Applied Intelligence (2022) 52:14025–14040
https://doi.org/10.1007/s10489-022-03311-y

BILEAT: a highly generalized and robust approach for unified aspect-based sentiment analysis

BILEAT

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

Accepted: 26 January 2022 / Published online: 1 March 2022

Abstract
Aspect-based sentiment analysis (ABSA) includes two subtasks, namely, aspect term extraction and aspect-level sentiment classification. Most existing works address these subtasks independently. Recently, many researchers have attempted to solve both the subtasks of ABSA with a unified framework. However, previous works have not focused on the generalization and robustness of such unified frameworks. This paper proposes a novel BERT-Based Interactive Learning with Ensemble Adversarial Training (BILEAT) to solve complete ABSA by using a unified tagging scheme. We build white-box adversarially post-trained domain knowledge BERT (WBDK-BERT) using a domain-specific dataset. During post-training, we regularize the training objective by adding perturbations in the embedding space to maximize the adversarial loss, enhancing the generalization and robustness of WBDK-BERT. BILEAT uses WBDK-BERT to generate contextualized embeddings and produce collaborative signals through interactive learning. Further, to build a highly reliable model, we generate adversarial examples using a black-box technique. These adversarial examples are grammatically fluent, semantically coherent with original input, and can mislead the neural network. Our proposed model is trained using original inputs and such adversarial examples in a combined way. Experimental results demonstrate that WBDK-BERT and black-box adversarial examples complement each other, and combining these two helps BILEAT become highly generalized and robust compared to existing methods. To the best of our knowledge, this is the first study that generates quality adversarial examples and evaluates the robustness of models for unified ABSA[1].

Keywords Unified ABSA · BERT · Deep neural network · Attention mechanism · Adversarial network · Black-box adversarial attack · White-box adversarial attack

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) focuses on identifying the aspect terms explicitly mentioned in sentences and detecting the sentiment polarities of the aspect-terms [1]. For example, in the review sentence *"Tasty food but the service was slow!"*, the user mentions two aspect-terms, namely, *"food"* and *"service"*, and conveys positive sentiment about the first, and negative sentiment for the second. Generally, the ABSA task can be broken into two sub-tasks: aspect-terms extraction and aspect-term sentiment classification. Aspect-term extraction aims to identify the aspect(s) mentioned in the text, and it has been broadly studied in [2–4]. The second sub-task, i.e., aspect-term sentiment classification, enhances the extracted aspect-term(s) usefulness by detecting its sentiment polarity. This sub-task has been also extensively studied in [5–7].

A unified approach that integrates both the subtasks has been adopted by previous researchers [8–10] to enhance the performance of ABSA. Despite the effectiveness of unified methods, we argue that most of the previous works have not given enough attention to the generalization and robustness of the model. A learned model is expected to perform well on unseen test examples and should be able to combat adversarial samples, which are created by

https://github.com/Raghu150999/BILEAT\protect_E2E\protect_ABSA

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

1 Birla Institute of Technology, Science, Pilani - Hyderabad, Hyderabad 500078, India

adding small perturbations to the original inputs [11]. These adversarial samples are un-noticeable to human judges and can mislead the neural networks to incorrect predictions. For example in the original review text *"Finally, I got sick of the* <u>*bad*</u> ***service****,obnoxious smirks, and snotty back talk."* the word *"bad"* can be replaced with *"terrible"* to generate a semantically coherent adversarial example *"Finally, I got sick of the* <u>*terrible*</u> ***service****,obnoxious smirks, and snotty back talk."*. A highly generalized and robust unified ABSA model is expected to detect aspect-term as *"service"* with associated *negative* sentiment in both original and adversarially generated review text. Adversarial training makes the neural network robust to such examples and helps the model generalize better.

Given the above point, we formulate the end-to-end ABSA as a single sequence labelling task with a unified tagging scheme[1] and propose a novel BERT-Based Interactive Learning with Ensemble Adversarial Training (BILEAT) for the same. BILEAT is a multi-layer unified framework that handles ABSA end-to-end, along with two auxiliary tasks AE and OE. Performance of main ABSA task improves by exchanging clues between AE and OE auxiliary tasks. We further add adversarial examples in an ensemble way to improve the generalization and robustness of our model.

The adversarial attack in BILEAT is two-folded. First, we build an adversarially post-trained White-Box Domain Knowledge BERT (WBDK-BERT) to efficiently capture the context-dependent meaning of the word in a sentence. WBDK-BERT is built by doing post-training of BERT [12] on the masked language model (MLM) task using a domain-specific dataset. During post-training, white-box adversarial training [13] is applied that augments the standard training objective with an additional term to maximize the adversarial loss via applying perturbation in the embedding space. BILEAT utilizes WBDK-BERT to generate a representation of words in a given sentence.

Second, to further enhance the robustness and reliability of our proposed model, we generate adversarial examples using a black-box [14] technique. We utilize the potential of BERT to create adversarial examples that are grammatically correct and semantically in line with the original input. Such adversarial examples can fool the neural network. Our method is inspired by Li et al. [15], but it differs from them in two ways: (1). we consider only the aspect and opinion terms in the sentence for replacement (2). And apply our scoring function along with BERT-MLM in a semantic-preserving way to produce substitutes for these words. As a perturbation generator, we use the masked language model and choose perturbations that maximize the likelihood of making the wrong prediction [11] by the model for the given sequence of words in a sentence. Our proposed model is trained using the original dataset along with generated adversarial examples.

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

– We propose a novel, BERT-Based Interactive Learning with Ensemble Adversarial Training (BILEAT) model for a unified ABSA task. BILEAT does interactive learning to understand the mutual relation between AE and OE auxiliary tasks and uses a domain-specific white-box adversarially trained WBDK-BERT built by us to generate context-aware word embeddings. Further, we apply a black-box attack to create fluent and semantically coherent adversarial examples. BILEAT is trained using both original inputs and such adversarial examples in a combined way.
– For a unified ABSA task, we create grammatically correct and semantically coherent adversarial datasets, which will be helpful for future research work.
– We do an ablation study of BILEAT for evaluating the impact of interactive learning between AE and OE, the usefulness of WBDK-BERT, and the effect of combined training of original inputs and generated adversarial examples.
– We utilize above mentioned adversarial test datasets to evaluate the robustness of various methods. Experimental results show BILEAT outperforms state-of-the-art methods. To the best of our knowledge, we are the first to perform such a detailed study about the robustness of unified ABSA methods. Our experimental results can serve as a benchmark for future research.

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

2 Related work

Existing Aspect-Based Sentiment Analysis approaches are broken down into two subtasks: Aspect Extraction (AE) [16–20] and Aspect-level Sentiment Classification (ASC) [5, 21–25]. The former refers to detecting aspect terms in a sentence, while the latter refers to detecting a review sentence's sentiment polarity towards a given aspect. These approaches have been studied extensively in previous works. Most existing methods solving the ASC assume that the aspects are already mentioned with the review sentence,

[1]{B, I}–{POS, NEG, NEU} denotes the beginning and inside of an aspect-term with the positive, negative, or neutral sentiment, respectively, and O denotes background words.

which limits the practical use of such methods. One way to employ these methods in practical settings is to use them in a pipelined manner. However, treating these tasks in a pipelined approach leads to error propagation across subtasks giving us poor results.

Some studies [8, 10] have a unified modelling approach to handle the above tasks in an end-to-end manner. These methods are modelled as sequence labelling tasks which fall into two types: collapsed tagging and joint training. The former uses shared features for each subtask, whereas the latter uses a multi-task learning framework that uses shared and private features. Li et al. [10] have identified auxiliary tasks such as boundary guidance and sentiment consistency for joint modelling. These tasks guide their model to learn the unified tagging scheme for a review sentence. He et al. [26] have tried to model the interaction using a message passing mechanism. They learn semantically related tasks (such as aspect-level sentiment classification and document-level sentiment classification) through joint training to get better results. However, these methods do not model the interactions between the subtasks to their full potential. Li et al. [27] have used contextual word embeddings instead of GloVe or Word2Vec embeddings to get better context-aware representations. Chen et al. [28] interact the semantic information between the encoded features generated from AE, ASC, and OE (opinion extraction) to enhance the sub-modules through mutual knowledge transfer. Liang et al. [29] have further introduced document-level sub-tasks (mainly domain classification and document-level sentiment classification) to infuse document-level information for enhancing the performance of the aspect-extraction sub-tasks. Luo et al. [30] have proposed a method called GRACE that uses post-trained BERT and applies a gradient harmonized method with virtual adversarial training to solve the ABSA adopting a cascaded labelling approach. Mao et al. [31] proposed a joint training framework that constructs machine reading comprehension tasks to solve AE, OE, and ASC problems using BERT-MRC models with parameter sharing. Lee et al. [32] have proposed a unified model for completing ABSA tasks by interacting signals between ATE and OE tasks. They also use self-supervised strategies such as pairwise relation masking, which help the model to better exploit the relations between aspects and opinions at a sentence level.

2.1 Adversarial training

Developing robust deep neural models for natural language processing continues to be a long-standing real-world problem. Attackers develop examples for inputs that can flip the prediction, thereby decreasing the model's accuracy. Adversarial training can enhance robustness, but past works have shown that it also affects generalization. There have been several studies for adversarial attacks on continuous data. In general, adversarial attacks are of two types (1). white-box and (2). black-box. In white-box attacks [13] model parameters can be accessed, while black-box attacks [14] work without accessing model parameters and only uses the input and output. However, generating adversarial examples for text continues to be a challenging task.

– ***White Box Attack:*** Xu et al. [33] proposed TextTricker for targeted and non-target attacks on classification model. These attacks have been implemented using two ways: $loss\text{-}based$ and $gradient\text{-}based$. Liu et al. [34] introduce Adversarial training for Large Neural Language Models, an algorithm that regularizes and improves both generalization and robustness of a deep neural network. Karimi et al. [35] add perturbations using gradients of the loss function to the encoded inputs and generate adversarial examples.
– ***Black Box Attack:*** Previous studies for generating adversarial examples rely on introducing error at the character level [36] or adding/deleting word [37] in a sentence. However, the added perturbations may result in a grammatically incorrect sentence, hence easily identifiable by a human. Rule-based approaches have been shown to come up with more natural-looking sentences. However, these approaches rely on external tools such as POS Tagger, NER Tagger, WordNet, etc., and do not generate semantically coherent sentences. Pruthi et al. [38] predict each word's correct substitution for all possibly misspelled words in a sentence using some back-off strategies. The predictions are passed to the downstream tasks for further training. Recent studies have used language models for adding perturbations to sentences. Li et al. [15] use $BERT\text{-}MLM$ for getting word substitutions in a sentence. The examples generated have word substitutions that are context-aware, and the overall sentences are semantically coherent. Following their work to extract important words from input sentences, Hofer et al. [39] use character-level adversarial attacks, which are inconspicuous to human observers. These attacks include replacing characters with visually similar-looking symbols, adding misspellings and irrelevant punctuation marks in a sentence.

Previous works have adopted various effective approaches to solve the ABSA in a unified way. We have presented the summary of the same in the Table 1. However, these previous works are effective but have not given enough focus on the generalization and robustness of the model. A learned model is expected to perform well on unseen test examples and should be able to combat adversarial samples, which are created by adding small perturbations to the original inputs. Our proposed model,

Table 1 Summary of previous works related to ABSA

Task	Approaches	Summary
Aspect Term Extraction (ATE) (A sub-task of pipeline approach)	Rule based [16]	Rule based method formed by modeling the relations using aspect and opinion terms in a sentence.
	Syntactic Features [17, 18]	Deep Learning methods exploit dependency tree relations to extract information about aspects and opinions in a sentence.
	Attention based [3, 20, 40]	Attention based models which generate opinion summarization vectors for a each aspect candidates.
Aspect Sentiment Classification (ASC) (A sub-task of pipeline approach)	Syntactic Features [5, 21, 22]	Neural models incorporate syntactic features which are extracted from the input sentence using a dependency parser.
	Attention based [23–25]	Neural models generate target-specific representation for a given input sentence to model relationships between the target and its context.
ABSA (A unified modelling approach)	Multi-task learning [10, 26–28, 30–32, 41]	Uses shared and private features of each symmetrically related subtask and learn unified tags for ABSA through joint training.

BILEAT, utilize interactive learning between auxiliary tasks to produce a collaborative signal and uses a domain-specific and white-box adversarially trained WBDK-BERT built by us to generate context-aware word embeddings. Further, we utilize BERT-MLM and apply a black-box attack to create fluent and semantically coherent adversarial examples. BILEAT is trained using both original inputs and such adversarial examples in a combined way, which makes our proposed model highly generalized and robust for the unified ABSA task.

3 Our method

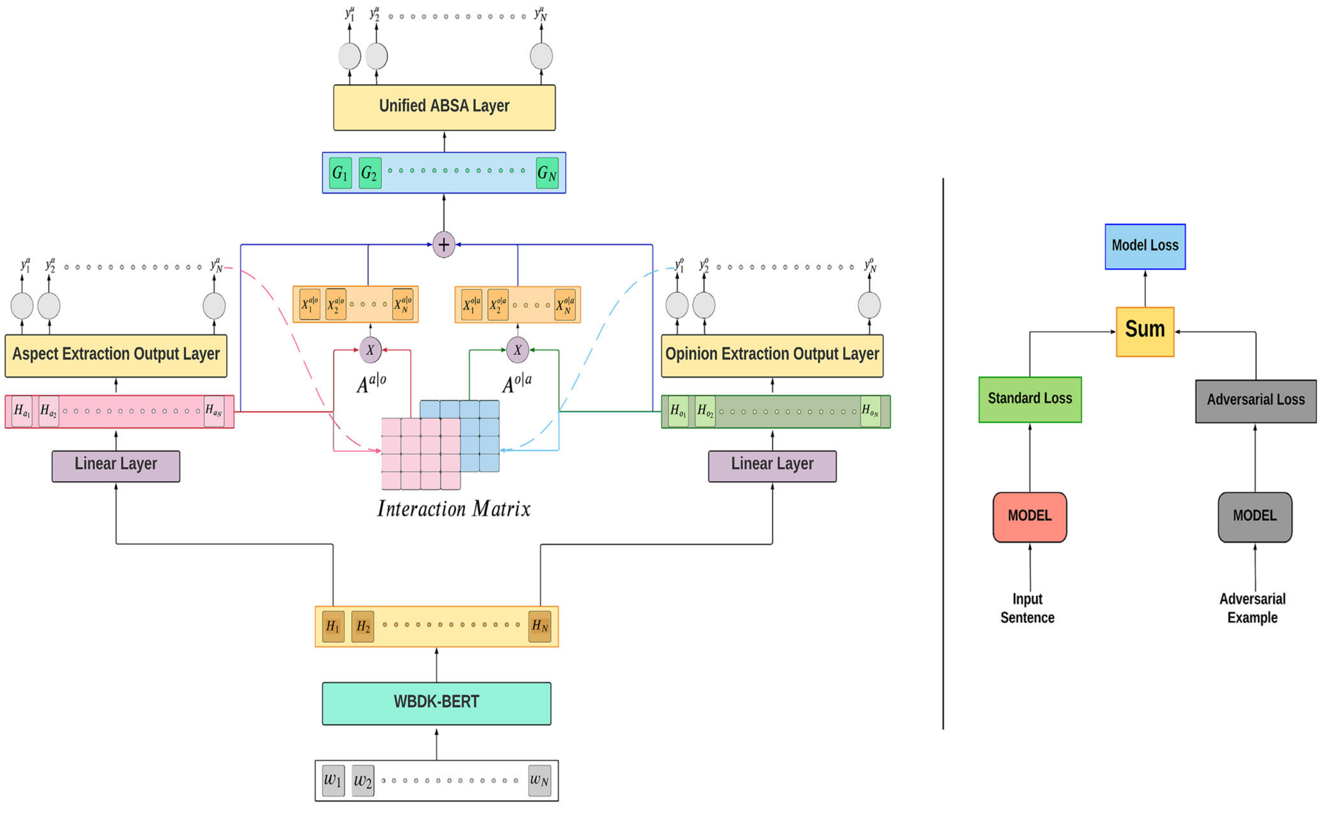
We formulate unified ABSA as sequence labelling problem and use a unified tagging scheme $\mathcal{Y} = \{B\text{-}POS, I\text{-}POS, B\text{-}NEG, I\text{-}NEG, B\text{-}NEU, I\text{-}NEU, O\}$, which consists of 7 tags. Each tag except O contains information about aspect-term and its associated sentiment. For example $B\text{-}POS$ denotes beginning of an aspect-term with positive sentiment. For a given a sentence $S = \{w_1, w_2, ..., w_N\}$, our ultimate goal is to is to predict a tag sequence $Y^u = \{y_1, y_2, ..., y_N\}$, where $Y_i^u \in \mathcal{Y}$.

3.1 Proposed model

In this section, we describe the architecture of BERT-Based Interactive Learning with Ensemble Adversarial Training (BILEAT). BILEAT is a highly generalized and robust model that uses both white-box [13] and black-box [14] adversarial training in a combined way. In white-box attacks, adversarial examples are generated by accessing model parameters, while black-box attacks create such examples using only the input and output without accessing model parameters. As illustrated in Fig. 1, BILEAT contains white-box adversarially trained domain knowledge BERT (WBDK-BERT), word encoding layer, interactive learning layer, adversarial perturbations generated through a black-box attack, and the objective function to be optimized. The ultimate goal is to solve ABSA in a unified way.

3.1.1 WBDK-BERT: white-box adversarially trained domain knowledge BERT

BERT is a pre-trained language representation model, which consists of a 12-layer bidirectional Transformer encoder [42]. Xu et al. in [43] have shown that inducing



(a) Core architecture of proposed model

(b) Adversarial training of proposed model

Fig. 1 The proposed BILEAT framework

domain-specific information to $BERT_{BASE}$ boosts the performance of ABSA. Following their work, we take $BERT_{BASE}$ and perform post-training using domain-specific datasets, where the aim of the standard objective function is to minimize the standard error on training data with the training objectives derived from the self-supervision MLM task.

In general, the training algorithm aims to learn a function $f(w;\beta)$: $w \rightarrow V$, parametrized by β. For MLM task, V is the vocabulary, and $f(w;\beta)$ attempts to predict the masked token u. During post-training, V becomes the task-specific label set, and $f(w;\beta)$ acts as the classifier. Given a training dataset $\mathcal{D}$ consisting of input-output pairs $(w;u)$ and the loss function $l(:;:)$ (e.g., cross entropy), $f(w;\beta)$ is trained to minimize the standard loss:

$$\min_{\beta} \mathbb{E}_{(w,u)\sim\mathcal{D}}[l(f(w;\beta),u)] \tag{1}$$

Liu et al. [34] have shown that white box adversarial training for large language models improves the generalization and robustness in downstream tasks. Inspired by their work, we build our white-box adversarially trained domain knowledge BERT (WBDK-BERT). To build WBDK-BERT, we randomly choose perturbation δ from a normal distribution and add the same to the embedding level. Subsequently, using the perturbated embedding, we compute adversarial gradient g_{adv}.

$$g_{adv} \leftarrow \nabla_{\delta} l(f(w;\beta), f(w+\delta;\beta)) \tag{2}$$

The motive behind adding the perturbation δ to the embedding level is to generate such an adversarial example that can maximize the adversarial loss of the model. Hence, we find the optimal value of δ by moving in the direction of increasing loss through the gradient ascent.

$$\delta \leftarrow \Pi_{||\delta||_{\infty}\leq\epsilon}(\delta + \eta g_{adv}) \tag{3}$$

Here, ϵ is the upper bound of perturbation δ.

Using the obtained value of perturbation δ adversarial examples are created. Finally, for MLM task we use both original input and adverserial examples and during training, we compute the overall loss of the model by combining the standard objective loss and adverserial loss:

$$\min_{\beta} \mathbb{E}_{(w,u)\sim\mathcal{D}}[l(f(w;\beta),u)+\alpha \max_{\delta} l(f(w+\delta;\beta), f(w;\beta))] \tag{4}$$

Finally, to minimize the oveall loss, the model parameters β are updated using the global learning rate τ through gradient descent.

$$g_\beta \leftarrow \nabla_\beta(f(w;\beta), y) + \alpha \nabla_\beta l(f(w;\beta), f(w+\delta;\beta)) \quad (5)$$

$$\beta \leftarrow \beta - \tau g_\beta \quad (6)$$

We have summarized the adversarial training of WBDK-BERT in Algorithm 1.

Algorithm 1 White box adversarial training of domain knowledge BERT.

for $t = 1, \ldots, T$ **do**
    /* T is the total number of iterations */
    for *each sentence from Training Dataset* **do**
        sample a random perturbation δ from a normal distribution
        and compute optimal adversarial perturbation using gradient ascent
        for $m = 1, \ldots, K$ **do**
            /* K is the number of iterations for perturbation estimation */
            compute adversarial gradient using model loss function w.r.t the perturbation δ as shown in (2)
            update perturbation by moving in the direction of adversarial gradient as shown in (3)
        end
        compute model loss by summing up the standard objective loss and adverserial loss as shown in (4)
        subsequently, update model parameters through gradient descent as shown in (5) & (6)
    end
end

3.1.2 Word encoding layer

Our proposed model employs $WBDK\text{-}BERT$ for generating contextual word representations. For a sentence $S = [w_1, w_2, \ldots, w_N]$, which consists of N words, $(w_1, w_2, \ldots, w_N)$ is passed to $WBDK\text{-}BERT$, to obtain hidden representations $H \in \mathbb{R}^{N \times d}$ for each word.

$$H = WBDK\text{-}BERT(S) \quad (7)$$

Here, d is the size of the hidden dimension of $WBDK\text{-}BERT$, and $H_i \in H$ is the hidden representation of i^{th} word of the sentence S.

3.1.3 Interactive learning

We pass H to two different linear layers to learn the two separate word representations corresponding to auxiliary tasks AE and OE.

$$H^a = HW_a^1 \quad (8)$$

$$H^o = HW_o^1 \quad (9)$$

Here, W_a^1 and W_o^1 are trainable parameters. $H^a \in \mathbb{R}^{N \times d}$ and $H^o \in \mathbb{R}^{N \times d}$ are the two learnt representations of the words in the given sentence S. Subsequently, H^a is passed to $Softmax$ layer to predict the probabilities $\hat{Y}^a \in \mathbb{R}^{N \times 3}$ of $\{BA, IA, OA\}$ tags for AE task. Likewise, H^o is passed to to $Softmax$ layer to predict the probabilities $\hat{Y}^o \in \mathbb{R}^{N \times 3}$ of $\{BO, IO, OO\}$ tags for OE task.

$$\hat{Y}^a = Softmax\left(H^a W_a^2\right) \quad (10)$$

$$\hat{Y}^o = Softmax\left(H^o W_o^2\right) \quad (11)$$

Here, W_a^2 and W_o^2 are trainable weights.

Usually, aspect-term and opinion-term are strongly correlated, and interaction between these two tasks can exchange important clues about the unified ABSA task. Hence, to learn the non-linear interactions between aspect and opinion words we define a score function $Q_{i,j}$:

$$Q_{i,j} = H_i^a W_c (H_j^o)^T * \frac{1}{|i-j|} \quad (12)$$

Here, W_c is a trainable weight matrix, which learns non-linear interactions between aspect and opinion words. We argue that aspect term and its corresponding opinion term occur in closer proximity. Thus, the second term in the (12) shows that scores are inversely proportional to the number of words between each other. Moreover, we define $Q_{i,i} = 0$ since a word cannot be both aspect and opinion word at the same time.

We make use of the above score function $Q_{i,j}$ and generate an interaction matrix $\mathcal{A}^{a|o}$ to capture the contribution of j^{th} word from OE-oriented features to the i^{th} word in the AE-oriented features.

$$\mathcal{A}_{i,j}^{a|o} = \frac{exp\left(Q_{i,j} * \hat{Y}^o_{j,\{BO,IO\}}\right)}{\sum_{j=1}^{n} exp\left(Q_{i,j} * \hat{Y}^o_{j,\{BO,IO\}}\right)} \quad (13)$$

Where, the term $\hat{Y}^o_{j,\{BO,IO\}}$ is the sum of probabilities of BO and IO output tag which denotes the predicted probability that the j-th token is part of any opinion term.

Similarly, we define interaction matrix $\mathcal{A}^{o|a}$ to determine the contribution of j^{th} word from AE-oriented features to the i^{th} word in the OE-oriented features.

$$\mathcal{A}_{i,j}^{o|a} = \frac{exp\left(Q_{i,j}^{T} * \hat{Y}_{j,\{BA,IA\}}^{a}\right)}{\sum_{j=1}^{n} exp\left(Q_{i,j}^{T} * \hat{Y}_{j,\{BA,IA\}}^{a}\right)} \tag{14}$$

Here, $\hat{Y}_{j,\{BA,IA\}}^{a}$ in (13) is the sum of probabilities of BA and IA output tag which is associated with the predicted probability that the j-th token is part of any aspect term.

Now, using the interaction matrices $\mathcal{A}^{o|a}$ we compute overall opinion representation of a word w_i with respect to each aspect word in the (14) and likewise $\mathcal{A}^{a|o}$ is used to compute overall aspect representation of the same word with respect to each opinion word in the (13).

$$X_i^{o|a} = \sum_{j=1}^{n}\left(\mathcal{A}_{i,j}^{o|a} * H_j^a\right) \tag{15}$$

$$X_i^{a|o} = \sum_{j=1}^{n}\left(\mathcal{A}_{i,j}^{a|o} * H_j^o\right) \tag{16}$$

At last, we compute final representation $G_i \in \mathbb{R}^{n\times 4d}$ of word $w_i \in S$ in the following way:

$$G_i = H_i^a \oplus X_i^{o|a} \oplus H_i^o \oplus X_i^{a|o} \tag{17}$$

Here, $\oplus$ is the concatenation operation. We use G for predicting unified labels for ABSA task.

$$\hat{Y}^u = Softmax(GW_g) \tag{18}$$

Here, W_g is a trainable weight.

3.1.4 Objective function

For every word in a sentence, we compute the loss for each of the three tasks (Unified, AE, and OE) using multi-margin loss or hinge loss l as given in (19)

$$l = \sum_{i;i\neq c}^{|\mathcal{Y}|} max(0, margin - p_c + p_i) \tag{19}$$

Here, $margin$ is a hyperparameter, p_c is the prediction logit of the correct label $c \in \mathcal{Y}$ and p_i is predicted logit of a wrong label. Standard loss $\mathcal{L}$ of our model is calculated by summing loss of unified ASBA l_u, loss of AE l_a, and loss of OE l_o.

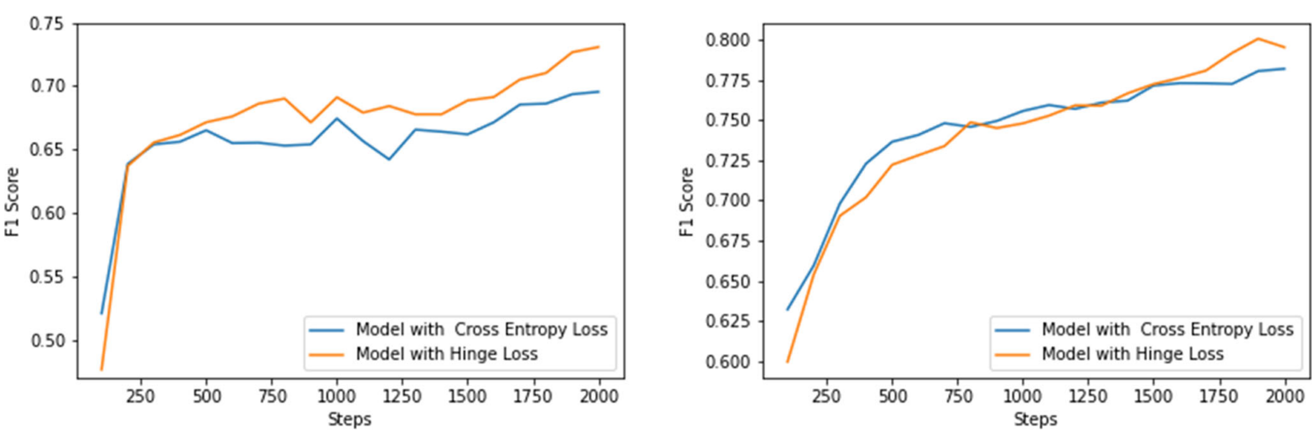
$$\mathcal{L} = l_u + \lambda(l_a + l_o) \tag{20}$$

Here, λ is a hyperparameter that controls the contribution of loss of auxiliary task in the overall loss $\mathcal{L}$.

In addition, we also built our model using cross-entropy loss and compared the model performance with hinge loss in Fig. 2b and a. This study shows that hinge loss helps better convergence of loss, which is getting translated in the performance of respective variants of BILEAT. Hence, we choose hinge loss over cross-entropy for building our model.

3.1.5 Adversarial examples generation using black-box attack

In this step, we apply black-box adversarial training to make our model highly generalized and robust. Unlike the white-box attack, the black-box attack doesn't have access to the model parameters, i.e., it treats the model as a black box. The attack is only allowed to query the model on input and retrieve the prediction probabilities. Using this prediction output, the attack tries to craft an adversarial example. We generate quality adversarial examples by transferring the perturbations from another model. Ensuring changes are unnoticeable to human judges yet capable of fooling



(a) F1 Score on Laptop dataset. (b) F1 Score on Restaurant dataset.

Fig. 2 Performances of our proposed model the neural network, maintaining grammatical fluency and semantic consistency with original inputs.

Recently Li. et al. [15] have utilized $BERT\text{-}MLM$ to generate the fluent and semantically consistent adversarial examples. Our adversarial sentence generation method is inspired by their work, but it differs from them in two ways: (1) We consider only the aspect and opinion terms to be vulnerable or **important words** in the given sentence (2) We replace these **important words** in the sentence using $BERT\text{-}MLM$ by applying our scoring function.

We take the original input sentence S and pass it to $BERT\text{-}MLM$ for finding the replacement candidates for important words $IW \subset S$ and then generating semantically coherent replacements.

$$P = BERT\text{-}MLM(S) \tag{21}$$

$$C_i = TopK(Filter(P_i)) \tag{22}$$

$BERT\text{-}MLM$ provides replacement probability score $P \in \mathbb{R}^{N \times |V|}$ where V is the vocabulary set used by BERT. We take all replacement candidates $P_i \in P$ of an important word $IW_i \in IW$ and pass it to $Filter$ and $TopK$ function in a sequence, where first all stops words, punctuation, and antonyms are removed and then top K words are selected as replacement candidates C_i based on its probability score.

We replace important words $IW_i \in IW$ of the sentence S with replacement candidates C_i, after each replacement, a modified sentence is created. We compute the semantic similarity score between each pair of the original sentence and a modified sentence using a Universal Sentence Encoder USE [44]. Modified sentences that carry a similarity score of more than a pre-defined threshold $sim\text{-}threshold$ are considered as candidate adversarial examples for the sentence S. We pass such candidate adversarial examples one by one to a pre-trained unified model $\mathcal{M}$ to calculate $Score$ by taking the sum of predicted probabilities corresponding to true labels of each important word. The lower $Score$ associated with a candidate adversarial example indicates its better ability to fool the model's prediction. Hence, among candidate adversarial examples, we consider the one which has the lowest $Score$ as an adversarial example for the sentence S.

Previous work [45] has shown that perturbations transferred from other models help the original model to become more robust to black-box attacks on image data. Motivated by this, we choose a strong baseline model $BERT\text{+}Linear$ [46] and use it as a pre-trained unified model $\mathcal{M}$ in the above-mentioned black-box attack to generate adversarial examples. Algorithm 2 describes the steps involved in the generation of adversarial examples.

Algorithm 2 Generate adversarial example.

for w_i *in* S **do**
    /* S is a input sentence */
    if *word* w_i *is in* $Important\text{-}Words$ **then**
        for *each replacement candidate* c_k *in* C_i **do**
            /* C_i are replacement candidates */
            create a new adversarial example S' by replacing w_i with c_k
            if $SemanticSim(S, S') < sim\text{-}threshold$ **then**
                /* If the adversarial example is not semantically similar to original sentence, drop the example */
                continue
            /* $Score$ function calculates the sum of predicted probabilities of a model for each important word of S' corresponding to true labels */
            finally, out of all possible adverserial examples select the one that has the least predicted probability calculated through the $Score$ function
        end
end

3.1.6 Adversarial training

We take original input and its corresponding adversarial example to train BILEAT, where the standard loss $\mathcal{L}$ for original input is calculated as mentioned in (20) and on the similar lines adversarial loss $\mathcal{L}_{adv}$ is calculated for the adversarial example. Total loss $\mathcal{L}_t$ of BILEAT is defined as follows:

$$\mathcal{L}_t = \mathcal{L} + \gamma \mathcal{L}_{adv} \tag{23}$$

Where, γ is a hyperparameter that controls the contribution of adversarial loss $\mathcal{L}_{adv}$. During training, model loss $\mathcal{L}_t$ is minimized.

4 Experiments

4.1 Datasets

We conduct experiments using Laptop and Restaurant review datasets taken from SemEval ABSA challenges. These datasets are re-prepared by Li et al. in [10]. The laptop

Table 2 Details of the laptop & restaurant datasets

Dataset		Train	Dev	Test	Total
Laptop	# POS # NEG # NEU	883 754 404	104 106 46	339 130 165	1326 990 615
Restaurant	# POS # NEG # NEU	2337 942 614	270 93 50	1524 500 263	4131 1535 927

dataset is prepared using SemEval ABSA challenge 2014 [47], which contains a train-test split same as the original dataset. Restaurant review dataset is union of SemEval ABSA challenge 2014, 2015 and 2016. The training dataset of Restaurant is created by merging training dataset of three years, and a similarly testing dataset is also built. For both the datasets, we take 10% randomly held-out of training data as the development set. Table 2 presents details of the datasets.

4.2 Baseline methods

We compare the performance of BILEAT with two groups of baseline. The first group has those models, wherein results are either copied or reproduced using original code from other papers.

- **CRF-Unified (Mitchell et al.** [48]): built this model by leveraging hand-crafted linguistic features with CRF to perform the sequence labeling task using a unified approach.
- **NN+CRF-Unified** (Mitchell et al. [48]): is an improved version of **CRF-Unified**, where target word embedding and context word embeddings are concatenated, and also hand-crafted linguistic features are used with CRF.

  We have taken the results of the above two models from [9].
- **LSTM+ CRF** : is the standard LSTM model that uses CRF layer for predicting the unified tags.
- **HAST-TNet** (Li et al. [3, 23]): is a pipeline approach based on two state-of-the-art models HAST [3] and TNet [23] on the tasks of aspect-term extraction and aspect sentiment classification respectively.
- **CMLA** (Wang et al. [40]): is a multi-layer coupled-attention architecture. Each layer of CMLA has two coupled GRUs that performs aspect and opinion terms co-extraction.
- **E2E-TBSA** (Li et al. [10]): is an end-to-end model that adopts unified tagging scheme to address complete ABSA task.

  We have taken results of HAST-TNet, and E2E-TBSA from [10].
- **IMN** (He et al. [26]): jointly learns multiple related tasks simultaneously using an iteratively message passing architecture.
- **DOER** (Luo et al. [9]): provides a framework that extracts aspect and its polarity simultaneously. It employs a dual RNN to extract the respective representation of each task, and a cross-shared unit help in understanding the relationship between each other.
- **E2E-Triplet-Unified** (Peng et al. [41]): provides an end-to-end framework that extracts triplet in the form of aspect-term, its sentiment, and associated opinion word. In one of the versions of their model, they also adopt a unified tagging scheme to only extract aspect-term and its sentiment. We call that version E2E-Triplet-Unified.

  Results of CMLA, and E2E-Triplet-Unified are taken from [41].
- **BERT+Linear** (Li et al. [46]): uses BERT to generate representations for tokens in a sentence, these representations are passes to a linear layer to address ABSA in a unified manner.
- **BERT+GRU** (Li et al. [46]): applies a stacked architecture of BERT with GRU to solve ABSA in a unified manner.
- **BERT+SAN** (Li et al. [46]): applies a stacked architecture of BERT with a self-attention network (SAN) to solve ABSA in a unified manner.

  We reproduce the results of BERT-Linear, BERT-GRU, and BERT-SAN using its original code.
- **GRACE** (Luo et al. [30]): provides a multi-head attention architecture with virtual adversarial training that uses a gradient harmonized method. We have taken the result of this model from the respective paper.

  The second group has variations of our **BILEAT** model. These models are also used in ablation studies.
- **BILEAT w/o BBT:** it doesn't use black-box adversarial training.
- **BILEAT w/o WBDK-BERT:** uses DK-BERT instead of white-box domain knowledge BERT (WBDK-BERT).
- **BILEAT w/o BBT & WBDK-BERT:** Both black-box adversarial training and WBDK-BERT are not used. For token representation generation DK-BERT is used in this model.
- **BILEAT w/o BBT, WBDK-BERT& DK-BERT:** Black-box adversarial training, WBDK-BERT, and DK-BERT are removed. $BERT_{BASE}$ is used to generate the token representation in this model.

- **BILEAT w/o BBT, WBDK-BERT, DK-BERT & A/O:** aspect-to-opinion interaction is switched off in this model, and also Black-box adversarial training, WBDK-BERT and DK-BERT are removed. $BERT_{BASE}$ is used to generate the token representation in this model.
- **BILEAT w/o BBT, WBDK-BERT, DK-BERT & O/A:** opinion-to-aspect interaction is switched off in this model, and also Black-box adversarial training, WBDK-BERT and DK-BERT are removed. $BERT_{BASE}$ is used to generate the token representation in this model.
- **BILEAT w/o BBT, WBDK-BERT, DK-BERT, A/O & O/A:** both opinion-to-aspect and aspect-to-opinion interactions are switched off in this model, and also Black-box adversarial training, WBDK-BERT, and DK-BERT are removed. $BERT_{BASE}$ is used to generate the token representation in this model.

4.3 Evaluation metrics

We evaluate the performance of a model using Precision (P), Recall (R), and F-score (F), which means that extracted aspect is considered to be correct when it exactly matches with the gold standard span of the mentioned aspect and its corresponding sentiment.

4.4 Settings

In order to build post-trained WBDK-BERT, we use Amazon laptop reviews and Yelp Dataset Challenge reviews provided by [49] and perform post-training of $BERT_{BASE}$. We do white-box adversarial training of WBDK-BERT by following [34], and set the gradient steps K to 1, the variance for initializing perturbation σ to 0.00001 and the step size η as 0.001. For generating a black-box adversarial example, we set *sim-threshold* to 0.8. BILEAT uses hidden size d as 768, learning rate of 4e-5 with *Adam* optimizer and batch size of 16 for both datasets. We train the model up to 2000 steps. After training for 1000 steps, we conduct model selection on the development set for every 100 steps using the F-score for comparison. For the auxiliary task of opinion term classification, we use the existing opinion lexicon[2] to provide opinion words. For the hinge loss function, we set the *margin* as 1, the loss contribution of auxiliary tasks λ is set to 0.1, and the loss contribution of adversarial example γ is set to 0.2.

[2] http://mpqa.cs.pitt.edu/

5 Results & discussion

- ***Generalization***: We discuss the generalization aspect of our proposed model by comparing its performance with various baseline models. Table 3 shows that our proposed model **BILEAT** performs the best on both Restaurant and Laptop datasets. CRF-based models perform quite poorly among all the baseline models. The performance of CRF depends on the quality of handcrafted features; also, CRF, like traditional statistical models, focuses more on learning explicit features and is unable to learn implicit features efficiently. This could be the primary reason for the poor performance of *CRF-Unified*. However, the performance of another CRF-based model *NN+CRF-Unified* enhances slightly by utilizing the pre-trained word embeddings. *LSTM+CRF* uses LSTM to encode the meaning of a word in the input text by learning latent features efficiently and utilizing CRF for classification. This architecture makes *LSTM+CRF* a strong classifier and helps in performing better than *CRF-Unified*. *HAST* and *TNet* are attention mechanisms based on two different models on the tasks of aspect extraction and sentiment classification, respectively. *HAST-TNet* is built by integrating both strong models in a pipeline. Thus, *HAST-TNet* perform better than *LSTM+CRF* on both the datasets with a good margin. *CMLA* is a multi-layer coupled-attention architecture that helps it performing slightly better than *HAST-TNet*. Instead of relying on the pipeline approach, *E2E-TBSA* is built by using two stacked recurrent neural networks to explore the inter-task dependency and predict the aspect-term and its related sentiment in a unified way. Learning inter-task dependency using gate mechanism helps *E2E-TBSA* to perform better than *HAST-TNet* and *CMLA*. It shows that a nicely designed integrated model can be more effective than pipeline-based methods. Unlike conventional multi-task learning methods *IMN* jointly learns common features for the different tasks using an iteratively message passing architecture. This unconventional architecture enables *IMN* to perform better than *E2E-TBSA*. *DOER* and *E2E-Triplet-Unified* are a joint model that learns from the interaction between the two relevant tasks. Learning through mutual influence through a cross-shared unit may be one potential reason for both these models to perform better than *IMN*.

  Language models based on Transformer architectures are very powerful in understanding the contextual meaning of the word. Hence, *BERT+Linear* that uses BERT embeddings to encode the meaning of the word with a linear output layer surprisingly outperforms many existing strong and complex architecture-based baselines (i.e., CMLA, IMN, and E2E-TABSA),

Table 3 Precision, Recall and F score of experimental results on original test datasets

Model	Laptop			Restaurant		
	P	R	F	P	R	F
CRF-Unified	-	-	49.24	-	-	59.52
NN+CRF-Unified	-	-	50.64	-	-	61.74
LSTM+CRF	58.61	50.47	54.24	66.10	66.30	66.20
HAST-TNet	56.42	54.20	55.29	62.18	73.49	67.36
CMLA	54.70	59.20	56.90	-	-	-
E2E-TBSA	61.27	54.89	57.90	68.64	71.01	69.80
IMN	-	-	58.37	-	-	-
DOER	-	-	60.35	-	-	72.78
E2E-Triplet-Unified	63.15	61.55	62.34		-	-
BERT-Linear	61.30	58.20	59.70	70.93	73.72	72.29
BERT-GRU	61.50	58.20	59.80	66.77	73.37	69.91
BERT-SAN	63.05	60.57	61.78	69.78	75.03	72.31
GRACE	72.38	69.12	70.71	75.95	80.31	78.07
BILEAT w/o BBT, WBDK-BERT, DK-BERT, O/A & A/O	65.05	61.36	63.14	70.45	75.69	72.97
BILEAT w/o BBT, WBDK-BERT, DK-BERT & O/A	64.69	61.83	63.22	70.69	75.51	73.02
BILEAT w/o BBT, WBDK-BERT, DK-BERT & A/O	65.38	61.67	63.47	71.96	74.95	73.42
BILEAT w/o BBT, WBDK-BERT & DK-BERT	65.85	63.88	64.85	71.82	77.35	74.48
BILEAT w/o BBT & WBDK-BERT	67.57	66.72	67.14	73.99	78.22	76.04
BILEAT w/o WBDK-BERT	68.02	69.93	68.96	73.84	79.10	76.37
BILEAT w/o BBT	69.65	68.77	69.20	75.08	80.06	77.49
BILEAT (Our Model)	***74.68***	***71.53***	***73.07***	***78.13***	***82.07***	***80.05***

Bold italic entries show the superiority of the performance of our proposed model (BILEAT) compared to other baseline models

which does not use BERT embeddings. The architecture of *BERT+GRU* and *BERT+SAN* are also based on BERT but use more powerful output layers like GRU and Self-Attention-Network, respectively. Such output layers lead both these models to achieve better performance than *BERT+Linear*. *GRACE* uses domain-specific post-trained BERT and applies a gradient harmonized method along with virtual adversarial training. The domain-specific BERT embedding and virtual adversarial training enhances the performance of *GRACE* and helps in performing better than all the baselines including *BERT+Linear*, *BERT+GRU*, and *BERT+SAN*. Our proposed model **BILEAT** outperforms better than the strongest baseline model *GRACE* on both the datasets. F-score comparison shows that **BILEAT** performs better than *GRACE* by a margin of **2.36%** and **1.98%** on Laptop and Restaurant datasets, respectively.

– ***Robustness***: We compare the robustness of **BILEAT** and its variants with some strong baseline models against adversarial attacks. We evaluate the performance of models using Laptop and Restaurant adversarial test datasets (details are mentioned in Section 3.1.5) generated by us. Table 4 shows that *GRACE* performs the best among all the baselines. *GRACE* is built using a post-pretraining BERT and virtual adversarial training (VAT), which makes this model robust and helps in performing better than other baseline models e.g. (*E2E-TABSA*, *BERT-Linear*, *BERT+GRU*, and *BERT+SAN*). Comparison of F1-score shows **BILEAT** outperform the strongest baseline model *GRACE*) by a margin of **3.63%** and **3.91%** on the Laptop and Restaurant adversarial test dataset, respectively. Ensemble adversarial training is the primary reason that makes **BILEAT** such a robust model and helps it to perform better than all baselines.

In terms of both generalization and robustness, the better performance of **BILEAT** over various baselines can be attributed to the following reasons:

- **BILEAT** utilizes interactive learning between AE and OE auxiliary tasks to produce a collaborative signal;
- **BILEAT** uses a domain-specific and white-box adversarially trained WBDK-BERT built by us to generate more effective context-aware word embeddings;
- By applying a black-box attack, quality adversarial examples are generated, **BILEAT** is trained using both original inputs and such adversarial examples in a combined way by doing task-specific fine-tuning of WBDK-BERT. This combined training makes **BILEAT** more robust and generalized.

5.1 Ablation study

To understand the effectiveness of different key components in improving the generalization and robustness of **BILEAT**, we conduct the ablation study.

In order to do an ablation study for generalization, we sequentially remove each component one after another and obtain six simplified variants. The second block of Table 3 has the results of all different variants of **BILEAT**. The result shows that each O/A and A/O component are individually contributing to improving the performance. However, when both these components are combined, the performance gets enhanced to a large extent. Result analysis also reveals that both white-box and black-box adversarial training individually contribute to improving the performance, and when both these training are combined, the performance gain is even better.

For a robustness ablation study, we evaluate the performance of our proposed model on the adversarial test datasets by sequentially removing both white-box and black-box adversarial components one after another and obtain three variants. The second block of Table 4 has the results of all different variants of **BILEAT**. The result shows the inclusion of white-box adversarial pre-trained WBDK-BERT, and black-box examples individually contribute to improving the robustness of **BILEAT**. Performance of *BILEAT w/o BBT* is marginally better than *BILEAT w/o WBDK-BERT*, which means the contribution of black-box examples is slightly larger than WBDK-BERT in enhancing the robustness of **BILEAT**. **BILEAT** performs better than both *BILEAT w/o WBDK-BERT* and *BILEAT w/o BBT*, which reveals combining WBDK-BERT with black-box adversarial examples further enhances the robustness of **BILEAT**.

5.2 Case study

In this subsection, we show the effectiveness of key components of BILEAT by presenting a case study. We pick some review sentences from our original datasets to study the contribution of important components in enhancing the performance of our proposed model. Table 5 presents the details of predicted aspect(s) and its related sentiment in a given sentence by various models. Actual aspects and their associated sentiment are shown in the bold italic font style under the *Sentence* column. In the first two sentences only *BILEAT* and *BILEAT w/o BBT* are able to make correct predictions. It shows that white-box adversarially post-trained WBDK-BERT contributes to improving the generalization of *BILEAT*. Similarly, the prediction results of the third and fourth sentences quantify the importance of black-box adversarial examples, as only *BILEAT w/o WBDK-BERT* and *BILEAT* make the correct prediction. In the result of the sixth sentence, expect *BILEAT w/o WBDK-BERT & BBT* all models make the correct prediction. It reveals that at the individual level,

Table 4 Precision, Recall and F-score of experimental results on adversarial test datasets

Model	Laptop			Restaurant		
	P	R	F	P	R	F
E2E-TBSA	57.51	49.53	53.22	63.81	60.12	61.90
BERT-Linear	46.02	60.47	52.26	64.95	71.23	67.94
BERT-GRU	47.11	66.28	55.07	67.34	71.58	69.39
BERT-SAN	58.60	55.36	56.93	66.86	70.75	68.74
GRACE	65.74	66.17	65.95	71.42	75.83	73.55
BILEAT w/o BBT & WBDK-BERT	65.07	61.99	63.48	71.40	74.90	73.11
BILEAT w/o WBDK-BERT	66.15	67.19	66.66	73.97	78.05	75.95
BILEAT w/o BBT	66.51	65.14	65.81	72.01	76.26	74.07
BILEAT (Our Model)	***68.40***	***70.82***	***69.58***	***75.70***	***79.31***	***77.46***

Bold italic entries show the superiority of the performance of our proposed model (BILEAT) compared to other baseline models

Table 5 Case study : Original Input sentences with predicted aspect-terms & its sentiment by various models

Sentence	BILEAT w/o WBDK-BERT & BBT	BILEAT w/o BBT	BILEAT w/o WBDK-BERT	BILEAT
Air has higher [***resolution***]$_{\mathbf{POS}}$ but the [***fonts***]$_{\mathbf{NEG}}$ are small.	[*resolution*]$_{NEG}$, [*fonts*]$_{NEG}$	[*resolution*]$_{POS}$, [*fonts*]$_{NEG}$	[*resolution*]$_{POS}$, [*fonts*]$_{POS}$	[*resolution*]$_{POS}$, [*fonts*]$_{NEG}$
The [***baterry***]$_{\mathbf{POS}}$ is very longer.	[*baterry*]$_{NEG}$	[*baterry*]$_{POS}$	[*baterry*]$_{NEG}$	[*baterry*]$_{POS}$
The [***staff***]$_{\mathbf{POS}}$ is unbelievably friendly, and I dream about their [***Saag gosht***]$_{\mathbf{POS}}$ good.	[*staff*]$_{POS}$, [*Saag*]$_{POS}$	[*staff*]$_{POS}$, [*Saag*]$_{POS}$	[*staff*]$_{POS}$, [*Saag gosht*]$_{POS}$	[*staff*]$_{POS}$, [*Saag gosht*]$_{POS}$
Having [***USB3***]$_{\mathbf{POS}}$ is why I bought this Mini.	[*USB3*]$_{NEU}$	[*USB3*]$_{NEU}$	[*USB3*]$_{POS}$	[*USB3*]$_{POS}$
The [***staff***]$_{\mathbf{NEG}}$ should be a bit more friendly.	*None*	[*staff*]$_{NEG}$	[*staff*]$_{NEG}$	[*staff*]$_{NEG}$
The [***battery life***]$_{\mathbf{POS}}$ is excellent 6-7 hours without charging.	[*battery life*]$_{POS}$, [*charging*]$_{NEU}$	[*battery life*]$_{POS}$, [*charging*]$_{NEU}$	[*battery life*]$_{POS}$, [*charging*]$_{NEU}$	[*battery life*]$_{POS}$
Great open and friendly [***ambience***]$_{\mathbf{POS}}$.	[*open*]$_{POS}$, [*ambience*]$_{POS}$	[*open*]$_{POS}$, [*ambience*]$_{POS}$	[*open*]$_{POS}$, [*ambience*]$_{POS}$	[*ambience*]$_{POS}$

both WBDK-BERT and black-box training are sufficient to make the correct prediction. Result analysis of the last two sentences is very interesting as except *BILEAT* no other model is able to make the correct prediction. It exhibits that the combination of WBDK-BERT and black-box adversarial examples complement each other, and a combination of these two helps *BILEAT* to achieve better performance.

We also pick a few review sentences from generated adversarial datasets (refer Section 3.1.5 for details) to study the impact of key components in enhancing the robustness of our proposed model. Table 6 provides the details of adversarial examples, which are generated by replacing underlined blue text in the corresponding original sentence. This table also gives information about the predicted aspect(s) and related sentiment in the given adversarial example by various models. It is evident that generated adversarial examples are grammatically fluent, semantically coherent with the original sentence, and have misled the *BILEAT w/o WBDK-BERT & BBT* to make an incorrect prediction. It shows the quality of these adversarial examples. In the first two adversarial examples, except for *BILEAT w/o WBDK-BERT & BBT*, all other models have made a correct prediction. It shows the individual potential of WBDK-BERT and black-box adversarial examples in improving the robustness of our proposed model. The prediction results of the third and fourth adversarial examples exhibit the importance of black-box adversarial training, as only *BILEAT w/o WBDK-BERT* and *BILEAT* can make the correct prediction. Result analysis of fifth and sixth adversarial examples reveals the resultant effect of combining WBDK-BERT with black-box adversarial examples in our proposed model, as except *BILEAT* all other models make the wrong prediction. It shows the combined effect of WBDK-BERT and black-box adversarial examples in enhancing the robustness of *BILEAT*.

5.3 Error analysis

In some of the review sentences of our original test datasets, BILEAT 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:

– **Multi-token aspect terms containing three or more words:** Some review sentences contain aspect terms that have three or more words. For example, in the review sentence *"I opted for the SquareTrade 3-Year Computer Accidental Protection Warranty $1500-2000 which also supports accidents like drops and spills that are NOT covered by AppleCare"* aspect term is *"SquareTrade 3-Year Computer Accidental Protection Warranty"*.

Table 6 Case study : Adversarial Input sentences with predicted aspect-terms & its sentiment by various models

Original sentence	Adversarial example	BILEAT w/o BBT & WBDK-BERT	BILEAT w/o BBT	BILEAT w/o WBDK-BERT	BILEAT
BTW, I really like Long Beach.	BTW, I really love Long Beach.	[Beach]$_{\text{POS}}$	NONE	NONE	NONE
Not only can the ***[selection]***$_{\textbf{POS}}$ be innovative, but there's a nice balance of traditional ***[sushi]***$_{\textbf{POS}}$ as well.	Not only can the ***[selection]***$_{\textbf{POS}}$ be innovative, but there's a good balance of traditional ***[sushi]***$_{\textbf{POS}}$ as well.	[selection]$_{\text{POS}}$, [traditional]$_{\text{POS}}$, [sushi]$_{\text{POS}}$	[selection]$_{\text{POS}}$, [sushi]$_{\text{POS}}$	[selection]$_{\text{POS}}$, [sushi]$_{\text{POS}}$	[selection]$_{\text{POS}}$, [sushi]$_{\text{POS}}$
Casablanca servces delicious ***[falafel]***$_{\textbf{POS}}$, ***[tabouleh]***$_{\textbf{POS}}$, ***[humus]***$_{\textbf{POS}}$ and other ***[Mediterranean delights]***$_{\textbf{POS}}$, which are all very inexpensive.	Casablanca servces delicious ***[falafel]***$_{\textbf{POS}}$, ***[tabouleh]***$_{\textbf{POS}}$, ***[humus]***$_{\textbf{POS}}$ and other ***[Mediterranean delights]***$_{\textbf{POS}}$, which are all very cheap.	[servces]$_{\text{POS}}$, [falafel]$_{\text{POS}}$, [tabouleh]$_{\text{POS}}$, [humus]$_{\text{POS}}$, [Mediterranean]$_{\text{POS}}$, [delights]$_{\text{POS}}$	[servces]$_{\text{POS}}$, [falafel]$_{\text{POS}}$, [tabouleh]$_{\text{POS}}$, [humus]$_{\text{POS}}$, [Mediterranean]$_{\text{POS}}$, [delights]$_{\text{POS}}$	[falafel]$_{\text{POS}}$, [tabouleh]$_{\text{POS}}$, [humus]$_{\text{POS}}$, [Mediterranean delights]$_{\text{POS}}$	[falafel]$_{\text{POS}}$, [tabouleh]$_{\text{POS}}$, [humus]$_{\text{POS}}$, [Mediterranean delights]$_{\text{POS}}$
Not only is this the best Thai restaurant I have been to, but it also ranks as one of my favorite places to dine.	Not only is this the top Thai restaurant I have been to, but it also ranks as one of my favourite places to dine.	[dine]$_{\text{POS}}$	[dine]$_{\text{POS}}$	NONE	NONE
Finally, I got sick of the bad ***[service]***$_{\textbf{NEG}}$, obnoxious smirks, and snotty back talk.	Finally, I got sick of the terrible ***[service]***$_{\textbf{NEG}}$, obnoxious smirks, and snotty back talk .	[service]$_{\text{NEG}}$, [smirks]$_{\text{NEG}}$, [back]$_{\text{NEG}}$,[talk]$_{\text{NEG}}$	[service]$_{\text{NEG}}$, [back]$_{\text{NEG}}$	[service]$_{\text{NEG}}$, [back]$_{\text{NEG}}$, [talk]$_{\text{NEG}}$	[service]$_{\text{NEG}}$
I complained to the ***[manager]***$_{\text{NEG}}$, but he was not even apologetic.	I complained to the ***[manager]***$_{\text{NEG}}$, but he was not really apologetic .	NONE	NONE	NONE	[manager]$_{\text{NEG}}$

- **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"*.
- **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.
- **Implicit opinion expressed about aspect:** Some review comments do not express direct sentiment towards the aspects. The example includes *"Overall, I would go back and eat at the restaurant again"*, where sentiment about the aspect *"restaurant"* is expressed implicitly.

6 Conclusion

In this work, we investigate the importance of interactive learning and the effectiveness of adversarial training for unified ABSA tasks. We build a white-box adversarially post-trained domain knowledge BERT (WBDK-BERT) and use the same to generate robust and contextualized embeddings in our proposed model. To enhance the sentence representation for unified ABSA, we introduced two auxiliary tasks. Interactive learning between these two tasks produces a collaborative signal that helps in improving the performance of our model. In order to make our model more generalized and robust, we generated adversarial examples using a black-box technique and trained our model using original inputs and such adversarial examples in a combined way. The experimental results show the superiority of our proposed model in terms of generalization and robustness compared to existing methods. Future work will focus on extending our proposed method to non-English languages. The multilingual language models mBERT [12], and XLM-R [50] will be used to investigate how cross-lingual transfer helps to solve unified ABSA tasks in multilingual settings.

Declarations

- The authors have no relevant financial or non-financial interests to disclose.
- The authors have no conflicts of interest to declare that are relevant to the content of this article.
- 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.
- The authors have no financial or proprietary interests in any material discussed in this article.

References

1. Liu B (2012) Sentiment analysis and opinion mining. Synthes Lect Human Lang Technol 5(1):1–167
2. Yin Y, Wei F, Dong L, Xu K, Zhang M, Zhou M (2016) Unsupervised word and dependency path embeddings for aspect term extraction. arXiv:1605.07843
3. Li X, Bing L, Li P, Lam W, Yang Z (2018) Aspect term extraction with history attention and selective transformation. arXiv:1805.00760
4. Xu H, Liu B, Shu L, Yu PS (2018) Double embeddings and cnn-based sequence labeling for aspect extraction. arXiv:1805.04601
5. Tang D, Qin B, Liu T (2016) Aspect level sentiment classification with deep memory network. arXiv:1605.08900
6. Xue W, Li T (2018) Aspect based sentiment analysis with gated convolutional networks. arXiv:1805.07043
7. Li Z, Wei Y, Zhang Y, Zhang X, Li X (2019) Exploiting coarse-to-fine task transfer for aspect-level sentiment classification. In: Proceedings of the AAAI Conference on artificial intelligence, vol 33, pp 4253–4260
8. Wang F, Lan M, Wang W (2018) Towards a one-stop solution to both aspect extraction and sentiment analysis tasks with neural multi-task learning. In: 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, pp 1–8
9. Luo H, Li T, Liu B, Zhang J (2019) Doer: Dual cross-shared rnn for aspect term-polarity co-extraction. arXiv:1906.01794
10. Li X, Bing L, Li P, Lam W (2019) A unified model for opinion target extraction and target sentiment prediction. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 6714–6721
11. Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv:1412.6572
12. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
13. Ebrahimi J, Rao A, Lowd D, Dou D (2017) Hotflip: White-box adversarial examples for text classification. arXiv:1712.06751
14. Ilyas A, Engstrom L, Athalye A, Lin J (2018) Black-box adversarial attacks with limited queries and information. arXiv:1804.08598
15. Li L, Ma R, Guo Q, Xue X, Qiu X (2020) Bert-attack: Adversarial attack against bert using bert. arXiv:2004.09984
16. Qiu G, Liu B (2011) Opinion word expansion and target extraction through double propagation. Comput Linguist 37(1):9–27
17. Wang W, Pan SJ, Dahlmeier D, Xiao X (2016) Recursive neural conditional random fields for aspect-based sentiment analysis. In: Conference on empirical methods in natural language processing
18. Wang W, Pan SJ, Dahlmeier D, Xiao X (2017) Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In: AAAI Conference on artificial intelligence
19. He R, Lee WS (2017) An unsupervised neural attention model for aspect extraction. In: Annual meeting of the association for computational linguistics
20. Kumar A, Veerubhotla AS, Narapareddy VT, Aruru V, Neti LBM, Malapati A (2021) Aspect term extraction for opinion mining using a hierarchical self-attention network. Neurocomputing 465:195–204

21. Vo D-T (2015) Target-dependent twitter sentiment classification with rich automatic features. In: International joint conference on artificial intelligence
22. Wang Y, Huang M (2016) Attention-based LSTM for aspect-level sentiment classification. In: Conference on empirical methods in natural language processing
23. Li X, Bing L, Lam W, Shi B (2018) Transformation networks for target-oriented sentiment classification. arXiv:1805.01086
24. Li Z, Wei Y, Zhang Y, Zhang X, Li X, Yang Q (2019) Exploiting coarse-to fine task transfer for aspect-level sentiment classification. In: AAAI, pp 2237–2243
25. Kumar A, Narapareddy VT, Srikanth VA, Neti LBM, Malapati A (2020) Aspect-based sentiment classification using interactive gated convolutional network. IEEE Access 8:22445–22453
26. He R, Lee WS, Ng HT, Dahlmeier D (2019) An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. arXiv:1906.06906
27. Li X, Bing L, Zhang W, Lam W (2019) Exploiting BERT for end-to-end aspect-based sentiment analysis. In: Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), pp 34–41
28. Chen Z, Qian T (2020) Relation-aware collaborative learning for unified aspect-based sentiment analysis. In: ACL, pp 3685–3694
29. Liang Y, Meng F, Zhang J, Xu J, Chen Y, Zhou J (2020) An iterative knowledge transfer network with routing for aspect-based sentiment analysis
30. Luo H, Ji L, Li T, Jiang D, Duan N (2020) Grace: Gradient harmonized and cascaded labeling for aspect-based sentiment analysis. In: Proceedings of the 2020 conference on empirical methods in natural language processing: findings, pp 54–64
31. Mao Y, Shen Y, Yu C, Cai L (2021) A joint training dual-mrc framework for aspect based sentiment analysis
32. Oh S, Lee D, Whang T, Park I, Gaeun S, Kim E, Kim H (2021) Deep context- and relation-aware learning for aspect-based sentiment analysis. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)
33. Xu J, Du Q (2020) Texttricker: Loss-based and gradient-based adversarial attacks on text classification models. Eng Appl Artif Intell 92:103641
34. Liu X, Cheng H, He P, Chen W, Wang Y, Poon H, Gao J (2020) Adversarial training for large neural language models. arXiv:2004.08994
35. Karimi A, Rossi L, Prati A, Full K (2020) Adversarial training for aspect-based sentiment analysis with bert
36. Ebrahimi J, Rao A (2018) Hotflip: White-box adversarial examples for text classification. In: Proceedings of the 56th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Melbourne, pp 31–36
37. Feng S, Wallace E, Iyyer M, Rodriguez P, Grissom II A, Boyd-Graber JL (2018) Right answer for the wrong reason: Discovery and mitigation. CoRR, arXiv:1804.07781
38. Pruthi D, Dhingra B, Lipton ZC (2019) Combating adversarial misspellings with robust word recognition
39. Hofer N, Schöttle P, Rietzler A, Stabinger S (2021) Adversarial examples against a bert absa model – fooling bert with l33t, misspellign, and punctuation, In: The 16th international conference on availability, reliability and security, ARES 2021. Association for Computing Machinery, New York
40. Wang W, Pan SJ, Dahlmeier D, Xiao X (2017) Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In: Thirty-first AAAI conference on artificial intelligence
41. Peng H, Xu L, Bing L, Huang F, Lu W, Si L (2020) Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In: AAAI, pp 8600–8607
42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
43. Xu H, Liu B, Shu L, Yu PS (2019) Bert post-training for review reading comprehension and aspect-based sentiment analysis. arXiv:1904.02232
44. Cer D, Yang Y, Kong S, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C et al (2018) Universal sentence encoder. arXiv:1803.11175
45. Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P (2017) Ensemble adversarial training: Attacks and defenses. arXiv:1705.07204
46. Li X, Bing L, Zhang W, Lam W (2019) Exploiting bert for end-to-end aspect-based sentiment analysis. arXiv:1910.00883
47. Pontiki M, Papageorgiou H, Galanis D, Androutsopoulos I, Pavlopoulos J, Manandhar S (2014) Semeval-2014 task 4: Aspect based sentiment analysis. SemEval:27
48. Mitchell M, Aguilar J, Wilson T, Van Durme B (2013) Open domain targeted sentiment. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 1643–1654
49. He R, Lee WS, Ng HT, Dahlmeier D (2019) An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics
50. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2019) Unsupervised cross-lingual representation learning at scale. arXiv:1911.02116