An Attribute Enhanced Domain Adaptive Model for Cold-Start Spam Review Detection

Zhenni You [1], **Tieyun Qian** [1*], **Bing Liu** [2]
[1] School of Computer Science, Wuhan University, Wuhan, China
[2] Department of Computer Science, University of Illinois at Chicago, IL, USA
[*] Contact author
{znyou,qty}@whu.edu.cn, liub@uic.edu

Abstract

Spam detection has long been a research topic in both academic and industry due to its wide applications. Previous studies are mainly focused on extracting linguistic or behavior features to distinguish the spam and legitimate reviews. Such features are either ineffective or take long time to collect and thus are hard to be applied to cold-start spam review detection tasks. Recent advance leveraged the neural network to encode the textual and behavior features for the cold-start problem. However, the abundant attribute information are largely neglected by the existing framework.

In this paper, we propose a novel deep learning architecture for incorporating entities and their inherent attributes from various domains into a unified framework. Specifically, our model not only encodes the entities of reviewer, item, and review, but also their attributes such as location, date, price ranges. Furthermore, we present a domain classifier to adapt the knowledge from one domain to the other. With the abundant attributes in existing entities and knowledge in other domains, we successfully solve the problem of data scarcity in the cold-start settings. Experimental results on two Yelp datasets prove that our proposed framework significantly outperforms the state-of-the-art methods.

1 Introduction

Online reviews and ratings are playing more and more critical roles in E-Commerce places. The Mintel flagship report showed that 69 percent of Americans seek out others' advice online before making a purchase [1]. This gives strong incentives for imposters to game the system. As a result, fraudulent reviews flood the e-market websites. Forbes news [2] reported that Amazon's fake review problem is now worse than ever in 2017. Fake reviews are perceived as threats to the ecosystem of the e-business sites, companies, and users, and thus it becomes an urgent task to detect fake or spam reviews.

Existing approaches in spam review detection mainly focused on exacting linguistic features and behavioral features. However, linguistic features are ineffective when they are used to detect the real-life fake reviews (Mukherjee et al., 2013b; Wang et al., 2017b), and it usually requires a large number of samples to make the observations on behavior features. When dealing with the cold-start problem, i.e., *a review is just posted by a new reviewer*, it is hard to construct effective behavioral features for the new reviewer. More recently, Wang et al. (2017b) proposed a neural network model to jointly encode the textual and behavioral information into the review embeddings. This was a good try. However, due to the lack of sufficient information, the results reported in (Wang et al., 2017b) are not promising enough (with an accuracy less than 65%).

[1] http://store.mintel.com/american-lifestyles-2015-the-connected-consumer-seeking-validation-from-the-online-collective-us-april-2015

[2] https://www.forbes.com/sites/emmawoollacott/2017/09/09/exclusive-amazons-fake-review-problem-is-now-worse-than-ever/#501eccb87c0f

Proceedings of the 27th International Conference on Computational Linguistics, pages 1884–1895
Santa Fe, New Mexico, USA, August 20-26, 2018.

Our study is inspired by the work (Wang et al., 2017b) but moves one step further, i.e., we incorporate not only the entities, but also the attributes into a unified framework. The rationale is that similar reviewers' comments on items with similar attributes will result in similar review contexts. For example, the review for luxury restaurants may focus on their atmosphere and personal services while for ordinary restaurants one cares about their tastes. Here the luxury or ordinary is the key property of an restaurant, and might be reflected by the *price range* attribute.

An attribute is the inherent characteristic of an entity and is available for all users (reviewers), items or reviews no matter whether it is in a cold-start setting. For example, when a new user joins a website, the location is automatically recorded as his/her attribute. Such a property makes attributes extremely suitable for our cold-start spam detection problem. Furthermore, while an entity may be new, its attributes may not be new. For example, the attributes for a hotel like "price range" or "having wifi or not" must have been appeared in many other hotels' attributes. Hence we can piece together the profile of new users, items, and reviews by using attributes from others.

To further alleviate the data scarcity in cold-start spam review detection, we turn to seek more available information from other domains. The basic idea is that the entities in different domains may have same attributes. For example, both a hotel and a restaurant have an attribute of "price range". Hence we aim to leverage the shared attributes across different domains. The challenge of this task arises from the discrepancy between two domains. For example, a price of 200 USD may be high for a restaurant but it is low for a hotel. Inspired by prior work on domain adaption (Ben-David et al., 2007; Ben-David et al., 2010; Tzeng et al., 2015), we extend the idea of knowledge transferring between domains to learn a generalized representation for the entity.

Based on the above analysis, we propose an attribute-enhanced domain adaptive (AEDA) embedding model in this paper. Our model not only exploits the inherent attributes of reviewers, items and reviews, but also captures the domain correlations. In summary, the main contributions of our work are as follows.

- We propose to leverage the attributes of entities and their relations to enhance the representations of entities. These attributes are inherent for the entities and thus do not need any experts' experience or time-consuming procedure to collect.
- we present a novel neural network to jointly encode the attributes, entities, and their relations, as well as to adapt the knowledge from one domain to the other. The abundant information greatly alleviate data scarcity problem in the cold-start scenario of spam review detection.
- Extensive experiments demonstrate that our model significantly outperforms the state-of-the-art baseline methods.

The rest of the paper is structured as follows. In Section 2, we present the related work. In Section 3, we introduce our methodology. In Section 4, we show the experimental evaluation. We conclude the paper in Section 5.

2 Related Work

2.1 Review Spam Detection

The problem of spam review detection has aroused great research interests in recent years. This problem was first introduced in (Jindal and Liu, 2008), and the focus of this work was to find effective features to represent the fake and real reviews. Later, a variety of linguistic features were introduced in the literature (Ott et al., 2011; Xu and Zhao, 2013; Harris, 2012; Feng et al., 2012a; Kim et al., 2015; Li et al., 2013; Li et al., 2014b; Fornaciari and Poesio, 2014; Li et al., 2014a; Hovy, 2016). However, an in-depth study (Mukherjee et al., 2013b) found that the linguistic features were insufficient for detecting fake reviews in real business website. Therefore, the features besides review contents were exploited, and the behavioral features of reviewers were intensively studied in (Lim et al., 2010; Jindal et al., 2010; Feng et al., 2012b; Mukherjee et al., 2012; Xie et al., 2012; Fei et al., 2013; Li et al., 2015; KC and Mukherjee, 2016; Wang et al., 2011; Akoglu et al., 2013; Mukherjee et al., 2013a; Mukherjee et al., 2013b). The intuition is that the reviewers with spammer-like behaviors are more likely to post fake reviews. Several methods (Li

et al., 2011; Rayana and Akoglu, 2015) were proposed to utilize multiple information mentioned above. Overall, the traditional features are manually constructed and depend heavily on the experts' knowledge.

More recently, deep learning techniques are applied to fake review detection. Ren and Zhang (Ren and Zhang, 2016) compared several neural networks and found that CNN was more effective than RNN on review text encoding in spam review detection task. Hai et al. (2016) implemented a semi-supervised multi-task learning method, which introduced a covariance matrix to capture the relation between tasks and a Laplacian regularizer to leverage the unlabeled data. Wang et al. (2016) employed a tensor decomposition based on the global behavioral information to learn the representation of the reviewer and item.

The cold-start problem in spam review detection was first introduced in (Wang et al., 2017b), where the authors proposed an embedding learning model to jointly utilize the behavioral information of reviewers and the textual information. While we aim to solve the same cold start problem as that in (Wang et al., 2017b), we propose a totally different framework which leverage both the new attribute and domain knowledge information.

2.2 Domain Adaption

Domain adaption has been applied to addressing the scarcity of labeled data in a variety of real-world applications. It has been well studied in many tasks in the area of computer vision, such as the image classification and object recognization. These methods mainly differ in network structure or objectives, including parameter sharing (Yosinski et al., 2014; Liu et al., 2017), minimizing the distance between source and target distributions (Baktashmotlagh et al., 2013; Tzeng et al., 2014; Long et al., 2015), maximizing of the loss of domain classifier (Ganin and Lempitsky, 2015; Tzeng et al., 2015; Tzeng et al., 2017; Gopalan et al., 2011), and applying semi-supervised or selective learning techniques (Ao et al., 2017; Tan et al., 2017).

Recent years also witnessed the applications of domain adaption technique in many other areas, especially in natural language processing tasks. To name a few, typical applications include question answering (Min et al., 2017), machine translation (Chu et al., 2017; Johnson et al., 2016; Wang et al., 2017a), sentiment analysis (Li et al., 2017), recommendation system (Man et al., 2017) and image-text retrieval (Huang et al., 2017).

While domain adaption technique was introduced in the above research fields, it was rarely adopted in spam review detection. We aim to exploit this technique to help alleviate the problem of data scarcity in the cold-start scenario of spam detection.

3 Methodology

In this section, we present the details of our attribute-enhanced domain adaptive embedding model (AEDA). The key idea is to leverage the relations between entities and attributes and adapt knowledge from one domain to the other. Motivated by recent advances in deep learning which has been proven to be powerful in learning nonlinear representations, we design a novel deep neural network to jointly encode the rich information in entities, attributes, and domains.

3.1 Attribute Enhanced Objective

When a reviewer comments on an item (or product), a review is posted along with a rating score. This process involves three kinds of entities: reviewers, items, and reviews. All these entities consist in a number of attributes. For example, when an item, e.g., a newly opened restaurant, first registers on Yelp, its location and other attributes are recorded in the website. The attributes are essential for legitimate users to learn about the entity. For example, a user on a business trip may care about whether the hotel is close to a metro station.

The attributes provide additional information of the entity and form the background of the comment action. Such information and background are particularly useful when dealing with the cold start spam review detection problem. Since these attributes are inherent components of an entity, we can simply piece together the profile for a new entity using the attributes in existing entities of the same type. Hence

we collect all the attributes and aim to incorporate them into our framework. Below is a brief introduction to these attributes.

- **Reviewer attributes** are recorded when the reviewer registers on the review system. The *YelpJoinDate* is the date when the reviewer joins the system, while the *Location* indicates where the reviewer comes from, which is filled by him/herself.
- **Item attributes** reflects the basic information of each item, such as its *Location*, *Average Rating*, *Price Ranges* (from 1 to 5). The *AcceptCreditCards*, *WiFi*, *WebSite*, and *PhoneNumber* are the attributes about whether the item provides the specific service or not. We convert them into boolean values like "hasWifi" or "ifAcc".
- **Review attributes** include the *Date* when the review is posted and the *Rating* score the reviewer rates on the item.

Overall, there are three kinds of entities and eight kinds of attributes, which further form three types of relations: entity-attribute, attribute-attribute, and entity-entity. We aim to leverage these relations between the attribute and the entity. We will give details below.

Relation 1: entity-attribute relation In order to exploit the rich information contained in the entity-attribute relation, we propose an attribute-enhanced objective L_{ea} to maximize the conditional probability $P(v|e)$ of an entity e's attribute a with a value of v. The intuition is that the attributes can be treated as the contexts of the entities, and the attribute information can be encoded into the representations of entities. We formalize it as follows:

$$\begin{aligned} L_{ea} &= P(v|e) \\ P(v|e) &= \frac{exp(V_v \cdot V_e)}{\sum_{u \in A} exp(V_u \cdot V_e)} \end{aligned} \tag{1}$$

where V_i indicates the embedding of i, including entities and attributes, and A is the corresponding value set of attribute a. Note that we employ negative sampling (Mikolov et al., 2013) to transform the objective. By maximizing the attribute-enhanced objective L_{ea}, the entities with similar contexts (attributes) tend to be similar and the attributes of similar entities become associated. In this way, the representations of attributes become informative and we obtain the attribute-enhanced entity embeddings.

Relation 2: attribute-attribute relation Several attributes, i.e., date, location, and rating, are shared by different entities. In order to capture the relation between these attributes, we take their difference as a new feature. Specifically, we subtract *Date* (review) from *YelpJoinDate* (reviewer) as the new *dateDif* feature, and subtract *Rating* (review) from *Average Rating* (item) as the new *ratingDif* feature, and treat the new *locDif* feature as 1 when the reviewer and item have same location otherwise 0.

Relation 3: entity-entity relation While the entity-attribute and attribute-attribute relations are proposed by us, the entity-entity relation has been examined before (Wang et al., 2017b). For a fair comparison, we follow the work of (Wang et al., 2017b) by applying the TransE model on our attribute-enhanced entity embedding. More formally, we propose an entity-interacted objective L_{ee} as follows:

$$\begin{aligned} L_{ee} = \sum_{(i,r,t) \in S} \sum_{(i',r,t') \in S'} & max\{0, 1 + d(i+r,t) - d(i'+r,t')\} \\ d(i+t,c) =& ||V_i + V_r - V_t||_2^2, \\ s.t.\ ||V_i||_2^2 =& ||V_r||_2^2 = ||V_t||_2^2 = 1 \end{aligned} \tag{2}$$

where S is a set of triples (i, r, t) in the training set of reviews, including the item $i \in I$ (item set) which the review talk about, the reviewer $r \in R$ (reviewer set) who posts this review, and the review text $t \in T$ (review set). S' denotes the corrupted set which is constructed by replacing the item *i* (or text *t*) with a random chosen i' (or t').

The entity-entity and entity-attribute relation can be encoded into the embedding of the entities and attributes in our model by the proposed attribute-enhanced and entity-interacted objective. To make full of them, we concatenate the review text embedding generated by CNN and all the attribute embeddings

of item, reviewer and review into a long vector V_{rcon} to represent the review. This is our basic attribute enhanced (AE) model. The representations of reviews can then be used as the features to train a spam review classifier and make a classification of the new reviews.

3.2 Domain Adaptive Objective

Our AE model is trained in the single domain, i.e., all the embeddings are trained using data merely from one hotel or restaurant domain. As proved in (Hai et al., 2016), the detection of spam review in different domains may have strong correlations. We are curious about whether the cold-start problem can be alleviated by borrowing knowledge from other domains. This is reasonable because entities in different domains may have same attributes. For example, both hotel and restaurant have an attribute of "price range". If we can confuse the domain difference and focus on the relations of attributes from different domains, we may complement the attribute embeddings and subsequently enhance entity embeddings.

Motivated by the recent advances in domain adaption (Ben-David et al., 2007; Ben-David et al., 2010; Tzeng et al., 2015), we add a domain classifier to perform domain classification of reviews in the training set based on the review embedding V_{rcon} mentioned before. Then an adversarial loss is adopted to intensify the domain confusion.

Domain Classifier The domain classifier is implemented by adding a dense layer to identify which domain a review belongs to according to its representation V_{rcon}. We aim to minimize the domain label prediction loss L_{dcla} (defined as the cross-entropy between the predicted label and the true label) by updating the parameters of the domain classifier using a softmax function.

$$q = softmax(W \cdot V_{rcon} + b) \tag{3}$$

$$L_{dcla} = -\sum_{d} \mathbf{1}[y_{dtrue} = d] \, log \, q_d \tag{4}$$

where q is the predicted domain distribution, d is the index of the corresponding domain, y_{dtrue} is the true domain where the review comes from, and W and b are the parameters of the domain classifier.

Adversarial Loss The target of the domain confusion is to adjust the representation to become too domain-invariant to distinguish its domain label. To this end, we minimize the domain confusion loss L_{dcon}, which is defined as the cross entropy between domain label distribution predicted by the domain classifier and a uniform distribution, by updating all parameters in our model except those of domain classifier. We formalize it as follows:

$$L_{dcon} = -\sum_{d} \frac{1}{D} \, log \, q_d \tag{5}$$

where D is the number of the domain in our training set.

After the adversarial training, we can reach a point at which the review representation can not be classified into the true domain label by the best domain classifier. This is what our attribute enhanced domain adaptive (AEDA) model does for training the representations.

Although the domain correlation problem was introduced in (Hai et al., 2016), our task is completely different from that in (Hai et al., 2016). We deal with the domain adaption problem while Hai et al. (Hai et al., 2016) address the multi-task learning problem. In addition, their method was heavily dependent on the labeled data, while our method is totally unsupervised when adapting the knowledge from other domains.

3.3 Model Architecture and Learning Procedure

We present the architecture of our AEDA model in Figure 1. It consists of three layers. The bottom layer of the architecture integrates various attributes into the model. The middle layer aims to capture three relations, i.e., entity-entity, entity-attribute, and attribute-attribute relations. The top layer is used to implement a domain classifier to capture the domain correlation.

We now discuss the learning procedure of our AEDA model in Algorithm 1. It works as follows. Line 1 initializes the parameters. Most parameters of model are randomly initialized, with the exception that the review texts E_t which are generated from the CNN based on the pre-trained word embedding by CBOW

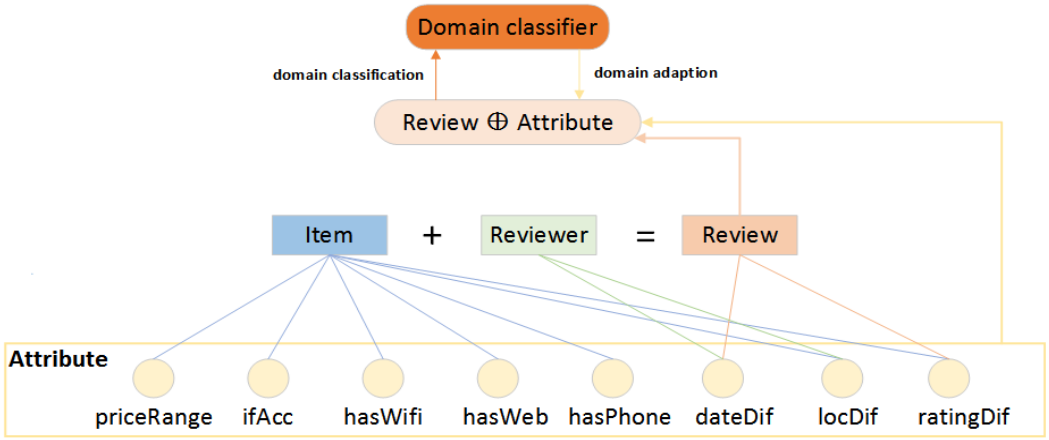

Figure 1: Architecture of AEDA model

(Mikolov et al., 2013). Then for each batch in the training set, we sequentially train entity-attribute relation in reviewer, item, and review, as shown in lines 4 to 6. Next, we use Eq.(2) to capture the entity-entity relation in line 7, which updates the parameters for the entities. Finally we minimize the Eq.(4) to train an accurate domain classifier, and minimize the Eq.(5) to confuse the review representations from different domains. Lines 8, 9 show this adversarial training procedure. After the model is properly trained, through the input of training sets and testing sets, we can obtain the representations of reviews in both sets. These embeddings of reviews can be fed into the traditional classification model like SVM to train the classifier and evaluate the performance of our model.

Algorithm 1 Attribute-enhanced domain adaptive (AEDA) model

Input: The combined training set of different domains, $X_{train} = \{x_i, y_{dtrue,i}\}_{i=1}^{n_{train}}$. Each x_i is composed of several parts including reviewer, item, review text, and their attributes, and $y_{dtrue,i}$ is the domain label of review i;

Output: The trained AEDA model

1: Initialize the parameters of the model, including embeddings of reviews E_t and their attributes E_{ta}, embeddings of reviewers E_r and their attributes E_{ra}, embeddings of items E_i and their attributes E_{ia}, the parameters of CNN Ψ_{tc} and those of domain classifier W, b
2: **repeat**
3: **for** each batch in X_{train} **do**
4: update E_r, E_{ra} by maximizing Eq.(1), where e is the reviewer, v is the value of reviewers' attributes.
5: update E_i, E_{ia} by maximizing Eq.(1), where e is the item, v is the value of items' attributes.
6: update E_t, Ψ_{tc}, E_{ta} by maximize Eq.(1), where e is the review generated from CNN, v is the value of reviews' attributes.
7: update E_r, E_i, E_t, Ψ_{tc} by minimizing Eq.(2)
8: update W, b by minimizing Eq.(4)
9: keep W, b fixed and update all other parameters in the model by minimizing Eq.(5)
10: **end for**
11: **until** converge or reach the predetermined number of epoches
12: **return** the trained AEDA model

4 Experiments

4.1 Experimental Settings

We verify the effectiveness of our proposed model on two cold-start datasets (Wang et al., 2017b), which are the subset of the Yelp datasets used in a number of previous studies (Mukherjee et al., 2013b; Rayana and Akoglu, 2015; Mukherjee et al., 2013a). To deal with the cold-start problem, the original datasets are divided into two parts (Wang et al., 2017b). The reviews posted before January 1, 2012 are used as the training data, and the first new reviews posted by the new reviewers after January 1, 2012 are used as the test data.

We use the SVM method to train the classifier on the training data and test it on the test data. We choose precision (P), recall (R), F1-Score (F1), and accuracy (Acc) as the evaluation metrics. Both the SVM method and the metrics are same as as those in (Wang et al., 2017b; Mukherjee et al., 2013b; Rayana and Akoglu, 2015; Mukherjee et al., 2013a).

4.2 Baseline Methods

We compare our model with eight state-of-the-art methods based on linguistic features and behavioral features, which are listed as follows.

LF (Mukherjee et al., 2013a) captures the linguistic features by extracting bigrams on the labeled review data.

Supervised-CNN uses the same textual information as LF but its features are trained in a supervised convolutional neural network.

LF+BF (Mukherjee et al., 2013a) is a concatenation of linguistic features (LF) and behavioral features (BF).

BF_EditSim+LF (Wang et al., 2017b) first calculates the edit distance between the current review and existing reviews and find the most similar one, then uses its reviewer's behavioral features as the approximation of the new reviewer.

BF_W2VSim+W2V (Wang et al., 2017b) is similar to the BF_EditSim+LF, but uses the similarity of averaged word embedding (pretrained by Word2vec (Mikolov et al., 2013)) to find the most similar review, and concatenates the behavioral features with the average word embedding instead of bigram.

RE* (Wang et al., 2017b) jointly utilizes the TransE (Bordes et al., 2013) to model the behavioral information of reviewers, CNN with same parameter settings of supervised-CNN to encode the textual information, and a constraint to preserve semantics of the sentiment polarity in rating.

RE+RRE+PRE* (Wang et al., 2017b) is an improved version of RE, which further concatenates the review embedding, the review's rating embedding and the item's average rating embedding into a long vector as the feature of the review.

ATT+LF uses the all the attribute values used in our model as the features and then directly concatenates them with bigrams.

For the baselines, we report the results in (Wang et al., 2017b) if they have been implemented, since we conduct experiments on the exactly same datasets and training/testing splits. In addition, we use the same hyper-parameters for our model as those in (Wang et al., 2017b) for a fair comparison.

4.3 Comparison with Baselines

We conduct the comparison experiments on two Yelp datasets. The results are shown in Table 1. AEDA denotes our proposed attribute-enhanced domain adaptive model, and AE refers to a variation of AEDA, which only trains in the single domain without domain-adaption.

It is clear that the proposed AEDA model and its variant AE significantly consistently outperform all the baseline methods on both hotel and restaurant datasets in terms of precision, F1, and accuracy. The slight decreases in recall can be due to the trade off between the precision and recall. The significant improvements of F1 values over baselines clearly demonstrate the effectiveness of our attribute-enhanced method (AE) plus domain adaption technique (AEDA) in cold-start spam review detection task. We have more observations for Table 1.

Table 1: Comparison with baselines.

Row	Features	Hotel				Restaurant			
		P	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**
1	LF	54.5	71.1	61.7	55.9	53.8	80.8	64.6	55.8
2	Supervised-CNN	61.2	51.7	56.1	59.5	56.9	58.8	57.8	57.1
3	LF+BF	63.4	52.6	57.5	61.1	58.1	61.2	59.6	58.5
4	BF_EditSim+LF	55.3	69.7	61.6	56.6	53.9	82.2	65.1	56.0
5	BF_W2Vsim+W2V	58.4	65.9	61.9	59.5	56.3	73.4	63.7	58.2
6	RE*	62.1	68.3	65.1	63.3	58.4	75.1	65.7	60.8
7	RE+RRE+PRE*	63.6	71.2	67.2	65.3	59.0	78.8	67.5	62.0
8	ATT+LF	71.1	74.7	72.8	72.1	64.0	73.2	68.3	66.0
9	AE (ours)	**76.7**	74.2	**75.4**	**75.8**	**80.3**	66.2	**72.6**	**75.0**
10	AEDA (ours)	**83.9**	74.2	**78.7**	**80.0**	**82.4**	65.1	**72.8**	**75.6**

(1) The LF and Supervised-CNN methods (rows 1-2) which only use linguistic features of reviews are the worst in terms of F1 or accuracy.

(2) Adding the behavioral features (rows 3-6) can improve the performance to some extents, but it is important to choose a good combination method. BF_EditSim+LF and BF_W2Vsim+W2V are both based on the idea of approximating the behavioral features of new reviewer with the existing one with the most similar text. RE* encodes the correlated behavioral information with a neural network framework and thus performs better than simple approximation.

(3) With the additional attribute information (rows 7-10), all the remaining methods outperform those without attributes. For example, our AE shows an increase of 10.3% in F1 and 12.5% in accuracy over RE(*) in hotel domain, and 6.9% in F1 and 14.2% in accuracy in restaurant domain.

(4) Among the methods with additional attributes (rows 7-10), our basic model AE is already significantly better than RE+RRE+PRE*. The improvements may be due to the fact that AE captures the relationship of the common attributes between two entities like the dateDif in Figure 1. Also note that both RE* and RE+RRE+PRE* take all the existing reviews posted before January 1, 2012 to train their models while we only use a small subset. More details about the effects of data size will be discussed in the next section. Furthermore, compared with ATT+LF, our AE leads to an improvement of 9.0% and 3.7% of accuracy, and 4.3% and 2.6% of F1 in restaurant and hotel domain, respectively. This shows that our neural framework can well captures the joint effects of attribute and behavioral information.

(5) Between our AEDA and its simple variant AE (rows 9-10), AEDA is better. The reason is that the attribute-attribute relation in different domains is captured by domain-adaptive objective in AEDA and thus knowledge in the restaurant domain can complement that in the hotel. We find that the performance of AEDA in restaurant does not improve too much. This is because the training set of hotel domain is too small to provide much information for restaurant domain.

4.4 With or Without Unlabeled Data

In the Yelp datasets, 99.18% and 92.58% reviews are unlabeled in hotel and restaurant domain. Wang et al. (2017b) exploited the large-scale unlabeled reviews into their neural network to encode the global behavioral information into the embeddings. Results shows that their model RE* benefits a lot from unlabeled data. However, this is at the expense of efficiency since the training set size increases dramatically when the unlabeled data are added. Our model is trained on the small subset consisting of labeled data. So the problem is how about the performance of RE* without unlabeled data and that of our AE

model with unlabeled data.

Table 2: Impacts of unlabeled data. * indicates that unlabeled data are included

Methods	Hotel				Restaurant			
	P	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**
RE	53.2	57.6	55.3	53.5	59.8	62.2	61.0	60.2
AE (ours)	76.7	74.2	75.4	75.8	80.3	66.2	72.6	75.0
RE*	62.1	68.3	65.1	63.3	58.4	75.1	65.7	60.8
AE* (ours)	82.4	71.4	76.5	78.1	80.0	67.7	73.3	75.4

We investigate the impacts of unlabeled data and the results are shown in Table 2. It can be seen that whether we use unlabeled data or not, our model is significantly better than that in (Wang et al., 2017b). On one hand, our AE model trained on the small labeled data significantly outperforms the RE* model trained on the whole dataset, let alone our enhanced AE* model. On the other hand, when RE* is reduced to RE, its performance drops a lot. For example, the F1 value decreases from 65.1% to 55.3%. This suggests that the performance of RE* is highly dependent on the size of training data. In contrast, our model is less sensitive to the training size than RE. We believe the reason is that we make good use of attributes. Our model can get discriminative representations even with a very small number of data with the help of attribute information.

4.5 Attribute Effects

To evaluate the effectiveness of each attribute, we remove one attribute from our model at a time, and the results are listed in Table 3.

Table 3: Attribute effects

Removed attribute	Hotel		Restaurant	
	F1	**Acc**	**F1**	**Acc**
location (locDiff)	-0.9	-0.9	0.0	-0.2
date (dateDif)	-22.5	-19.8	-12.1	-12.1
rating (ratingDif)	-1.2	-0.9	-1.0	-0.6
price range	0.5	0.7	0.0	-0.3
ifAcc	0.1	0.0	-0.3	-0.4
hasWifi	1.2	1.2	-0.1	-0.2
hasWeb	-0.3	-0.5	-0.2	-0.4
hasPhone	-0.5	-0.7	0.4	0.1

We find that three attributes (location, date, and rating) are the most important features in both domain. This is reasonable because they are not only the components of the entity itself, but also they are shared by two entities. In particular, we find that the date (dateDif) feature plays a critical role in spam review detection. This is an interesting finding which indicates that spammers may write reviews right after they register at the website. The location difference is more important in hotel than in restaurant. The reason can be that a legitimate user needs an accommodation usually when he/she is out for travelling or business, while a spammer just makes comments and does not care about the location of the hotel.

Among the single attributes, removing "hasWeb" results in the decrease of performance in both hotel and restaurant domains. We suppose that users prefer to search the website of the hotel or restaurant for further information before they make a decision to put it into the schedule. However, these information do not make any sense for spammers. Hence the attribute of "hasWeb" is a good indicator of spam review detection.

5 Conclusion

We introduce an attribute-enhanced domain adaptive embedding (AEDA) model to cope with the cold-start problem in spam review detection. With a carefully designed neural network, our model can jointly encode the inherent attributes of entities, behavioral information, and domain correlations into the review representations. The learnt embeddings are then fed into the traditional SVM machine learning framework to train a classifier and to detect the spam reviews. Experimental results demonstrate the superiority of our proposed AEDA model over all the baseline methods. In the future, we will exploit more available information to enhance the review embedding.

Acknowledgments

The work described in this paper has been supported in part by the NSFC projects (61572376, 91646206), and the 111 project(B07037).

References

Leman Akoglu, Rishi Chandy, and Christos Faloutsos. 2013. Opinion fraud detection in online reviews by network effects. *ICWSM*, 13:2–11.

Shuang Ao, Xiang Li, and Charles X Ling. 2017. Fast generalized distillation for semi-supervised domain adaptation. In *AAAI*, pages 1719–1725.

Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. 2013. Unsupervised domain adaptation by domain invariant projection. In *ICCV*, pages 769–776.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In *NIPS*, pages 137–144.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of simple domain adaptation methods for neural machine translation. *arXiv preprint arXiv:1701.03214*.

Geli Fei, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Exploiting burstiness in reviews for review spammer detection. *ICWSM*, 13:175–184.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012a. Syntactic stylometry for deception detection. In *ACL*, pages 171–175.

Song Feng, Longfei Xing, Anupam Gogar, and Yejin Choi. 2012b. Distributional footprints of deceptive product reviews. *ICWSM*, 12:98–105.

Tommaso Fornaciari and Massimo Poesio. 2014. Identifying fake amazon reviews as learning from crowds. In *ACL*, pages 279–287.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189.

Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. 2011. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, pages 999–1006.

Zhen Hai, Peilin Zhao, Peng Cheng, Peng Yang, Xiao Li Li, and Guangxia Li. 2016. Deceptive review spam detection via exploiting task relatedness and unlabeled data. In *EMNLP*, pages 1817–1826.

Christopher Harris. 2012. Detecting deceptive opinion spam using human computation. In *Workshops at AAAI*, pages 87–93.

Dirk Hovy. 2016. The enemy in your own camp: How well can we detect statistically-generated fake reviews–an adversarial study. In *ACL*, pages 351–356.

Xin Huang, Yuxin Peng, and Mingkuan Yuan. 2017. Cross-modal common representation learning by hybrid transfer network. *arXiv preprint arXiv:1706.00153*.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *WSDM*, pages 219–230.

Nitin Jindal, Bing Liu, and Ee-Peng Lim. 2010. Finding unusual review patterns using unexpected rules. In *CIKM*, pages 1549–1552.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Santosh KC and Arjun Mukherjee. 2016. On the temporal dynamics of opinion spamming: Case studies on yelp. In *WWW*, pages 369–379.

Seongsoon Kim, Hyeokyoon Chang, Seongwoon Lee, Minhwan Yu, and Jaewoo Kang. 2015. Deep semantic frame-based deceptive opinion spam analysis. In *CIKM*, pages 1131–1140.

Fangtao Li, Minlie Huang, Yi Yang, and Xiaoyan Zhu. 2011. Learning to identify review spam. In *IJCAI*, page 2488.

Jiwei Li, Claire Cardie, and Sujian Li. 2013. Topicspam: a topic-model based approach for spam detection. In *ACL*, volume 2, pages 217–221.

Huayi Li, Bing Liu, Arjun Mukherjee, and Jidong Shao. 2014a. Spotting fake reviews using positive-unlabeled learning. *Computación y Sistemas*, 18(3):467–475.

Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. 2014b. Towards a general rule for identifying deceptive opinion spam. In *ACL*, pages 1566–1576.

Huayi Li, Zhiyuan Chen, Arjun Mukherjee, Bing Liu, and Jidong Shao. 2015. Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In *ICWSM*, pages 634–637.

Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *IJCAI*, page 2237.

Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. 2010. Detecting product review spammers using rating behaviors. In *CIKM*, pages 939–948.

Jiaming Liu, Yali Wang, and Yu Qiao. 2017. Sparse deep transfer learning for convolutional neural network. In *AAAI*, pages 2245–2251.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105.

Tong Man, Huawei Shen, Xiaolong Jin, and Xueqi Cheng. 2017. Cross-domain recommendation: an embedding and mapping approach. In *IJCAI*, pages 2464–2470.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. *arXiv preprint arXiv:1702.02171*.

Arjun Mukherjee, Bing Liu, and Natalie Glance. 2012. Spotting fake reviewer groups in consumer reviews. In *WWW*, pages 191–200.

Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie Glance. 2013a. Fake review detection: Classification and analysis of real and pseudo reviews. *Technical Report UIC-CS-2013–03, University of Illinois at Chicago, Tech. Rep.*

Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie S Glance. 2013b. What yelp fake review filter might be doing? In *ICWSM*.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *ACL*, pages 309–319.

Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *SIGKDD*, pages 985–994.

Yafeng Ren and Yue Zhang. 2016. Deceptive opinion spam detection using neural network. In *COLING*, pages 140–150.

Ben Tan, Yu Zhang, Sinno Jialin Pan, and Qiang Yang. 2017. Distant domain transfer learning. In *AAAI*, pages 2604–2610.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous deep transfer across domains and tasks. In *ICCV*, pages 4068–4076.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *CVPR*, page 4.

Guan Wang, Sihong Xie, Bing Liu, and S Yu Philip. 2011. Review graph based online store review spammer detection. In *ICDM*, pages 1242–1247.

Xuepeng Wang, Kang Liu, Shizhu He, and Jun Zhao. 2016. Learning to represent review with tensor decomposition for spam detection. In *EMNLP*, pages 866–875.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017a. Sentence embedding for neural machine translation domain adaptation. In *ACL*, pages 560–566.

Xuepeng Wang, Kang Liu, and Jun Zhao. 2017b. Handling cold-start problem in review spam detection by jointly embedding texts and behaviors. In *ACL*, pages 366–376.

Sihong Xie, Guan Wang, Shuyang Lin, and Philip S Yu. 2012. Review spam detection via temporal pattern discovery. In *SIGKDD*, pages 823–831.

Qiongkai Xu and Hai Zhao. 2013. Using deep linguistic features for finding deceptive opinion spam. In *COLING*, pages 1341–1350.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *NIPS*, pages 3320–3328.