation extraction by the Siamese network can be executed very efficiently.

4.2 Factor Modeling of Polarity Relations

An example of a GML factor graph for ATSA is shown in Figure 2(d). S-GML models polarity relations as binary factors to enable gradual knowledge conveyance from labeled instances to unlabeled ones.

Formally, the constructed factor graph G defines a joint probability distribution over its variables V by

$$P_w(V = v) = \frac{1}{Z_w} \prod_{f \in F} \phi_f(v_i, v_j), \qquad (8)$$

where v_i denotes a Boolean variable indicating the polarity of an aspect unit, $F = F_c \cup F_k \cup F_s$ denotes the set of all binary factors corresponding to context-based, KNN-based and Siamese-based relational features respectively, and the binary factor $\phi_f(v_i, v_j)$ is formulated as

$$\phi_f(v_i, v_j) = \begin{cases} e^{w_f} & if\ v_i = v_j; \\ 1 & otherwise; \end{cases} \qquad (9)$$

where v_i and v_j denote the two variables sharing the binary feature f, and w_f denotes the weight of f. It is noteworthy that a factor function, which aims to measure the correlation between variables, is usually defined as an exponential function (Kschischang et al., 2001). It should take non-negative values, and have larger values if its correlated variables take desired values. Therefore, in Eq. 9, the weight of a *similar* factor is positive, or $w_f > 0$, while the weight of an *opposite* factor is negative, or $w_f < 0$. It can be observed that such way of encoding would force two variables sharing a *similar* factor to hold the same polarity, while forcing two variables sharing an *opposite* factor to hold the opposite polarities.

In S-GML, we have five types of relational factors, two modeling explicit relations (*similar* and *opposite*), one modeling implicit relations detected by polarity classifier (only *similar*), and the remaining two modeling implicit relations detected by Siamese network (*similar* and *opposite*). The factors of the same type are supposed to have the same weight. In our implementation, the weights of *similar* factors are initially set to 2 while the weights of *opposite* factors are set to -2. However, all the five factor weights have to be continuously learned in the process of gradual inference.

5 Empirical Evaluation

In this section, we empirically evaluate the performance of the proposed approach, denoted by S-GML, on real benchmark data. We compare S-GML with the existing GML solution as well as the state-of-the-art DNN models. Even though the focus of this paper is on ATSA, the proposed approach can also be applied to the task of ACSA. Therefore, we also compare S-GML with its alternatives on ACSA.

The rest of this section is organized as follows: Section 5.1 describes the experimental setup. Section 5.2 presents the evaluation results on ATSA. Section 5.3 presents the evaluation results of parameter sensitivity. Section 5.4 presents the evaluation results on ACSA.

5.1 Experimental Setup

We have used benchmark datasets in three domains (restaurant, laptop, and neighborhoods) from the SemEval-2014 Task 4,[2] SemEval-2015 Task 12,[3] SemEval-2016 Task 5,[4] and SentiHood.[5] This paper considers both ATSA and ACSA as binary classification tasks. Note that we use the annotated labels provided by Wang et al. (2021) when aspect terms are not specified in ATSA. In all the datasets, we ignore neutral instances and label aspect polarities as *positive* or *negative*.

For performance evaluation, as usual, we randomly split the default training data of each benchmark dataset into two parts by the ratio of $8:2$, which specifies the proportions of training and validation data, respectively. Since we run each approach multiple times, we leverage validation data to pick the best model in each run. On Sentihood, we use the default partition of training and validation data. We use the classical metrics of *Accuracy* and *Macro-F1* to measure performance, and conduct *pairwise t-test* on both metrics to verify whether the achieved improvements are statistically significant.

[2] https://alt.qcri.org/semeval2014/task4.
[3] https://alt.qcri.org/semeval2015/task12.
[4] https://alt.qcri.org/semeval2016/task5/.
[5] https://github.com/HSLCY/ABSA-BERT-pair/tree/master/data/sentihood.

Compared Approaches. For ATSA, the compared GML solutions include:

- **Unsupervised Lexicon-based GML (Wang et al., 2021).** The first unsupervised solution relies on sentiment lexicons and explicit polarity relations indicated by discourse structures for knowledge conveyance.
- **Unsupervised DNN-based GML (Ahmed et al., 2021).** As an improvement of lexicon-based GML, it leverages an unsupervised attention-based neural network to automatically extract sentimental features for knowledge conveyance.
- **Hybrid GML (Ahmed et al., 2021).** Built upon the unsupervised DNN-based GML, it leverages labeled training data in a naive way by simply integrating the outputs of supervised DNN as unary factors into a factor graph to give a hybrid prediction. It is noteworthy that for fair comparison, the hybrid approach uses the same labeled data as S-GML to train DNN models. The original solution used the DNN model of PH-SUM. Since RoBERTa+MLP has been empirically shown to outperform PH-SUM, we implement the hybrid solution with RoBERTa+MLP as its DNN model in this paper.

Since the deep models for ATSA based on the pre-trained language models have been empirically shown to outperform their earlier alternatives, we compare S-GML with these state-of-the-art models, which include:

- **BERT-SPC (Song et al., 2019).** It feeds the sequences of ''[CLS] + context + [SEP] + target + [SEP]'' into the basic BERT model for sentence pair classification.
- **AEN-BERT (Song et al., 2019).** It uses an Attentional Encoder Network (AEN) to model the correlation between context and target.
- **LCF-BERT (Zeng et al., 2019).** It uses a Local Context Focus (LCF) mechanism based on Multi-head Self-Attention (MHSA) to pay more attention to local context words.
- **BERT-PT (Xu et al., 2019).** It uses post-trained BERT on task-aware knowledge to enhance BERT fine-tuning.
- **BAT (Karimi et al., 2020a).** It uses adversarial training to fine-tune BERT for ATSA.
- **PH-SUM (Karimi et al., 2020b).** It uses two simple modules named Parallel Aggregation and Hierarchical Aggregation on the top of BERT for ATSA.
- **RGAT (Bai et al., 2021).** It uses a novel relational graph attention network to integrate typed syntactic dependency information for ATSA.
- **RoBERTa+MLP (Dai et al., 2021).** It uses RoBERTa to generate context-based word embeddings of explicit aspect terms, and then leverages an MLP layer for polarity output.

Implementation Details. We have implemented S-GML based on the open-sourced GML solution for ATSA (Wang et al., 2021). To extract neighborhood-based polarity similarity relations, we use the model of RoBERTa+MLP, whose performance has been empirically shown to be state of the art. In the implementation of RoBERTa+MLP, we use the split set of default training data and the default parameter settings as presented in Dai et al. (2021). Specifically, the size of hidden layer is set at 768, batch size at 32, learning rate at $2e-5$, dropout at 0.5, and the number of epoches at 40. In the implementation of Siamese network, we use the post-trained BERT (Xu et al., 2019), which was trained using an uncased version of BERT-base on the domains of restaurant and laptop. To generate training data for the Siamese network, for each labeled instance in the training set, we randomly select totally 80 polarity relations, 40 of which are *similar* while the remaining 40 are *opposite*. With regard to Siamese network, we set the size of hidden layer at 768, the maximum length of inputs at 80, learning rate at $3e-5$ and batch size at 32.

In the default setting of S-GML, we select top-5 nearest neighbors from labeled training data and unlabeled test data for each unlabeled instance based on the learned embedding of RoBERTa+MLP (or $k_n = 5$ in Subsection 4.1.1), and randomly select 3 instances from both labeled training data and unlabeled test data to extract polarity relations based on the Siamese network (or $k_s = 3$ in Subsection 4.1.2). Our sensitivity evaluation results presented in Subsection 5.3

Model	RES14		RES15		RES16	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
BERT-SPC	93.61%	90.47%	85.80%	83.75%	92.06%	87.60%
AEN-BERT	91.77%	88.17%	87.52%	85.88%	91.22%	87.11%
LCF-BERT	93.94%	90.87%	85.99%	84.23%	91.89%	87.05%
BERT-PT	95.50%	93.24%	88.52%	87.37%	93.62%	90.27%
BAT	95.45%	93.12%	89.17%	87.86%	94.76%	91.67%
PH-SUM	95.87%	93.69%	89.44%	88.21%	94.56%	91.49%
RGAT	95.45%	92.89%	85.70%	83.60%	92.53%	88.90%
RoBERTa+MLP	95.74%	93.53%	89.51%	88.04%	94.37%	91.05%
Unsupervised Lexicon-based GML	83.83%	79.34%	80.22%	78.94%	85.64%	80.33%
Unsupervised DNN-based GML	87.05%	82.93%	81.19%	79.92%	86.31%	81.15%
Hybrid GML(RoBERTa+MLP)	95.92%	93.86%	89.70%	88.29%	94.88%	91.64%
S-GML	**96.90%**	**95.33%**	**90.83%**	**89.70%**	**96.00%**	**93.65%**
S-GML vs RoBERTa+MLP (p-value)	$7.98e-8$ †	$5.56e-8$ †	0.0183 †	0.0192 †	$1.52e-5$ †	$1.86e-5$ †
S-GML vs Hybrid GML (p-value)	$1.28e-6$ †	$6.65e-7$ †	0.0247 †	0.0252 †	0.0007 †	0.0008 †
Model	**LAP14**		**LAP15**		**LAP16**	
	Acc	**Macro-F1**	**Acc**	**Macro-F1**	**Acc**	**Macro-F1**
BERT-SPC	91.68%	89.84%	89.45%	88.53%	86.64%	85.37%
AEN-BERT	93.39%	91.65%	90.99%	90.32%	86.64%	84.74%
LCF-BERT	91.90%	90.20%	88.93%	88.30%	86.85%	85.29%
BERT-PT	93.22%	91.78%	93.10%	92.72%	87.72%	86.38%
BAT	93.13%	91.55%	93.51%	93.07%	87.93%	86.20%
PH-SUM	92.84%	91.30%	92.12%	91.67%	87.61%	86.15%
RGAT	93.12%	90.76%	90.75%	90.51%	87.55%	86.09%
RoBERTa+MLP	94.24%	92.83%	93.06%	92.62%	88.73%	87.44%
Unsupervised Lexicon-based GML	82.25%	79.41%	82.42%	81.52%	80.31%	78.74%
Unsupervised DNN-based GML	85.84%	83.33%	84.05%	83.26%	81.62%	80.07%
Hybrid GML(RoBERTa+MLP)	94.46%	93.12%	93.39%	92.93%	88.94%	87.68%
S-GML	**95.10%**	**93.95%**	**93.70%**	**93.26%**	**89.77%**	**88.49%**
S-GML vs RoBERTa+MLP (p-value)	0.0016 †	0.0006 †	0.0002 †	0.0006 †	0.0053 †	0.0016 †
S-GML vs Hybrid GML (p-value)	0.0261 †	0.0098 †	0.0111 †	0.0227 †	0.0236 †	0.0106 †

Table 2: Comparative Evaluation Results on ATSA: 1) RES and LAP stand for Restaurant and Laptop domains respectively; 2) the best accuracies are highlighted in **bold**; 3) the marker † indicates p-value < 0.05.

demonstrate that the performance of S-GML is very robust w.r.t. the parameters of k_n and k_s provided that their values are set to be within a reasonable range (between 3 and 9). Our implementations of S-GML have been available at our website.[6]

5.2 Comparative Evaluation on ATSA

The detailed comparative results on ATSA are presented in Table 2, in which *Hybrid GML(RoBERTa+MLP)* denotes the Hybrid GML solution with RoBERTa+MLP as its DNN model. The reported results are the averages over 25 runs. To verify statistical significance of S-GML's performance advantage, we have conducted pairwise t-test between S-GML and its best alternatives, RoBERTa+MLP and Hybrid GML.

It can be observed that S-GML consistently achieves the state-of-the-art performance on all the datasets. It outperforms the best DNN model by the margins between 1% and 2% on most datasets. For instance, on RES14, RES15, and RES16, the improvements are close to 2.0% in terms of Macro-F1. On LAP14 and LAP16,

[6] https://chenbenben.org/sgml.html.

Model	RES14		RES15		RES16	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
S-GML(w/o knn)	96.44%	94.68%	89.02%	87.58%	95.66%	93.19%
S-GML(w/o Siamese)	95.14%	92.55%	88.25%	86.77%	93.00%	88.85%
S-GML(w/o context)	96.68%	95.00%	90.55%	89.38%	95.80%	93.35%
S-GML	**96.90%**	**95.33%**	**90.83%**	**89.70%**	**96.00%**	**93.65%**
Model	**LAP14**		**LAP15**		**LAP16**	
	Acc	**Macro-F1**	**Acc**	**Macro-F1**	**Acc**	**Macro-F1**
S-GML(w/o knn)	93.60%	92.16%	93.06%	92.55%	88.10%	86.69%
S-GML(w/o Siamese)	91.90%	89.69%	90.51%	89.95%	89.35%	88.05%
S-GML(w/o context)	94.67%	93.48%	93.66%	93.23%	88.10%	86.73%
S-GML	**95.10%**	**93.95%**	**93.70%**	**93.26%**	**89.77%**	**88.49%**

Table 3: The evaluation results of ablation study on ATSA.

the improvements are more than 1% in terms of Macro-F1. S-GML also beats previous unsupervised GML solutions by large margins; in terms of accuracy, S-GML outperforms Unsupervised DNN-based GML by the margins between 8% and 10% across all the test workloads. It is noteworthy that S-GML consistently beats the Hybrid GML, which achieves overall better performance than the state-of-the-art deep model (RoBERTa+MLP). For instance, in terms of Macro-F1, the improvement margins are around 1.5%, 1.5%, and 2% on RES14, RES15, and RES16, respectively. Due to the widely recognized challenge of ATSA, the achieved improvements can be considerable.

It can also be observed that with regard to pairwise t-test, the p-values of S-GML against RoBERTa+MLP and Hybrid GML are all below 0.05, which means the performance improvements are statistically significant. These experimental results clearly demonstrate the efficacy of S-GML.

Ablation Study. The evaluation results are presented in Table 3, where *S-GML(w/o knn)*, *S-GML(w/o Siamese)*, and *S-GML(w/o context)* denote the ablated models with the components of knn-based, Siamese-based and context-based relational features removed, respectively. It can be observed that without either KNN relations or Siamese relations, the performance of S-GML drops on all the test workloads. This observation clearly indicates that KNN and Siamese relations are complementary to each other and their combined modeling in GML achieves better performance than either of them. However, it can also be observed that compared with knn relations, the performance of GML drops more considerably without Siamese relations. The KNN relations capture only similarity features, while the Siamese relations can capture both similarity and opposite, or more diverse, relations. It is noteworthy that these experimental results are consistent with the expected characteristic of GML that more diverse features can usually facilitate knowledge conveyance more effectively.

An Illustrative Example. We illustrate the efficacy of S-GML by the examples extracted from RES14, which are shown in Figure 3. Based on GML, the instance t_1 has the most evidential support, followed by t_2, t_3, and finally t_4. Meanwhile, the instances t_1 and t_2 have less evidential conflict than t_3 and t_4. Therefore, S-GML labels them in the order of t_1, t_2, t_3, and t_4. In spite of the noisy relations of t_4, S-GML can correctly label t_4 because after t_1, t_2, and t_3 are labeled, the majority of evidence neighbors provide correct polarity hints.

5.3 Sensitivity Evaluation

To evaluate sensitivity, we vary the values of the parameters k_n and k_s, which denote the number of nearest neighbors selected by polarity classifier and the number of relations randomly selected based Siamese network, respectively, within the range between 3 and 9. Since polarity relation detection between two arbitrary instances is generally more challenging than polarity similarity detection between close neighbors in an embedding space, we set $k_n \geq k_s$. The detailed evaluation results are presented in Table 4. It

Id	Text	Aspect	RoBERTa+MLP	S-GML
t_1	I was highly disappointed by their service and food.	food	NEG	NEG
t_2	The service is great, my soup always arrives nice and hot.	soup	POS	POS
t_3	If your looking for nasty high priced food with a dash of ghetto scenery cheap BX A$$ this is the place to be!!	priced	POS	NEG
t_4	If your looking for nasty high priced food with a dash of ghetto scenery cheap BX A$$ this is the place to be!!	food	POS	NEG

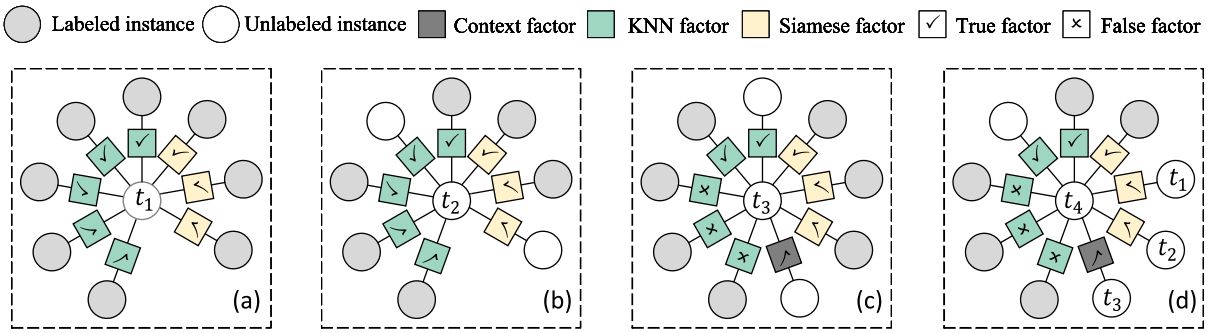

Figure 3: The illustrated examples of S-GML: the four subfigures show the extracted relational features of four instances respectively, in which a **true factor** (resp. **false factor**) means that its corresponding polarity relation is **true** (resp. **false**).

k_n	k_s	RES14		RES15		RES16	
		Acc	**Macro-F1**	**Acc**	**Macro-F1**	**Acc**	**Macro-F1**
5	3	96.90%	95.33%	90.83%	89.70%	96.00%	93.65%
5	5	97.04%	95.55%	90.96%	89.86%	95.93%	93.53%
7	3	96.61%	94.87%	90.79%	89.63%	95.80%	93.28%
7	5	96.66%	94.94%	90.68%	89.50%	95.83%	93.34%
7	7	96.61%	94.86%	90.77%	89.62%	95.83%	93.36%
9	3	96.61%	94.87%	90.87%	89.74%	95.76%	93.22%
9	5	96.63%	94.90%	90.70%	89.54%	95.68%	93.11%
9	7	96.61%	94.86%	90.74%	89.59%	95.76%	93.23%
9	9	96.57%	94.81%	90.74%	89.59%	95.74%	93.19%
k_n	k_s	LAP14		LAP15		LAP16	
		Acc	**Macro-F1**	**Acc**	**Macro-F1**	**Acc**	**Macro-F1**
5	3	95.10%	93.95%	93.70%	93.26%	89.77%	88.49%
5	5	94.88%	93.70%	93.70%	93.26%	89.35%	88.05%
7	3	94.46%	93.18%	93.40%	92.96%	89.98%	88.84%
7	5	94.67%	93.46%	93.32%	92.89%	89.56%	88.34%
7	7	94.46%	93.21%	93.28%	92.85%	89.56%	88.38%
9	3	94.88%	93.70%	93.32%	92.87%	89.77%	88.62%
9	5	95.10%	93.95%	93.36%	92.92%	89.56%	88.34%
9	7	95.10%	93.95%	93.43%	93.00%	89.56%	88.38%
9	9	94.88%	93.70%	93.47%	93.04%	89.35%	88.13%

Table 4: Sensitivity evaluation results on ATSA.

can be observed that the performance of S-GML fluctuates very marginally with different value combinations of k_n and k_s. These experimental results clearly demonstrate that the performance of S-GML is very robust w.r.t. to the parameter setting of k_n and k_s. They bode well for S-GML's applicability in real scenarios.

5.4 Comparative Evaluation on ACSA

For ACSA, we compare performance on all the RES and LAP workloads except LAP14 because it does not provide implicit aspect categories. Additionally, we compare performance on the benchmark dataset of SentiHood, which is usually considered as a task of targeted aspect-based sentiment analysis. In SentiHood, aspect category consists of two parts: explicit entity (e.g., location 1) and implicit category (e.g., safety).

We compare S-GML with the following BERT-based models specifically targeting ACSA: 1) BERT-pair-QA-M (Sun et al., 2019). It converts ACSA to a sentence-pair classification task, where the auxiliary sentence is a question. 2) BERT-pair-NLI-M (Sun et al., 2019). It converts ACSA to a sentence-pair classification task and learns aspect-specific representations by pseudo-sentence natural language inference. 3) QACG-BERT (Wu and Ong, 2021). As an improved variant of CG-BERT model (Context-Guided BERT), it learns quasi-attention weights in a compositional manner to enable subtractive attention lacking in softmax-attention. Since many deep models proposed for ATSA, e.g., BRET-SPC, AEN-BERT, LCF-BERT, BERT-PT, BAT, and PH-SUM, can be directly applied to the task of ACSA, we also compare S-GML with these models. However, we do not compare S-GML with RoBERTa+MLP and Hybrid GML because they cannot directly handle implicit aspects.

In the implementation of S-GML for ACSA, we extract neighborhood-based polarity similarity based the model of BAT (Karimi et al., 2020a), whose performance has been empirically shown to be state of the art. We use the same Siamese network proposed for ATSA to extract binary relations between arbitrary instances.

The detailed comparative results on ACSA are presented in Table 5. We have also conducted pairwise t-test between S-GML and its best alternative, BAT, over 25 runs. It can be observed that similar to what have been reported on ATSA, S-GML outperforms the best alternatives by the margins between 1% and 2% on all the test workloads. For instance, in terms of Macro-F1, S-GML beats BAT by around 2.0%, 1.5%, and 1.5% on RES14, RES15, and RES16, respectively. With regard to pairwise t-test, it can be observed that the p-values of S-GML against BAT are all well below 0.05, which means the achieved improvements are statistically significant. These experimental results clearly demonstrate the efficacy of S-GML on ACSA.

6 Conclusion and Future Work

In this paper, we have proposed a novel supervised GML approach for ATSA that can effectively exploit labeled examples to improve gradual learning. It leverages both polarity classification DNN and Siamese network to extract implicit polarity relations between instances, and then instills them into a factor graph to enable supervised knowledge conveyance. Our extensive empirical study has validated its efficacy. Our work has demonstrated clearly that in collaboration with DNN for feature extraction, GML can outperform pure DNN solutions.

For future work, it can be observed that even though the proposed solution is built upon the specific polarity classifier and Siamese network for aspect-level sentiment analysis, similar classifiers and Siamese networks are readily available or can be constructed for other binary classification tasks, especially NLP tasks. Therefore, the proposed collaboration approach of DNN and GML can be potentially generalized to other binary classification tasks.

Generalization to Multi-class Classification Tasks. It is worth pointing out that even though this paper focuses on binary classification, the proposed approach can be potentially generalized to multi-class classification tasks. In principle, instead of binary values, a variable in a factor graph can take one out of multiple values, each of which corresponds to a specific class. Relational factors can also be similarly constructed to indicate similar or different label relations between variables. We briefly illustrate the generalization by the example of three-class aspect-based sentiment analysis, whose candidate polarities include *positive*, *negative*, and *neutral*. The technical details however need further investigation in the future.

Model	RES14		RES15		RES16	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
BERT-SPC	93.90%	91.87%	88.41%	88.12%	90.71%	88.39%
AEN-BERT	94.68%	92.86%	87.09%	86.78%	91.20%	89.01%
LCF-BERT	94.74%	93.02%	88.86%	88.56%	91.98%	89.95%
BERT-PT	94.51%	92.75%	87.14%	86.79%	92.14%	90.04%
BAT	95.22%	93.58%	88.80%	88.51%	93.62%	92.05%
PH-SUM	94.99%	93.36%	89.02%	88.70%	93.25%	91.53%
BERT-pair-QA-M	94.81%	93.17%	88.25%	88.00%	92.64%	90.80%
BERT-pair-NLI-M	95.13%	93.57%	88.58%	88.30%	92.27%	90.16%
QACG-BERT	94.31%	92.47%	87.31%	86.93%	91.17%	88.96%
S-GML	**96.72%**	**95.71%**	**90.14%**	**89.86%**	**94.87%**	**93.52%**
S-GML vs BAT (p-value)	$4.22e-8$ †	$1.01e-7$ †	0.0007 †	0.0016 †	$2.06e-9$ †	$4.66e-9$ †
Model	**SentiHood**		**LAP15**		**LAP16**	
	Acc	**Macro-F1**	**Acc**	**Macro-F1**	**Acc**	**Macro-F1**
BERT-SPC	92.06%	91.14%	89.95%	89.35%	87.03%	86.26%
AEN-BERT	91.45%	90.33%	90.92%	90.38%	87.88%	86.73%
LCF-BERT	93.08%	92.29%	91.18%	90.65%	88.74%	87.68%
BERT-PT	91.53%	90.40%	91.51%	90.92%	89.13%	88.29%
BAT	93.16%	92.29%	92.56%	92.15%	89.51%	88.71%
PH-SUM	91.68%	90.59%	91.15%	90.65%	89.03%	88.20%
BERT-pair-QA-M	93.36%	92.56%	90.83%	90.28%	88.12%	87.19%
BERT-pair-NLI-M	92.85%	91.97%	91.13%	90.61%	88.47%	87.51%
QACG-BERT	92.85%	91.91%	90.44%	89.86%	87.22%	86.21%
S-GML	**93.83%**	**93.05%**	**93.74%**	**93.36%**	**90.52%**	**89.79%**
S-GML vs BAT (p-value)	0.0077 †	0.0084 †	$2.24e-5$ †	$2.56e-5$ †	$1.08e-6$ †	$2.56e-7$ †

Table 5: Comparative Evaluation Results on ACSA: the marker † indicates p-value < 0.05.

r_i	s_{ij}	Text	Aspect polarities
r_1	s_{11}	The **manager** then told us we could order from whatever **menu** we wanted but by that time we were so annoyed with the **waiter** and the resturant that we let and went some place else.	(manager, neutral), (menu, neutral), (waiter, negative)
r_2	s_{21}	Even when the **chef** is not in the house, the **food** and **service** are right on target.	(chef, neutral), (food, positive), (service, positive)
r_3	s_{31}	My friend had a **burger** and I had these wonderful **blueberry pancakes**.	(burger, neutral), (blueberry pancakes, positive)
r_4	s_{41}	It's about $7 for **lunch** and they have **take-out** or **dine-in**.	(lunch, neutral), (take-out, neutral), (dine-in, neutral)

Table 6: Illustrative examples of three-class aspect-based sentiment analysis.

Since the open-source GML inference engine[7] can effectively support gradual inference on multi-class factor graphs and modeling relational features as binary factors in a factor graph is straightforward, we focus on how to extract relational features for the task of three-class sentiment analysis. Similar to the case of binary sentiment analysis, we can extract explicit relations by analyzing discourse structures, and implicit ones by supervising a classification deep

[7] https://github.com/gml-explore/gml.

model and a Siamese network separately as follows:

- For explicit relations, we can similarly extract *opposite* relations based on the presence of shift words, because they can reliably indicate polarity shift regardless of actual sentiments. For instance, as shown in Table 6, the shift words ''but'' and ''even'' in s_{11} and s_{21} shift polarity from *neutral* to *negative* and *positive* respectively. However, identifying *similar* relations may be more subtle. Since the *neutral* polarity usually does not involve any opinion word, two aspect polarities can be reasoned to be *similar* if no shift word exists between them, and both of them contain opinion words or neither of them does. As shown in Table 6, the two aspect polarities in s_{41} can be reasoned to be *similar* due to the absence of shift words and opinion words, while the two aspect polarities in s_{31} cannot because its second part contains the opinion word of ''wonderful''.

- For implicit relations, we can similarly leverage the SOTA polarity classifiers (e.g., RoBERTa) and Siamese network for their detection. Since the SOTA polarity classifiers can naturally support three-class classification, they can be trained to detect polarity similarity based on vector neighborhood as in binary classification. As for a Siamese network, it can be similarly trained to detect similar and dissimilar relations between polarities provided that training data sufficiently cover different combinations of polarities.

Acknowledgments

Our work has been supported by National Natural Science Foundation of China (62172335, 61732014, and 61672432). We would also like to thank the action editor, and anonymous reviewers for their insightful comments and suggestions, which have significantly strengthened the paper.

References

Murtadha H. M. Ahmed, Qun Chen, Yanyan Wang, Youcef Nafa, Zhanhuai Li, and Tianyi Duan. 2021. DNN-driven gradual machine learning for aspect-term sentiment analysis. In *Findings of the Association for Computational Linguistics, ACL/IJCNLP*, pages 488–497.

Xuefeng Bai, Pengbo Liu, and Yue Zhang. 2021. Investigating typed syntactic dependencies for targeted sentiment classification using graph attention neural network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 503–514. https://doi.org/10.1109/TASLP.2020.3042009

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 452–461. https://doi.org/10.18653/v1/D17-1047

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, pages 539–546.

Junqi Dai, Hang Yan, Tianxiang Sun, Pengfei Liu, and Xipeng Qiu. 2021. Does syntax matter? A strong baseline for aspect-based sentiment analysis with roberta. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 1816–1829. https://doi.org/10.18653/v1/2021.naacl-main.146

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL*, pages 49–54. https://doi.org/10.3115/v1/P14-2009

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, EMNLP*, pages 3433–3442. https://doi.org/10.18653/v1/D18-1380

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Effective attention modeling for aspect-level sentiment classification. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING*, pages 1121–1131.

Boyi Hou, Qun Chen, Yanyan Wang, Youcef Nafa, and Zhanhuai Li. 2022. Gradual machine learning for entity resolution. *IEEE Transactions on Knowledge and Data Engineering*, 34(4):1803–1814. https://doi.org/10.1109/TKDE.2020.3006142

Binxuan Huang, Yanglan Ou, and Kathleen M. Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In *Social, Cultural, and Behavioral Modeling - 11th International Conference, SBP-BRiMS*, pages 197–206. https://doi.org/10.1007/978-3-319-93372-6_22

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 562–570. https://doi.org/10.18653/v1/P17-1052

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2020a. Adversarial training for aspect-based sentiment analysis with BERT. In *Proceedings of the 25th International Conference on Pattern Recognition, ICPR*, pages 8797–8803. https://doi.org/10.48550/arXiv.2010.11731

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2020b. Improving BERT performance for aspect-based sentiment analysis. *arXiv preprint arXiv:2010.11731*.

Mahmut Kaya and Hasan Sakir Bilge. 2019. Deep metric learning: A survey. *Symmetry*, 11(9):1066. https://doi.org/10.3390/sym11091066

Frank R. Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. 2001. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519. https://doi.org/10.1109/18.910572

Zeyang Lei, Yujiu Yang, and Yi Liu. 2018. LAAN: A linguistic-aware attention network for sentiment analysis. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW*, pages 47–48. https://doi.org/10.1145/3184558.3186922

Gaël Letarte, Frédérik Paradis, Philippe Giguère, and François Laviolette. 2018. Importance of self-attention for sentiment analysis. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP*, pages 267–275. https://doi.org/10.18653/v1/W18-5429

Qiao Liu, Haibin Zhang, Yifu Zeng, Ziqi Huang, and Zufeng Wu. 2018. Content attention model for aspect based sentiment analysis. In *Proceedings of the 2018 Web Conference, WWW*, pages 1023–1032. https://doi.org/10.1145/3178876.3186001

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. A cognition based attention model for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 462–471. https://doi.org/10.18653/v1/D17-1048

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level aentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI*, pages 4068–4074.

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 V. Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez Zafra, and Gülsen Eryigit. 2016. SemEval-2016 Task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT*, pages 19–30. https://doi.org/10.18653/v1/S16-1002

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@NAACL-HLT*, pages 486–495. https://doi.org/10.18653/v1/S15-2082

Qiao Qian, Minlie Huang, Jinhao Lei, and Xiaoyan Zhu. 2017. Linguistically regularized LSTM for sentiment classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 1679–1689. https://doi.org/10.18653/v1/P17-1154

Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46. https://doi.org/10.1016/j.knosys.2015.06.015

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-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 3980–3990. https://doi.org/10.18653/v1/D19-1410

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 12th Language Resources and Evaluation Conference, LREC*, pages 4933–4941.

Emanuel H. Silva and Ricardo M. Marcacini. 2021. Aspect-based sentiment analysis using BERT with disentangled attention. In *Proceedings of the LatinX in AI (LXAI) Research workshop at ICML 2021*.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*. https://doi.org/10.48550/arXiv.1902.09314

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, NAACL-HLT*, pages 380–385.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective LSTMs for target-dependent sentiment classification. In *Proceedings of the 26th International Conference on Computational Linguistics, COLING*, pages 3298–3307.

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, ACL*, pages 6578–6588. https://doi.org/10.18653/v1/2020.acl-main.588

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, NAACL-HLT*, pages 2910–2922. https://doi.org/10.18653/v1/2021.naacl-main.231

Bailin Wang and Wei Lu. 2018. Learning latent opinions for aspect-level sentiment classification. In *Proceedings of the 32nd Conference on Artificial Intelligence, AAAI, the 30th innovative Applications of Artificial Intelligence, IAAI, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI*, pages 5537–5544. https://doi.org/10.1609/aaai.v32i1.12020

Yanyan Wang, Qun Chen, Jiquan Shen, Boyi Hou, Murtadha Ahmed, and Zhanhuai Li. 2021. Aspect-level sentiment analysis based on gradual machine learning. *Knowledge-Based Systems*, 212:106509. https://doi.org/10.1016/j.knosys.2020.106509

Zhengxuan Wu and Desmond C. Ong. 2021. Context-guided BERT for targeted aspect-based sentiment analysis. In *Proceedings of 35th AAAI Conference on Artificial Intelligence, AAAI, 33rd Conference on Innovative Applications of Artificial Intelligence, IAAI, The 11th Symposium on Educational Advances in Artificial Intelligence, EAAI*,

pages 14094–14102. https://doi.org/10.1609/aaai.v35i16.17659

Bowen Xing and Ivor W. Tsang. 2022. Understand me, if you refer to aspect knowledge: Knowledge-aware gated recurrent memory network. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(5):1092–1102. https://doi.org/10.1109/TETCI.2022.3156989

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 2324–2335.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 2514–2523. https://doi.org/10.18653/v1/P18-1234

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. LCF: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9(16):3389. https://doi.org/10.3390/app9163389

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Syntax-aware aspect-level sentiment classification with proximity-weighted convolution network. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*, pages 1145–1148. https://doi.org/10.1145/3331184.3331351

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems, NIPS*, pages 649–657.

Pinlong Zhao, Linlin Hou, and Ou Wu. 2020. Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. *Knowledge-Based Systems*, 193:105443. https://doi.org/10.1016/j.knosys.2020.106292