IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 24, 2020, accepted October 13, 2020, date of publication October 16, 2020, date of current version October 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3031665

SEML: A Semi-Supervised Multi-Task Learning Framework for Aspect-Based Sentiment Analysis

NING LI, CHI-YIN CHOW, (Senior Member, IEEE), AND JIA-DONG ZHANG, (Member, IEEE)
Department of Computer Science, City University of Hong Kong, Hong Kong

Corresponding authors: Chi-Yin Chow (chiychow@cityu.edu.hk) and Jia-Dong Zhang (jzhang26@cityu.edu.hk)

This work was supported in part by the Innovation and Technology Fund (ITF) under Grant UIM/376.

ABSTRACT Aspect-Based Sentiment Analysis (ABSA) involves two sub-tasks, namely Aspect Mining (AM) and Aspect Sentiment Classification (ASC), which aims to extract the words describing aspects of a reviewed entity (e.g., a product or service) and analyze the expressed sentiments on the aspects. As AM and ASC can be formulated as a sequence labeling problem to predict the aspect or sentiment labels of each word in the review, supervised deep sequence learning models have recently achieved the best performance. However, these supervised models require a large number of labeled reviews which are very costly or unavailable, and they usually perform only one of the two sub-tasks, which limits their practical use. To this end, this paper proposes a SEmi-supervised Multi-task Learning framework (called SEML) for ABSA. SEML has three key features. (1) SEML applies Cross-View Training (CVT) to enable semi-supervised sequence learning over a small set of labeled reviews and a large set of unlabeled reviews from the same domain in a unified end-to-end architecture. (2) SEML solves the two sub-tasks simultaneously by employing three stacked bidirectional recurrent neural layers to learn the representations of reviews, in which the representations learned from different layers are fed into CVT, AM and ASC, respectively. (3) SEML develops a Moving-window Attentive Gated Recurrent Unit (MAGRU) for the three recurrent neural layers to enhance representation learning and prediction accuracy, as nearby contexts within a moving-window in a review can provide important semantic information for the prediction task in ABSA. Finally, we conduct extensive experiments on ABSA over four review datasets from the SemEval workshops. Experimental results show that SEML significantly outperforms the state-of-the-art models.

INDEX TERMS Aspect-based sentiment analysis, semi-supervised learning, multi-task learning, end-to-end learning, cross-view training, moving-window attention.

I. INTRODUCTION

Product and service reviews posted by their users have been drawn a lot of attentions from both industry and academic communities. Document-level or sentence-level sentiment analysis tells an overall opinion about a review or sentence, whereas Aspect-Based Sentiment Analysis (ABSA) provides more fine-grained information by mining aspects and analyzing aspect-level opinions for a discussed entity [1], [2]. For instance, a user posts a review on a laptop: *"I love the operating system but not the preloaded software"* which contains two aspects, i.e., ''operating system'' with a positive sentiment and ''preloaded software'' with a negative sentiment.

Generally, ABSA can be divided into two sub-tasks, namely Aspect Mining (AM) and Aspect Sentiment Classification (ASC) [1]. The AM sub-task extracts the aspect words from each sentence of reviews, which has been extensively studied by applying unsupervised models [3]–[5], supervised models [6]–[13], or semi-supervised techniques [14]–[19]. The ASC sub-task that aims to predict the sentiment polarities on these aspects also has been increasingly discussed recently [20]–[25]. However, these works [3]–[25] only focus on one of the sub-tasks. As a result, it is required to train two different models and pipeline them together for ABSA. Nonetheless, the literatures [26], [27] show that the pipeline method is usually not the best solution for highly related tasks in Natural Language Processing (NLP) and an integrated method is more effective by jointly training different and related tasks. Thus, increasing attention has been paid on this integration direction [28]–[30]. As aspect words and

The associate editor coordinating the review of this manuscript and approving it for publication was Firooz B. Saghezchi.

sentiment words often co-appear and can help find each other, AM and ASC are strongly related sub-tasks; a jointly trained method for the two sub-tasks in ABSA is promising. Moreover, most existing works [6]–[13], [21]–[25] adopt supervised learning for the AM or ASC sub-tasks and require a large amount of labeled reviews. The manual labeling on training data is very costly, especially for domain-dependent aspects, i.e., different domains may have different aspect spaces. Researchers are motivated to develop more effective semi-supervised learning models for ABSA [31]. Thus, our two main concerns are: (1) whether we can fully use both labeled and unlabeled reviews and (2) whether we can perform both AM and ASC sub-tasks in an end-to-end architecture at the same time. Our previous work [19] has addressed the first concern, in which the proposed model can leverage both labeled and unlabeled reviews only for the AM sub-task in the unified framework.

In this paper, we propose a new SEmi-supervised Multi-task Learning framework (called **SEML**) to enhance ABSA on user reviews. SEML follows the method in our previous work [19] to alternately learn a model on a mini-batch of labeled reviews and a mini-batch of unlabeled reviews from the same domain based on Cross-View Training (CVT) [32] to enable semi-supervised learning. In the CVT, one primary prediction module for either AM or ASC is trained with the standard supervised learning on labeled reviews and four auxiliary prediction modules with different views on unlabeled reviews are trained to agree with the AM or ASC primary prediction module. CVT switches training on labeled and unlabeled reviews to improve both review representations and prediction modules.

Meanwhile, as AM and ASC are highly coupled together, SEML applies multi-task learning by sharing the representation learning in different layers for performing AM and ASC in the same framework. More specifically, three stacked bidirectional recurrent neural layers are employed to learn representations of reviews, in which the representations from the first layer are fed into the four auxiliary prediction modules of CVT to leverage unlabeled reviews, the representations from the second layer are fed into the primary prediction module for AM, and the representations from the third layer are fed into the primary prediction module for ASC. Moreover, each upper layer uses the representations from lower layer as inputs, so SEML enables multi-task learning and interaction between different sub-tasks to improve the aspect and sentiment prediction.

Further, SEML considers a significant observation that nearby contexts of a word in a sentence provide important semantic information for a prediction task in ABSA. For instance, the past nearby aspect words (e.g., ''operating system'') should be more significant than other words to guide the extraction of subsequent aspects (e.g., ''preloaded software''), and a closer sentiment word is more likely to be the corresponding opinion for the aspect (e.g., ''love'' for ''operating system'' and ''not'' for ''preloaded software''). Therefore, SEML devises a Moving-window Attentive Gated Recurrent Unit (MAGRU) as the neural unit in the three bidirectional recurrent neural layers; MAGRU extends Gated Recurrent Unit (GRU) [33] with an attention mechanism to encode the information within a moving-window.

In general, the contributions of this paper can be summarized as below.

- We propose the first semi-supervised deep multi-task learning framework for both AM and ASC sub-tasks in ABSA, which introduces CVT to use unlabeled reviews to improve the representation learning within a unified end-to-end architecture.
- We enable multi-task learning to perform AM and ASC sub-tasks in the same framework with three stacked bidirectional recurrent neural layers and corresponding prediction modules.
- We develop a moving-window attention mechanism within the GRU, i.e., MAGRU, to capture significant past nearby information for the aspect and sentiment prediction.
- We conduct extensive experiments to evaluate the performance of SEML for AM, ASC and complete ABSA based on the four review datasets from the SemEval workshops. Experimental results show that SEML is significantly better than the state-of-the-art models.

The reminder of this paper is organized as follows. Section II discusses related works. Then, we present our SEML framework in Section III. Section IV shows the experimental results. Finally, Section V concludes this paper.

II. RELATED WORKS

A. ABSA AS SEQUENCE LABELING

Sequence labeling is a very common problem in NLP (e.g., part-of-speech tagging and named-entity recognition) and aims to assign a label to each element in a sequential input. Both AM and ASC can be formulated as a sequence labeling problem, in which a label is given to each word in the review. Formally, AM predicts a label sequence $\{y_1^A, \ldots, y_T^A\}$ for a given sentence with T words $\{x_1, \ldots, x_T\}$, where $y_t^A \in \{ASPECT, NONASPECT\}$, and the label space changes to $y_t^S \in \{SENTIMENT\ POLARITIES\}$ for ASC. For instance, the reference [6] defines a set of labels to distinguish feature aspects, component aspects and function aspects, and trains hidden Markov models to label each word in a review. Further, the researchers [7] simplify these labels and apply $\{B, I, O\}$ scheme, where B, I and O identify the beginning of an aspect, the continuation of the aspect, and other words, respectively. The $\{B, I, O\}$ scheme can well handle aspects expressing in phrases and has been applied for AM [9], [18] and aspect-opinion term co-extraction [11], [12]. In the ASC sub-task, as aspects are assumed to be known, the prediction model only needs to assign a sentiment polarity to each aspect with $\{POS, NEG, NEU, O\}$ scheme [24], [25], where POS, NEG and NEU denote positive, negative and neutral sentiment, respectively, and O for other words. Recently, a collapsed labeling scheme is applied to perform ABSA as a single sequence labeling task [29],

in which aspects and sentiments labels are combined as {*B*-{*POS*, *NEG*, *NEU*}, *I*-{*POS*, *NEG*, *NEU*}, *O*} scheme. We do not follow this collapsed scheme in our SEML, as we consider the interaction between two sub-tasks can improve the aspect and sentiment prediction. Thus, our SEML uses the {*B*, *I*, *O*} and {*POS*, *NEG*, *NEU*, *O*} labeling schemes.

B. SEMI-SUPERVISED APPROACHES

Most existing semi-supervised methods for ABSA are proposed only for AM. One direction is to use prior domain knowledge to guide an unsupervised topic model (e.g., Latent Dirichlet Allocation). For instance, some methods manually choose domain specified seed words [14]–[17] for topic modeling. However, this kind of methods often need manually defined domain knowledge and do not fully use labeled reviews. Another direction takes full advantage of unlabeled reviews in the same domain to improve supervised models. The idea of pre-training has been applied in the AM model [18] to learn domain-specific word embeddings from unlabeled reviews in advance which then fed into normal supervised models. However, instead of pre-training, our previous work [19] learns both task- and domain-specific representations of reviews in a unified framework, which improves the AM sub-task. For the sentiment classification problem, some researchers [20], [28], [31], [34], [35] prefer to use external linguistic resources to catch the affective information of words, which can be considered as the special case of semi-supervised approaches. For example, people propose commonsense knowledge networks to perform concept-level sentiment analysis [20], [31], [34], [35]; the authors of the work [28] try to encode commonsense knowledge into their attentive neural network for ASC. The literature [36] uses data augmentation method to generate more labeled training data to achieve semi-supervised learning for ABSA. Nonetheless, supervised deep learning models currently have achieved great successes when applied to AM [9]–[13], ASC [21]–[25] and complete ABSA [29]. To the best of our knowledge, we are the first to propose an end-to-end semi-supervised deep learning framework that can leverage labeled and unlabeled reviews for both the AM and ASC sub-tasks.

C. CROSS-VIEW TRAINING

Normally, a deep learning model works best when it is trained on a large amount of data with reliable labels. However, for domain-dependent aspects, manual labeling could be a huge investment. One solution is to apply effective semi-supervised learning to leverage a plenty of unlabeled reviews. Current semi-supervised learning models [18] separate the training process into two phases: pre-training and supervised learning. A key disadvantage of such models is that the first phase on representation learning does not benefit from any labeled reviews. More sophisticatedly, CVT [32] implements semi-supervised learning by alternately switching the training process on labeled data and unlabeled data, which is the meaning of Cross-View. Note that, the term *Cross-View* may refer to multi-view learning in some works [37], [38], in which the models learn from multi-view data (e.g., images taken from different viewpoints) instead of switching between labeled and unlabeled data. Our previous work [19] has showed that CVT can well leverage both labeled and unlabeled reviews. Thus, the SEML framework is also based on CVT, but it has a more task-specific architecture that combines CVT with multi-task learning to jointly train multiply models at the same time.

D. MULTI-TASK LEARNING

Extensive works [26], [27] show that the models jointly trained for closely related tasks can outperform those models for a single task. For ABSA, a multi-task supervised model with two coupled GRU layers is proposed to co-extract aspects and sentiment words [12]. The authors of the paper [9] employ a neural network with three Long Short-Term Memory (LSTM) layers to perform the multi-task learning for AM. Further, they add an attention mechanism into the AM model [13]. The multi-task learning is also applied for ASC. For example, people propose the Sentic LSTM (a two-step attentive LSTM) [28] to perform aspect categorization and ASC; the researchers of the literature [29] propose the E2E-TBSA model for ASC based on two bidirectional LSTM layers and the collapsed labeling scheme; and an interactive multi-task learning network (IMN) [30] is proposed to learn the model from the token-level AM and ASC sub-tasks as well as the document-level tasks. However, all the aforementioned studies are based on supervised learning that relies on large amounts of labeled reviews to guarantee good performance. In contrast, our SEML framework can leverage unlabeled reviews to enhance supervised models and alleviate the costly demand on data labeling.

III. THE FRAMEWORK SEML

In this section, we present our semi-supervised deep learning framework SEML for ABSA. First, we formulate the two sub-tasks AM and ASC into sequence labeling problems. Then, we present the technical details of the key components in SEML.

A. PROBLEM STATEMENT

Suppose there are one set (D_u) of unlabeled reviews from a domain (or an entity) and two sets (D_l^{AM} and D_l^{ASC}) of labeled reviews from the same domain which are annotated for AM and ASC, respectively. The AM sub-task is to learn a classifier from the reviews in D_l^{AM} and D_u to extract a set of aspects, while the ASC sub-task is to train a model from the reviews in D_l^{ASC} and D_u to predict the sentiment polarities for the aspects. These two sub-tasks can be formulated as different sequence labeling problems by using different tagging schemes. Specifically, we use the {*B*, *I*, *O*} scheme for AM, where *B*, *I*, and *O* indicate the beginning of, the continuation of, and the out of the aspect, respectively (refer to Section II-A). For the ASC sub-task, the {*POS*, *NEG*, *NEU*, *O*} scheme is applied, where *POS*, *NEG*, and *NEU* express the

TABLE 1. An example on AM and ASC as sequence labeling problems. Y^A shows the aspect labels for each word, and Y^S means the sentiment polarities for the aspect.

X	I	love	the	operating	system	but	not	the	preloaded	software	which	slow	down	the	booting	a	lot
Y^A	O	O	O	B	I	O	O	O	B	I	O	O	O	O	B	O	O
Y^S	O	O	O	POS	POS	O	O	O	NEG	NEG	O	O	O	O	NEG	O	O

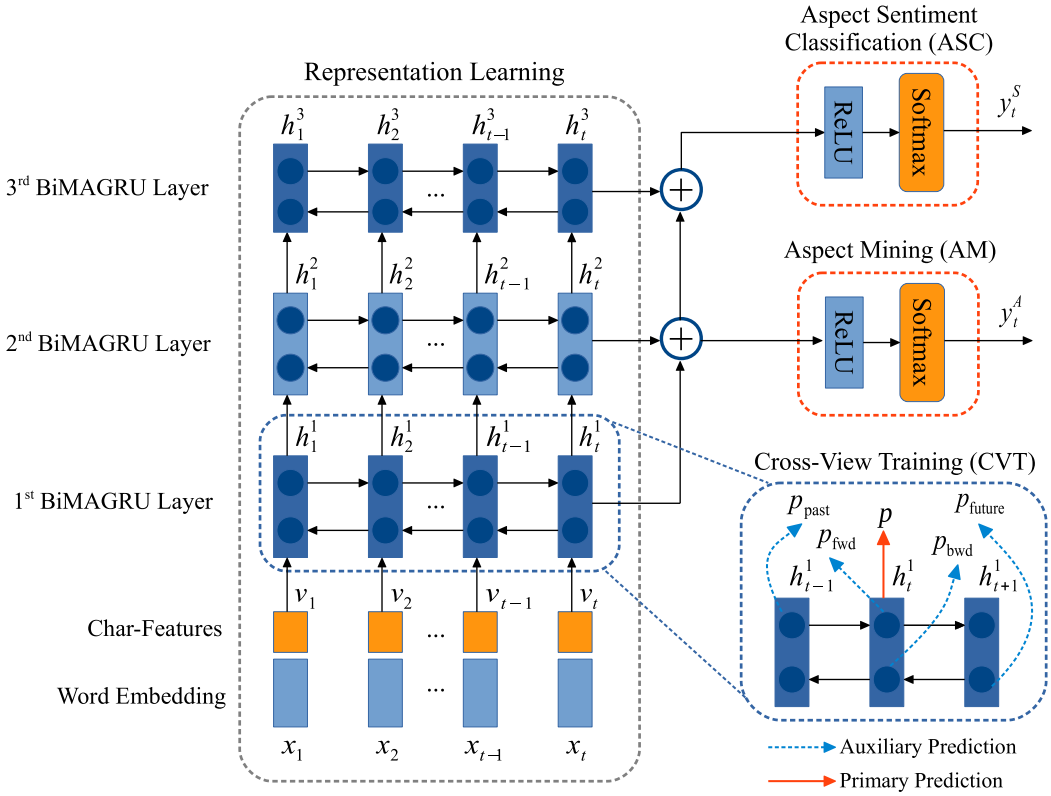

FIGURE 1. The architecture of our SEML framework. Refined word embedding and char-features are fed into three stacked BiMAGRU layers. The first BiMAGRU layer is shared with CVT to leverage unlabeled reviews, in which four auxiliary prediction modules are trained to agree with both AM and ASC (i.e., the primary prediction modules). The second BiMAGRU layer trains the AM model to extract aspects, and the third layer trains the ASC model to predict sentiment polarities.

positive, negative, and neural sentiment respectively, and O means the NULL sentiment for a word. Then, each word x_t in the review sentence $X = \{x_1, \ldots, x_T\}$ should be assigned as one of $Y^A \in \{B, I, O\}$ and one of $Y^S \in \{POS, NEG, NEU, O\}$ (see TABLE 1 for instance).

B. FRAMEWORK ARCHITECTURE

Our motivation is to fully use the labeled and unlabeled reviews and simultaneously perform both AM and ASC within an end-to-end framework. As shown in FIGURE 1, our SEML framework consists of four components including representation learning, AM, ASC and CVT. Since recurrent neural networks (RNNs) can naturally represent the sequential information, our framework employs deep RNNs as the basic architecture to build the shared contextualized representation learning component for both AM and ASC sub-tasks. Specifically, three stacked bidirectional recurrent neural layers with MAGRU are employed to build the shared memory; MAGRU extends GRU with moving-window attention mechanism to encode nearby semantic significances. We give a detailed design of MAGRU in Section III-C.

Moreover, each stacked Bidirectional MAGRU (BiMAGRU) layer is designed to learn representations for different tasks. Specifically, the first layer is shared with CVT to train four auxiliary prediction modules for AM or ASC by leveraging unlabeled reviews; the second layer uses the representations from the first layer as inputs and trains one primary prediction module for AM; and the third layer inputs the representations from the second layer and trains the other primary prediction module for ASC. As each upper layer uses the outputs from the lower layer as inputs, our SEML enables not only multi-task learning but also the interaction between different sub-tasks to improve the aspect extraction and sentiment prediction. The detailed representation learning process is presented in Section III-D.

To enable semi-supervised learning, our SEML framework trains on both labeled and unlabeled reviews for two sub-tasks

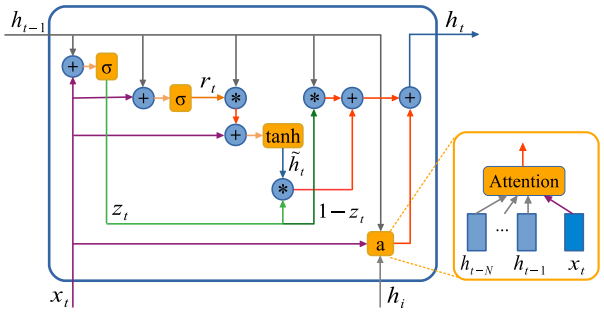

FIGURE 2. An illustration of the proposed Moving-window Attentive GRU (MAGRU).

(AM and ASC) in ABSA by applying CVT. While performing CVT, the primary prediction modules for AM and ASC are trained with the standard supervised learning on labeled reviews; on unlabeled reviews, four auxiliary prediction modules (namely p_{past}, p_{fwd}, p_{bwd}, and p_{future}) with different views on the input data are trained to agree with the primary prediction modules. We discuss the specific multi-task CVT in Sections III-E and III-F.

C. MOVING-WINDOW ATTENTIVE GRU

As introduced above, our framework employs deep RNNs to build the shared representation learning component. However, in ABSA, the information from past nearby steps provide useful clues for a prediction, e.g., the aspect label "*I*" cannot follow "*O*", and the previous aspects can guide the extraction of subsequent aspects. Though RNNs with (LSTM) [39] or GRU [33] can well encode long period of sequential information, they are difficult to pay attention to exactly useful nearby contexts at each time step. To this end, our framework extends GRU with a Moving-window Attention mechanism (called MAGRU) that can capture past nearby significances.

We prefer extending GRU as it has a simpler structure and less parameters than LSTM but shows competitive performance in many NLP tasks [40]. Specifically, as shown in FIGURE 2, MAGRU has three gates, namely reset gate r, update gate z, and attention gate a. The update gate z_t at time step t is obtained as follow:

$$z_t = \sigma(U_z x_t + W_z h_{t-1}), \tag{1}$$

where h_{t-1} is the previous hidden state, x_t is the input of current step, U_z and W_z indicate gate parameters, and σ is the sigmoid activation function. At the same time, the reset gate r_t is computed by:

$$r_t = \sigma(U_r x_t + W_r h_{t-1}). \tag{2}$$

Thus, the new candidate hidden state $\tilde{h}_t$ without any attentions for current time step can be obtained by using *tanh* activation function:

$$\tilde{h}_t = tanh(U_h x_t + W_h(r_t * h_{t-1})). \tag{3}$$

The above update and reset gates are the same with GRU. However, we add a new attention gate to encode past nearby significances. Specifically, the moving-window attention considers the most recent N (moving-window size) hidden states. At step t, we calculate the normalized significance score s_i^t of each cached past state h_i ($i \in [t-N, t-1]$) as follow:

$$s_i^t = Softmax(U_a \cdot tanh(W_a^1 h_i + W_a^2 x_t)), \tag{4}$$

where *tanh* is the activation function, U_a, W_a^1, and W_a^2 are the attention parameters. Then the attention gate a_t is given by:

$$a_t = ReLU\left(\sum_{i=t-N}^{t-1} s_i^t h_i\right), \tag{5}$$

where we compute the weighted sum of the cached previous N hidden states h_i with the score weights s_i^t, and apply the *ReLU* activation function.

Finally, to calculate current moving-window attentive hidden state h_t at step t, our framework considers all the three gates:

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t + a_t, \tag{6}$$

in which $\tilde{h}_t$ is determined by the reset gate to combine the new input with the previous hidden state. The update gate defines how much of the previous information to keep, and the attention gate gives the past nearby significance.

D. REPRESENTATION LEARNING

Pre-trained general word embeddings (e.g., GloVe [41]) have been widely used in recent NLP models, which became a essential component to convert texture information into contextualized vectors for later computation. However, the researchers [42] discover that these general word embeddings often represent opposite sentiment words (e.g., *good* and *bad*) with similar vectors, which affects the final sentiment classification. Thus, they propose an adjusting method to use a sentiment lexicon that refines the embeddings of sentiment words to be closer to sentimentally similar words and farther away from sentimentally different ones, and improves the classification performance on many sentiment related tasks. SEML also applies the same method [42]. Moreover, because combining general embeddings with char-features can help handle misspelling words [43], SEML represents each word in the input sequence as the concatenation of the refined embedding vector and char-features from a character-level Convolutional Neural Network (CNN) [43]. Then the concatenation vectors are fed into the deep bidirectional RNN.

The RNN employs three stacked BiMAGRU layers to build the shared memory for both AM and ASC sub-tasks, in which each upper layer uses the hidden states from the lower layer as inputs. Specifically, we feed the inputs forwardly and backwardly to MAGRUs and combine them as one BiMAGRU layer, because both forward and backward information is important for the prediction on the current position.

Formally, let $V = \{v_1, \ldots, v_T\}$ be the concatenation vectors of refined word embeddings and char-features. The hidden representations for each layer are derived by concatenating the outputs of both forward $\overrightarrow{MAGRU}$ and backward $\overleftarrow{MAGRU}$ as follows:

$$h_t^1 = [\overrightarrow{MAGRU}(v_t) \oplus \overleftarrow{MAGRU}(v_t)], \quad (7)$$

$$h_t^2 = [\overrightarrow{MAGRU}(h_t^1) \oplus \overleftarrow{MAGRU}(h_t^1)], \text{ and} \quad (8)$$

$$h_t^3 = [\overrightarrow{MAGRU}(h_t^2) \oplus \overleftarrow{MAGRU}(h_t^2)], \quad (9)$$

in which $t \in [1, T]$ and $\oplus$ denotes the concatenation operation, h_t^1 is the hidden representations from the first BiMAGRU layer at t time step, h_t^2 is from the second layer, and h_t^3 is from the third layer.

E. PREDICTION MODULES

SEML trains models on both labeled reviews and unlabeled reviews for two sub-tasks (AM and ASC) in ABSA. SEML learns one primary prediction module from labeled reviews and four auxiliary prediction modules from unlabeled reviews with restricted views of inputs for each sub-task (AM or ASC). Suppose y_t^A is the aspect label for the word $x_t \in X$. The primary prediction module for AM determines the probability distribution $p(y_t^A|x_t)$ over the aspect labels $\{B, I, O\}$ from the representations (h_t^1 and h_t^2) from the first and second MAGRU layers with a simple one-hidden-layer neural network, given by:

$$p(y_t^A|x_t) = Softmax(U_p^A \cdot ReLU(W_p^A(h_t^1 \oplus h_t^2)) + b^A), \quad (10)$$

in which U_p^A, W_p^A and b^A are the model parameters.

Further, since ASC relies on the position of aspects, the aspect boundary information from the primary module for AM is delivered into the third BiMAGRU layer for ASC. Therefore, the moving-window attention in the third layer can help the primary module for ASC focus on the corresponding sentiment words and maintain the consistency of sentiment labels assigned to multi-word aspects. The primary prediction module for ASC adopts the similar architecture as in AM, given by

$$p(y_t^S|x_t) = Softmax(U_p^S \cdot ReLU(W_p^S(h_t^1 \oplus h_t^2 \oplus h_t^3)) + b^S), \quad (11)$$

where $y_t^S \in \{POS, NEG, NEU, O\}$.

As mentioned above, SEML shares the first BiMAGRU layer with the auxiliary prediction modules that have restricted views of unlabeled reviews. There are four different auxiliary prediction modules (p_{past}, p_{fwd}, p_{bwd}, and p_{future}) in the framework for each sub-task (AM or ASC), where p_{past} means, for the prediction of current word, this module only has a view of all past words on the left of current word in the sentence; p_{fwd} has a view of past (left) and current words; p_{bwd} observes current and words on the future (right); and p_{future} only observes all future words on the right, as shown in FIGURE 1. BiMAGRU can easily provide these restricted views without additional computation as follows:

$$p_{\text{past}}(y_t^k|x_t) = nn_{\text{past}}(\overrightarrow{h}^1_{t-1}), \quad (12)$$

$$p_{\text{fwd}}(y_t^k|x_t) = nn_{\text{fwd}}(\overrightarrow{h}^1_{t}), \quad (13)$$

$$p_{\text{bwd}}(y_t^k|x_t) = nn_{\text{bwd}}(\overleftarrow{h}^1_{t}), \quad (14)$$

$$p_{\text{future}}(y_t^k|x_t) = nn_{\text{future}}(\overleftarrow{h}^1_{t+1}), \quad (15)$$

where $k \in \{A, S\}$, nn_{past}, nn_{fwd}, nn_{bwd}, and nn_{future} denote the neural network with the structure given in Equation (10) or (11). Since the second and third BiMAGRU layers have already seen all words, we can only feed the hidden representations $\overrightarrow{h}^1$ and $\overleftarrow{h}^1$ from the first BiMAGRU layer to the auxiliary prediction modules in order to restrict their view on an input sequence.

F. MULTI-TASK CROSS-VIEW TRAINING

The key idea of CVT is to use unlabeled reviews from the same domain of labeled reviews to enhance the representation learning and alternately learn primary and auxiliary prediction modules on a mini-batch of labeled reviews or unlabeled reviews. In order to perform multi-task learning, i.e., to train one primary module and four auxiliary modules for AM or ASC, we randomly choose a sub-task (AM or ASC) with its labeled reviews at first. Then the Cross-Entropy (CE) loss is utilized to train the corresponding primary prediction module $p(y_t^A|x_t)$ or $p(y_t^S|x_t)$:

$$L_{\text{SUP}}^k = \frac{1}{|D_l^k|} \sum_{x_t, y_t \in D_l^k} CE(y_t, p(y_t^k|x_t)), \; k \in \{A, S\}. \quad (16)$$

For the unlabeled reviews D_u, the framework first infers $p(y_i^A|x_i)$ as well as $p(y_i^S|x_i)$ ($x_i \in D_u$) based on the primary modules for AM and ASC and then trains the auxiliary prediction modules to match two primary prediction modules by using the Kullback-Leibler (KL) divergence function as the loss:

$$L_{\text{CVT}} = \frac{1}{|D_u|} \sum_{x_i \in D_u} \sum_k \sum_j KL(p(y_i^k|x_i), p_j(y_i^k|x_i)), \quad (17)$$

where $j \in \{\text{left, fwd, bwd, right}\}$, $k \in \{A, S\}$, and the parameters of the primary modules are fixed during training. The auxiliary prediction modules can learn to enhance the shared representations, because the new words that are not in labeled reviews may have been encoded into the model and be useful for making predictions on aspects and sentiments. Reviews labeled across both tasks are useful for multi-task models, but most publicly available labeled reviews are only for one particular task (e.g., either AM or ASC). SEML utilizes unlabeled reviews for both sub-tasks and actually constructs all-tasks-labeled examples from unlabeled reviews.

Finally, we combine the supervised and CVT losses and minimize the total loss L with stochastic gradient descent:

$$L = L_{\text{SUP}}^A + L_{\text{SUP}}^S + L_{\text{CVT}}. \quad (18)$$

In particular, we randomly choose a sub-task and alternately minimize L_{SUP}^A or L_{SUP}^S over a mini-batch of corresponding

labeled reviews and L_{CVT} over a mini-batch of unlabeled reviews.

IV. EXPERIMENTS

In this section, we evaluate the performance of our proposed SEML framework and compare it with the state-of-the-art approaches for both AM and ASC sub-tasks in ABSA. Moreover, we test SEML to perform complete ABSA and compare it to those pipeline and unified approaches.

A. EXPERIMENTAL SETTINGS

1) DATASETS

We conduct experiments over four benchmark datasets from the SemEval workshops [44], [45]. TABLE 2 shows their statistics. $\mathbf{D}^{\mathbf{AM}}_{\mathbf{laptop}}$ and $\mathbf{D}^{\mathbf{AM}}_{\mathbf{rest}}$ contain reviews of the laptop and restaurant domain for the AM sub-task, while $\mathbf{D}^{\mathbf{ASC}}_{\mathbf{laptop}}$ and $\mathbf{D}^{\mathbf{ASC}}_{\mathbf{rest}}$ are for the ASC sub-task. In the AM datasets, the sentiment polarities for aspects are not given. In the ASC datasets, both aspects and their sentiment polarities are known. As some testing sentences in one sub-task may appear in the other sub-task's training dataset, we simply remove those sentences from the training dataset for fair comparison.

TABLE 2. Statistics of labeled datasets.

AM	$\mathbf{D}^{\mathbf{AM}}_{\mathbf{laptop}}$		$\mathbf{D}^{\mathbf{AM}}_{\mathbf{rest}}$	
	Training	Testing	Training	Testing
#Sentences	3,045	800	2,000	676
#Aspects	2,358	654	1,743	622
ASC	$\mathbf{D}^{\mathbf{ASC}}_{\mathbf{laptop}}$		$\mathbf{D}^{\mathbf{ASC}}_{\mathbf{rest}}$	
	Training	Testing	Training	Testing
#Positive	987	341	2,164	728
#Negative	866	128	805	196
#Neutral	460	169	633	196

Moreover, SEML needs unlabeled reviews for CVT (semi-supervised learning). We collect unlabeled reviews corresponding to two domains (laptop and restaurant) of labeled datasets to train the model, which include laptop reviews from Amazon Review Dataset (230,373 sentences) [46] and restaurant reviews from Yelp Review Dataset (2,677,025 sentences) [47]. For comparison, we also train the model on a general unlabeled dataset (One Billion Word Language Model Benchmark) [48] to see whether perform CVT on general texts can improve the supervised model for AM and ASC. As some sentences in the testing dataset may also appear in unlabeled reviews, we remove these sentences in unlabeled reviews to make the comparison fair.

2) COMPARED MODELS

We first compare SEML with the state-of-the-art models for the AM sub-task, including:

- **CMLA** [12] applies a multi-layer architecture with coupled-attentions to locate aspect words.
- **MIN** [9] consists of three LSTM layers for multi-task learning, in which a sentiment lexicon (to find opinion words) and dependency rules are used to extract corresponding aspects.
- **DE-CNN** [18] is based on CNNs and utilizes both general word embeddings and domain-specific embeddings learned from unlabeled reviews.
- **EMOVA** [19] uses the CVT and moving-window attention mechanism to leverage both labeled and unlabeled reviews.

Then, we compare SEML with the state-of-the-art models for the ASC sub-task, including:

- **RAM** [49] employs the multiple attentions on multi-layer RNNs to combine hidden word features in each layer.
- **TNet** [24] utilizes a CNN layer instead of an attention layer to extract the salient features from the representations learned by deep RNNs.
- **MGAN** [25] applies transfer learning to leverage knowledge learned from a rich-resource source domain to improve the learning in a low-resource target domain.

In addition, since BERT [50] is one of the key innovations in the recent progress of language modeling and achieves the state-of-the-art performance on many NLP tasks, we fine-tune the pre-trained BERT model on the datasets for both AM and ASC as a baseline:

- **BERT** [50] can learn better representations by training a deep language model on large amounts of texts, we apply $\text{BERT}_{\text{BASE}}$ on the datasets as the baseline to perform AM and ASC as well as complete ABSA.

We also investigate the performance of important variants of SEML:

- **SEML-SUP** is our supervised model but without CVT on unlabeled reviews, so it is a purely supervised multi-task learning model.
- **SEML-GNL** is the full framework but only performing CVT on the general unlabeled text (One Billion Word Language Model Benchmark) [48] which is not specific to the laptop or restaurant domain.
- **SEML-AM** is the single task model for AM with CVT on unlabeled reviews.
- **SEML-ASC** is the single task model for ASC with CVT on unlabeled reviews.

Finally, our goal is to perform complete ABSA within an end-to-end framework, but the baselines above are for either the AM or ASC sub-task. While performing ASC, the testing datasets in $\mathbf{D}^{\mathbf{ASC}}_{\mathbf{laptop}}$ and $\mathbf{D}^{\mathbf{ASC}}_{\mathbf{rest}}$ show the golden aspects. In order to achieve complete ABSA, these aspects labels are removed from the testing datasets, denoted as $\mathbf{D}^{\mathbf{ABSA}}_{\mathbf{laptop}}$ and $\mathbf{D}^{\mathbf{ABSA}}_{\mathbf{rest}}$, correspondingly. We compare SEML with the following baselines on the new testing datasets:

- **DE-CNN-MGAN** is the pipeline method which combines two state-of-the-art methods DE-CNN[1] for AM and MAGAN[2] for ASC.
- **LM-LSTM-CRF** [51] is a competitive model on some sequence labeling tasks in NLP. We train the model[3] for complete ABSA in a collapsed labeling scheme.
- **E2E-TBSA** [29] is the state-of-the-art supervised model to perform complete ABSA in an unified framework with a collapsed labeling scheme.

3) TRAINING SETTINGS

We use pre-trained GloVe 840B 300-dimension vectors [41] and refine the sentiment vectors [42] to initialize the word embeddings, and the char-feature size is 50. All of the weight matrices except those in BiMAGRU are initialized from the uniform distribution $U(-0.2, 0.2)$. For the initialization of the matrices in BiMAGRU, we adopt the Glorot Uniform strategy [52]. We apply dropout and the rates are set as 0.5 for labeled reviews and 0.8 for unlabeled reviews. The hidden state size is set to 1,024, and the learning rate is 0.05. We set the mini-batch size as 30 sentences, and the moving-window size (i.e., the number of cached past nearby hidden states in MAGRU) N is 5.

B. EXPERIMENTAL RESULTS

We report the results of CMLA, MIN, DE-CNN, EMOVA, RAM, TNet and MGAN in their original works, since we use exactly the same datasets. For the other models, we average the evaluation results of five runs. We follow the standard evaluation metrics of SemEval workshops to report the F1 score for AM and the accuracy and Macro-F1 (MF1) score for ASC.

1) MAIN RESULTS

Results on AM. TABLE 3 depicts the results of all the evaluated models for AM, in which SEML performs the best.

TABLE 3. Comparison results on F1 for AM.

		D^{AM}_{laptop}	D^{AM}_{rest}
	Models	**F1**	**F1**
1	**CMLA**	77.80	72.77
	MIN	77.58	73.44
	DE-CNN	81.59	74.37
	EMOVA	81.72	75.18
	BERT	78.70	73.23
2	**SEML-SUP**	78.53	73.91
	SEML-GNL	79.62	74.21
	SEML-AM	81.86	75.28
	SEML	**83.37**	**78.24**

[1]https://github.com/howardhsu/DE-CNN
[2]https://github.com/hsqmlzno1/MGAN
[3]https://github.com/LiyuanLucasLiu/LM-LSTM-CRF

For example, compared to the competitive models including CMLA, MIN, DE-CNN and EMOVA, SEML achieves absolute gains of 5.57%, 5.79%, 1.78% and 1.65% on $\mathbf{D^{AM}_{laptop}}$, and 5.47%, 4.80%, 3.87% and 3.06% on $\mathbf{D^{AM}_{rest}}$, respectively. Even our pure supervised SEML-SUP (without CVT) can perform better than CMLA and MIN. The main reason is the effectiveness of MAGRU which can derive the significant information of nearby contexts of the aspects. Moreover, SEML-GNL with general unlabeled texts improves SEML-SUP, which verifies the advantage of semi-supervised learning. While comparing to the two-phase semi-supervised approaches including DE-CNN and BERT, SEML shows the great superiority; the two-phase training (i.e., pre-training and supervised learning) cannot take advantage of labeled reviews for learning representations in the pre-training step; however, SEML learns domain- and task-specific representations alternately over labeled and unlabeled reviews within an unified end-to-end framework. Finally, EMOVA also employs CVT but only performs the single AM task, so SEML records better results than EMOVA by enabling the multi-task learning.

Results on ASC. TABLE 4 depicts the results of all the evaluated models for ASC, where SEML also achieves the best accuracy and MF1. More specifically, SEML-ASC, i.e., the variant of SEML for the single ASC task already outperforms all the supervised models including RAM, TNet and MGAN, which shows that semi-supervised learning can improve the prediction performance by taking full advantage of unlabeled reviews. Interestingly, BERT gives a slightly better accuracy (0.03%) than SEML-ASC on $\mathbf{D^{ASC}_{rest}}$, our explanation is that BERT learns representations by training on much more domain-free texts than SEML-ASC and the ASC sub-task is more domain-independent than the AM sub-task, i.e., aspect words are more dependent on domains than sentiment words. Fortunately, while performing multi-task learning, the shared representations in SEML can get significantly improved and then enhance the final prediction results.

TABLE 4. Comparison results on Accuracy and MF1 for ASC.

		D^{ASC}_{laptop}		D^{ASC}_{rest}	
	Models	**Acc.**	**MF1**	**Acc.**	**MF1**
1	**RAM**	74.49	71.35	80.23	70.80
	TNet	76.54	71.75	80.69	71.27
	MGAN	76.21	71.42	81.49	71.48
	BERT	75.19	70.76	81.67	72.01
2	**SEML-SUP**	76.01	71.32	81.33	71.70
	SEML-GNL	76.63	71.58	81.52	71.91
	SEML-ASC	76.87	71.83	81.64	72.10
	SEML	**77.54**	**73.02**	**83.21**	**72.76**

Results on ABSA. FIGURE 3 reports the F1 score for ABSA based on the exact match, i.e., a joint labeling result is considered to be correct only if it matches with both aspect and sentiment labels. SEML obtains consistent improvement

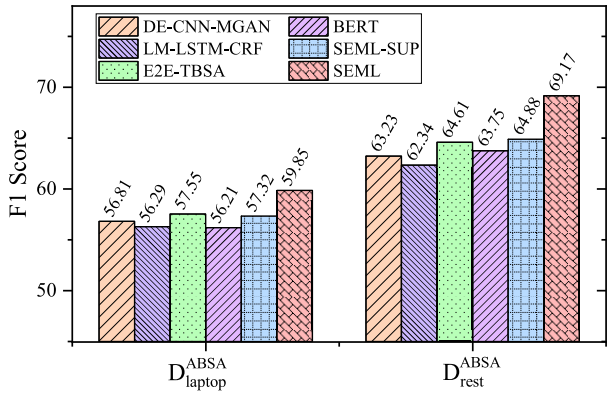

FIGURE 3. Comparison results on F1 for ABSA. Note that the testing datasets D^{ABSA}_{laptop} and D^{ABSA}_{rest} are the same as in D^{ASC}_{laptop} and D^{ASC}_{rest} but without the aspect labels.

over the pipeline model (DE-CNN-MGAN) and unified models (LM-LSTM-CRF and E2E-TBSA). The reason is that SEML leverages a more integrated way for multi-task learning for highly coupled tasks (e.g., AM and ASC) than the pipeline model. Further, compared to the unified models with a collapsed labeling scheme, SEML also shows the effectiveness of a joint model that considers the interaction between two related sub-tasks in ABSA.

2) ABLATION STUDY

The key components of SEML include char-features, refined word embeddings and auxiliary prediction modules, as shown in FIGURE 1. To show the significance of each key component, we disable each of them and evaluate the F1 score for AM and MF1 for ASC, as depicted in TABLE 5. Firstly, we disable the char-features and the result shows only slight effect in the row for **w/o char-features**. Then, we do not refine the word embedding with sentiment lexicon before training, the result drops slightly for AM but drops more for ASC in the row for **w/o refining**, which shows the essentiality of word embedding refining for the sentiment-related task. To explore which auxiliary prediction modules are more important, we only enable two of them (p_{fwd} and p_{bwd}, or p_{left} and p_{right}) at each time. We find that SEML **w/o fwd & bwd** that do not see the current word is better than SEML **w/o left & right**, which may be caused by the more restricted view on the unlabeled input.

TABLE 5. Ablation study on the key components of SEML.

	AM (F1)		ASC (MF1)	
Models	D^{AM}_{laptop}	D^{AM}_{rest}	D^{ASC}_{laptop}	D^{ASC}_{rest}
SEML	**83.37**	**77.04**	**73.02**	**72.76**
w/o char-features	-0.06	-0.06	-0.04	-0.05
w/o refining	-0.09	-0.06	-0.27	-0.20
w/o fwd & bwd	-0.32	-0.27	-0.19	-0.21
w/o left & right	-0.43	-0.60	-0.45	-0.52

3) VISUALIZATION OF MOVING-WINDOW ATTENTION

We use an example to visualize the significance score in Equation (4) for the moving-window attention with the window size $N = 5$. FIGURE 4 shows the visualization results in the second BiMAGRU for AM and the third BiMAGRU layer for ASC. SEML pays more attention on "software" and "system" to identify the aspect label of "preloaded", and greatly attends on "not" and "slow" to predict the sentiment polarity of "preloaded".

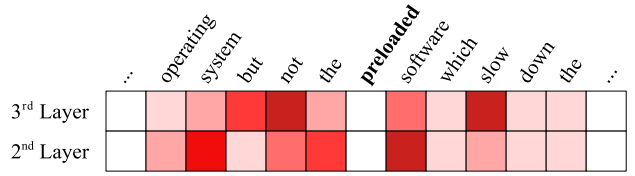

FIGURE 4. An example on the moving-window attention for "preloaded".

4) EFFECTS OF MOVING-WINDOW SIZE

We also evaluated the effects of the size of moving-window in the MAGRU of our SEML framework, the results are shown in FIGURE 5. It is hard to improve the overall performance by simply increasing the moving-window size, i.e., SEML can achieve better AM and ASC accuracy by focusing attention on a certain number of nearby words. To reduce the computation cost, the moving-window size N is set to 5 in our experiments.

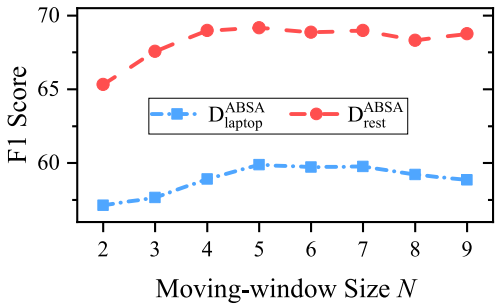

FIGURE 5. Effects of the moving-window size N.

5) EFFECTS OF MODEL SIZE

Most supervised models for ABSA use RNNs (e.g., LSTM and GRU) with small hidden state sizes around 300 [12], [13], [29], as a larger hidden state size may not surely improve the performance of supervised model [53]. We exam the effects of the hidden state size on our semi-supervised SEML and supervised SEML-SUP. FIGURE 6 shows that SEML-SUP without CVT also do not gain much from having a larger model size. However, as SEML can learn from unlabeled reviews by using CVT, the performance benefits from the increase of the model size. As the consequence, SEML enables the development of larger and more accurate models for the domain with limited amounts of labeled reviews but large numbers of unlabeled reviews, by using a large model size, e.g., 1,024 in our previous experiments.

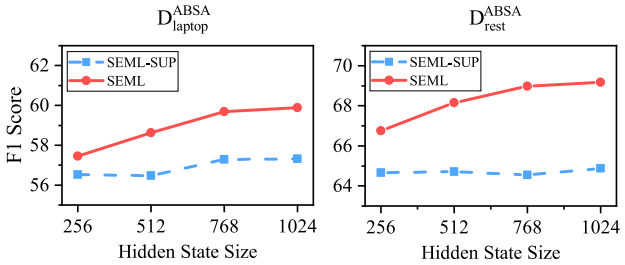

FIGURE 6. Effects of the model (hidden state) size.

6) LESS LABELED TRAINING DATA

A very common situation in aspect mining is some domains (or products) may not have large volumes of labeled data. To this end, we explore how SEML scales with less data by only feeding a subset (25%, 50%, and 75%) of the labeled training datasets, as presented in FIGURE 7. SEML with half of the training data can perform as well as SEML-SUP without CVT that sees all the training data. Thus, SEML is particularly useful when only a small set of labeled reviews is available, which greatly reduces the cost on manual labeling.

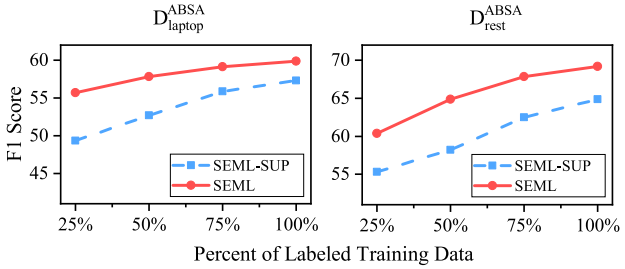

FIGURE 7. Performance vs. percent of the labeled training set.

V. CONCLUSION AND FUTURE WORK

In this paper, we have proposed the first end-to-end SEmi-supervised Multi-task Learning framework (SEML) for ABSA on customer reviews. The two related sub-tasks, namely AM and ASC in ABSA are jointly learned in an end-to-end fashion. Moreover, SEML derives the shared representations of reviews based on three stacked and bidirectional neural layers with Moving-window Attentive Gated Recurrent Units (MAGRU); MAGRU extends GRU with the moving-window attention mechanism to capture significant nearby semantic contexts. Further, SEML employs CVT to train auxiliary prediction modules on unlabeled reviews to improve the representation learning in an unified end-to-end architecture. Finally, we have conducted experiments for AM and ASC sub-tasks as well as complete ABSA over four datasets from the SemEval workshops and the experimental results show that SEML significantly outperforms the state-of-the-art models, even on much smaller labeled training datasets.

We consider two future research directions. First, as SEML directly delivers hidden representations between sub-tasks that may bring inconsistency of AM and ASC results (e.g., the ASC predictor may label sentiment polarities on non-aspect words), we will design more constraints to enforce stronger consistency between two sub-tasks in the future. Second, in addition to labeled and unlabeled reviews, we will try to encode linguistic knowledge (e.g., commonsense knowledge bases) into the framework to improve the performance.

REFERENCES

[1] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.
[2] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar. 2016.
[3] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in *Proc. 4th ACM Int. Conf. Web Search Data Mining WSDM*, 2011, pp. 815–824.
[4] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2010, pp. 56–65.
[5] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "An unsupervised neural attention model for aspect extraction," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2017, pp. 388–397.
[6] W. Jin and H. H. Ho, "A novel lexicalized HMM-based learning framework for Web opinion mining," in *Proc. 26th Annu. Int. Conf. Mach. Learn. ICML*, 2009, pp. 465–472.
[7] N. Jakob and I. Gurevych, "Extracting opinion targets in a single-and cross-domain setting with conditional random fields," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2010, pp. 1035–1045.
[8] Z. Toh and W. Wang, "DLIREC: Aspect term extraction and term polarity classification system," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 235–240.
[9] X. Li and W. Lam, "Deep multi-task learning for aspect term extraction with memory interaction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2886–2892.
[10] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowl.-Based Syst.*, vol. 108, pp. 42–49, Sep. 2016.
[11] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Recursive neural conditional random fields for aspect-based sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 616–626.
[12] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Coupled multi-layer attentions for co-extraction of aspect and opinion terms," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 3316–3322.
[13] X. Li, L. Bing, P. Li, W. Lam, and Z. Yang, "Aspect term extraction with history attention and selective transformation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 4194–4200.
[14] Z. Chen and B. Liu, "Topic modeling using topics from many domains, lifelong learning and big data," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 703–711.
[15] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting domain knowledge in aspect extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1655–1667.
[16] A. Mukherjee and B. Liu, "Aspect extraction through semi-supervised modeling," in *Proc. Annu. Meeting Assoc. for Comput. Linguistics*, 2012, pp. 339–348.
[17] N. Li, C.-Y. Chow, and J.-D. Zhang, "Seeded-BTM: Enabling biterm topic model with seeds for product aspect mining," in *Proc. IEEE 21st Int. Conf. High Perform. Comput. Commun., IEEE 17th Int. Conf. Smart City; IEEE 5th Int. Conf. Data Sci. Syst. (HPCC/SmartCity/DSS)*, Aug. 2019, pp. 2751–2758.
[18] H. Xu, B. Liu, L. Shu, and P. S. Yu, "Double embeddings and CNN-based sequence labeling for aspect extraction," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Short Papers)*, vol. 2, 2018, pp. 592–598.
[19] N. Li, C.-Y. Chow, and J.-D. Zhang, "EMOVA: A semi-supervised end-to-end moving-window attentive framework for aspect mining," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, 2020, pp. 811–823.
[20] R. Y. K. Lau, C. Li, and S. S. Y. Liao, "Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis," *Decis. Support Syst.*, vol. 65, pp. 80–94, Sep. 2014.
[21] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 606–615.

[22] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 214–224.
[23] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," in *Proc. Int. Conf. Comput. Linguistics*, 2016, pp. 3298–3307.
[24] X. Li, L. Bing, W. Lam, and B. Shi, "Transformation networks for target-oriented sentiment classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2018, pp. 946–956.
[25] Z. Li, Y. Wei, Y. Zhang, X. Zhang, and X. Li, "Exploiting coarse-to-fine task transfer for aspect-level sentiment classification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 4253–4260.
[26] J. R. Finkel and C. D. Manning, "Joint parsing and named entity recognition," in *Proc. Hum. Lang. Technol., Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics - NAACL*, 2009, pp. 326–334.
[27] M. Miwa and Y. Sasaki, "Modeling joint entity and relation extraction with table representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1858–1869.
[28] Y. Ma, H. Peng, T. Khan, E. Cambria, and A. Hussain, "Sentic LSTM: A hybrid network for targeted aspect-based sentiment analysis," *Cognit. Comput.*, vol. 10, no. 4, pp. 639–650, Aug. 2018.
[29] X. Li, L. Bing, P. Li, and W. Lam, "A unified model for opinion target extraction and target sentiment prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6714–6721.
[30] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "An interactive multi-task learning network for end-to-end aspect-based sentiment analysis," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 504–515.
[31] A. Hussain and E. Cambria, "Semi-supervised learning for big social data analysis," *Neurocomputing*, vol. 275, pp. 1662–1673, Jan. 2018.
[32] K. Clark, M.-T. Luong, C. D. Manning, and Q. Le, "Semi-supervised sequence modeling with cross-view training," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1914–1925.
[33] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.
[34] E. Cambria, J. Fu, F. Bisio, and S. Poria, "Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 508–514.
[35] E. Cambria, S. Poria, R. Bajpai, and B. Schuller, "Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives," in *Proc. Int. Conf. Comput. Linguistics*, 2016, pp. 2666–2677.
[36] Z. Miao, Y. Li, X. Wang, and W.-C. Tan, "Snippext: Semi-supervised opinion mining with augmented data," in *Proc. Web Conf.*, Apr. 2020, pp. 617–628.
[37] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J. T. Zhou, and S. Yang, "Structured AutoEncoders for subspace clustering," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5076–5086, Oct. 2018.
[38] X. Peng, Z. Huang, J. Lv, H. Zhu, and J. T. Zhou, "Comic: Multi-view clustering without parameter selection," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5092–5101.
[39] S. Hochreiter and J. J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1980, 1997.
[40] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. Neural Inf. Process. Syst. Workshop Deep Learn.*, 2014, pp. 2–10.
[41] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[42] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining word embeddings for sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 534–539.
[43] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2016, pp. 1064–1074.
[44] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, and V. Hoste, "Semeval-2016 task 5: Aspect based sentiment analysis," in *Proc. Int. Workshop Semantic Eval.*, 2016, pp. 19–30.
[45] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 task 4: Aspect based sentiment analysis," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 27–35.
[46] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proc. 25th Int. Conf. World Wide Web WWW*, 2016, pp. 507–517.
[47] Y. Dataset. *Yelp Dataset Challenge*. Accessed: Mar. 5, 2019. [Online]. Available: https://www.yelp.com/dataset/challenge
[48] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson, "One billion word benchmark for measuring progress in statistical language modeling," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 2635–2639.
[49] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 452–461.
[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2018, pp. 4171–4186.
[51] L. Liu, J. Shang, X. Ren, F. F. Xu, H. Gui, J. Peng, and J. Han, "Empower sequence labeling with task-aware neural language model," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5253–5260.
[52] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statistic*, 2010, pp. 249–256.
[53] N. Reimers and I. Gurevych, "Optimal hyperparameters for deep LSTM-networks for sequence labeling tasks," 2017, *arXiv:1707.06799*. [Online]. Available: http://arxiv.org/abs/1707.06799

NING LI received the M.Sc. degree in computing and security from the King's College London, London, U.K., in 2011. He is currently pursuing the Ph.D. degree with the Department of Computer Science, City University of Hong Kong, Hong Kong. His research interests include text mining, sentiment analysis, deep learning, and recommender systems. He received the Best Paper Award in IEEE DSS 2019.

CHI-YIN CHOW (Senior Member, IEEE) received the M.S. and Ph.D. degrees from the University of Minnesota, Twin Cities, MN, USA, in 2008 and 2010, respectively. He is currently an Associate Professor with the Department of Computer Science, City University of Hong Kong. His research interests include big data analytics, data management, GIS, mobile computing, location-based services, and data privacy. He was a recipient of the VLDB 10-year Award in 2016. He received best paper awards at ICA3PP 2015 and the IEEE MDM 2009. From 2012 to 2016, he was a Co-Founder and a Co-Chair of the ACM SIGSPATIAL MobiGIS. He is an Editor of the ACM SIGSPATIAL Newsletter.

JIA-DONG ZHANG (Member, IEEE) received the M.Sc. degree from Yunnan University, China, in 2009, and the Ph.D. degree from the City University of Hong Kong, in 2015. He is currently a Research Fellow with the Department of Computer Science, City University of Hong Kong. His research has been published in premier conferences, including the ACM SIGIR, CIKM, and SIGSPATIAL, in transactions, including the *ACM TIST*, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, the IEEE TRANSACTIONS ON SERVICES COMPUTING, and the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, and in journals, including IEEE ACCESS, *Pattern Recognition*, and *Information Sciences*. His research interests include deep learning, data mining, recommender systems, and location-based services.