DNN-driven Gradual Machine Learning for Aspect-Term Sentiment Analysis

Murtadha Ahmed, Qun Chen, Yanyan Wang, Youcef Nafa, Zhanhuai Li and Tianyi Duan
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
{a.murtadha@mail.,chenbenben@,wangyanyan@mail., youcef.nafa@mail., lizhh@,tianyiduan@mail.}nwpu.edu.cn

Abstract

Recent work has shown that Aspect-Term Sentiment Analysis (ATSA) can be performed by Gradual Machine Learning (GML), which begins with some automatically labeled easy instances, and then gradually labels more challenging instances by iterative factor graph inference without manual intervention. As a non-i.i.d learning paradigm, GML leverages shared features between labeled and unlabeled instances for knowledge conveyance. However, the existing GML solution extracts sentiment features based on pre-specified lexicons, which are usually inaccurate and incomplete and thus lead to inadequate knowledge conveyance.

In this paper, we propose a Deep Neural Network (DNN) driven GML approach for ATSA, which exploits the power of DNN in feature representation for gradual learning. It first uses an unsupervised neural network to cluster the automatically extracted features by their sentiment orientation. Then, it models the clustered features as factors to enable implicit knowledge conveyance for gradual inference in a factor graph. To leverage labeled training data, we also present a hybrid solution that fulfills gradual learning by fusing the influence of supervised DNN predictions and implicit knowledge conveyance in a unified factor graph. Finally, we empirically evaluate the performance of the proposed approach on real benchmark data. Our extensive experiments have shown that the proposed approach consistently achieves the state-of-the-art performance across all the test datasets in both unsupervised and supervised settings and the improvement margins are considerable.

1 Introduction

Aspect-Term Sentiment Analysis (ATSA) aims at inferring the sentiment polarity towards a particular aspect in a sentence (Hu and Liu, 2004; Pontiki et al., 2016). ATSA is important for many applications (e.g., e-commerce and social media), where the sentimental opinions in reviews can be leveraged to create value for businesses and customers. In ATSA, an aspect-term, also called target, is explicitly mentioned in a review. For instance, consider the running example shown in Table 1, r_1 evaluates the restaurant through two explicit aspects *service* and *staff*. The goal of ATSA is then to detect the respective sentiment polarities expressed towards these two aspects.

r_i	s_{ij}	text
r_1	$s_{1.1}$	**service** was awful mostly because **staff** were overwhelmed.
	$s_{1.2}$	The **staff** should be a bit more friendly.
r_2	$s_{2.1}$	We ordered **lamb** which was perfectly cooked and tasted awesome.
	$s_{2.2}$	The **food** was well-prepared and presented.

Table 1: A running example: r_i denotes a review and $s_{i.j}$ denotes a sentence.

Up to now, the state-of-the-art solutions for ATSA have been built upon various DNN models. The earlier solutions were usually equipped with an attention mechanism (Tang et al., 2016b; Wang et al., 2016; Tang et al., 2016a; Ma et al., 2017; Chen et al., 2017; Li et al., 2018; Wang et al., 2018; Tang et al., 2019). They mostly attempted to learn aspect-related semantic representation of an input sentence. Recently, ATSA has experienced a considerable shift towards pre-trained language models (Sun et al., 2019; Tang et al., 2019; Karimi et al., 2020). Despite the effectiveness of these approaches, unfortunately their efficacy heavily relies on large quantities of accurately labeled data, which require intensive human labor.

To alleviate the burden of manual labeling, a solution based on the paradigm of Gradual Machine Learning (GML) has recently been proposed for

ATSA (Wang et al., 2021). First proposed for entity resolution in (Hou et al., 2019, 2020), GML can enable effective machine labeling without the requirement for manual intervention. Given a classification task, GML begins with some easy instances, which can usually be automatically labeled by the machine with high accuracy, and then gradually reasons about the labels of its more challenging instances by factor graph inference. As a non-i.i.d (Independent and Identically Distributed) learning paradigm, GML leverages shared features between labeled and unlabeled instances for knowledge conveyance . However, the existing GML solution for ATSA relies on pre-specified lexicons to extract sentiment features. Its limitation is twofold: 1) sentiment lexicons may be inaccurate and incomplete; 2) a shared feature must explicitly appear in both instances. However, explicit features cannot capture the implicit similarity between instances and thus lead to inadequate knowledge conveyance. Consider the running example in Table 1. Unfortunately the word *well-prepared* is not included in most of the existing lexicons. It can also be observed that the instances $s_{2.1}$ and $s_{2.2}$ do not share any explicit feature, while *perfectly cooked* and *well-prepared* have very similar meanings and can thus serve as an implicit common feature.

Recently, DNN models have been proven to be very powerful in feature representation for many NLP tasks, where the features with the same semantic context are mapped to close points in the latent space (Devlin et al., 2018). For instance, the words *"cooked"* and *"well-prepared"* are usually represented by two points close to each other because they are semantically very close. Unfortunately, the existing embedding models are designed to map the features semantically, regardless of their sentiment orientation. Therefore, they may map two features with opposite polarities (e.g., *"good"* and *"bad"*) to two close points in the embedding space, which raises a challenge to be directly applied to feature extraction for ATSA.

In this paper, we propose a novel DNN-driven GML approach for ATSA. It essentially exploits DNN to sentimentally map the features of aspect-terms into different *polarity indicators*, and models them as shared factors in a factor inference graph to enable implicit knowledge conveyance. To this end, we first combine the sentiment lexicon and dependency parser-based relations, which are readily available, to generate aspect-opinion words. Secondly, we use an unsupervised neural network to filter the aspect-irrelevant and unsentimental words from an input sentence. Finally, the resulting weighted sentences, which can be considered to be purely sentimental, are used to learn polarity indicators. The model is trained to reconstruct the weighted sentence through a linear combination from polarity indicators. To leverage labeled training data, we also present a hybrid GML solution that fulfills gradual learning by fusing the influence of supervised DNN predictions and implicit knowledge conveyance in a unified factor graph.

Our main contributions can be summarized as follows:

1. We propose a DNN-driven GML approach for ATSA, which can effectively exploit the power of DNN in feature representation for GML;

2. We present an unsupervised attention-based neural network to cluster the features of aspect-terms by their sentimental orientation;

3. We present a hybrid GML solution for ATSA, which fulfills gradual learning by fusing the influence of supervised DNN predictions and implicit knowledge conveyance in a unified factor graph.

4. We empirically validate the efficacy of the proposed approach on benchmark data. Our extensive experiments have shown that the proposed approach consistently achieves the state-of-the-art performance across all the test datasets in both unsupervised and supervised settings and the improvement margins are considerable.

2 Related work

Aspect-Term Sentiment Analysis is a sub-task of aspect-based sentiment analysis, which aims to detect the sentiment polarity in response to a particular aspect in a sentence (Hu and Liu, 2004; Pontiki et al., 2016). Traditional machine learning techniques (Kiritchenko et al., 2014; Castellucci et al., 2014) proposed to learn SVM classifiers based on different features such as n-grams, non-contiguous n-grams and lexicon features. In comparison, the DNN-based models equipped with an attention mechanism have been shown to be more effective on ATSA (Tang et al., 2016b; Wang

et al., 2016; Tang et al., 2016a). Following this trend, researchers have resorted to more sophisticated attention mechanisms to refine neural ATSA models (Ma et al., 2017; Chen et al., 2017; Li et al., 2018; Wang et al., 2018; Tang et al., 2019). To improve performance, they essentially attempted to explicitly capture the importance of each context word by learning aspect-related representation of an input sentence. SenHint (Wang et al., 2019) proposed to integrate DNN predictions and linguistic hints in a joint framework. Recently, ATSA has experienced a considerable shift towards pre-trained language models (Sun et al., 2019; Tang et al., 2019; Karimi et al., 2020). Unfortunately, the efficacy of these models heavily relies on labeled training data, which may not be readily available in real-scenario.

From unsupervised perspective, earlier solutions (Alvarez-López et al., 2016; Hutto and Gilbert, 2014) proposed to detect the polarities of aspect-terms based on lexicon rules. The authors of (Schouten et al., 2017) proposed a mechanism of spread activation for aspect-based polarity detection. More recently, the authors of (Wang et al., 2021) proposed an unsupervised solution based on GML for ATSA. However, the existing GML solution extracts features based on sentiment lexicons, which may not be accurate nor complete and thus lead to inadequate knowledge conveyance.

The idea of mapping features into different clusters has been investigated with different purposes. The authors of (Iyyer et al., 2016) proposed to learn a set of descriptors representing the fictional relationship between two characters changes over time, and (He et al., 2017) proposed to learn a set of aspect representatives from the corpora. Unfortunately, none of them investigated how to cluster implicit features by their polarity orientation.

3 Preliminaries

3.1 Task Definition

We formulate the task of aspect-term sentiment analysis as follows:

Definition 3.1 *[Aspect-Term Sentiment Analysis]. Let $x = (r, s, t)$ be a target unit, where r denotes a review, s a sentence in the review and t an aspect-term associated with the sentence. Given a set of target units, X, the goal of ATSA is to infer the sentiment polarity of each target unit in X.*

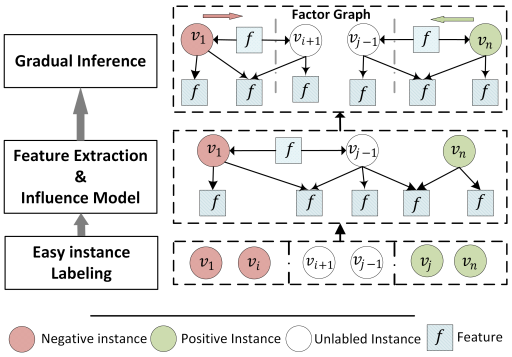

Figure 1: GML Paradigm Overview

3.2 GML Paradigm Overview

Our solution is built upon the non-i.i.d learning paradigm of GML (Hou et al., 2019, 2020). As shown in Figure 1, GML consists of the following three steps:

3.2.1 Easy Instance Labeling.

Given a classification task, it is usually very challenging to accurately label all the instances in the task without good-coverage training examples. However, the work can become much easier if we only need to automatically label some easy instances in the task. In real scenarios, easy instance labeling can be performed based on the simple user-specified rules or the existing unsupervised learning techniques. GML begins with the observations provided by the labels of easy instances. Therefore, high accuracy of automatic machine labeling on easy instances is critical for its ultimate performance on a given task.

For ATSA, this paper uses the unsupervised algorithm of spread activation (Schouten et al., 2017) to label easy instances. An instance is considered as easy if its resulting dominate label meets a pre-specified threshold.

3.2.2 Feature Extraction and Influence Modeling.

Feature serves as the medium for knowledge conveyance. This step extracts the common features shared by labeled and unlabeled instances. To facilitate effective knowledge conveyance, it is desirable that a wide variety of features are extracted to capture as much information as possible. For each extracted feature, this step also needs to model its influence over the labels of its relevant instances.

For ATSA, we extract two types of features: sentimental feature and relational feature. Relational feature, which has been well studied in (Wang et al.,

2021), represents the explicit sentimental connection between sentences within the same review. In the running example, due to the absence of any shift word between s_{11} and s_{12}, their polarities can be supposed to be similar. In this paper, we focus on how to enable implicit knowledge conveyance by leveraging DNN for automatic extraction of sentimental features.

3.2.3 Gradual Inference.

This step gradually labels the instances with increasing hardness in a task. Since the scenario of gradual learning does not satisfy the i.i.d assumption, gradual learning is fulfilled from the perspective of evidential certainty. Gradual learning is conducted over a factor graph, which consists of the labeled and unlabeled instances and their common features, by iterative inference. At each iteration, it chooses to label the unlabeled instance with the highest degree of evidential certainty. The iteration is repeatedly invoked until all the instances in a task are labeled.

Given a factor graph, G, GML defines the probability distribution over its variables V as follows:

$$P_w(V) = \frac{1}{Z_w} \prod_{v \in V} \prod_{f \in F_v} \phi_f(v) \prod_{f' \in F'} \phi_{f'}(v_i, v_j), \tag{1}$$

where F_v denotes the set of sentimental features associated with the variable v, F' denotes the set of relational features, $\phi_f(v)$ denotes the factor associated with v and f, $\phi_{f'}(v_i, v_j)$ denotes the factor associated with the relational feature f', and Z is a partition function, i.e. normalization constant. To effectively learn the factor weights without access to the true labels of unlabeled variables, V_I, GML minimizes the negative log marginal likelihood given the observed labels of labeled variables, Λ, as follows:

$$\hat{w} = arg \min_{w} -log \sum_{V_I} P_w(\Lambda, V_I). \tag{2}$$

A scalable approach for gradual inference on ATSA has been presented in (Wang et al., 2021). First, the unlabeled variables are sorted according to their evidential support. Then, the top-m unlabeled variables are considered as the candidates for probability inference. To reduce the invocation frequency of factor graph inference, an efficient algorithm is used to approximate entropy estimation on m candidates and select the top-k most promising variables for factor graph inference. Finally, the probabilities of the selected k variables are inferred in the subgraphs of G. Since the inference process of the DNN-driven GML is very similar to what was presented in (Wang et al., 2021), its technical details are omitted here due to space limit.

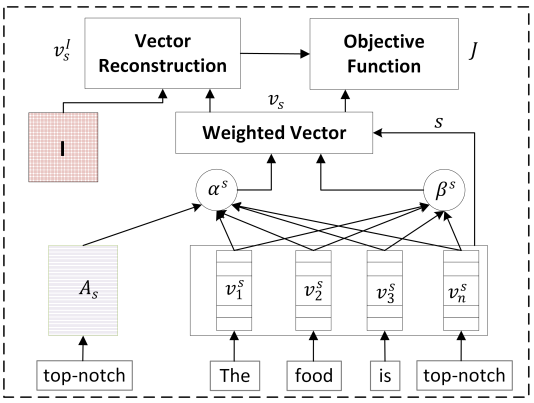

Figure 2: DNN for Implicit Feature Extraction.

4 DNN-driven GML

In this section, we first present an unsupervised neural network to extract implicit sentimental features. Then, we describe the unsupervised DNN-driven GML that integrates implicit features into the process of gradual inference. Finally, we describe the hybrid GML solution that fuses the influence of DNN predictions and implicit features for gradual learning.

4.1 Implicit Feature Extraction by DNN

The purpose of implicit feature extraction is to learn a set of polarity indicator embeddings $\mathrm{I} \in \mathbb{R}^{k \times d}$, where k is the number of indicators, which can be leveraged to capture the similar features between instances. Each indicator represents a set of features that often occur in the contexts with the same polarity.

Specifically, for each input sentence s and its aspect term t, we first generate a set of aspect-opinion words, denoted by A_s. Then, we use A_s to construct a weighted vector $v_s \in \mathbb{R}^d$ that can be read as the sentimental representation of the input sentence given the target of t. To this end, we propose an attention-based unsupervised neural network to filter the sentence by down-weighting aspect-irrelevant and unsentimental information. The model is trained by reconstructing v_s as a linear combination of indicator embeddings from I. The architecture of the proposed DNN has been presented in Figure 2.

4.1.1 The Input

The input to our model is a couple of sentence s and its aspect-opinion words A_s. We use dependency-based parse tree to generate aspect-opinion words (*modifiers*) (Hu and Liu, 2004), then leverage the adjective words and those detected by the lexicon to construct A_s. Considering the running example in Table 1, the sentiment words of s_{11} are *awful* and *overwhelmed*. Suppose that we have a feature embedding matrix $\mathrm{L} \in \mathbb{R}^{c \times d}$, where c is the vocabulary size and d is the embedding dimension. Each word is then associated with a real-valued embedding $v_i \in \mathbb{R}^d$ from L representing its feature vector (Mikolov et al., 2013):

$$s = \{v_1^s, v_2^s, ..., v_n^s\}, \tag{3}$$

$$A_s = \{v_1^a, v_2^a, ..., v_m^a\}, \tag{4}$$

where $s \in \mathbb{R}^{n \times d}$ and n is the sentence length, while the input sentiment $A_s \in \mathbb{R}^{m \times d}$ and m is the number of aspect-opinion words.

4.1.2 Attention-based Sentimental Representation

For each input sentence s, we construct a weighted vector v_s to capture the sentimental information in response to the aspect t. To this end, we apply two attention mechanisms to filter away the irrelevant information. The first one attempts to down-weight non-sentimental words, while the second one is a self-attention to attend to aspect-relevant information (He et al., 2017).

Specifically, the first attention layer takes both the sentence s and its opinion words A_s as an input. Conceptually, we first compute the global sentiment vector v^a by averaging the word embeddings of A_s, and then use it to weight each word embedding v_i^s in s as follows:

$$v^a = \frac{1}{m} \sum_{v_i^a \in A_s} v_i^a, \tag{5}$$

$$o_i = {v_i^s}^\top \cdot \mathrm{U} \cdot v^a, \tag{6}$$

where the symbol $\cdot$ stands for element-wise dot product, while $\mathrm{U} \in \mathbb{R}^{d \times d}$ is the transformation matrix (i.e., to be learned during training) between the global sentiment vector v^a and the input sentence s. Next, we apply a softmax layer to yield a non-negative weight for each word in s as follows:

$$\alpha_i^s = \frac{\exp(o_i)}{\sum_{j=1}^{n} \exp(o_j)}, \tag{7}$$

where the value of α_i^s can be read as the probability of each word in the sentence s being a sentiment word.

Although we have computed the sentimental importance for each word in s, but not all the sentiment words are contextually related to the aspect. Therefore, we apply another self-attention mechanism that takes only the sentence s as input. To compute each word's probability of being aspect-relevant information, namely β_i^s, we follow the same steps in the first attention layer. The only difference is that the global sentiment v^a in Equation 6 is replaced by the global context, which is simply computed by averaging the word embeddings of the input sentence s itself.

Finally, we sum both attention layer outputs, $\alpha^s \in \mathbb{R}^n$ and $\beta^s \in \mathbb{R}^n$, and use it to construct the weighted vector v_s as follows:

$$v_s = s^\top \cdot (\alpha^s + \beta^s), \tag{8}$$

in which the resulting weight vector v_s can be read as the aspect-relevant sentiment representation of the input sentence s.

4.1.3 Unsupervised Training

Now that we have obtained the aspect-relevant sentiment representation of an input sentence s, we explain how to learn its polarity indicators using a variant of dictionary learning. Considering the matrix of indicators I as a dictionary, we attempt to approximate v_s as a linear combination of items from I.

Formally, for each aspect-specific vector v_s, we compute a corresponding vector v_s^k over k polarity indicators by simply reducing v_s from d dimensions to k dimensions through a softmax layer as follows:

$$v_s^k = softmax(W \cdot v_s + b), \tag{9}$$

where $W \in \mathbb{R}^{k \times d}$ denotes a weight matrix and b denotes the bias, both of which are supposed to be learned during training. Note that v_s^k can be read as the probability that the input sentence s belongs to each indicator. Then, we reconstruct the representation vector by taking a weighted average over the polarity indicators as follows:

$$v_s^I = \mathrm{I}^\top \cdot v_s^k. \tag{10}$$

Since the objective is to make v_s^I similar to v_s, we apply the widely used contrastive max-margin

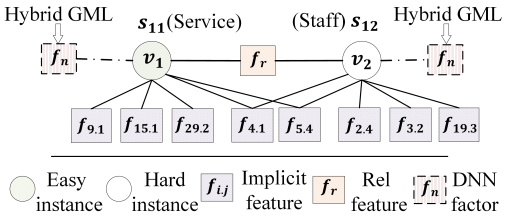

Figure 3: Factor graph of r_1 in the running example.

objective function (Iyyer et al., 2016; Ahmed et al., 2020). To that end, we randomly sample some sentences from the training data as negative samples, and compute the weighted vector v_s for each sample using Equation 8.

Specifically, the objective is a hinge loss that minimizes the inner product between the representation vector of v_s and the reconstructed vector v_s^I for the negative samples, while simultaneously maximizes their inner product for other samples in the training data. Formally, the hinge loss is defined as:

$$J(\theta) = \sum_{s \in D} \sum_{s^- \in D^-} max\{0, 1 - v_s^I \cdot v_s + v_s^I \cdot v_{s^-}\}, \tag{11}$$

where θ represents the model parameters, D represents the set of training data and D^- the subset of negative samples. Note that $\theta = \{I, U, W, b\}$.

To discourage the model from learning similar indicators, we add a regularization term to the objective function J that penalizes redundancy in the matrix of polarity indicators (Iyyer et al., 2016):

$$M(\theta) = ||\mathrm{I} \cdot \mathrm{I}^\top - \mathrm{Y}||, \tag{12}$$

where Y denotes the identity matrix. The final training objective L is then represented by the weighted sum of J and M as follows:

$$L(\theta) = J(\theta) + \lambda M(\theta), \tag{13}$$

where λ is a hyper-parameter that controls the magnitude of the regularization term.

4.2 Unsupervised GML Solution

Now that we have already learned a set of polarity indicators, we describe how they can serve as implicit features for gradual inference. Given a sentence, we first estimate its aspect-specific vector v_s using Equation 8, then compute the cosine similarity between v_s and each polarity indicator in I. It thus results in a list of scores in the form (*indicator index, similarity score*) representing how the aspect-term's features are close to each indicator. We sort the scores and use top-k corresponding indicators as representative features. We scale up the similarity score to 10 to augment the number of features and meanwhile avoid polarity conflict between features. As shown in Figure 3, $f_{5.4}$ represents an indicator feature with the index of 5 and the similarity scale of 4.

Considering the instance of s_{11} in Figur 3, s_{11}'s top-5 indicators and their similarity scores are (5, 0.44), (29, 0.2), (15, 0.16),(4, 0.13) and (9, 0.13). Then, its representative features are $F_{s_{11}} = \{f_{5.4}, f_{29.2}, f_{15.1}, f_{4.1}, f_{9.1}\}$. In gradual inference, we restrict two instances to share an implicit feature if and only if they are similar to the same indicator with the same score scale. For instance, in Figure 3, s_{11} and s_{12} share the indicators 4 and 5 with the same score scale of 1 and 4 respectively.

Given ATSA task, each aspect-term within the same review is represented by a variable. The evidence variables are assigned constant values 0 or 1 representing their polarity labeling, while the values of the inference ones are inferred based on G. The factor of an implicit feature f_e in Equation 1 is defined by:

$$\varphi_{f_e}(v_i) = \begin{cases} 1 & v_i = 0; \\ e^{w_{f_e}} & v_i = 1; \end{cases} \tag{14}$$

where v_i denotes a variable having the feature f_e, and w_{f_e} denotes the weight of f_e. Note that the weight w_{f_e} is initialized to zero, but needs to be learned in the process of gradual inference.

4.3 Hybrid GML Solution

In the hybrid solution, we model the influence of DNN outputs by DNN factors, denoted by f_n, as shown in Figure 3. In this paper, we have implemented the hybrid solution by the state-of-the-art BERT-based DNN of HP-SUM for ATSA (Karimi et al., 2020). However, other DNN models can be fused in the same way. Since supervised learning is usually more accurate than unsupervised learning, we also label easy instances by supervised DNN predictions. In other words, we consider the instances with the most extreme probabilities predicted by HP-SUM as easy ones to kick-start gradual inference.

In factor graph, the DNN factor f_n of a variable corresponding to the aspect-term unit, (r, s, t) is defined by:

$$\varphi_{f_n}(v_i) = \begin{cases} 1 & v_i = 0; \\ e^{w_{f_n}} & v_i = 1; \end{cases} \tag{15}$$

	Method	Rest 14	Rest 15	Rest 16	Lap 14	Lap 15	Lap 16
Unsupervised	VADER	79.65	76.21	75.18	69.72	74.31	68.31
	LEX-SYN	80.84	75.82	76.77	70.17	75.81	69.64
	SPD-ACT	81.89	76.02	81.06	74.84	77.15	73.77
	Lexicon GML	83.83	80.22	85.64	82.25	82.42	80.31
	DNN-driven GML	**87.05**	**81.19**	**86.31**	**85.84**	**84.05**	**81.62**
Attentive	ATAE-LSTM	88.56	76.72	81.19	79.45	77.11	73.13
	IAN	87.98	77.09	78.37	76.68	77.36	74.87
	RAM	90.0	76.26	87.72	81.87	80.61	76.27
	GCAE	88.55	78.64	87.87	80.81	80.82	80.83
	AEN-Glove	89.86	79.14	86.82	83.79	80.73	77.17
	TNet-LF	90.36	80.74	87.95	83.22	81.17	76.75
Bert-based	AEN-BERT	92.74	84.64	90.07	90.11	90.74	84.89
	BERT-SPC	93.76	83.49	91.72	90.62	88.74	86.64
	HP-SUM	93.39	88.29	94.76	93.6	90.34	87.85
	Hybrid GML	**95.51**	**88.67**	**95.6**	**94.91**	**91.74**	**88.88**

Table 2: Comparative Evaluation Results. Rest and Lap stand for Restaurant and Laptop domains respectively. The respective best accuracies in the unsupervised and supervised setting are highlighted in **bold**.

in which w_{f_n} denotes factor weight. The value of w_{f_n} is defined as

$$w_{f_n} = \ln(\frac{p}{1-p}), \tag{16}$$

where p is the probability output of DNN (i.e., estimated by HP-SUM) of a target t being positive in the sentence s. It can be observed that $W_{f_n} > 0$ if $p > 0.5$; otherwise, if $p < 0.5$, then $w_{f_n} < 0$.

5 Empirical Evaluation

We evaluate our solution on six benchmark datasets provided by the SemEval ABSA task across the years 2014, 2015 and 2016 for the Restaurant and Laptop domains (Pontiki et al., 2016). Note that the original datasets are three-way labels (i.e., positive, negative and neutral). Since this paper focuses on binary polarity classification, we only include the reviews with *positive* or *negative* labels in our experiments. Furthermore, we have trained the polarity indicators for the restaurant and laptop domains on unlabeled corpus collected from Citysearch and Amazon, which have also been widely used in previous work (Zhao et al., 2010; Ahmed et al., 2020).

For unsupervised training, we initialized word vectors by word2vec. We implemented GML inference using the Numbskull library [1], a Python NUMBA-based Gibbs sampler. Our GML implementation optimizes the parameters by Adam with the learning rate of 0.001. On all the test datasets, we set the number of polarity indicators k to 50, and the number of negative samples to 20. In the spread activation algorithm for easy instance labeling, the easiness threshold is set to 0.7 for all datasets. For the hybrid GML solution, the easy instances are the top-30% ones with most extreme probabilities as predicted by supervised DNN. For each instance, the associated implicit features are the top-5 polarity indicators' scores scaled to 10.

The compared unsupervised techniques include: (1) **LEX-SYN** (Alvarez-López et al., 2016). It infers polarity based on lexicon and syntactic dependency analysis; (2) **VADER** (Hutto and Gilbert, 2014). A rule-based approach; (3) **SPD-ACT** (Schouten et al., 2017). It infers polarity by spread activation; (4) **Lexicon-based GML** (Wang et al., 2021). The GML solution built upon sentiment lexicons.

The compared supervised DNN models include the latest BERT-based models as well as traditional attention-based models: (1) **ATAE-LSTM** (Wang et al., 2016). An attention-based LSTM; (2) **IAN** (Ma et al., 2017). An interactive attention model; (3) **RAM** (Chen et al., 2017). A deep memory model; (4) **TNet-LF** (Li et al., 2018). A target-specific transformation network; (5) **GCAE** (Xue and Li, 2018). A gated convolutional network; (6) **AEN** (Song et al., 2019). An attentional encoder network. AEN has two variants: AEN-Glove that uses Glove as feature embedding input, and

[1] https://github.com/HazyResearch/numbskull

AEN-BERT based on the pre-trained model BERT fine-tuning; (7) **BERT-SPC** (Song et al., 2019). A pseudo-sentence (i.e., sentence and aspect) BERT-based approach; (8) **HP-SUM** (Karimi et al., 2020). A BERT-based model equipped with parallel aggregation and hierarchical aggregation modules.

Note that among the listed DNN models, the last 3 models (i.e. AEN-BERT, BERT-SPC, and HP-SUM) were built upon the latest pre-trained BERT.

5.1 Main Results

We average the three runs' performances and report the detailed evaluation results in Table 2. We have the following observations: (1) **the unsupervised DNN-driven GML consistently gives the best accuracy compared to the unsupervised alternatives across all datasets.** The performance advantage of the DNN-driven GML over the lexicon-based GML suggests that a carefully-designed implicit feature mechanism can effectively perform better than lexicon-based explicit features for ATSA; (2) the unsupervised DNN-driven GML is even competitive with the traditional supervised attention-based models; (3) the supervised BERT-based approaches indeed achieve better performance than both traditional attention-based DNNs and unsupervised GML. However, their efficacy depends on the fine-tune phase that requires an access to the labeled training data, which are not available in the unsupervised setting; (4) **The hybrid GML solution consistently achieves the state-of-the-art performance across all datasets.** It improves the best performance by almost 2% on two datasets and 1%-2% on four out of six datasets. In light of the well recognized challenge of ATSA, these improvements are indeed considerable.

Illustrative Examples. To illustrate the effectiveness of implicit features, we present the features of the running example in Table 3. It can be observed that *overwhelmed*, *well-prepared*, and *presented* in r_{11} and r_{22} respectively are not captured by the lexicon, and r_{12} contains the context misunderstanding of *friendly*. Even though r_{11} and r_{12} do not share any explicit information, the negative context of *friendly* is very close to *overwhelmed*; they thus share the implicit features $f_{4.1}$ and $f_{5.4}$. Likewise, *well-prepared* in r_{22} is very close to *perfectly cooked* in r_{21}, and they share the implicit features $f_{7.4}$ and $f_{15.3}$.

r_{ij}	Features		GML Labeling	
	LEX	DDN-based	LEX	DNN
r_{11}	Awful	$\{f_{9.1}, f_{15.1}, f_{29.2}, f_{4.1}, f_{5.4}\}$	False	True
r_{12}	Friendly	$\{f_{2.4}, f_{3.2}, f_{4.1}, f_{5.4}, f_{19.3}\}$	False	True
r_{21}	Perfectly, Awesome	$\{f_{7.4}, f_{15.3}, f_{21.3}, f_{27.4}, f_{30.2}\}$	True	True
r_{22}	-	$\{f_{7.4}, f_{15.3}, f_{12.1}, f_{1.1}, f_{19.1}\}$	False	True

Table 3: Illustrative examples of implicit features.

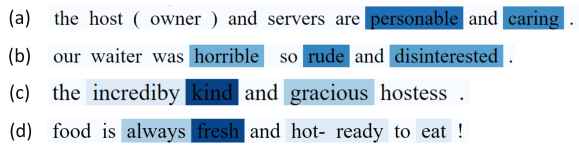

Figure 4: Visualization of the attention weights.

5.2 Effectiveness of Sentiment Weighting

We illustrate the effectiveness of the designed attention mechanisms in terms of attending to aspect-relevant sentiment information, and understanding the context. We retrieve samples from the datasets and visualize their attention weights in Figure 4, in which the deeper the color, the more importance a word has. It can be observed that the aspect-opinion words are weighted among the others and the model effectively attends to the sentiment words that are not in the lexicon (e.g., personable, gracious in (a) and (c) respectively). Since the sentiment words dominate the sentence representation in Equation 8, this indeed encourages the model to sentimentally learn the representations of polarity indicators.

6 Conclusion

In this work, we propose a novel DNN-driven GML approach for ATSA that can effectively leverage common implicit features for knowledge conveyance. Our extensive experiments have shown that the proposed approach consistently achieves the state-of-the-art performance in both unsupervised and supervised setting. For future work, it is noteworthy that the DNN-driven GML approach is potentially applicable to other classification tasks; the technical solutions however need further investigation.

Funding: The work was supported by the National Key Research and Development Program of China (2018YFB1003400), National Natural Science Foundation of China (61732014, 61672432,

61472321 and 61502390), Fundamental Research Funds for the Central Universities (Program No. 3102019DX1004) and Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2018JM6086).

References

Murtadha Ahmed, Qun Chen, and Zhanhuai Li. 2020. Constructing domain-dependent sentiment dictionary for sentiment analysis. *Neural Computing and Applications*, pages 1–14.

Tamara Alvarez-López, Jonathan Juncal-Martínez, Milagros Fernández-Gavilanes, Enrique Costa-Montenegro, and Francisco Javier González-Castano. 2016. Gti at semeval-2016 task 5: Svm and crf for aspect detection and unsupervised aspect-based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 306–311.

Giuseppe Castellucci, Simone Filice, Danilo Croce, and Roberto Basili. 2014. Unitor: Aspect based sentiment analysis with structured learning. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 761–767.

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*, pages 452–461.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397.

Boyi Hou, Qun Chen, Jiquan Shen, Xin Liu, Ping Zhong, Yanyan Wang, Zhaoqiang Chen, and Zhanhuai Li. 2019. Gradual machine learning for entity resolution. In *The World Wide Web Conference*, pages 3526–3530.

Boyi Hou, Qun Chen, Yanyan Wang, Youcef Nafa, and Zhanhua Li. 2020. Gradual machine learning for entity resolution. *IEEE Transactions on Knowledge and Data Engineering, (accepted online available)*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544.

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

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland. Association for Computational Linguistics.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956.

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

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *10th International Workshop on Semantic Evaluation (SemEval 2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Kim Schouten, Onne Van Der Weijde, Flavius Frasincar, and Rommert Dekker. 2017. Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. *IEEE transactions on cybernetics*, 48(4):1263–1275.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint 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 NAACL-HLT*, pages 380–385.

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

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

Jialong Tang, Ziyao Lu, Jinsong Su, Yubin Ge, Linfeng Song, Le Sun, and Jiebo Luo. 2019. Progressive self-supervised attention learning for aspect-level sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 557–566.

Shuai Wang, Sahisnu Mazumder, Bing Liu, Mianwei Zhou, and Yi Chang. 2018. Target-sensitive memory networks for aspect sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 957–967.

Yanan Wang, Qun Chen, Murtadha Ahmed, Boyi Hou, and Zhanhuai Li. 2019. Joint inference for aspect-level sentiment analysis by deep neural networks and linguistic hints. *IEEE Transactions on Knowledge and Data Engineering, (accepted, online available)*.

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*, Volume 212.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

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 (Volume 1: Long Papers)*, pages 2514–2523.

Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 56–65, Cambridge, MA. Association for Computational Linguistics.