SN Computer Science (2023) 4:756
https://doi.org/10.1007/s42979-023-02229-7

ORIGINAL RESEARCH

IAN-BERT: Combining Post-trained BERT with Interactive Attention Network for Aspect-Based Sentiment Analysis

Sharad Verma[1] · Ashish Kumar[1] · Aditi Sharan[1]

Received: 16 February 2023 / Accepted: 7 August 2023

Abstract

Aspect-based sentiment analysis (ABSA), a task in sentiment analysis, predicts the sentiment polarity of specific aspects mentioned in the input sentence. Recent research has demonstrated the effectiveness of Bidirectional Encoder Representation from Transformers (BERT) and its variants in improving the performance of various Natural Language Processing (NLP) tasks, including sentiment analysis. However, BERT, trained on Wikipedia and BookCorpus dataset, lacks domain-specific knowledge. Also, for the ABSA task, the Attention mechanism leverages the aspect information to determine the sentiment orientation of the aspect within the given sentence. Based on the abovementioned observations, this paper proposes a novel approach called the IAN-BERT model. The IAN-BERT model leverages attention mechanisms to enhance a post-trained BERT representation trained on Amazon and Yelp datasets. The objective is to capture domain-specific knowledge using BERT representation and identify the significance of context words with aspect terms and vice versa. By incorporating attention mechanisms, the IAN-BERT model aims to improve the model's ability to extract more relevant and informative features from the input text, ultimately leading to better predictions. Experimental evaluations conducted on SemEval-14 (Restaurant and Laptop dataset) and MAMS dataset demonstrate the effectiveness and superiority of the IAN-BERT model in aspect-based sentiment analysis.

Keywords Aspect-based sentiment analysis · Attention mechanism · BERT · Post-trained BERT

Introduction

The Web 2.0 era witnessed a drastic change in online shopping culture as customers are likely to share their opinions by writing reviews on online platforms. These reviews are subjective and vital for new customers willing to purchase the product. Moreover, it is advantageous for both product manufacturers and sellers as they gain insight into the preferences and dislikes of their product, allowing them to enhance it in the future.

Unlike document-level and sentence-level sentiment analysis, Aspect based Sentiment Analysis (ABSA) is a more specific type of sentiment analysis that concentrates on identifying the sentiment towards a particular aspect or characteristic in a given review. ABSA provides a finer-grained understanding of sentiment than the more broad-level approaches of document-level and sentence-level sentiment analysis. For example, given a review, "Excellent screen but the battery life could be better", the aspect "screen" is positive, and the aspect "battery life" is of negative sentiment in the given review. By considering aspect information, ABSA can avoid inaccurate sentiment analysis at the sentence level and better facilitate the comprehension of users' emotional expressions in various aspects. Jiang et al. [1] stressed the significance of aspects in sentiment classification tasks and stated that most errors could be attributed to neglecting the aspect information.

Methods to solve ABSA tasks can be broadly classified into three categories: Rule-based, Machine Learning, and

This article is part of the topical collection "Research Trends in Computational Intelligence" guest edited by Anshul Verma, Pradeepika Verma, Vivek Kumar Singh and S. Karthikeyan.

✉ Sharad Verma
sharadlnx@gmail.com

Ashish Kumar
ashishkumar2912@gmail.com

Aditi Sharan
aditisharan@mail.jnu.ac.in

[1] School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi 110067, New Delhi, India

Published online: 29 September 2023

Deep Learning models [2]. Rule-based models are based on criteria designed by considering the domain knowledge and linguistic patterns. These models have limited learning abilities because the rules they are built with are static, meaning they are applied to the data without being updated or changed. Machine learning models necessitate the labeling or annotating of training data for guidance. Machine learning models are constructed using feature extraction techniques (feature engineering) and continually refining the model parameters based on the training data. Deep learning models came into the picture to bypass the manual feature extraction and let the model learn the significant features by itself. The deep learning model is a Neural Network with a fixed architecture. One of the benefits of utilizing a Neural Network is its inherent adaptability, as it can adjust its parameters through errors generated during training. In recent years, deep learning has made breakthroughs in the fields of computer vision [3], speech recognition [4], NLP [5], and medical domain [6, 7].

In NLP tasks, the length of an input sentence can vary, necessitating the development of neural networks capable of accommodating such variability. This led to the creation of sequential models, which capture the meaning of a word dependent on the context within the word sequence. Rather than treating words independently, these models capture sequential information during training. Recurrent Neural Networks (RNNs) [8] were developed to capture sequence and sentence length. Long Short-Term Memory (LSTM) [9] is a sophisticated type of RNN designed to overcome the vanishing and exploding gradient issues [10] common in traditional RNNs. LSTMs utilize gates that selectively remember and forget learned information. Many variants of LSTMs, such as Gated Recurrent Unit (GRU) and Bidirectional LSTM (BiLSTM), were developed and used for sentiment analysis subtasks. For example, Target dependent LSTM (TD-LSTM) and Target Connection LSTM (TC-LSTM) [11] utilize LSTM models and aspect targets for sentiment analysis, and ATE-SPD [12] uses BiLSTM and CRF for ABSA tasks. A significant drawback of LSTM-based sequence models is that they cannot extract semantic information in parallel, resulting in substantial training time overhead.

Attention [13] is a crucial concept in deep learning. It enables the model to concentrate on a specific part of the sentence provided. It has been widely used in Sentiment Analysis tasks. Previous studies [14–16] have demonstrated that the efficiency of neural network models can be enhanced by focusing on specific aspect terms of input sentences and incorporating attention mechanisms. ATAE-LSTM [14] used Attention and LSTM for sentiment analysis. IAN [15] applies an attention mechanism between the aspect vector and its corresponding context vector, obtained from an LSTM using Glove embeddings[16]. Experimental results demonstrate that utilizing the interactive attention mechanism enhances sentiment classification performance. However, GloVe, a static encoding method used as word embedding in the IAN model, cannot perform dynamic differential encoding according to the context. In addition, the model employs different LSTM to separately learn the word semantic information of the aspect and the context.

Bidirectional Encoder Representation from Transformers(BERT) [17], a self-supervised masked language model, significantly enhances performance when fine-tuned for specific NLP tasks. BERT can model bidirectional context by utilizing a multi-layer self-attention mechanism. Variations of the BERT model have been employed as the state-of-the-art approach in ABSA [18–20].

The training of BERT is based on datasets from Wikipedia and BookCorpus. It aims to learn general-purpose knowledge from the extensive corpus data. Recent studies [21–23] show that learning domain-specific knowledge would be more beneficial since it can capture long-tailed information, which could be important for domain-specific end tasks. Post-trained BERT, trained on a specific end task dataset, performs better than general-purpose BERT trained on the Wikipedia and BookCorpus datasets.

To this end, an IAN-BERT aspect sentiment analysis model is proposed. It leverages post-trained BERT to dynamically encode word vectors that utilize transformer encoder [24] to extract semantic features in a parallel manner. It alleviates the problems of static word encoding and significant time overhead in the LSTM sequence sentiment analysis model. The contextualized representation is refined using an attention mechanism between the aspect and context vectors. Finally, the sentiment classification layer determines the sentiment orientation for the aspect. Tests conducted on the SemEval-14 Restaurant and Laptop and MAMS datasets demonstrate that our model is more precise than the baseline model.

The following are the key contributions of this work:

1. We present a new approach, IAN-BERT, which leverages the BERT representations and recognizes the mutual influence of aspects and context.
2. We tried to create a model incorporating an attention mechanism that interacts between the aspect and context representation obtained through the self-attention mechanism.
3. We utilized a post-trained BERT model trained on Yelp and Amazon review datasets to obtain representation with domain-specific knowledge.

The rest of the paper is organized as follows: “Related Work” highlights BERT, post-trained BERT, and attention-based approaches for sentiment analysis. “Proposed Work” provides a comprehensive explanation of our proposed method, IAN-BERT. “Experiments and Results” presents the experiment setup and result analysis. Finally, the paper is concluded in “Discussion and Conclusion”.

Related Work

Significant advancements have been made in NLP, particularly in tasks such as Question Answering, Sentiment Analysis, and Named Entity Recognition, due to the utilization of large pre-trained language models. In this work, we have used post-trained BERT and Attention for the ABSA task. BERT shows significant improvement over models using static encoding-based representation. Further, the post-trained BERT model on domain-specific datasets performs better than the standard BERT. Prior work based on BERT, post-trained BERT, and Attention mechanism for ABSA are discussed in the following subsections:

BERT-Based ABSA

Li et al. [18] investigated the utilization of BERT-derived contextualized embeddings for the ABSA tasks. They have designed neural baselines to deal with ABSA. Semeval laptop and restaurant dataset is used in this work. BERT output is combined with GRU, self-attention (SAN), Transformer, and CRF for comparison purposes. The outcome indicates that BERT-GRU excels in the laptop dataset while BERT-SAN is superior in the restaurant dataset. The author has claimed that the BERT-based model can also give comparable results with simple linear layers.

Li et al. [19] introduced a new unified approach for two subtasks of E2E-ABSA. They utilized two stacked recurrent networks for the task. The first RNN finds the target boundary for the aspect, which the second RNN further uses for predicting unified tags as the final output. They have explicitly modeled the constrained transition from the first task to the second for inter-task dependency. They have utilized a gate mechanism for sentiment consistency that captures the relation between contiguous words. Experiments were performed on Semeval laptop, Restaurant, and Twitter datasets.

Hu et al. [20] have proposed a span-based extract-then-classify framework for Open-domain targeted sentiment analysis. Instead of using a sequence tagging scheme, they experimented with finding all the spans containing aspect words, and it is further used for the polarity detection task. They have investigated three approaches: pipeline, joint, and collapsed models. The proposed framework is divided into two parts, a multi-target extractor and a polarity classifier which are utilized for two subtasks. Experiments were performed on Semeval Laptop, restaurant dataset, and Twitter dataset. The pipelined method outperforms the other two techniques.

Post-trained BERT for Downstream Tasks

Xu et al. [21] have introduced a new task called Review Reading Comprehension (RRC), where they utilize review sentences as a source of knowledge to answer questions from users. They build the dataset ReviewRC by taking data from the popular benchmarks for ABSA. BERT is used as the base model for this work. They proposed a post-training approach since the dataset is limited and standard BERT lacks domain-specific knowledge. The joint post-training process enhances domain and task knowledge that improves the performance of BERT for RRC. Experiments have shown that utilizing the post-training technique for aspect extraction and aspect sentiment classification tasks leads to enhanced performance than BERT.

Xu et al. [23] examined the hidden representation acquired from BERT for the ABSA task. The author claimed Masked LM(MLM) learns fine-grained features and treats each word/token equally. While learning aspect representation, it focuses on aspect features rather than opinions and vice versa. This method has proven advantageous for extracting the feature but not for sentiment classification. Many end task examples are required to map BERT feature space. They conclude with the requirement of alternative learning tasks besides MLM. The major drawback of MLM is equal treatment for all words. Being a sentiment word or aspect word does not affect MLM. Aspect representation may be obtained by grouping reviews for the same item so that model gets the clue for the aspect of the item. Sentiment representation may be obtained by considering rating.

To bridge the gap between standard language models (such as ELMo and BERT) and domain understanding, Xu et al. [22] designed a language model capable of capturing the domain knowledge guided by the end tasks. It tries to combine the standard language model trained on large and mixed domain datasets with low-resource domain-specific knowledge. The authors have introduced DomBERT, a language model that expands upon BERT by incorporating domain knowledge. It facilitates learning domain knowledge-enhanced language models while using low resources. The results from the SemEval Laptop and Restaurant datasets showed that incorporating domain knowledge into BERT results in improved performance across various tasks.

Attention-Based ABSA

Song et al. [25] have introduced an attentional encoder network (AEN) that employs attention mechanisms rather than the traditional RNN architecture to capture the aspect and context interrelationship. The model consists of two attention layers: the Attentional Encoder and Target-Specific Attention layers. The Attentional Encoder layer encodes the interaction between words of a sentence. In the first layer, multi-head attention is employed to analyze context and context-aware aspects. A convolution operation (PCT) transforms the information obtained from the Multi-Head Attention (MHA), and the same transformation is applied to each token. The target-specific attention layer, the second attention layer, employs another multi-head attention mechanism to acquire a context representation specifically tailored to the target. They used label Smoothing regularization (LSR) for label unreliability issues. The experiments were conducted on three datasets: SemEval-14 restaurant and laptop dataset, and the ACL-14 Twitter dataset, having three sentiment polarities: positive, negative, and neutral. Pretrained BERT has been employed in this work.

Wu et al. [26] have used an attention mechanism to generate representations of aspect and context for aspect-based sentiment analysis. The proposed model employs attention for sequence modeling, ensuring that the context and aspect are aware of each other during the modeling process. It eliminates the need to model aspects and context separately and can have an interactive aspect context representation. The model can deal with multiple aspects and sentiments in the given sentence. Experiments were conducted on Twitter and Semeval14 datasets.

Ma et al. [15] divided the input sentence into aspect and context, utilizing Long Short-Term Memory (LSTM) to learn the sequential representation of the aspect target separately from its context and vice versa. A new Interactive Attention Network (IAN) has been suggested to determine the mutual significance of context and aspect terms and create two different representations. Together, these depictions reflect the target aspect and its surroundings, enabling the identification of the sentiment orientation of the specified aspect. The author has designed a variety of models for performance comparison. The idea behind the model is to extract important information from both aspects and context independently and then utilize the resulting combined representation for sentiment identification. Experimental results on the SemEval-14 dataset indicate that the IAN model effectively captures relevant features and provides crucial information for the sentiment classification task.

Ambartsoumian et al. [27] investigated the efficacy of the Self Attention Network (SAN) for various tasks in sentiment analysis. Experiments conducted on six datasets demonstrated the superior performance of SAN as it outperforms RNN and CNN variants. The SAN model competes on benchmark metrics, including training time, memory consumption, and accuracy. Several modifications, such as the number of heads and sequence position information, have been explored in this work. Experiments were conducted on word embeddings obtained from the Word2Vec algorithm. This works emphasize the importance of relative position representation and claim that it performs better than other variants of position encodings.

BERT is trained on a general-purpose dataset; it possesses a context-aware representation but lacks domain-specific knowledge. Post-trained BERT on domain-specific datasets can effectively capture the domain knowledge, resulting in word representations specific to that domain. It motivates us to use post-trained BERT instead of standard BERT for contextualized representation. Also, the core of BERT is centered around self-attention. Self-attention aims to comprehend the importance of words with one another. Each word is treated equally regardless of whether it is an aspect word. So it cannot use explicit knowledge in the form of aspect information fed to the model. It motivates us to use attention mechanism on top of self-attention-based contextualized representation.

Proposed Work

This section will provide a brief overview of the components of our model, followed by an in-depth explanation of our proposed model.

BERT Model

BERT [17] is the encoder component of the Transformer [24] architecture that has been specifically designed for natural language processing. The given input is transformed into a contextualized representation through a self-attention mechanism. Contrary to earlier language models such as Word2Vec [28], Glove [16], and ELMo [29], BERT learns the representation by considering the context from both directions, hence the name bidirectional. BERT uses a fine-tuning approach that makes it appropriate for end tasks.

Unlike Word2Vec, which generates static embeddings, BERT considers the context and relates each word (token) to all other words (tokens) in the given sentence. Hence it creates dynamic embeddings, aware of its context. BERT utilizes a multi-headed attention mechanism to make a context-based representation for every word in the input sentence. It further uses residual connection and layer normalization to obtain the final representation.

Along with token embedding, BERT takes segment embedding and position embedding as input. Segment embedding emphasizes a sentence out of two sentences given as input. Also, since the BERT model eliminates the need for recurrence required in previous sequential models and processes all words simultaneously, it uses position embedding to sense the relative position of words in the given sentence.

BERT has two standard configurations, viz. BERT-base and BERT-large as shown in Table 1

Table 1 Parameters of BERT

	BERT-base	BERT-large
No. of encoder layers	12	24
Hidden dimension	768	1024
No. of attention heads	12	16
No. of parameters	110 Million	330 Million
Vocabulary size	30 K	30 K

In this work, we have used bert-base-uncased and bert-large-uncased (here uncased signifies words have been lowercased before tokenization)

Post-trained BERT Model

The BERT model was trained using the Wikipedia and BookCorpus datasets. These datasets contain information from various domains. It is beneficial for learning the word vectors since words become aware of different possible contexts while learning the vector representation. However, the training dataset belongs to a specific domain for a downstream task like ABSA. Post-training on the BERT model is done to utilize the domain knowledge so that words become aware of domain-specific context. Post-training BERT model, trained on diverse and extensive datasets, equips the language model with restricted domain-specific expertise. In this study, we utilized a post-trained BERT model that was trained on a combination of Amazon and Yelp datasets [21].

IAN Model

Ma et al. [15] created an Interactive Attention Model (IAN) to capture context's impact on aspect terms and vice versa. Glove word vectors are fed to the LSTM model for context-aware representation. Further, context and aspect are separated, and attention mechanisms are applied. The two representations obtained from different attention mechanisms are concatenated. Finally, a dense layer containing three nodes corresponding to positive, negative, and neutral sentiment and a nonlinear activation function is applied for sentiment classification.

In our research, we have applied an interactive attention mechanism that facilitates the exchange of information between the aspect and context vectors, both of which were derived from post-trained BERT.

Proposed Model

In this study, we have integrated post-trained BERT with interactive attention to determine the sentiment polarity of a specified aspect within a sentence. The design of our model is illustrated in Fig. 1.

The BERT model initially processes the sentence for a given sentence and aspect to obtain a context-aware representation. Further, the context and aspect representations are separated. Attention is applied to these two representations to find the contribution of these representations to each other. Finally, the resultant vectors obtained after applying the attention mechanism are concatenated, and the sentiment classification layer is used to find the sentiment polarity of the aspect in the given sentence. The following subsections describe these steps in detail:

Contextualized Representation Using BERT

In this work, we have used BERT and post-trained BERT to obtain the contextualized representation of input words/tokens. Unlike earlier language models, BERT focuses on the entire input sentence to compute the vector representation of tokens. Given an input sentence and aspect as

$$S = [w_1, w_2, \dots, w_n] \tag{1}$$

$$A = [a_1, a_2, \dots, a_m] \tag{2}$$

where $A \in S$, n and m are sentence and aspect length, respectively. The input vector fed to BERT is represented as

$$E = [e_1, e_2, \dots, e_n] \tag{3}$$

where

$$e_t = T_t + S_t + P_t \tag{4}$$

Here, T_t represents token embedding, S_t represents segment embedding, and P_t represents position embedding corresponding to the input token w_t. The L layer transformer processes the input vector E to produce refined context-aware token features. Specifically, the final representation of input after passing by all L-layers of the transformer is represented in Eq. 5:

$$H_L = \text{Transformer}_L(E) \tag{5}$$

For a more detailed understanding of how BERT works, readers should refer to [24]. We consider $H_L = [h_1, h_2, \dots, h_n]$ as the context-aware representations of the input tokens and utilize them for prediction in subsequent tasks.

Attention Mechanism

In this step, context-aware word representation obtained from the BERT model is utilized to ascertain the contribution of context words on aspect terms and vice versa. The contextualized representation is separated into context and aspect. Suppose we have an input sentence of n tokens and an aspect of m tokens, and the aspect appears at k^{th} position

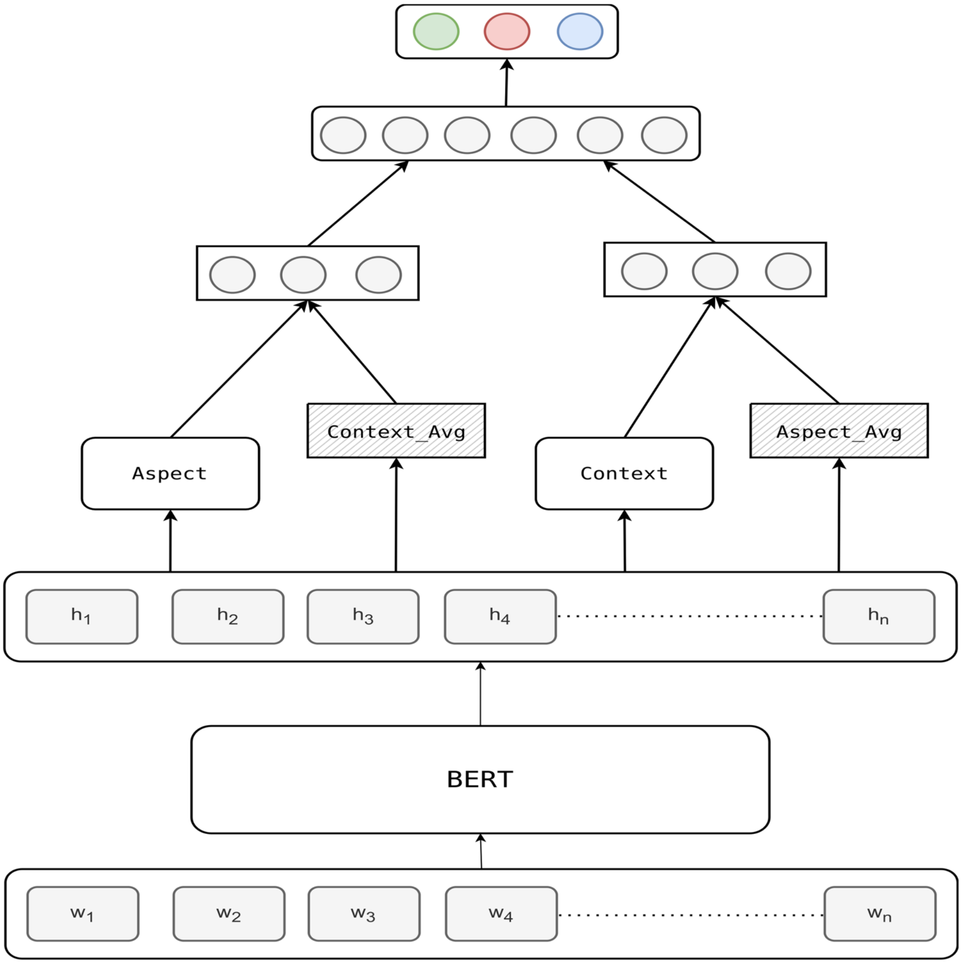

Fig. 1 Proposed architecture of IAN-BERT model

in the input sentence. Then, we can extract two representations as follows:

$$C = [h_1, \dots, h_{k-1}, h_{k+m}, \dots, h_n] \tag{6}$$

$$A = [h_k, h_{k+1}, \dots, h_{k+m-1}] \tag{7}$$

Two additional representations are acquired by taking the average of the aspect and context representations:

$$C_{\text{avg}} = \frac{\sum_{i=1}^{k-1} h_i + \sum_{i=k+m}^{n} h_i}{(n-m)} \tag{8}$$

$$A_{\text{avg}} = \frac{\sum_{j=k}^{k+m-1} h_j}{m} \tag{9}$$

We leverage the information from Eq. 6 to Eq. 9 and utilize an attention mechanism to identify the key information in determining the sentiment polarity of the aspect in the sentence. Our approach considers both the impact of the aspect on the context and the impact of the context on the aspect, providing a deeper understanding of relevant sentiment features. We use a pair of contexts and aspects and apply the attention mechanism, as shown in Fig. 1. The attention mechanism generates an attention vector α_i through the use of the aspect representation A_{avg} and the context word representations C_i, as expressed as follows:

$$\alpha_i = \frac{\exp(\beta(C_i, A_{\text{avg}}))}{\sum_{j=1}^{n-m} \exp(\beta(C_j, A_{\text{avg}}))} \tag{10}$$

where β is score function that signifies how much context word C_i attends to the aspect A_{avg}. The score function β is calculated as

$$\beta(C_i, A_{\text{avg}}) = \tanh(C_i.W_a.A_{\text{avg}}^T + b_a) \tag{11}$$

where W_a and b_a are trainable parameters.

Similarly, the attention vector γ_i is calculated by taking into account both the average context representation C_{avg} and the aspect representation A_i as follows:

$$\gamma_i = \frac{\exp(\beta(A_i, C_{\text{avg}}))}{\sum_{j=1}^{m} \exp(\beta(A_j, C_{\text{avg}}))} \tag{12}$$

The context and aspect representations are derived from Eqs. 13 and 14, respectively, as

$$C_r = \sum_{i=1}^{n-m} \alpha_i C_i \tag{13}$$

$$A_r = \sum_{i=1}^{m} \gamma_i A_i \tag{14}$$

Finally, the aspect representation A_r and context representation C_r are combined into a single vector S, and the final representation is obtained as

$$S = \text{Concat}(A_r, C_r) \tag{15}$$

Sentiment Classification

The vector S is passed through a fully connected layer with three output neurons, each corresponding to a sentiment class: positive, negative, and neutral. Softmax is applied to find the output probability distribution over the sentiment classes for the given aspect:

$$O = \text{Softmax}(W * S + b) \tag{16}$$

where parameters W and b are trainable.

The loss function utilized in this work is cross-entropy, represented by Eq. 17:

$$L = -\sum_{i=1}^{c} Y_i \log(O_i) \tag{17}$$

where Y_i is a true polarity vector, and c signifies polarity classes.

Experiments and Results

This section explores the dataset used for experiments for Targeted sentiment analysis, parameter settings, and model training, followed by performance evaluation.

Dataset

We conducted experiments using three publicly available datasets: the SemEval-14 Restaurant and Laptop dataset [30] and the Multi-Aspect Multi-Sentiment (MAMS) dataset [31]. The SemEval-14 dataset includes labeled customer reviews for the Laptops and Restaurants categories. The MAMS dataset contains review sentences that have at least two aspects. The sentiment labels are categorized into three predefined classes in all three datasets: positive, negative, and neutral. The statistics of the dataset are given in Table 2.

Model Hyperparameters

Experiments were performed on three review datasets. The "bert-base-uncased" and "bert-large-uncased" pre-trained models and the "bert-base-uncased" post-trained models were evaluated. The number of encoder layers in the pre-trained "bert-base-uncased" and post-trained "bert-base-uncased" models is 12, while in the "bert-large-uncased" model, it is 24. The learning rate is 2e-5. The batch size is set to 32 for both Laptop and Restaurant datasets. We train the model for 25 epochs. As per the settings outlined in Table 3, we train five models with different random seeds and present the average of the results.

Table 3 Hyperparameters

Hyperparameters	Value
Batch size	32
Learning rate	2e−5
Optimizer	ADAM
Epoch	25
Dropout	0.2

Table 2 Statistics of dataset

	Restaurant 14		Laptop 14		MAMS	
	Train	Test	Train	Test	Train	Test
No. of sentences	1980	599	1454	409	4297	500
No. of target aspects	3608	1119	2282	632	11,186	1336
No. of positive aspects	2164	727	976	337	3380	400
No. of negative aspects	807	196	851	128	2764	329
No. of neutral aspects	637	196	455	167	5042	607

Evaluation Metrics

For performance evaluation, accuracy and f1-score metrics were used for targeted sentiment analysis tasks. Accuracy is defined to measure the correct predictions among total predictions as given in Eq. 18:

$$\text{Accuracy} = \frac{\text{number of correct predictions}}{\text{total predictions}} \tag{18}$$

F1-score is defined as the harmonic mean of precision and recall as given in Eq. 19:

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{19}$$

Performance Evaluation

Variation of Our Model

We conducted experiments using various BERT variations, namely bert-base-uncased, bert-large-uncased, and post-trained BERT. In Table 4, we present six different models, where the first three (bert_base, bert_large, and bert_base_PT) use bert-base-uncased, bert-large-uncased, and post-trained BERT, respectively, to generate contextualized representations. To obtain the aspect vector, we applied an aspect mask over the contextualized representation and then used the mean of the aspect vector to determine the sentiment orientation.

On the other hand, the last three models (IAN-BERT_BB, IAN-BERT_BL, and IAN-BERT_BB_PT) apply interactive attention over the contextualized representation obtained from the different BERT variants (bert-base-uncased, bert-large-uncased, and post-trained BERT, respectively). Interactive attention is used to get the aspect and context vectors concatenated and passed through a softmax layer for polarity detection.

Our experiment findings demonstrate that bert-base achieved the lowest accuracy and f1-score for the Restaurant and MAMS dataset, whereas bert-large showed the lowest accuracy for the Laptop dataset. However, a combination of post-trained BERT and interactive attention performed best on all three datasets, as depicted in Table 4.

Base Models

In this section, we have discussed the base models used to compare our method.

1. *TD-LSTM* [11] employs two LSTMs that simultaneously analyze the context on either side of the targeted aspect. The output of these LSTMs is combined to form the final representation, which is then used for sentiment classification.
2. *ATAE-LSTM* [14] uses an attention-based LSTM for sentiment analysis of aspect terms and aspect categories. The target embedding is combined with the hidden states of the LSTM to calculate the attention weights.
3. *MemNet* [32] treats the ABSA as a Question Answering task and combines the context word vectors by summing up linearly transformed aspect vectors with the attention output.
4. *IAN* [15] resorts to an attention mechanism that interactively learns the score between the target word and its context. This attention score is further used to create a context-aware target and target-aware context representation.
5. *RAM* [33] utilizes a 2-layer BiLSTM to generate memory, incorporating position information and producing a custom-made set of memory. It employs a recurrent attention network to extract sentiment features related to the designated target.
6. *MGAN* [34] employs fine-grained and coarse-grained attention for aspect term and aspect category sentiment classification tasks. Fine-grained attention is utilized for word-level interaction between aspects and their context.
7. *CDT* [35] employs BiLSTM over the input and further learns the aspect representation using Graph Convolution Network(GCN) over the dependency tree.
8. *AOA-MultiACIA* [26] employs interactive attention between aspect and context for classifying aspects

Table 4 Different variations of the proposed model for various datasets

	Restaurant 14		Laptop 14		MAMS	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
bert_base	84.08%	77.16%	79.94%	75.97%	80.13%	79.58%
bert_large	85.06%	78.57%	78.69%	75.44%	82.42%	81.71%
bert_base_PT	86.67%	79.86%	81.46%	78.21%	81.84%	81.47%
IAN-BERT_BB	84.68%	78.26%	80.32%	76.57%	82.19%	81.65%
IAN-BERT_BL	85.06%	79.48%	80.47%	77.52%	82.83%	82.13%
IAN-BERT_BB_PT	86.95%	81.52%	81.92%	78.65%	84.38%	83.24%

present in the input. Multiple groups of key and value pairs of aspect and context are utilized to generate aspect and context representation, respectively. The result is obtained through a combination of these two representations.

9. *AEN-BERT* [25] uses an attention mechanism to model the interaction between aspects and context for targeted sentiment analysis and incorporates label smoothing regularization to address the issue of fuzzy labels.
10. *BERT-SPC* [25] uses input sentence and aspect as sentence pairs and feeds them to pre-trained BERT. Aspect classification into predefined classes is done using pooled embedding.

Result Analysis

Table 5 demonstrates that our IAN-BERT_BB_PT model outperforms other models on Laptop, Restaurant and MAMS datasets. These datasets contain domain-specific terms that require specialized word embeddings, which can be more effectively captured by post-trained BERT models trained on relevant sources such as Amazon and Yelp datasets. Furthermore, our model leverages interactive attention between aspect and context vectors, enhancing the quality of representations and augmenting aspect classification accuracy. Hence, our IAN-BERT_BB_PT model provides superior performance on these challenging datasets thanks to its ability to leverage domain-specific knowledge and attention mechanisms.

BiLSTM models outperform LSTM models by effectively capturing complete contextual information for every word, resulting in superior performance. Consequently, models utilizing LSTM, such as TD-LSTM and ATAE-LSTM, exhibit inferior performance compared to those using BiLSTM, such as RAM and IAN. However, BERT-based models outperform both BiLSTM and attention models due to their bidirectional capability. Therefore, models that incorporate BERT, such as AEN-BERT, IAN-BERT_BB, IAN-BERT_BL, and IAN-BERT_BB_PT, demonstrate superior performance compared to other models.

Standard BERT vs Post-trained BERT

The idea behind applying post-trained BERT is to capture the domain-specific knowledge during word(token) representation, which needs to be improved in standard BERT trained on the general dataset. Table 6 illustrates instances where standard BERT (IAN-BERT_BB) falls short in comparison to post-trained BERT(IAN-BERT_BB_PT). The table exhibits sentences and their predicted sentiment polarities by both BERT and post-trained BERT models. It can be observed from the table that in comparison to BERT, a model with post-trained BERT efficiently predicts the correct sentiment of sentences having multiple aspect terms also.

Attention vs Interactive Attention

As a transformer model, BERT uses self-attention to learn the contextualized representation. With self-attention, every word or token is given equal consideration, and an effort is made to understand the influence of a word or token on the other words or tokens in the sentence. In the ABSA task, we have both a sentence and an aspect, and the goal is to determine the sentiment toward the aspect within the sentence. To exploit this additional aspect information, we have added interactive attention on top of the post-trained contextualized word representation.

Table 5 Accuracy of various techniques on the SemEval-14 dataset

	Restaurant 14		Laptop 14		MAMS	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
TD-LSTM	75.6%	–	68.1%	–	–	–
ATAE-LSTM	77.2%	–	68.7%	–	–	–
MemNet	78.2%	65.8%	70.3%	64.1%	–	–
IAN	78.6%	–	72.1%	–	76.6%	–
RAM	80.3%	70.8%	74.5%	71.4%	–	–
MGAN	81.5%	71.5%	76.2%	71.4%	–	–
CDT	82.3%	74 %	77.2%	72.9%	80.7%	79.8%
AOA-MultiACIA	82.6%	72.1%	75.3%	70.2%	–	–
AEN-BERT	83.1%	73.8%	79.9%	76.3%	–	–
BERT-SPC	84.5%	76.9%	78.9%	75%	82.8%	81.9%
IAN-BERT_BB	84.7%	78.3%	80.3%	76.6%	82.2%	81.7%
IAN-BERT_BL	85.1%	79.5%	80.5%	77.5%	82.8%	82.1%
IAN-BERT_BB_PT	86.9%	81.5%	81.9%	80.3%	84.4%	83.2%

Table 6 Result comparison of BERT(IAN-BERT_BB) vs post-trained BERT(IAN-BERT_BB_PT)

Sentence	Target aspect term	Actual sentiment	Predicted sentiment	
			IAN-BERT_BB	IAN-BERT_BB_PT
1. However, I chose 2 days shipping, and it took over a week to arrive	Shipping	Negative	Neutral	Negative
2. I use it mostly for content creation (audio, video, photo editing) and it is reliable	Content creation	Positive	Neutral	Positive
3. Air has a higher resolution, but the fonts are small	Resolution	Positive	Negative	Positive
4. It is just good food, nothing more, and that is all we want!	Food	Positive	Negative	Positive
5. This place has the strangest menu, and the restaurants try too hard to make fancy food	Food	Positive	Negative	Positive
6. It took about 2 1/2 h to be served our 2 courses	Courses	Neutral	Negative	Neutral

Discussion and Conclusion

This work integrates interactive attention with cutting-edge context-sensitive representation to perform aspect-based sentiment analysis. Post-trained BERT trained on Amazon, and Yelp datasets are used instead of generalized BERT trained on Wikipedia and BookCorpus datasets. Further, an interactive Attention mechanism between aspect and context is applied on top of Domain Knowledge enhanced context-aware BERT representation. Experiments were performed on SemEval-14 Laptop, Restaurant, and MAMS datasets. The results demonstrate superior performance compared to other existing works. Applying Graph attention over BERT would be interesting future work.

Funding The authors have no funding to disclose.

Data availability The dataset used in this article is publicly available.

Declarations

Conflict of Interest Not applicable.

Ethics Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Financial Interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Jiang L, Yu M, Zhou M, Liu X, Zhao T. Target-dependent twitter sentiment classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies; 2011. p. 151–160.
2. Liu H, Chatterjee I, Zhou M, Lu XS, Abusorrah A. Aspect-based sentiment analysis: a survey of deep learning methods. IEEE Trans Comput Soc Syst. 2020;7(6):1358–75.
3. Wang C-Y, Bochkovskiy A, Liao H-YM. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv:2207.02696. 2022.
4. Gulati A, Qin J, Chiu C-C, Parmar N, Zhang Y, Yu J, Han W, Wang S, Zhang Z, Wu Y, et al. Conformer: Convolution-augmented transformer for speech recognition. arXiv:2005.08100. 2020.
5. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, He H, Li A, He M, Liu Z, et al. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. arXiv:2304.01852. 2023.
6. Adeniji OD, Adeyemi SO, Ajagbe SA. An improved bagging ensemble in predicting mental disorder using hybridized random forest-artificial neural network model. Informatica. 2022;46(4):543–50.
7. Ajagbe SA, Amuda KA, Oladipupo MA, Oluwaseyi FA, Okesola KI. Multi-classification of Alzheimer disease on magnetic resonance images (MRI) using deep convolutional neural network (dcnn) approaches. Int J Adv Comput Res. 2021;11(53):51.
8. Elman JL. Finding structure in time. Cogn Sci. 1990;14(2):179–211.
9. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
10. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw. 1994;5(2):157–66.
11. Tang D, Qin B, Feng X, Liu T. Effective LSTMs for target-dependent sentiment classification. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan; 2016. p. 3298–3307. https://www.aclweb.org/anthology/C16-1311.
12. Kumar A, Verma S, Sharan A. ATE-SPD: simultaneous extraction of aspect-term and aspect sentiment polarity using bi-LSTM-CRF neural network. J Exp Theor Artif Intell. 2021;33(3):487–508.
13. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473. 2014.
14. Wang Y, Huang M, Zhu X, Zhao L. Attention-based lSTM for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016. p. 606–615.
15. Ma D, Li S, Zhang X, Wang H. Interactive attention networks for aspect-level sentiment classification. arXiv:1709.00893. 2017.

16. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014. p. 1532–1543.
17. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. 2018.
18. Li X, Bing L, Zhang W, Lam W. Exploiting bert for end-to-end aspect-based sentiment analysis. arXiv:1910.00883. 2019.
19. Li X, Bing L, Li P, Lam W. A unified model for opinion target extraction and target sentiment prediction. Proc the AAAI Conf Artif Intell. 2019;33:6714–21.
20. Hu M, Peng Y, Huang Z, Li D, Lv Y. Open-domain targeted sentiment analysis via span-based extraction and classification. arXiv:1906.03820. 2019.
21. Xu H, Liu B, Shu L, Yu PS. Bert post-training for review reading comprehension and aspect-based sentiment analysis. arXiv:1904.02232. 2019.
22. Xu, H., Liu, B., Shu, L., Yu, P.S.: Dombert: Domain-oriented language model for aspect-based sentiment analysis. arXiv:2004.13816. 2020.
23. Xu H, Shu L, Yu PS, Liu B. Understanding pre-trained bert for aspect-based sentiment analysis. arXiv:2011.00169. 2020.
24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30:5998–6008.
25. Song Y, Wang J, Jiang T, Liu Z, Rao Y. Attentional encoder network for targeted sentiment classification. arXiv:1902.09314. 2019.
26. Wu Z, Li Y, Liao J, Li D, Li X, Wang S. Aspect-context interactive attention representation for aspect-level sentiment classification. IEEE Access. 2020;8:29238–48.
27. Ambartsoumian A, Popowich F. Self-attention: A better building block for sentiment analysis neural network classifiers. arXiv:1812.07860; 2018.
28. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds). Advances in neural information processing systems, vol. 26. Curran Associates, Inc; 2013. p. 3111–3119. https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf
29. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (long papers). Association for Computational Linguistics, New Orleans, Louisiana; 2018. p. 2227–2237. https://doi.org/10.18653/v1/N18-1202. https://www.aclweb.org/anthology/N18-1202.
30. Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S. Semeval-2014 task 4: Aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). ACL; 2014, p. 27–35.
31. Jiang Q, Chen L, Xu R, Ao X, Yang M. A challenge dataset and effective models for aspect-based sentiment analysis. 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); 2019. p. 6280–6285.
32. Tang D, Qin B, Liu T. Aspect level sentiment classification with deep memory network. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas; 2016. p. 214–224. https://doi.org/10.18653/v1/D16-1021. https://aclanthology.org/D16-1021.
33. Chen P, Sun Z, Bing L, Yang W. Recurrent attention network on memory for aspect sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark; 2017. p. 452–461. https://doi.org/10.18653/v1/D17-1047. https://www.aclweb.org/anthology/D17-1047.
34. Li Z, Wei Y, Zhang Y, Zhang X, Li X. Exploiting coarse-to-fine task transfer for aspect-level sentiment classification. Proc AAAI Conf Artif Intell. 2019;33:4253–60.
35. Sun K, Zhang R, Mensah S, Mao Y, Liu X. Aspect-level sentiment analysis via convolution over dependency tree. 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); 2019. p. 5679–5688.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.