Neural Processing Letters (2021) 53:4207–4224
https://doi.org/10.1007/s11063-021-10596-6

BERT Based Semi-Supervised Hybrid Approach for Aspect and Sentiment Classification

Avinash Kumar[1] · Pranjal Gupta[1] · Raghunathan Balan[1] · Lalita Bhanu Murthy Neti[1] · Aruna Malapati[1]

Accepted: 12 July 2021 / Published online: 23 July 2021

Abstract
Aspect-based sentiment analysis (ABSA) includes two sub-tasks namely, aspect extraction and aspect-level sentiment classification. Most existing works address these sub-tasks independently by applying a supervised learning approach using labeled data. However, obtaining such labeled sentences is difficult and extremely expensive. Hence, it is important to solve ABSA without taking a dependency on labeled sentences. In this work, we propose a three-step semi-supervised hybrid approach that jointly detects an aspect and its associated sentiment in a given review sentence. The first step of our approach takes a small set of seed words for each aspect and sentiment class to construct respective semantically coherent class vocabularies. The second step makes use of these constructed vocabularies along with POS tags to label a subset of sentences from the training corpus. As we adopt a semi-automated approach to label the data, this process may induce noise in the labels during the annotation. In the last step, we use such labeled sentences to build a noise-robust deep neural network for aspect and sentiment classification. We conduct experiments on two real data sets to verify the effectiveness of our model (https://github.com/Raghu150999/UnsupervisedABSA).

Keywords ABSA · Language model · BERT · Deep neural network · Attention mechanism · Unsupervised learning

1 Introduction

In the era of the digital revolution, social media has become one of the preferred platforms for people to express their opinion. Online reviews play a key role in influencing everyone's decision-making process. Sentiment analysis aims to identify polarity (e.g., positive vs. negative) of opinion expressed in a given piece of text. Aspect Based Sentiment Analysis (ABSA) [1–3] takes one step further to analyze people's opinions and sentiments about a product, or an event in a fine-grained manner. ABSA enables business owners to take appropriate decisions to grow their business. For example, a restaurant owner can know the reason

✉ Avinash Kumar
p20150507@hyderabad.bits-pilani.ac.in

[1] Birla Institute of Technology & Science, Pilani—Hyderabad, Hyderabad 500078, India

behind the recent drop in popularity of his restaurant by analyzing the sentiment related to various aspects of a restaurant like room, staff, location, food quality, etc. Opinion without knowing the fine-grained target or topic provides limited insight of the user's viewpoint [4]. Aspect category detection is a key task of ABSA that helps in identifying the generic topics or category (i.e., characteristics of the product or service) discussed in the opinionated text. For example, in the given sentences *"The food is yummy."* and *"Price was reasonable"*, a positive sentiment is conveyed using opinion term *"yummy"* and *"reasonable"* about the aspect categories *"Food"* and *"Price"*, respectively. Identifying the aspect category helps to do aspect-based sentiment analysis.

Previous work for aspect category detection can primarily be categorized into two approaches: supervised, and unsupervised. Supervised methods require a significant number of labeled sentences to train the model. Moreover, obtaining such labeled sentences is difficult and extremely expensive. Unsupervised methods are used to avoid dependency on labeled data. Among previous unsupervised approaches, Latent Dirichlet Allocation (LDA) [5] based topic modeling has been widely used for identifying the hidden aspect categories. Many variants of LDA [6–8] have shown fairly good results for this task. Recently, neural network topic models [9–11] have been used for aspect category detection tasks and have shown better performance than LDA based approaches. He et al. in [12] present a state-of-the-art neural network-based approach that explicitly encodes the tokens into word-embeddings and uses dimension reduction to extract the most important aspect categories. Tulkens et al. [13] introduces a Radial Basis Function (RBF) kernel-based single-head attention approach that uses a POS tagger and domain-specific word embeddings. Both these models [12,13] utilize Word2Vec [14] word embeddings, which provides a general static semantic meaning of the word, and use the same to encode the global context of the sentence.

Generally, the ABSA task consists of two sub-tasks, namely, aspect extraction and aspect sentiment classification. In order to solve the ABSA task completely without using labeled data, it is not only important to detect aspect categories but also to identify its associated sentiment. Some researchers [7,15–17] have adopted a joint modeling approach by integrating both the subtasks and built unsupervised models to solve the ABSA task completely. Despite the effectiveness of joint modeling methods, existing methods have ignored the contextual meaning of the word that has impacted the performance of aspect and sentiment classification tasks. A word can have a different meaning in different sentences. For example, in the review sentences *they delivered us food really fast* and *"this restaurant is known for its yummy fast food"* , word *"fast"* is used in two different contexts. In the first sentence, it is referring to *"speed of service"* while in the second sentence it is used for *"food item"*. Hence, considering the contextual meaning of the word becomes important to identify the aspect category and sentiment precisely. Pre-trained language model (LM) BERT [18] makes use of the contextual meaning of a word in the given sentence to generate the word embedding. Xu et al. in [19] show that adding domain-specific information in BERT can boost its performance in ABSA.

Given the above-discussed point, our work is motivated to provide end-to-end solutions to ABSA tasks without taking a dependency on labeled sentences. In this paper, we leverage power of post-trained, domain knowledge BERT (DK-BERT) and present a simple and highly efficient semi-supervised hybrid *CASC* approach for Context aware Aspect category and Sentiment Classification. Our model is built in a simple three-step process: (1) We take a small set of seed words for each aspect and sentiment class and then construct a semantically coherent vocabulary for each class using BERT masked language model (MLM). (2) A set of sentences in the training corpus are labeled with aspect and sentiment using POS tags and class vocabularies constructed in the previous step. (3) In the last step, we make use of labeled sentences of the training corpus and jointly train the BERT-based deep neural network for aspect and sentiment classification. Since we apply a semi-automated approach to label the sentences it may induce noise in the annotation process. Hence, to address this we use a *noise-robust loss function* to train our deep neural network. We have evaluated the performance of our proposed method on two real-life datasets and results show that it outperforms existing methods.

The main contributions of our work can be summarized as follows:

- We propose a semi-supervised hybrid aspect and sentiment classification model *CASC* using domain-knowledge BERT and a noise-robust loss function. Our proposed model does not need any labeled documents but only a set of small seed words for each class.
- We investigate the effectiveness of domain-knowledge BERT embedding, MLM, deep neural network, and a small number of seed words in the performance of *CASC*.
- We conduct extensive experiments on two real-world datasets, and results show *CASC* outperforms the existing methods.

The rest of this paper is organized as follows, after discussing related work in Sect. 2, we define task definition in Sect. 3 and present a detailed description of our proposed model, in Sect. 4. In Sect. 5 and 6, we discuss the details of our extensive experiments and do the analysis of results. Finally, we summarize our work in Sect. 7.

2 Related Work

Aspect Based Sentiment Analysis comprises of two sub-tasks: Aspect Extraction and Sentiment Classification. Previous works have dealt with them individually [12] and integrated them in a pipelined approach [20]. We will first review the works on aspect extraction and then review joint modeling approaches.

2.1 Aspect Extraction

Aspect category detection is a key task of ABSA that helps in identifying the generic topics or aspect category (i.e., characteristics of the product or service) discussed in the opinionated text. Previous works in aspect category detection can be categorized into three ways: rule-based, supervised and unsupervised methods. Recent methods have used neural models for aspect extraction task. He et al. [12], proposed the ABAE model where sentence-level information is encoded through the use of attention mechanism. Aspect categories are learned by reconstructing the sentence embedding in a way similar to that of autoencoders. Tulkens et al. [13], introduced a Radial Basis Function (RBF) kernel attention method that employs POS Tagger and domain-specific word embeddings for detecting aspect categories. These methods lack in two ways: first, these methods are hindered by the fact that the learned aspects do not align themselves with the user-defined aspect category labels, and hence, additional human efforts is required for mapping topics to aspects. Second, all these methods use the general static semantic meaning of words to encode sentence information. These encodings lack contextual and positional information of words and thus can degrade performance of the method.

Several semi-unsupervised approaches solve the first issue (i.e. alignment between topic and aspect) by using seed words as a form of regularisation in the training process. Angelidis et al. [21] extends ABAE and uses a weighted average of seed word embeddings to initialize aspect embeddings. Karamanolakis et al. [22] co-trains a bag-of-words classifier and an embedding based neural classifier to generalize the seed word supervision. However, these methods do not learn aspect-specific sentiment words which can provide more fine-grained information and thus improve the performance of aspect extraction.

2.2 Joint Learning of Aspect and Sentiment

Joint modeling of aspect category and sentiment classification helps in predicting the aspect category and the sentiment towards that aspect category. Previous studies that jointly extract aspect and sentiments are mostly LDA-based methods. Xu et al. [23] adapt LDA by introducing sentiment-related variables and integrating sentiment prior knowledge. Zhao et al. [24] include aspect-specific opinion models along with aspect models in a generative process. Wang et al. [25] introduced a Boltzmann Machine-based approach to extract aspect and sentiment.

Recent methods use a weakly-supervised setting for the joint extraction task. Garcıa-Pablos et al. [16] uses brown clustering to separate aspect and opinion terms and construct biased hyperparameters α and β by embedding similarity. Meng et al. JASA [26] extends ABAE model to learn aspect/sentiment representations. To enhance their performance, they interact with the aspect and sentiment representations to learn aspect-specific opinion words (such as delicious). Meng et al. [17] propose JASen model, which first learns joint topic embedding by imposing regularizations to encourage topic distinctiveness, and then use neural models to generalize the word-level discriminative information by pre-training the classifiers with embedding-based predictions and self-training them on unlabeled data.

3 Task Definition

We formulate our semi-supervised hybrid aspect and sentiment identification approach as a multi-class classification problem. The input contains corpus $\mathcal{D} = [X_1, X_2, \ldots, X_N]$ of N unlabeled review sentences from a domain (e.g., restaurant or laptop). Given the set of aspect categories A, a small list of seed words L_a for each aspect category $a \in A$, and sentiment polarities S along with a small list of seed words L_s for each sentiment polarity $s \in S$. The aim is to predict a pair of (a, s) for a given unseen review sentence.

4 Our Method

In this section, we will describe our *CASC* method. Our method involves a three-step process: (1) Class vocabulary construction using seed words, (2) Labeled data preparation, and (3) Building a joint deep neural network for aspect and sentiment identification. Figure 1 describes the first two steps of our method where labels are assigned to the sentences using a semi-automated approach. Figure 2 shows the architecture of a deep neural network that is built in the last step using the labels of the sentences generated in the previous steps. We will introduce these steps in detail in further sections.

4.1 Class Vocabulary Construction Using Seed Words

We take a small set of seed words L_a associated with $a \in A$ aspect and find the set of sentences $X_a \subset \mathcal{D}$ which contains any of the given seed words. Intuitively, in a given sentence, a word can be replaced with another word that carries the same contextual meaning. Motivated with

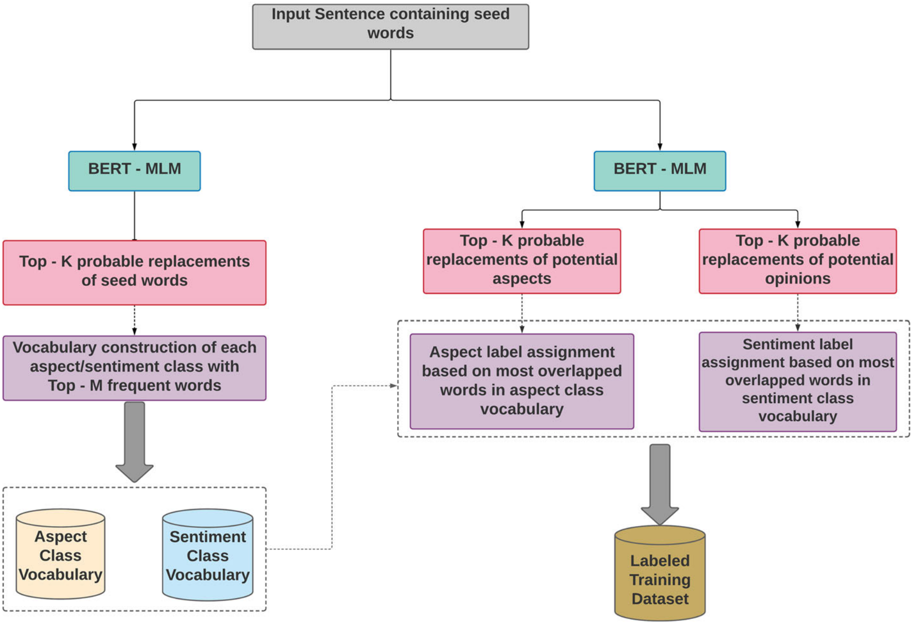

Input Sentence containing seed words
BERT - MLM
BERT - MLM
Top - K probable replacements of seed words
Top - K probable replacements of potential aspects
Top - K probable replacements of potential opinions
Vocabulary construction of each aspect/sentiment class with Top - M frequent words
Aspect label assignment based on most overlapped words in aspect class vocabulary
Sentiment label assignment based on most overlapped words in sentiment class vocabulary
Aspect Class Vocabulary
Sentiment Class Vocabulary
Labeled Training Dataset

Fig. 1 Labeled data preparation

this, we pass the sentence $X_i \in X_a$ to the post-trained Domain Knowledge BERT masked language model(MLM) to predict the words for finding replacement candidates of all the tokens present in the sentence.

$$P_i = BERT\text{-}MLM(X_i) \tag{1}$$

$$R_i = TopK(Filter(P_i)) \tag{2}$$

BERT-MLM provides replacement probability scores $P \in \mathbb{R}^{|X_i| \times |V|}$ where V is the vocabulary set used by BERT. Replacements generated by BERT-MLM are highly contextual and coherent with the original token. We take replacement candidates of only those tokens in the sentence which are present in L_a. The replacement candidates are then passed through the $Filter$ and $TopK$ function in sequence to remove stop words and punctuation, and then top K words are selected as replacement candidates R_i based on their probability scores. We do a union of all replacement words R_i and create a frequency table by sorting the words in decreasing order of their frequency. Finally, we take the top M words to construct vocabulary V_a for an aspect category a. We construct the vocabulary V_s for each sentiment polarity $s \in S$ in the same way as that of aspect category vocabulary construction. Further, to ensure all aspect and sentiment vocabularies contain a disjoint set of words, we remove common word(s) that are present in more than one vocabulary.

4.2 Labeled Data Preparation

In the previous work [1], it has been observed that noun(s) that are frequently associated with sentiment-bearing adjectives are likely to be aspect(s). Similarly, opinions are usually expressed using adjectives or adverbs in the sentence. Hence, we consider nouns to be

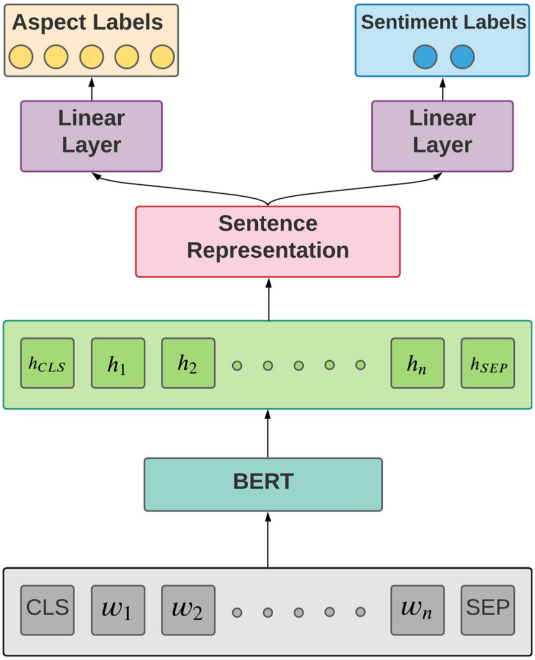


Fig. 2 Joint ABSA model

potential-aspects and adjectives or adverbs to be *potential-opinions*, i.e. these terms are likely to be aspect and opinion words, respectively. We experiment with this notion and use Spacy [27] for *POS* tagging to get nouns, adjectives, and adverbs in a sentence.

We take the set of all input sentences ($X_q \subset \mathcal{D}$) in which each sentence contains at least one *potential-aspect* and one *potential-opinion*. A sentence $X \in X_q$ is passed through post-trained Domain Knowledge BERT MLM to get replacement candidates with probability scores for *potential-aspects*. Using these replacement candidates, the list G_{aspect} is created by taking top K replacement words based on the probability scores. In the given sentence X, we find overlap score S_a of each aspect $a \in A$ by counting the common words between its corresponding aspect vocabulary V_a and the list G_{aspect}. In the case of multiple *potential-aspects*, we take the average of overlap scores. Overlap scores for the given sentences X_q with all aspect categories A is stored in a score matrix $\mathcal{M} \in \mathbb{R}^{|X_q| \times |A|}$. Our aspect vocabularies are constructed with the help of Domain Knowledge BERT, which is post-trained using a real-world domain-specific dataset. Usually, in a real-world dataset presence of review sentences related to various aspect categories are not evenly distributed. Hence, the constructed vocabulary of certain aspect categories contains a relatively larger number of semantically coherent words as compared to other aspect categories. Such inherent differences in quality of aspect category vocabularies lead to large variations in overlap scores that can unduly favor the classification of certain aspect categories. Thus, in order to normalize the overlap scores of aspect categories, we compute the mean (μ_a) and standard deviations (σ_a) of each aspect $a \in A$ using score matrix $\mathcal{M}$.

$$\mu_a = \frac{1}{|X_q|} \sum_{j=1}^{|X_q|} \mathcal{M}_{ja} \tag{3}$$

$$\sigma_a^2 = \frac{1}{|X_q|}\sum_{j=1}^{|X_q|}(\mathcal{M}_{ja} - \mu_a)^2 \tag{4}$$

Subsequently, normalized score S'_a of each aspect a corresponding to a given sentence X is computed as follows:

$$S'_a = \frac{S_a - \mu_a}{\sigma_a} \tag{5}$$

Here normalized score S'_a shows how many standard deviations away a raw overlap score S_a lies from the mean value. For a given sentence X, we assign aspect category a as a label if the normalized score S'_a is the largest and above a pre-defined threshold λ. We follow the same process as aspect label assignment for assigning a sentiment label s to the given sentence. At the end of this process, we get labeled dataset $\mathcal{D}_{\mathcal{L}} \subset \mathcal{D}$, which is a subset of the original corpus. The hyperparameter λ plays an important role in improving the quality of assigned labels. In order to show its effect, we built our model using different values of λ and capture the macro-F1 score on our datasets. Figure 4 shows the variations in the results.

4.3 Building a Joint Deep Neural Network for Aspect and Sentiment Identification

In the previous step, we get labeled dataset $\mathcal{D}_{\mathcal{L}}$ from the unlabeled training corpus. However, a sizable chunk of corpus still remains unlabeled due to the following reasons:

- In some sentences, the aspect, and its associated sentiment are implicitly expressed.
- There exist sentences that have overlap scores below the pre-defined threshold λ.

In order to do aspect and sentiment classification of all unseen sentences including the above-mentioned sentences, we build a BERT-based joint deep neural network. This neural network learns complex latent features using the labeled data $\mathcal{D}_{\mathcal{L}}$. We use post-trained Domain Knowledge BERT for encoding sentences and assume that the output of BERT's last hidden layer could be very close to its own target function (i.e. Next Sentence Prediction and Masked Language Model). Thus, we use the output of the second-to-last hidden layer (i.e. eleventh encoder) as the embeddings of all the tokens in the sentence. For a sentence $X_l \in \mathcal{D}_{\mathcal{L}}$, we compute the input sentence representation X_l as:

$$X_l = ([CLS], w_1, \ldots, w_n, [SEP]) \tag{6}$$

Here [CLS] and [SEP] are special tokens that indicate the beginning and end of the sequence, respectively. Constructed sequence of length $n + 2$ is passed through BERT to obtain hidden representation $H \in \mathbb{R}^{(n+2)\times d}$ for each token in X_l.

$$H = BERT(X_l) \tag{7}$$

These hidden representations contain contextual and positional information of each token in the sentence. We apply mean-pooling technique to obtain the global contextual sentence representation $T \in \mathbb{R}^{1\times d}$:

$$T = \frac{1}{n}\sum_{j=1}^{n} H_j \tag{8}$$

Here, we ignore the hidden representations of [CLS] and [SEP] token while generating the sentence representation. Subsequently, we pass T to two different linear layers for aspect and sentiment classification:

$$\hat{a} = SoftMax(TW_a + b_a) \tag{9}$$

$$\hat{s} = SoftMax(TW_s + b_s) \tag{10}$$

where, $W_a \in \mathbb{R}^{d\times|A|}$ and $W_s \in \mathbb{R}^{d\times|S|}$ are the weight matrix parameters, b_a and b_s are the bias vectors. $\hat{a}$ and $\hat{s}$ are the predicted probabilities for aspect and sentiment classification, respectively.

4.4 Generalized Cross Entropy Loss for Noisy Labels

We adopt a semi-automated approach to label the dataset that induces noise in the annotation process. Hence, to tackle this, we use a noise-robust loss function. The categorical cross-entropy (CCE) loss emphasizes more on difficult samples during training. CCE is usually faster to converge but overfits the noise present in the dataset. On the contrary Mean Absolute Error (MAE) gives equal importance to all the samples in the dataset. Hence, MAE is more robust to the noise but takes more time to converge. To solve this problem Zhang et al. [28] proposed $\mathcal{L}_q$ loss that utilizes the advantage of both CCE and MAE loss functions. During training $\mathcal{L}_q$ loss function gives lesser emphasis on difficult samples compared with CCE but pays more attention to the same when compared with MAE.

We compute aspect category classification loss L_a and sentiment polarity classification loss L_s using $\mathcal{L}_q$ loss as follows:

$$\mathcal{L}_a = \frac{1-(\hat{a}_{y_a})^q}{q} \tag{11}$$

$$\mathcal{L}_s = \frac{1-(\hat{s}_{y_s})^q}{q} \tag{12}$$

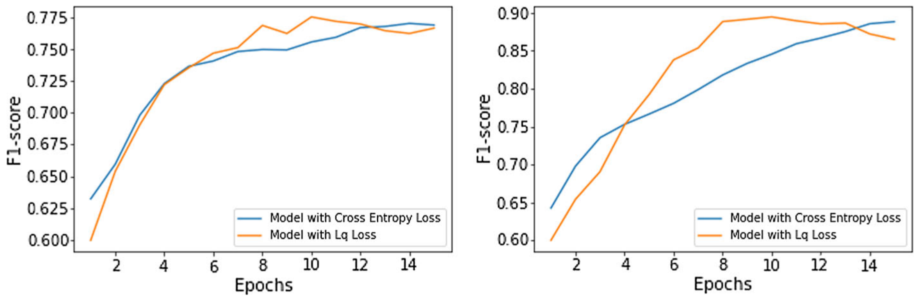
Here $\hat{a}_{y_a}$ and $\hat{s}_{y_s}$ denotes the predicted probabilites against the true aspect and sentiment labels, $q \in (0, 1)$ is a hyperparameter. Overall loss $\mathcal{L}$ of our model is calculated by summing over both the losses $\mathcal{L}_a$ and $\mathcal{L}_s$, which is minimized during the training process.

$$\mathcal{L} = \mathcal{L}_a + \mathcal{L}_s \tag{13}$$

In addition, we also built our model using cross-entropy loss and compared the model performance with $\mathcal{L}_q$ loss in Figure 3. This study shows that $\mathcal{L}_q$ loss helps to achieve the better performance of *CASC*. Hence, we choose $\mathcal{L}_q$ loss over cross-entropy for building our model.

4.5 Datasets

We conduct experiments on *Laptop* and *Restaurant* domain reviews. Training datasets contain Yelp and Amazon reviews for restaurant and laptop domains, respectively. Labeled test data is collected from SemEval [3]. All training and testing datasets are re-prepared by Huang et al. in [17]. We follow their experimental settings and use the single-label sentences for evaluation to avoid ambiguity. Details of the datasets are presented in Table 1.



(a) Aspect classification task. **(b)** Sentiment classification task.

Fig. 3 Performance of our model with cross entropy and $\mathcal{L}_q$ loss

Table 1 Dataset statistics

Dataset	#Training reviews	#Test reviews
Restaurant	17,027	643
Laptop	14,683	307

5 Experiments

5.1 Compared Models

We perform a comparison of our framework with two groups of baselines. The first group of models is inspired by the previous works. Description of the models are given below:

- *CosSim*: The topic representation is averaged by the embedding of seed words trained by *Word2Vec* on the training corpus. Cosine similarity is computed between a test sample and the topics to classify the sentence.
- *ABAE-Extended*: ABAE [12] is an attention-based model to extract aspects in an unsupervised manner using autoencoder. We use the official code[1] of the model and build an extended version of this model that uses two autoencoders to learn both aspect and sentiment embeddings. We reproduce the results of this extended model.
- *CAt-Extended*: CAt [29] is an unsupervised method for aspect extraction which uses Radial Basis Function(RBF) kernel as a scoring function to calculate single head attention and generate the sentence summary. Further, using $Cosine\ Similarity$ between sentence summary and class labels, aspect category classification is done. We extended their model and followed the same process to do sentiment polarity classification. We use their official code[2] to reproduce the results of the extended model.
- *W2VLDA* [16]: A state-of-the-art topic modeling-based method that leverages keywords for each aspect/sentiment to jointly do aspect and sentiment classification.
- *JASen* [17]: is a very recent weakly supervised model for aspect and sentiment classification that jointly learns the representation of sentiment and aspect topic in the embedding space. We copy the results of JASen and *W2VLDA* method mentioned in the same paper [17].

[1] https://github.com/ruidan/Unsupervised-Aspect-Extraction.

[2] https://github.com/clips/cat.

The second group has variations of our *CASC* model. These models are also used in ablation studies.

- *CASC*: Our full model framework using post-trained domain knowledge BERT-MLM and a small set of seed words for labeled data preparation and then build a robust neural network using labeled data for aspect and sentiment prediction. A test sentence is passed to the neural network for aspect and sentiment classification.
- *CASC w/o DK*: We remove post-trained domain knowledge BERT and use pre-trained base BERT in both BERT-MLM and neural network training steps.
- *CASC w/o DL*: Neural network training step is skipped. Aspect and sentiment class vocabulary is constructed as per the steps given in the Sect. 4.1. and then the classification of a test sentence is done by following the steps mentioned in the Sect. 4.2.
- *CASC w/o MLM*: In this model, BERT-MLM is not used. Class vocabulary construction is done using only a set of seed words provided by the user. Aspect labeling is done by simply finding such aspect vocabulary that contains the *potential-aspects* of the given sentence. Likewise, sentiment labeling is done by finding such sentiment vocabulary that contains the *potential-opinions* of the sentence. The neural network is trained using this labeled dataset for aspect and sentiment classification.

5.2 Evaluation Metrics

We evaluate the performance of a model for aspect and sentiment classification, respectively. For both the aspect and sentiment classification task, we evaluate the performance by using Macro-precision (Macro-P), Macro-recall (Macro-R), and Macro-F Score (Macro-F).

5.3 Settings

CosSim does not require any adjustable parameters to be tuned. For *ABAE-Extended* and *CAt-Extended*, we have performed grid-search to find the optimal values of hyperparameters. For vocabulary construction of our proposed model, we take the aspect and sentiment seed words given in Tables 2 and 3, respectively. We have experimented with various values of the hyperparameters and the best performance of our model we get after setting K as 20 and M as 100 for generating the vocabulary of each class. During the *"labeled data preparation"* step, overlap score computation is done using the same value of K to ensure a reasonable number of hits in the class vocabularies. We set the threshold λ as 0.5 for labeling the data. Our deep neural network uses hidden size d as 768, learning rate of 1e-5 with Adam [30] optimizer, q as 0.4 in $\mathcal{L}_q$ loss and batch size of 32 for both datasets. We train our model for 15 epochs. The results reported for all models are the average over 5 runs.

6 Results and Discussion

The overall performance of all the models for aspect category and sentiment classification task are shown in Tables 4 and 5, respectively. The comparison of the Micro-F score on the aspect category classification task shows that *CASC* outperforms all other models on both Restaurant and Laptop datasets. CosSim model classifies a given sentence by computing cosine similarity between representation of the test sentence and the aspect category representations, its performance is lower as compared to the remaining models. CAt is a very recent

Table 2 Seed word sets for aspect categories

Domain	Aspect	Seed words
Restaurant	Food	food, spicy, sushi, pizza, taste, delicious
	Ambience	Ambience, atmosphere, seating, surroundings, environment, location
	Service	Service, tips, manager, waitress, rude, staff
Laptop	Support	support, service, warranty, coverage, replace
	OS	OS, windows, ios, mac, system, linux
	Display	Display, screen, led, monitor, resolution
	Battery	Battery, life, charge, last, power
	Company	Company, product, hp, toshiba, dell, apple, lenovo
	Mouse	Mouse, touch, track, button, pad
	Software	Software, programs, applications, itunes, photo
	Keyboard	Keyboard, key, space, type, keys

Table 3 Seed word sets for sentiment categories

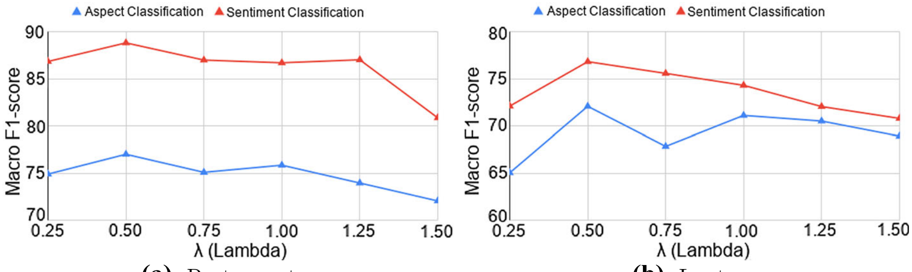
Domain	Sentiment	Seed words
Restaurant	Positive	good, great, nice, fresh, warm, friendly, delicious, fast, clean
	Negative	Bad, terrible, horrible, awful, smelled, disappointment, spoiled, cold, slow
Laptop	Positive	good, great, nice, perfect, impressed, best, thin, cheap, fast
	Negative	Bad, disappointed, terrible, small, slow, broken, complaint, virus, junk, crap, cramp

Table 4 Performance of various methods on the aspect classification task

Model	Restaurant Macro-P	Macro-R	Macro-F	Laptop Macro-P	Macro-R	Macro-F
CosSim	38.80	49.86	43.64	50.15	55.86	52.85
ABAE-Extended	43.60	51.31	47.14	55.62	58.91	57.22
CAt-Extended	46.91	51.79	49.23	57.40	60.04	58.69
W2VLDA	58.82	57.44	51.40	67.78	65.79	63.44
JASen	64.73	72.95	66.28	69.55	71.31	69.69
CASC w/o DK	65.70	72.88	69.10	61.84	62.72	62.27
CASC w/o DL	62.20	70.68	66.16	64.65	65.67	65.15
CASC w/o MLM	62.45	65.68	64.02	65.86	60.94	63.30
CASC	76.01	79.12	77.53	71.26	73.24	72.23

Table 5 Performance of various methods on the sentiment classification task

Model	Restaurant Macro-P	Macro-R	Macro-F	Laptop Macro-P	Macro-R	Macro-F
CosSim	64.53	60.12	62.25	70.53	63.86	67.03
ABAE-Extended	65.09	68.74	66.87	65.08	70.15	67.52
CAt-Extended	67.07	73.45	70.12	69.29	71.48	70.37
W2VLDA	75.66	70.52	67.23	71.62	71.37	71.22
JASen	82.85	78.11	79.44	74.69	74.65	74.59
CASC w/o DK	81.26	83.16	82.19	70.51	67.70	69.07
CASC w/o DL	75.41	72.89	74.12	67.60	65.71	66.64
CASC w/o MLM	75.24	67.63	71.23	71.52	69.49	70.49
CASC	88.48	90.51	89.48	76.59	77.31	76.94



(a) Restaurant **(b)** Laptop

Fig. 4 Performance of our model with different values of λ

RBF kernel-based single-head attention model, it performs better than other attention-based baselines like ABAE on both datasets. CAt uses a domain-specific general static meaning of the word for generating the sentence representations. W2VLDA is a topic modeling approach that leverages keywords for each aspect for doing aspect classification. The performance of W2VLDA is better than CAt on both datasets. However, JASen that jointly learns the representation of sentiment and aspect topic in the embedding space perform better than W2VLDA on both the datasets. JASen performs the best among all the baselines for the aspect classification task. Macro-F score comparison shows that our proposed model *CASC* outperforms JASen by a margin of 11.25% and 2.54% on Restaurant and Laptop datasets, respectively.

For sentiment classification, our proposed model *CASC* outperforms all other baselines on both datasets. Following the similar trend of aspect category classification task, JASen performs the best among all the baselines in the sentiment classification task as well. In this task, *CASC* perform better than JASen by a margin of 10.04% and 2.35% on Restaurant and Laptop datasets, respectively.

CASC perform better than various baselines on both aspect and sentiment classification task. This enhanced performance might be attributed to the following reasons: (1) *CASC* emphasizes the contextual meaning of the word and uses BERT MLM for constructing the vocabulary for each aspect and sentiment class; (2) *CASC* creates a quality-labeled dataset by using POS tag, constructed vocabularies, and an effective scoring mechanism; (3) *CASC* make use of labeled dataset and builds a BERT based neural network using a noise-robust loss function to jointly classify aspect and sentiment. During training neural network jointly learns latent features of aspect and its associated sentiment. Thus, it helps our neural network in classifying such sentences where aspect and sentiment-related keywords are not present in the constructed vocabularies.

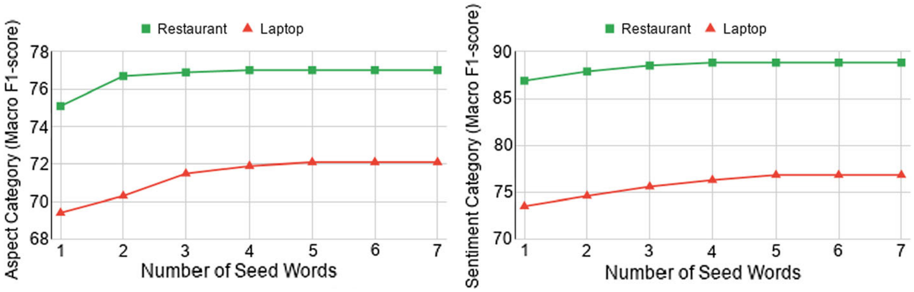
6.1 Ablation Study

To understand the effectiveness of different key components in *CASC*, we conduct the ablation study. We remove each component one after another and obtain three simplified variants. The second block of Table 4 & 5 has the results of all different variants of *CASC* on the aspect and sentiment classification tasks, respectively. The result shows that domain knowledge BERT enhances the performance of CASC on both tasks. Performance comparison of *CASC* and *CASC w/o DL* reveals that the deep learning model learns latent features using labeled sentences and enhances the performance of our proposed network significantly on both tasks. In *CASC w/o MLM*, vocabulary for each aspect and sentiment class is limited to only user-guided seed words. Thus, its performance is significantly inferior to *CASC*. This shows the importance of the MLM task as it helps in constructing a rich vocabulary for each class considering the contextual meaning of the word. Result analysis also reveals that all the important three components (e.g. DK, DL, and MLM) individually contribute to improving the performance of *CASC*, and when all three components are combined the performance gain is even better.

Further to understand the importance of *number of seed words*, we take the list of 10 seed words for each aspect and sentiment on both data sets. We randomly pick *p* seed words from each list, and run *CASC* for 5 times and take the average of macro-F1 scores. We vary *p* from 1 to 7 and plot the results of aspect and sentiment classification tasks in Fig. 5. Results show that the performance of *CASC* improves when the number of seed words is less than five and gradually saturates when this number exceeds more than five. It shows that our model needs only a small set of seed words to perform well. Table 6 presents the details of constructed aspect vocabularies for Restaurant domain.

6.2 Case Study

We pick some review sentences from our test datasets to study the significance of various components in enhancing the performance of our proposed model. Table 7 shows the details of



(a) Aspect classification task. **(b)** Sentiment classification task.

Fig. 5 Performance of our model with different number of seed words

Table 6 Restaurant domain constructed aspect vocabularies

Aspect	Constructed vocabulary
food	food, foods, meal, cuisine, meat, good, meals, menu, dinner, breakfast, chicken, rice, delicious, taste, bread, eat, drink, flavor, wonderful, delightful, dish, cooking, fabulous, drinks, sandwich, …
ambience	ambience, atmosphere, place, joint, spot, places, restaurant, stuff, seating, house, location, environment, space, surroundings, shop, establishment, area, bar, idea, club, venue, night, town, way, business, …
service	service, services, staff, server, servers, waitress, experience, music, host, waiter, vibe, environment, hostess, management, manager, rude, served, people, coffee, attitude, serve, personnel, bartender, presentation, manner, …

the predicted aspect category and sentiment class corresponding to the examples. The actual aspect category and sentiment class are shown in bold font style in the sentence column. In the first example, *CASC w/o DL* recognizes the word *"great"* and predicts *positive* sentiment, but it fails to recognize *"hot dogs"* as *"Food"* item and does wrong aspect category prediction. Similarly, in the second example *CASC w/o DL* is unable to identify correct sentiment. However, in both the sentences *CASC* does correct aspect and sentiment classification. It shows the importance of deep learning, which helps our model to recognize latent features in sentences. The third example shows the importance of the MLM component, wherein all models correctly classify the aspect except *CASC w/o MLM*. As *CASC w/o MLM* works on limited vocabulary, the model takes the literal meaning of the word *"Software"* and wrongly classifies the sentence under *"Software"* aspect category. Analysis of the fourth and fifth examples is very interesting wherein all other models except *CASC* fail to do a correct prediction for both aspect and sentiment. It exhibits the combination of DK-BERT, MLM, and DL components help *CASC* to achieve enhanced performance.

6.3 Error Analysis

In some of the review test sentences, our proposed model is unable to identify either the aspect terms or its associated sentiment correctly. We analyze those errors and classify the same into the following categories:

- *Sentence without context:* Some review sentences are very short and use opinion words without any context. Our model is unable to detect the correct aspect category in such

Table 7 Case analysis: Input sentences with predicted aspect category & sentiment polarity by various models

Sentence	CASC w/o DK	CASC w/o MLM	CASC w/o DL	CASC
great hot dogs! *[Food, Pos]*	Food, Pos	Food, Pos	Service, Pos	Food, Pos
windows 7 rocks *[OS, Pos]*	OS, Pos	OS, Pos	OS, Neg	OS, Pos
also, many software programs are not compatible with the ios platform. *[OS, Neg]*	OS, Neg	Software, Neg	OS, Neg	OS, Neg
the mac app store is your one-stop shop for mac software, though you can also download software online at your own risk. *[Software, Pos]*	OS, Neg	OS, Pos	Software, Neg	Software, Pos
the only beverage we did receive was water in dirty glasses ! *[Service, Neg]*	Food, Neg	Food, Neg	Food, Neg	Service, Neg

sentences. For example, in the sentence, *"don't leave the restaurant without it."*, the word *"it"* is not clear.

- *Sentence without aspect & sentiment:* There are some review sentences, which do not contain any aspect and its associated sentiment. For example, in the comment *"Besides, the Apple stocks have been falling due to lack of sales"*, no aspect term and associated sentiment exist, but *"sales"* is detected as aspect term with *negative* sentiment.
- *Use of idioms in review sentence:* Few review comments use idioms to express sentiment about aspect term. For example in the comment *"The two waitresses looked like they had been sucking on lemons"* sentiment about aspect *"waitress"* is expressed using *"sucking on lemons"*.

7 Conclusion

In this work, we proposed a semi-supervised hybrid model *CASC* for aspect and sentiment classification to solve the ABSA task end-to-end. The extensive experiments on two real-life datasets show that *CASC* outperforms the state-of-the-art models. Our proposed method has a couple of advantages over existing approaches that help it to perform better than others; (1) Existing unsupervised methods ignores the contextual meaning of the word but *CASC* emphasizes the contextual meaning of the word and adopt a masked language model-driven semi-automated approach to label the sentences. (2) *CASC* uses a noise-robust loss function to train a deep neural network that jointly learns latent features of aspect and its associated sentiment. The combination of these two factors helps *CASC* to classify an unseen sentence in a better way and present a complete solution to the ABSA task. In future work, we will focus on low-resource non-English languages as getting labeled data for such languages is a challenging task. We will use multilingual masked language models (MLM) like mBERT [18] and XLM-R [31] to investigate how cross-lingual transfer helps our semi-supervised hybrid approach to solve aspect and sentiment classification tasks in multilingual settings.

Funding All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Declarations

Conflict of interest The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. Hu M, Bing L (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 168–177
2. Liu B, Lei Z (2012) A survey of opinion mining and sentiment analysis. In: Mining text data, pp 415–463. Springer
3. Pontiki M, Dimitrios G, Haris P, Ion A, Suresh M, Al-Smadi M, Al-Ayyoub M, Yanyan Z, Bing Q, De Clercq O, et al (2016) Semeval-2016 task 5: aspect based sentiment analysis. In: 10th international workshop on semantic evaluation (SemEval 2016)
4. Guang Qiu, Liu Bing Bu, Jiajun Chen Chun (2011) Opinion word expansion and target extraction through double propagation. Comput Linguist 37(1):9–27
5. Blei DM, Ng AY, Jordan MT (2003) Latent dirichlet allocation. J Mach Learn Res 3(Jan):993–1022
6. Brody S, Noemie E (2010) An unsupervised aspect-sentiment model for online reviews. In Human language technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics. Association for Computational Linguistics, pp 804–812
7. Wayne Xin Z, Jing J, Hongfei Y, Xiaoming L (2010) Jointly modeling aspects and opinions with a maxent-lda hybrid. In: Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 56–65
8. Zhiyuan C, Arjun M, Bing L (2014) Aspect extraction with automated prior knowledge learning. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (vol 1, long papers), pp 347–358
9. Miao Y, Lei Y, Phil B (2016) Neural variational inference for text processing. In: International conference on machine learning, pp 1727–1736
10. Akash S, Charles S (2017) Autoencoding variational inference for topic models. arXiv preprint arXiv:1703.01488
11. Ming L, Jing L, Haisong Z, Lingzhi W, Xixin W, Kam-Fai W (2019) Coupling global and local context for unsupervised aspect extraction. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 4571–4581
12. He R, Wee Sun L, Hwee Tou N, Daniel D (2017) An unsupervised neural attention model for aspect extraction. In: Proceedings of the 55th annual meeting of the association for computational linguistics (vol 1, long papers). Association for Computational Linguistics, Vancouver, Canada, pp 388–397
13. Tulkens S, van Cranenburgh A (2020) Embarrassingly simple unsupervised aspect extraction. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics, pp 3182–3187 (Online, July 2020)
14. Mikolov T, Kai C, Greg C, Jeffrey D (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
15. Wang L, Kang L, Zhu C, Jun Z, De Melo G (2015) Sentiment-aspect extraction based on restricted boltzmann machines. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (vol 1, long papers), pp 616–625

16. Aitor G-P, Montse C, German R (2018) W2vlda: almost unsupervised system for aspect based sentiment analysis. Exp Syst Appl 91:127–137
17. Jiaxin H, Meng Y, Fang G, Heng J, Jiawei H (2020) Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding. arXiv preprint arXiv:2010.06705
18. Devlin J, Ming-Wei C, Kenton L, Kristina T (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
19. Xu H, Bing L, Lei S, Yu Philip S (2019). Bert post-training for review reading comprehension and aspect-based sentiment analysis. arXiv preprint arXiv:1904.02232
20. He R, Wee Sun L, Hwee Tou N, Daniel D (2018) Exploiting document knowledge for aspect-level sentiment classification. In: Proceedings of the 56th annual meeting of the association for computational linguistics. Association for Computational Linguistics
21. Angelidis S, Mirella L (2018) Summarizing opinions: aspect extraction meets sentiment prediction and they are both weakly supervised. CoRR, abs/1808.08858
22. Karamanolakis G, Daniel H, Luis G (2019) Leveraging just a few keywords for fine-grained aspect detection through weakly supervised co-training. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 4611–4621
23. Xu X, Songbo T, Yue L, Xueqi C, Zheng L (2012) Towards jointly extracting aspects and aspect-specific sentiment knowledge. In: Proceedings of the 21st ACM international conference on information and knowledge management, CIKM ’12. Association for Computing Machinery, New York, NY, USA, pp 1895–1899
24. Zhao X, Jiang J, Hongfei Y, Xiaoming L (2010) Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics, Cambridge, MA, pp 56–65
25. Wang L, Kang L, Zhu C, Jun Z, de Melo G (2015) Sentiment-aspect extraction based on restricted boltzmann machines. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (vol 1: long papers). Association for Computational Linguistics, Beijing, China, pp 616–625
26. Zhuang H, Fang G, Chao Z, Liyuan L, Jiawei H (2020)Joint aspect-sentiment analysis with minimal user guidance. In: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, SIGIR ’20. Association for Computing Machinery, New York, NY, USA, pp 1241–1250
27. Honnibal M, Ines M, Van Landeghem S, Adriane B (2020) spaCy: industrial-strength natural language processing in python
28. Zhang Z, Sabuncu Mert R (2018) Generalized cross entropy loss for training deep neural networks with noisy labels. arXiv preprint arXiv:1805.07836
29. Tulkens S, van Cranenburgh A (2020) Embarrassingly simple unsupervised aspect extraction. arXiv preprint arXiv:2004.13580
30. Kingma Diederik P, Jimmy B (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
31. Conneau A, Kartikay K, Naman G, Vishrav C, Guillaume W, Francisco G, Edouard G, Myle O, Luke Z, Veselin S (2019) Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116