Reducing the Need for Manual Annotated Datasets in Aspect Sentiment Classification by Transfer Learning and Weak-Supervision

Ermelinda Oro[1,2(✉)], Massimo Ruffolo[1,2], and Francesco Visalli[2]

[1] High Performance Computing and Networking Institute of the National Research Council (ICAR-CNR), Via Pietro Bucci 8/9C, 87036 Rende, CS, Italy
{ermelinda.oro,massimo.ruffolo}@icar.cnr.it
[2] altilia.ai, Piazza Vermicelli, Technest, University of Calabria, 87036 Rende, CS, Italy
{ermelinda.oro,massimo.ruffolo,francesco.visalli}@altiliagroup.com

Abstract. Users' opinions can be greatly beneficial in developing and providing products and services and improving marketing techniques for customer recommendation and retention. For this reason, sentiment analysis algorithms that automatically extract sentiment information from customers' reviews are receiving growing attention from the computer science community. Aspect-based sentiment analysis (ABSA) allows for a more detailed understanding of customer opinions because it enables extracting sentiment polarities along with the sentiment target from sentences. ABSA consists of two steps: Aspect Extraction (AE) that allows recognizing the target sentiment; Aspect Sentiment Classification (ASC) that enables to classify the sentiment polarity. Currently, most diffused sentiment analysis algorithms are based on deep learning. Such algorithms require large labeled datasets that are extremely expensive and time consuming to build. In this paper, we present two approaches based on transfer learning and weak supervision, respectively. Both have the goal of reducing the manual effort needed to build annotated datasets for the ASC problem. In the paper, we describe the two approaches and experimentally compare them.

Keywords: Deep learning · Aspect sentiment classification · Aspect based sentiment analysis · Sentiment analysis · Natural language processing · Transfer learning · Post-trained language model · Fine-tuning · Post-training · Transformers · BERT · Weak-supervision · Data programming

This work was supported by Horizon 2020 - Asse I – PON I&C 2014–2020 FESR - Fondo per la Crescita Sostenibile - Sportello Fabbrica Intelligente DM 05/03/2018 – DD 20/11/2018. "Validated Question Answering" Project.

© Springer Nature Switzerland AG 2021
A. P. Rocha et al. (Eds.): ICAART 2020, LNAI 12613, pp. 445–464, 2021.
https://doi.org/10.1007/978-3-030-71158-0_21

1 Introduction

The volume of social Web content is rapidly growing. Social media, e-commerce websites, and booking platforms enable users and customers to share their comments, opinions and preferences about products and services with the rest of the world. Opinions influence single and group of Web users, which potentially are customers for products and services providers. Users' opinions can be of great help in developing and providing new products and services [66], modifying existing ones [39], as well as improving marketing techniques for recommending their products to customers [34,35]. The enormous amount of available reviews posted on the Web is attracting both the research community and business companies. Many applications can exploit sentiment analysis for obtaining insights from reviews [19].

Sentiment analysis concerns the development of algorithms that can automatically extract sentiment information from a set of reviews. It operates at the intersection of information retrieval, natural language processing, and artificial intelligence [39]. Because sentiment is related to the human emotions communicated by natural language expressions, it can be hard to write, measure, and quantify. For these reasons, most sentiment analysis approaches have the objective of classifying the sentiment polarity of a text, such as *positive, negative, neutral*. As discussed in [25], sentiment analysis has been studied mainly at three levels of classification: document level, sentence level, or target level. Document and sentiment levels are useful if we assume that the text taken as input expresses sentiment about only one topic. More interesting is the case that considers the target of the sentiment. In general, sentiment analysis can be formally defined as finding the quadruple (s, g, h, t) where s represents the sentiment, g represents the *target* object for which the sentiment is expressed, h represents the holder (i.e., the one expressing the sentiment), and t represents the time at which the sentiment was expressed. Most approaches focus on finding only the pair (s, g) [25]. The target can have different granularities: (i) The whole entity considered in the review (e.g., a laptop, a restaurant). (ii) An aspect of the entity, i.e., a characteristic or property of an entity (e.g., battery, food).

Aspect-based sentiment analysis (ABSA) aims to find sentiment-target (s, g) pairs in a given text where targets are aspects of the reviewed entity. It allows for a more detailed analysis that utilizes more of the information provided by the textual review. We use the two following sentences to explain some further details: "*The **battery** gets so **not** it is scary*" and "*All the **food** was **not** tasty*". These sentences are related to the laptops and restaurants domains respectively. ABSA can be split down into two major problem sets [45]:

(i) Aspect Extraction (AE), or aspect detection, has the objective to recognize the target object g in sentences. For instance, considering our previous sentences, ***battery*** and ***food*** respectively. Given a review, sentences can be easily extracted from reviews with a sentence splitter. Whereas, there exist different approaches to retrieve aspects [15,42].
(ii) Aspect Sentiment Classification (ASC) has the objective of recognizing and classifying the sentiment polarity of terms related to aspects and expressing

a sentiment/opinion. The sentiment can depend on the knowledge domain. For instance, in both previous sentences, the sentiment is associated to the word ***hot***, but in the laptop domain such word is related to the aspect *battery* and takes a *negative* meaning, whereas in the second restaurant the same word is referred to the aspect food and takes a *positive* connotation. In fact, a battery that gets hot is not desirable, while a hot tasty food is good. ASC is a challenging problem because it is difficult to model the semantic relatedness of a target with its surrounding context words. It is necessary to understand semantic connections between the target word and the context words, which influence the sentiment polarity (e.g., the adjective related to an aspect).

Recently, and particularly since the introduction of BERT [11], deep transfer-learning methods have been applied successfully to a myriad of Natural Language Processing (NLP) tasks, including ABSA [54,76]. Deep learning automates the feature engineering process typical of classical machine learning approaches.

However, deep learning, in particular when applied to unstructured data, needs very large training sets to learn and generalize the parameters, and avoid overfitting [50]. But, building large labeled datasets is an extremely expensive and time-consuming task that often requires domain expertise [36]. To deal with this problem, different approaches that reduce the need for labeled data, such as transfer learning [59] and weak supervision [52], have been presented in the literature. Transfer learning methods like [70,79] rely on the fact that a model trained on a specific source domain can be exploited to do the same task in another target domain, thus reducing the need for labeled data in the target domain. Recently, transfer learning exploits pre-trained language models (PTMs), which can learn accurate general language representation from huge unlabeled corpora. One of the most important works of semi-supervised learning is BERT [11]. PTMs can be fine-tuned in a supervised way to learn various down-stream NLP tasks. The idea behind fine-tuning is that most of the features are already learned and represented in the PTM. Then, the model just needs to be specialized for the specific NLP task. For instance, BERT can be fine-tuned on the ASC problem requiring less annotated data than learning the entire task from scratch. Instead, weak supervision simplifies the annotation process to make it more automatic and scalable, even less accurate and noisier. Weak supervised methods rely on several different data annotation techniques such as heuristics, distant learning, pattern matching, weak classifiers, etc. Recently, [52] proposed data programming as a paradigm for semi-automatic datasets labeling, and Snorkel [51] the system that implements it. Data programming is based on the concept of Labeling Functions (LFs), where LFs are procedures that automatically assign labels to data on the base of domain knowledge embedded in the form of annotation rules. At Google, Batch et al. [3] extended Snorkel in order to achieve better scalability and knowledge base re-usability for enterprise-grade training sets labeling.

In this paper, we define and compare a transfer learning approach with a weakly-supervised approach [60] for ASC task, assuming that aspects have already been recognized in the input text.

The main contributions of this work are:

- We present a brief overview of sentiment analysis research. In particular, first, we indicate recent related surveys and datasets. Then, we provide a short review with a categorization of the existing literature related to deep learning, transfer learning, and weak supervision methods for sentiment analysis.
- We define two approaches, respectively based on cross-domain transfer learning and weak supervision, aimed at reducing the need for manually annotated datasets for ASC in a specific domain. In particular, the weak supervision method we propose is grounded on the paper [60].
- We provide the experimental evaluation of the two presented approaches for solving the ASC task considering two disjoint domains, laptops, and restaurants. In particular we used the datasets of SemEval task 4 subtask 2 [45].

The rest of this paper is organized as follows. Section 2 briefly reviews related work of sentiment analysis classifying it by adopted approaches and objectives. Section 3 describes our transfer learning approach and our weak-supervision method for ASC. In Sect. 4, we experimentally compare the two presented approaches used to reduce the need for annotated training sets in the target domain to perform ASC. Finally, Sect. 5 concludes the work.

2 Related Work

In recent years, many papers related to sentiment analysis have been presented. There exists different surveys that address various characteristics of the sentiment analysis [2,4,8,20,31,46,53,62,81]. In addition, many datasets have been created for sentiment analysis, for instance Amazon product reviews [6,22], tweets [1,16,18,56,80], IMDB movie review [29,37,38,40], news [10,74], Stanford Sentiment Treebank [61], Yelp dataset[1], SemEval Aspect-Based sentiment analysis dataset [43–45].

In this section, we briefly review related work classifying in different categories: (i) Methods based on different architectures of deep neural networks for ASC, opinion expression extraction, and sentiment classification. (ii) Recent papers exploit transformers with transfer learning techniques to perform ABSA, in particular, BERT-based approaches. (iii) Papers that apply weak-supervision methods to perform ABSA.

2.1 Deep Neural Network for ABSA

Aspect Sentiment Classification. Many deep learning architectures and techniques have been defined for aspect-level sentiment analysis before transformers. To the best of our knowledge, there were no dominating techniques in the literature. Dong et al. [12] present an adaptive recursive neural network (AdaRNN). They apply their method to perform a target-dependent sentiment analysis of

[1] https://www.yelp.com/dataset/challenge.

tweets. Vo and Zhang [69] use rich automatic features to perform aspect-based Twitter sentiment classification. The authors prove that rich sources of feature information help achieve better performance considering multiple embeddings, multiple pooling functions, and sentiment lexicons. Zhang et al. [82] present two-gated neural networks. Tang et al. [63] extend LSTM to consider the target of sentiment defining target-dependent LSTM (TD-LSTM) and target-connection LSTM (TC-LSTM). In these models, the target is given as feature and it is concatenated with the context features. Ruder et al. [58] use hierarchical and bidirectional LSTM model. They consider both intra- and inter-sentence relations. Wang et al. [73] present an attention-based LSTM method with target embedding (ATAE-LSTM). Yang et al. [78] present a two attention-based bidirectional LSTMs. Liu and Zhang [26] extend the attention modeling considering different attention to the left and right context of the given target. Tang et al. [64] use an end-to-end memory network for aspect-level sentiment classification, adding attention mechanisms. The method is able to understand the importance of each context word for the sentiment. Lei et al. [23] use a neural network method to extract pieces of input text as rationales (reasons) for review ratings. Li et al. [24] present an end-to-end approach. Ma et al. [28] present an interactive attention network (IAN) that considers attention on target and context. Chen et al. [9] use recurrent/dynamic attention network. Tay et al. [65] present a dyadic memory network (DyMemNN) that models dyadic interactions between aspect and context.

Opinion Expression Extraction. Different deep neural networks have also been introduced to addresses the problem of extracting opinion expressions. Yang and Cardie [77] use traditional shallow RNNs Irsoy and Cardie [21] use deep bi-RNN that outperformed [77]. Liu et al. [27] defined a model based on RNNs and word embedding. Wang et al. [71] combine RecNNs and CRF to extract aspect and opinion terms. Successively, Wang et al. [72] defined the CMLA method.

Sentiment Classification. To understand the polarity of the sentiment, it is often necessary to combine textual expressions with individually different polarities that influence each other. In addition, the polarity of terms can be dependent on the specific context and domain. Socher et al. [61] defined a Tree-based neural network based on RecNNs. Irsoy and Cardie [21] use deep RecNNs. Zhu et al. [83] present a neural network to combine compositional and non-compositional sentiment.

2.2 Transfer Learning for ABSA

With BERT [11], which obtained outstanding performances in multiple NLP tasks, pre-training with fine-tuning has become one of the most effective and used methods to solve NLP related problems. Compared to the word-level vectors (e.g. Word2Vec [32] released in 2013 and still quite popular, Glove [41], and FastText [7]) BERT trains sentence-level vectors and gets more information from context. BERT uses a bi-directional Transformer. Transformers were introduced from

Vaswani et at. [67]. After introducing BERT, many approaches based on it have been proposed by the natural language processing and understanding community [55].

Pre-trained language models (PTMs) can learn a good general representation from huge unlabeled corpora. Transfer learning enables for adapting the knowledge from a source task (or domain) to a target task (or domain) [57].

Xu et al. [76] propose a review reading comprehension (RRC) task and investigate the use of reviews for answering questions about sentiments of aspects, perform AE, and ASC. They adopt BERT as a base model and propose a joint post-training and fine-tuning approach to add both domain and task knowledge.

Rietzler et al. [54] extend the work by Xu et al. by further investigating the behavior of BERT models with post-training and fine-tuning in and cross domains, focusing on the ASC problem and considering the SemEval 2014 datasets [45].

In this paper, we follow the approaches presented in [54,76] and compare two approaches, transfer learning and weak-supervision, for addressing the ASC task.

2.3 Weak-Supervision for ABSA

Many papers have been presented to address the ABSA problem. To the best of our knowledge, few of these papers exploit weak-supervision methods.

García Pablos et al. [13] use some variations of [48] and [49] to perform AE and ASC. In particular, they use a double-propagation approach, and they model the obtained terms and their relations as a graph. Then, they apply the PageRank algorithm to score the obtained terms.

After, García Pablos et al. [14] perform AE task by bootstrapping a list of candidate domain aspect terms and using them to annotate the reviews of the same domain. The polarity detection is performed using a polarity lexicon exploiting the Word2Vec model [33] for each domain[2].

Then, García Pablos et al. [15] present a fully "almost unsupervised" ABSA system. Starting from a customer reviews dataset and a few words list of aspects they extract a list of words per aspect and two lists of positive and negative words for every selected aspect. It is based on a topic modeling approach combined with continuous word embeddings and a Maximum Entropy classifier.

Purpura et al. [47] perform the AE phase with a topic modeling technique called Non-negative Matrix Factorization. It allows the user to embed a list of seed words to guide the algorithm towards a more significant topic definition. The ASC is done by using a list of positive and negative words, with a few sentiment terms for each topic. This list is then extended with the Word2Vec model [33].

In [42], aspect and opinion lexicons are extracted from an unsupervised dataset belonging to the same domain as the target domain. The process is

[2] In [14] the addressed task is a bit different from ASC: the authors classify *entity-attribute* pair, where *entity* and *attribute* belong to predefined lists, e.g. food, price, location for *entity* and food-price, food-quality for *attribute*.

initialized with a seed lexicon of generic opinion terms. New aspect and opinion terms are extracted by using the dependency rules proposed in [49]. The opinion lexicon is then filtered and scored while the aspect lexicon can be modified by hand in a weakly supervised manner. ASC is performed on a target domain by detecting a direct or second-order dependency relation of any type between aspect-opinion pairs.

Unlike previous works, the proposed weak-supervision approach [60] simplifies and automates the sentiment terminology annotation making the ASC approach easily applicable in multiple domains. In addition, it enables the use of any discriminative and deep learning models.

3 Transfer Learning and Weak-Supervision Approaches for ASC

Insufficient supervised training data significantly limits the performance of the ASC task. To reduce the cost of human annotation in creating training sets, transfer learning exploits the knowledge obtained in other domains to avoid training parameters of deep learning algorithms from scratch, while weak-supervision provides semi-automatic data labeling methods to lower the manual effort. In this section, we present two approaches based on transfer learning and weak-supervision, respectively.

ASC aims to classify the sentiment polarity (*positive*, *negative*, or *neural*) related to an aspect from a review. The input in a couple $\langle d, g \rangle$ where d is a sentence and g is an aspect mentioned in d, The output is the polarity of the sentiment associated to the aspect g.

Our approach is based on BERT [11] and follows the pipeline defined in [76] and [54] for ASC. We define a standard training procedure as follows:

$$D_{post\text{-}training} \rightarrow D_{fine\text{-}tuning} \rightarrow D_{testing} \tag{1}$$

where $D_{post\text{-}training}$, $D_{fine\text{-}tuning}$, and $D_{testing}$ are three consecutive steps:

(i) **post-training**, the BERT pre-trained language model is post-trained on the specific domain dataset $D_{post\text{-}training}$ to obtain a language model with knowledge obtained from the considered dataset.
(ii) **fine-tuning**, the obtained model is fine-tuned on the specific labeled training dataset $D_{fine\text{-}tuning}$ to add task knowledge.
(iii) **testing**, the model is tested on the target test dataset $D_{testing}$.

BERT is a deep learning architecture built on Transformers [68] and it provides a language model pre-trained on a non-review knowledge, i.e., on Wikipedia and BooksCorpus dataset [84]. Post-training enables injecting into BERT the missing domain knowledge training the language model on an unsupervised corpus (i.e., without any manual annotations). BERT learns needed information from data and produces a new, more domain-related, language model. Xu et al. [76] show that post-training the model on a specific domain contributes to performance improvement.

The model resulting from the post-training is fine-tuned for the down-stream task to inject the task knowledge into the model. In fact, the idea behind BERT is to provide a pre-trained language model that can be further fine-tuned requiring almost no specific architecture for each end task. In our case, as describe and like in [76], we obtain a language model further post-trained on an additional unsupervised corpus. BERT learns needed information from data. To fine-tune BERT on the specific ASC end-task we just extend BERT with one extra task-specific layer, detailed in the following.

Let $x = (\texttt{[CLS]}, g_1, \ldots, g_m, \texttt{[SEP]}, d_1, \ldots, d_n, \texttt{[SEP]})$, where $, g_1, \ldots, g_m$ is the aspect (goal of our ABSA) having m tokens, and $d_1, \ldots, d_n$ is our document taken as input, which correspond to a review sentence containing that aspect g. Tokens of the input sentence and aspect are tokenized by the WordPiece algorithm [75]. `[CLS]` and `[SEP]` are two special tokens. The `[SEP]` token is used to separate two different inputs. Applying BERT, we obtain $h = \text{BERT}(x)$. `[CLS]` is used for classification problem and $h_{\texttt{[CLS]}}$ is the aspect-aware representation of the whole input through BERT [11]. The distribution of polarity is predicted with the added task-specific layer as $l = softmax(W \cdot h_{\texttt{[CLS]}} + b)$, where $W \in \mathbb{R}^{3*r_h}$ and $b \in \mathbb{R}^3$, with 3 the number of polarities (*positive*, *negative*, or *neural*). Softmax is applied along the dimension of labels on `[CLS]` : $l \in [0, 1]^3$.

The training of loss function is the cross-entropy on the polarities. Formula (2) is the cross-entropy function used to measures the discrepancy between a true distribution p and an estimated one q output of a classifier.

$$H(p, q) = -\sum_{x \in X} p(x) \log q(x) \tag{2}$$

It is noteworthy that, depending on the nature of the training set, the true distribution p can be a one-hot vector when we know the correct class as input or a probabilistic vector. In general, machine learning/deep learning methods use a human-annotated dataset. Therefore we are in the first case where the true distribution is a one-hot vector having the bit set to one to the correct class of the training example. Instead, as described in the following, for the weak-supervision approach, we will have a probability distribution over the classes as input.

3.1 Transfer Learning for ASC

A standard evaluation of machine learning models exploits train set and test set in the same domain, i.e., *in-domain*, and in particular standard approach uses the same distribution of the sets. In order to evaluate the capabilities and robustness of transfer learning approaches, models can be evaluated *cross-domain*. Thus, we apply cross-domain transfer learning for ASC and we use BERT language model fine-tuning, according to [54].

Following the standard training procedure defined in (1), unlike the in-domain case, in cross-domain settings we have $D_{fine\text{-}tuning}$ and $D_{testing}$ belonging to different domains. A special case of cross-domain is when post-training and

testing are in the same domain (for instance, both on Laptops or both on Restaurants). This case is named cross-domain adaptation.

3.2 Weak Supervision for ASC

Our proposed weak-supervision method [60], specifically designed for the ASC task of the ABSA problem, is grounded on data programming [52] that is a weak-supervised paradigm based on the concept of Labeling Functions (LFs). LFs are procedures, designed by data scientists and/or subject matter experts, that automatically assign labels to data based on domain knowledge embedded in the form of annotation rules.

More in detail, our method consists of a set of predefined, easy-to-use, and flexible LFs capable of automatically assigning sentiment to sentence-aspect pairs. The method is based on the ideas that: (i) it must require minimum NLP knowledge to the user, and (ii) it must be reusable in multiple domains with minimal effort for domain adaptation.

One of the most important characteristics of data programming is that LFs are noisy (e.g., different LFs may label the same data in different ways, LFs can label false positive examples). In this paper, to deal with the ASC task of ABSA problems by data programming, we use the Snorkel system [51] that enables handling the entire life-cycle of data programming tasks. Once LFs have been written, Snorkel applies them to data and automatically learns a generative model over these LFs, which estimates their accuracy and correlations, with no ground-truth data needed. It is noteworthy that Snorkel applies LFs as a generative process, which automatically de-noises the resulting dataset by learning the accuracy of the LFs along with their correlation structure. Thus, this process's output is a training set composed of probabilistic labels that can be used as input for deep learning algorithms that have to use a noise-aware loss function.

In our weak-supervised method LFs take as input pairs having the form $\langle d, g \rangle$ where d is a sentence and g is an aspect, and return triples having the form $\langle d, g, s \rangle$ where d and g are the sentence and the aspect respectively, and $s \in \{Positive, Negative, Neutral\} \cup \{Abstain\}$ is the sentiment obtained in output. A LF can choose to abstain if it doesn't have sufficient information to label a sentiment.

The proposed method recognizes and exploits chunks in the input text, in particular noun phrase (NP) verb phrase (VP). To recognize chunks the Stanford CoreNLP Parser [30] is used. The aspects g are recognized and assigned to chunks. To compute sentiment polarity s of chunks, external sentiment analyzers are exploited, such as Stanford CoreNLP [30], TextBlob[3], NLTK [5], and Pattern[4]. To assign s to the pair $\langle d, g \rangle$ two different simple and intuitive strategies are applied. The first computes the polarity of every single chunk. For instance, if all chunks have the same polarity, the method returns the corresponding label. If chunks have mixed *Positive* and *Negative* polarity the method returns the

[3] https://textblob.readthedocs.io/.
[4] https://www.clips.uantwerpen.be/pages/pattern-en/.

Abstain label. When there are *Neutral* chunks mixed with at least one *Positive* chunk the method returns the *Positive* label, and the same happens for mixed *Neutral* and *Negative* chunks. The second strategy computes the global polarity of the resulting text string. Considering the two different strategies and the four exploited sentiment analyzers, the approach includes a total of 8 LFs. For more details, refer to [60]. It is noteworthy that created labeling function templates are simple and powerful. They enable to reuse of existing knowledge embedded in already available NLP tools to create new training sets.

Table 1 and Table 2 show statistics about the LFs when applied to laptops and restaurants datasets, respectively. In particular, columns of the tables represent coverage, overlaps, conflicts, and empirical accuracy of LFs when they are executed to a small number (150 in our case) of manually labeled examples called dev set. Rows in the tables correspond to the eight LFs we defined by using NLP tools and strategies described above, where each row of the tables contains values computed for a specific labeling function.

Table 1. LFs application stats on laptops domain. Source [60].

LF srategy	Coverage	Overlaps	Conflicts	Emp. Acc.
$LF_{(StanfordCoreNLP,\ FirstStrategy)}$	0.7667	0.7667	0.5267	0.5478
$LF_{(StanfordCoreNLP,\ SecondStrategy)}$	0.7933	0.7933	0.5333	0.5042
$LF_{(TextBlob,\ FirstStrategy)}$	0.7400	0.7400	0.4867	0.4775
$LF_{(TextBlob,\ SecondStrategy)}$	0.7933	0.7933	0.5333	0.5042
$LF_{(NLTK,\ FirstStrategy)}$	0.7533	0.7533	0.4933	0.4956
$LF_{(NLTK,\ SecondStrategy)}$	0.7933	0.7933	0.5333	0.4874
$LF_{(Pattern.en,\ FirstStrategy)}$	0.7467	0.7467	0.4933	0.4821
$LF_{(Pattern.en,\ SecondStrategy)}$	0.7933	0.7933	0.5333	0.4790

Experiments on laptops in Table 1 show that LFs have about 77% of coverage and 50% of empirical accuracy, while Table 2 shows that restaurants have a

Table 2. LFs application stats on restaurants domain. Source [60].

LF strategy	Coverage	Overlaps	Conflicts	Emp. Acc.
$LF_{(StanfordCoreNLP,\ FirstStrategy)}$	0.6200	0.6200	0.3867	0.4946
$LF_{(StanfordCoreNLP,\ SecondStrategy)}$	0.6800	0.6800	0.4200	0.5882
$LF_{(TextBlob,\ FirstStrategy)}$	0.6667	0.6667	0.4067	0.5100
$LF_{(TextBlob,\ SecondStrategy)}$	0.6800	0.6800	0.4200	0.5098
$LF_{(NLTK,\ FirstStrategy)}$	0.6667	0.6667	0.4133	0.5000
$LF_{(NLTK,\ SecondStrategy)}$	0.6800	0.6800	0.4200	0.5098
$LF_{(Pattern.en,\ FirstStrategy)}$	0.6667	0.6667	0.4067	0.5100
$LF_{(Pattern.en,\ SecondStrategy)}$	0.6800	0.6800	0.4200	0.5000

coverage of about 69% and empirical accuracy of 52%. Results on coverage and empirical accuracy suggest that defined LFs work properly and can be used to annotate the two datasets.

The result of the labeling process is a matrix of labels $\Lambda \in (\{Positive,\ Negative,\ Neutral\} \cup \{Abstain\})^{mxn}$, where m is the cardinality of the training set and n is the number of LFs. This matrix is the input of the Snorkel generative model [51]. Such model produces a list of probabilistic training labels $\tilde{Y} = (\tilde{y}_1, \ldots, \tilde{y}_m)$, where each $\tilde{y}_i \in \mathbb{R}^3$ is a probability distribution over the classes $\{Positive,\ Negative,\ Neutral\}$.

This probabilistic dataset is the input of discriminative models that use noise-aware loss functions described in Formula (2), where p is a probability distribution for a training example over the classes obtained as output from the Snorkel generative model.

4 Experiments

In this section, we compare the two presented approaches, i.e., transfer learning and weak-supervision, used to reduce the need for annotated training sets in the target domain to perform aspect sentiment classification (ASC). In particular, we describe the used datasets, the applied parameters, and the performed experiments for ASC task.

4.1 Dataset

For domain knowledge post-training, we used those presented in [76] consisting of around 1 Million of examples derived from Amazon laptop reviews [17], and around 2.5 Million of examples derived from 700K Yelp Dataset Challenge reviews[5].

To perform fine-tuning, we use two disjoint domains, laptops and restaurants, of SemEval task 4 subtask two datasets [45], which we simply call SemEval from now. Each dataset is already split into training, development, and test set. Statistics with the number of examples for each set are shown in Table 3. Each partition was hand-labeled by subject matter experts with labels within the set $\{Positive,\ Negative,\ Neutral\}$.

Table 3. Number of examples for each dataset. Source [60].

Domain	Train set	Dev set	Test set
Laptops	2163	150	638
Restaurants	3452	150	1120

[5] https://www.yelp.com/dataset.

In transfer learning experiments, we use ground truth labels, i.e., human-annotated labels of training and test set provided by SemEval. Instead, we replace original annotations in the training sets with probability distributions computed by Snorkel generative models in weak-supervision experiments. We use the dev set for tuning the LFs and the hyper-parameters of the generative models. We calculate metrics on the test set.

4.2 Experimental Settings

Cross-domain Transfer Learning Models. In our experiments we consider the two domains *laptops* and *restaurants*, and we test the different combination of our procedure $D_{post\text{-}training} \rightarrow D_{fine\text{-}tuning} \rightarrow D_{testing}$, defined in (1). So, for instance, *laptops* → *laptops* → *restaurants* is a model post-trained and fine-tuned on the laptops domain and tested on the restaurants domain. The best results can be obtained when in-domain or joint-domains [54]. In this paper, because the purpose is to reuse data on different domains, we have tried all possible cross-domain combinations (i.e., *laptops* → *laptops* → *restaurants*, *laptops* → *restaurants* → *laptops*, *restaurants* → *restaurants* → *laptops*, and *restaurants* → *laptops* → *restaurants*).

The backbone architecture for the post-training step is uncased BERT-base [11]. Models were post-trained by using the following settings: the maximum length of an example is 320, batch size 16 for each type of knowledge, the learning rate set to $3e{-}5$ leveraging the Adam optimizer, the number of post-training steps were 70.000 and 140.000 for the laptops and restaurants domains respectively, which corresponds about an epoch for each domain.

Generative Model. The hyperparameters of the generative model for the weak-supervision approach are searched through a grid search. A search configuration can be formally defined as a triple $\langle e, lr, o \rangle$ where $e \in \{100, 200, 500, 1000, 2000\}$ is the number of epochs, $lr \in \{0.01, 0.001, 0.0001\}$ is the learning rate, and $o \in \{sgd, adam, adamax\}$ is the optimizer. To tune the hyperparameters, we use the dev sets of SemEval.

The best configuration for laptops domain we found is $\langle 100, 0.01, adamax \rangle$. With these settings, the generative model applied to SemEval produced a dataset of 1702 probabilistic examples on laptops. The best configuration for restaurants domain we obtained is $\langle 100, 0.001, adamax \rangle$. The cardinality of the probabilistic dataset produced by Snorkel on restaurants is 2471.

Table 4 shows for each considered training set the number of examples obtained by Snorkel classified in *Positive*, *Negative* or *Neutral* labels. Because the labels are probabilistic, we consider for each training example the class with highest probability.

Table 4. Number of (the most probable) examples for each label. Source [60].

Train set	Positive	Negative	Neutral
Laptops	627	448	627
Restaurants	1371	301	792
Sampled restaurants	700	301	700

The restaurants' training set obtained by Snorkel is unbalanced. Therefore, we limit the number of *Positive* and *Neutral* examples to 700. The results discussed discussed in the next section are computed by averaging the metrics of 10 models trained with *Positive* and *Neutral* examples obtained by different sampling strategies.

During the fine-tuning step of both considered approaches, like in [76], we trained a simple softmax classifier whose output belongs to $\mathbb{R}^3$, where 3 is the number of polarities, on top of BERT post-trained models. We fine-tuned the discriminative models for 4 epochs using a batch size of 32 and the Adam optimizer with a learning rate of 3e-5. Results are obtained by averaging 10 runs sampling the mini-batches differently.

4.3 Results and Discussion

Table 5 shows results of transfer learning (Subsect. 3.1) with focus on cross-domain adaptation case considering laptops and restaurants SemEval dataset.

Table 5. Results of the experiments on laptops and restaurants using transfer learning methodology.

Test Set	**Laptops**		**Restaurants**	
Training Set for fine-tuning	**Restaurants**		**Laptops**	
	Accuracy	**Macro F1**	**Accuracy**	**Macro F1**
Xu_post-training Laptops	73.31	67.34	78.73	71.00
Xu_post-training Restaurants	73.95	70.74	81.00	71.79

In particular, in Table 5, accuracy and Macro F1 are computed for cross-domain configurations and considering both language models fine-tuned on the target or different domain. The cells with gray background correspond to the case of a cross-domain adaptation, where the language model is post-trained on the same target domain. Xu_post-training Laptops and Xu_post-training Restaurants are models post-trained with unsupervised data on laptops and restaurants respectively. So, for instance, the model post-trained on the laptops, then fine-tuned on the restaurants, and finally tested on the laptops, i.e., *laptops* → *restaurants* → *laptops*, obtains accuracy 73.31 and Macro F1 67.34. It is noteworthy that for *laptops* domain, we obtain better results performing post-training and

fine-tuning on the different domain *restaurants* wrt using the same dataset for pre-training *laptops*. This is explainable by considering that *restaurants* is a wider dataset, and it seems to be sufficient to learn language model that generalizes well cross-domain.

Table 6 shows results of accuracy and macro F1 on laptops and restaurants domains with the weak-supervision approach (Subsect. 3.2) compared with the approach proposed in [76] but limiting the size of the datasets used for fine-tuning.

Table 6. Results of the experiments on laptops and restaurants respectively, by using weak-supervision methodology.

Model type	Laptops		Restaurants	
	Accuracy	Macro F1	Accuracy	Macro F1
Xu_150	57.77	52.60	48.62	39.15
Xu_300	71.59	68.00	69.49	60.79
Xu_450	—	—	77.93	67.57
Xu_Weak	69.36	65.37	75.39	67.33

More in detail, Xu_Weak, shown in Table 6, is the model trained with probabilistic labels computed by our weak-supervision method. The weak-supervision approach uses 150 examples belonging to SemEval dev sets in order to fine-tune the Snorkel generative model. Xu_150, Xu_300, Xu_450 are the models obtained by fine-tuning [76] with a different number of hand-labeled examples (150, 300, and 450 respectively) belonging to the in-domain SemEval training set. To obtain balanced datasets, examples are randomly extracted. In particular, we sampled the train set 10 times averaging the results. The weak-supervision approach reaches better results in the restaurants domain. It could be due to the availability of more examples obtained in post-trained for the restaurant domain.

Summarizing, as shown in these experiments, when different domains with commons features are available, it is convenient to use cross-domain transfer learning. In fact, a lot of syntactic features, structural information, and semantic language relationships can be learned and exploited cross-domains. Instead, weak-supervision remains a useful way to deal with real-world use cases where hand-labeled training sets with specific knowledge are needed and may become a bottleneck for implementing deep learning models. The main advantage of discriminative models for automatically label probabilistic examples is the easy capability to scale the number of labeled examples needed to get better performances.

5 Conclusions

In this paper, we described two approaches, based on transfer learning and weak supervision, both having the goal to reduce the manual effort needed to build an

annotated dataset for the ASC problem. We adopted a general transfer learning approach that exploits the knowledge obtained in other domains to avoid training parameters of deep learning algorithms from scratch. At the same time, we presented a weak-supervision approach, specifically designed for the ASC task of the ABSA problem, that provides semi-automatic data labeling methods to lower the manual effort. To test effectiveness and applicability, we extensively tested proposed approaches on the laptops and restaurants dataset of SemEval task 4 subtask 2. Experiments have shown that when different domains with commons features are available, it is convenient to use cross-domain transfer learning. In fact, a lot of syntactic features, structural information, and semantic language relationships can be learned and exploited cross-domains. Instead, weak-supervision remains a useful way to deal with real-world use cases where hand-labeled training sets with specific knowledge are needed and may become a bottleneck for implementing deep learning models.

References

1. Abdul-Mageed, M., Ungar, L.: EmoNet: fine-grained emotion detection with gated recurrent neural networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers, vol. 1, pp. 718–728 (2017)
2. Al-Moslmi, T., Omar, N., Abdullah, S., Albared, M.: Approaches to cross-domain sentiment analysis: a systematic literature review. IEEE Access **5**, 16173–16192 (2017)
3. Bach, S.H., et al.: Snorkel drybell: a case study in deploying weak supervision at industrial scale. In: Proceedings of the 2019 International Conference on Management of Data, pp. 362–375. ACM (2019)
4. Balazs, J.A., Velásquez, J.D.: Opinion mining and information fusion: a survey. Inf. Fusion **27**, 95–110 (2016)
5. Bird, S.: NLTK: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive presentation sessions, pp. 69–72. Association for Computational Linguistics (2006)
6. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 440–447 (2007)
7. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
8. Borele, P., Borikar, D.A.: A survey on evaluating sentiments by using artificial neural network (2016)
9. Chen, P., Sun, Z., Bing, L., Yang, W.: Recurrent attention network on memory for aspect sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 452–461 (2017)
10. Deng, L., Wiebe, J.: MPQA 3.0: an entity/event-level sentiment corpus. In: HLT-NAACL, pp. 1323–1328 (2015)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2019)

12. Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., Xu, K.: Adaptive recursive neural network for target-dependent twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Volume 2, Short Papers, vol. 2, pp. 49–54 (2014)
13. García-Pablos, A., Cuadros, M., Rigau, G.: V3: unsupervised generation of domain aspect terms for aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 833–837 (2014)
14. Garcia-Pablos, A., Cuadros, M., Rigau, G.: V3: unsupervised aspect based sentiment analysis for SemEval2015 task 12. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 714–718 (2015)
15. García-Pablos, A., Cuadros, M., Rigau, G.: W2VLDA: almost unsupervised system for aspect based sentiment analysis. Expert Syst. Appl. **91**, 127–137 (2018)
16. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford 1(12) (2009)
17. He, R., McAuley, J.: Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. In: Proceedings of the 25th International Conference on World Wide Web, pp. 507–517 (2016)
18. Hltcoe, J.: SemEval-2013 task 2: sentiment analysis in twitter, Atlanta, Georgia, USA, p. 312 (2013)
19. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177. ACM (2004)
20. Hussein, D.M.E.D.M.: A survey on sentiment analysis challenges. J. King Saud Univ. Eng. Sci. **30**(4), 330–338 (2016)
21. Irsoy, O., Cardie, C.: Deep recursive neural networks for compositionality in language. In: Advances in Neural Information Processing Systems, pp. 2096–2104 (2014)
22. Jo, Y., Oh, A.H.: Aspect and sentiment unification model for online review analysis. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 815–824. ACM (2011)
23. Lei, T., Barzilay, R., Jaakkola, T.: Rationalizing neural predictions. arXiv preprint arXiv:1606.04155 (2016)
24. Li, C., Guo, X., Mei, Q.: Deep memory networks for attitude identification. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, pp. 671–680. ACM (2017)
25. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
26. Liu, J., Zhang, Y.: Attention modeling for targeted sentiment. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, vol. 2, pp. 572–577 (2017)
27. Liu, P., Joty, S., Meng, H.: Fine-grained opinion mining with recurrent neural networks and word embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1433–1443 (2015)
28. Ma, D., Li, S., Zhang, X., Wang, H.: Interactive attention networks for aspect-level sentiment classification. arXiv preprint arXiv:1709.00893 (2017)
29. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 142–150. Association for Computational Linguistics (2011)

30. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60 (2014)
31. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014)
32. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
33. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
34. Oro, E., Pizzuti, C., Procopio, N., Ruffolo, M.: Detecting topic authoritative social media users: a multilayer network approach. IEEE Trans. Multimedia **20**(5), 1195–1208 (2017)
35. Oro, E., Pizzuti, C., Ruffolo, M.: A methodology for identifying influencers and their products perception on twitter. In: ICEIS (2018)
36. Oro, E., Ruffolo, M., Pupo, F.: A cognitive automation approach for a smart lending and early warning application. In: Poulovassilis, A., et al. (eds.) Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference, Copenhagen, Denmark, March 30, 2020. CEUR Workshop Proceedings, vol. 2578. CEUR-WS.org (2020). http://ceur-ws.org/Vol-2578/DARLIAP6.pdf
37. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, p. 271. Association for Computational Linguistics (2004)
38. Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 115–124. Association for Computational Linguistics (2005)
39. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retrieval **2**(1–2), 1–135 (2008)
40. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical Methods in Natural Language Processing-Volume 10, pp. 79–86. Association for Computational Linguistics (2002)
41. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. EMNLP **14**, 1532–1543 (2014)
42. Pereg, O., Korat, D., Wasserblat, M., Mamou, J., Dagan, I.: ABSApp: a portable weakly-supervised aspect-based sentiment extraction system. arXiv preprint arXiv:1909.05608 (2019)
43. Pontiki, M., et al.: SemEval-2016 task 5: aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 19–30 (2016)
44. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: SemEval-2015 task 12: aaspect based sentiment analysis. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 486–495 (2015)

45. Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.: SemEval-2014 task 4: aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 27–35. Association for Computational Linguistics, Dublin, Ireland (2014). https://doi.org/10.3115/v1/S14-2004
46. Poria, S., Hazarika, D., Majumder, N., Mihalcea, R.: Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. arXiv preprint arXiv:2005.00357 (2020)
47. Purpura, A., Masiero, C., Susto, G.A.: WS4ABSA: an NMF-based weakly-supervised approach for aspect-based sentiment analysis with application to online reviews. In: Soldatova, L., Vanschoren, J., Papadopoulos, G., Ceci, M. (eds.) Discovery Science. Lecture Notes in Computer Science, vol. 11198, pp. 386–401. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01771-2_25
48. Qiu, G., Liu, B., Bu, J., Chen, C.: Expanding domain sentiment lexicon through double propagation. In: Twenty-First International Joint Conference on Artificial Intelligence (2009)
49. Qiu, G., Liu, B., Bu, J., Chen, C.: Opinion word expansion and target extraction through double propagation. Comput. Linguist. **37**(1), 9–27 (2011)
50. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100, 000+ questions for machine comprehension of text. In: EMNLP (2016)
51. Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S., Ré, C.: Snorkel: rapid training data creation with weak supervision. Proc. VLDB Endow. **11**(3), 269–282 (2017)
52. Ratner, A.J., De Sa, C.M., Wu, S., Selsam, D., Ré, C.: Data programming: creating large training sets, quickly. In: Advances in Neural Information Processing Systems, pp. 3567–3575 (2016)
53. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl.-Based Syst. **89**, 14–46 (2015)
54. Rietzler, A., Stabinger, S., Opitz, P., Engl, S.: Adapt or get left behind: domain adaptation through bert language model finetuning for aspect-target sentiment classification. arXiv preprint arXiv:1908.11860 (2019)
55. Rogers, A., Kovaleva, O., Rumshisky, A.: A primer in bertology: what we know about how bert works. arXiv preprint arXiv:2002.12327 (2020)
56. Rosenthal, S., Farra, N., Nakov, P.: SemEval-2017 task 4: sentiment analysis in twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 502–518 (2017)
57. Ruder, S.: Neural transfer learning for natural language processing. Ph.D. thesis, NUI Galway (2019)
58. Ruder, S., Ghaffari, P., Breslin, J.G.: A hierarchical model of reviews for aspect-based sentiment analysis. arXiv preprint arXiv:1609.02745 (2016)
59. Ruder, S., Peters, M.E., Swayamdipta, S., Wolf, T.: Transfer learning in natural language processing. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, pp. 15–18 (2019)
60. Ruffolo, M., Visalli, F.: A weak-supervision method for automating training set creation in multi-domain aspect sentiment classification. In: ICAART (2020)
61. Socher, R., et al.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), vol. 1631, p. 1642. Citeseer (2013)
62. Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.F., Pantic, M.: A survey of multimodal sentiment analysis. Image Vis. Comput. **65**, 3–14 (2017)

63. Tang, D., Qin, B., Feng, X., Liu, T.: Effective lstms for target-dependent sentiment classification. arXiv preprint arXiv:1512.01100 (2015)
64. Tang, D., Qin, B., Liu, T.: Aspect level sentiment classification with deep memory network. arXiv preprint arXiv:1605.08900 (2016)
65. Tay, Y., Tuan, L.A., Hui, S.C.: Dyadic memory networks for aspect-based sentiment analysis. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 107–116. ACM (2017)
66. Van Kleef, E., Van Trijp, H.C., Luning, P.: Consumer research in the early stages of new product development: a critical review of methods and techniques. Food Qual. Prefer. **16**(3), 181–201 (2005)
67. Vaswani, A.,et al.: Attention is all you need. In: NIPS (2017)
68. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
69. Vo, D.T., Zhang, Y.: Target-dependent twitter sentiment classification with rich automatic features. In: IJCAI, pp. 1347–1353 (2015)
70. Wang, W., Pan, S.J.: Recursive neural structural correspondence network for cross-domain aspect and opinion co-extraction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 2171–2181 (2018)
71. Wang, W., Pan, S.J., Dahlmeier, D., Xiao, X.: Recursive neural conditional random fields for aspect-based sentiment analysis. arXiv preprint arXiv:1603.06679 (2016)
72. Wang, W., Pan, S.J., Dahlmeier, D., Xiao, X.: Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In: AAAI, pp. 3316–3322 (2017)
73. Wang, Y., Huang, M., Zhao, L., et al.: Attention-based lstm for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 606–615 (2016)
74. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Lang. Resour. Eval. **39**(2), 165–210 (2005)
75. Wu, Y., et al.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
76. Xu, H., Liu, B., Shu, L., Yu, P.S.: Bert post-training for review reading comprehension and aspect-based sentiment analysis. arXiv preprint arXiv:1904.02232 (2019)
77. Yang, B., Cardie, C.: Extracting opinion expressions with semi-markov conditional random fields. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1335–1345. Association for Computational Linguistics (2012)
78. Yang, M., Tu, W., Wang, J., Xu, F., Chen, X.: Attention based LSTM for target dependent sentiment classification. In: AAAI, pp. 5013–5014 (2017)
79. Ying, D., Yu, J., Jiang, J.: Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction (2017)
80. Zagibalov, T., Carroll, J.: Automatic seed word selection for unsupervised sentiment classification of Chinese text. In: Proceedings of the 22nd International Conference on Computational Linguistics, vol. 1, pp. 1073–1080. Association for Computational Linguistics (2008)
81. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: a survey. Data Mining and Knowledge Discovery. Wiley Interdisciplinary Reviews. Wiley, Hoboken, New Jersey (2018)
82. Zhang, M., Zhang, Y., Vo, D.T.: Neural networks for open domain targeted sentiment. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 612–621 (2015)

83. Zhu, X., Guo, H., Sobhani, P.: Neural networks for integrating compositional and non-compositional sentiment in sentiment composition. In: Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, pp. 1–9 (2015)
84. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 19–27 (2015)