IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 16, 2021, accepted September 23, 2021, date of publication September 28, 2021,
date of current version October 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3116053

A Multi-Layer Network for Aspect-Based Cross-Lingual Sentiment Classification

KALIM SATTAR[1], QASIM UMER[2], DINARA G. VASBIEVA[3], SUNGWOOK CHUNG[4], ZOHAIB LATIF[5], AND CHOONHWA LEE[5], (Member, IEEE)

[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518000, China
[2]Department of Computer Science, COMSATS University Islamabad, Vehari 61100, Pakistan
[3]Department of English Language for Professional Communication, Financial University under the Government of the Russian Federation, 125167 Moscow, Russia
[4]Department of Computer Engineering, Changwon National University, Changwon 51140, South Korea
[5]Department of Computer Science, Hanyang University, Seoul 04763, South Korea

Corresponding author: Choonhwa Lee (lee@hanyang.ac.kr)

This work was supported by Korea Meteorological Administration Research and Development Program under Grant KMI 2021-01310.

ABSTRACT In the recent era, the advancement of communication technologies provides a valuable interaction source between people of different regions. Nowadays, many organizations adopt the latest approaches, i.e., sentiment analysis and aspect-oriented sentiment classification, to evaluate user reviews to improve the quality of their products. The processing of multi-lingual user reviews is a key challenge in Natural Language Processing (NLP). This paper proposes a multi-layer network with divided attention to perform aspect-based sentiment classification for cross-lingual data. It extracts the Part-of-Speech (POS) tagging information of the given reviews, preprocesses them, and converts them into tokens. Furthermore, bi-lingual dictionaries are leveraged to map the converted tokens from one language to another. Given the preprocessed and mapped reviews, vectors are generated by leveraging the multi-lingual BERT and passed to the proposed deep learning classifier. The 10351 *restaurant* reviews from *SemEval-2016 Task 5* dataset are exploited for the prediction of aspect-based sentiment. The results of cross-lingual validation suggest that the proposed approach significantly outperforms the state-of-the-art approaches and improves the precision, recall, and F1 by more than 23%, 20%, and 22%, respectively.

INDEX TERMS Natural language processing, cross-lingual, divided attention, aspect-based sentiment classification.

I. INTRODUCTION

In a globalized world, the rapid growth of web technologies, i.e., social media and digital marketing, generates vast amounts of data and sets new trends. For the evaluation and analysis of such data, researchers and organizations exploit data mining techniques to extract meaningful patterns from the collected data [1]. Among them, NLP is the field that deals with the processing of human-generated text. Sentiment analysis is one of the core tasks of NLP that aims to predict opinion polarity. It often divides the predicted sentiment into three categories (positive, negative, and neutral) that apply to nearly every domain, e.g., customer product reviews, political predictions, healthcare, and financial services. Sentiment analysis of text can be performed on document-level, sentence-level, and aspect-level [2]–[4].

Businesses focus on aspect-level product analysis to understand the impact and limitation(s) of products. Such analysis helps in making plans to meet the requirements of consumers. Based on target words of aspects, the aspect-level study predicts sentiments of consumers from the product reviews [5]. The target words of aspects may have explicit/implicit inclusion in the text of product reviews. If the target words of an aspect physically exist in the text of the product review, such aspect is called explicit aspect, otherwise implicit aspect. For example, a review "The food quality of this restaurant is excellent but very costly" contains a clear aspect "food quality" and has positive sentiment because of an opinion word "excellent". At the same time, the review contains an implicit aspect "food price" and has negative sentiment

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

because of an opinion word ''costly'' [6]. Notably, we focus on explicit aspects in this paper.

The recent advancements in technology provide communication channels among people of different regions and countries having different social values, cultures, and languages. Therefore, international companies receive reviews of their products in multiple languages. Since each language has a different grammatical structure, word sense, and background history, aspect-based sentiment analysis of such reviews is challenging for companies [7].

Heretofore, most of the efforts for aspect-based sentiment analysis are related to monolingual [7]. The use of multi-lingual word embedding for the data mapping from one language to another handles cross-lingual data efficiently. However, the performance of multi-lingual word embedding gradually decreases from rich-resource language to poor-resource language because of data unavailability [8]. To this end, different neural network-based models are presented to perform multiple NLP tasks [3], [6], [7], [9]–[11], OpenAI [12], ULM-FiT [13], ELMo [14], and BERT [15]. Notably, BERT provides multi-lingual word embedding for more than 100 languages [16] and outperforms the above models for the prediction of sentiments even for poor-resource languages. Similarly, some researches [17]–[19] exploit attention mechanism among different neural networks to extract targeted context. Although many studies are conducted for the aspect-based sentiment analysis, the grammatical properties (Parts-of-Speech (POS)), i.e., nouns, adjectives, and verbs, are ignored that have strong connectivity with aspects and their sentiments. For example, a sentence with its POS tags ''The-DET food-NOUN seemed-VERB pretty-ADV fresh-ADJ and-CCONJ the-DET service-NOUN impeccable-ADJ'' contains nouns, verbs, and adjectives indicating that the food is pretty fresh and service is impeccable. Moreover, homographic words are not targeted in multi-lingual aspect-based sentiment classification due to the small vocabulary size of cross-lingual data.

In this perspective, we propose a deep learning approach for cross-lingual aspect-based sentiment classification by exploiting the multi-lingual BERT and bi-lingual dictionaries. The proposed approach first extracts the Part-of-Speech (POS) tagging information of the given reviews. Second, it preprocesses the reviews and converts them into tokens. Third, it leverages bi-lingual dictionaries to map the tokens from one language to another. Fourth, given the preprocessed and mapped reviews, it generates vectors by exploiting the multi-lingual BERT. Fifth, the generated vectors are pass to the proposed deep learning classifier for training. Finally, the proposed approach is evaluated with a multi-lingual dataset for the aspect-based sentiment classification. The results of cross-lingual validation suggest that the proposed approach is significant and improves the precision, recall, and F1 by more than 23%, 20%, and 22%, respectively.

This paper makes the following contributions:

- A data modeling technique is introduced to map cross-lingual data from one language to another by exploiting bi-lingual dictionaries.
- A deep learning-based neural network is proposed for aspect-based sentiment classification to utilize the proposed data modeling technique for cross-lingual data effectively. It uses a divided attention mechanism for paying attention to POS tagging information.
- The proposed approach is compared with state-of-the-art approaches for the performance evaluation. The evaluation results of cross-lingual validation suggest that the proposed approach is accurate and outperforms the state-of-the-art approaches.

The rest of the paper is divided as follows. Section II presents the related work. Section III explains the proposed approach. Section IV describes the evaluation process of the proposed approach, compares the performance results of the proposed approach with the baseline approaches, and explains the threats. Section V concludes the paper.

II. RELATED WORK

Sentiment analysis has a significant impact on digital marketing and customer reviews. Therefore, most of its applications target such domains [20]. Cambria *et al.* [21] argued that most of the researchers consider sentiment analysis as a simple task, but in reality, it is a complex problem. The authors further explained that the primary aim of NLP is to achieve human-like performance in NLP tasks and identified fifteen issues that need to be solved to achieve human-like performance in the field of sentiment analysis. Such problems are divided into three main layers: a syntactic layer that deals with the pre-processing of the text, i.e., micro text normalization, sentence boundary disambiguation, POS tagging, text chunking, and lemmatization; semantics layer that deals with the deconstruction and normalization of the text, i.e., word sense disambiguation, concept extraction, named entity recognition, anaphora resolution, and subjectivity detection; and pragmatics layer that solves problems (i.e., personality recognition, sarcasm detection, metaphor understanding, aspect extraction, and polarity detection) by using syntactic and semantics layers.

Tang *et al.* [22] explained some of the essential applications of sentiment detection that include product comparison, opinion summarization, and opinion reason mining. They also mentioned other valuable tasks that can be performed using sentiment analysis, i.e., political discussion group posting, message sentiment filtering, email sentiment classification, attitude analysis, and sentiment with search engines.

Medhat *et al.* [23] described the taxonomy of sentiment analysis techniques and divided it into two main categories: machine learning approaches and lexicon-based approaches. Liu [24] further described feature selection as a critical factor for machine learning-based techniques and identified the most common feature selection methods: terms and their

frequency, POS tagging, identification of sentiment words and phrases, sentiment shifters, and syntactic dependencies.

Traditionally, a one-hot vector represents textual information that suffers from high dimensionality and poor co-relation problems. To avoid such issues, Bengio *et al.* [25] replaced the one-hot vector with a low-dimensional distributed representation that becomes a standard technique. Notably, several pre-trained word embedding techniques are used to capture syntactic and semantic information from text, e.g., Word2Vec [26] and Glove [27].

A. MONO-LINGUAL ASPECT-BASED SENTIMENT CLASSIFICATION

The goal of aspect-based sentiment classification is to predict the sentiment polarity to a particular aspect or target word. Schouten and Frasincar [5] categorized aspect-based sentiment analysis into three significant categories: aspect detection, sentiment analysis, and joint-aspect detection and sentiment analysis. Earlier methods for aspect-based sentiment classification are based on rule-based methods, i.e., Hu and Liu [20] used frequent nouns as product features using association rule mining. They exploited WordNet to get the synonym words and extract the sentiment of extracted nouns. Nasukawa and Yi [28] identified the importance of the semantic relationship between sentiment expression and targeted subject. They used syntactic parser and sentiment lexicon to achieve better performance.

In supervised-based machine learning techniques, Jiang *et al.* [9] introduced the target-dependent features to perform better target sentiment classification on Twitter data using a support vector machine. other researchers [10], [11], [17], [18] also adopted neural network-based models for targeted sentiment classification. Among them, Tang *et al.* [29] proposed the target-dependent long short-term memory network to perform aspect-based sentiment analysis tasks. They applied forward and backward LSTM to capture bi-directional information for target words, whereas Huang and Carley [30] constructed parameterized filters and gates to incorporate aspect information in CNN. However, Wang *et al.* [17], Tang *et al.* [31], and Chen *et al.* [11] exploited attention mechanism for aspect-based sentiment classification, attention mechanism with explicit memory to perform aspect-based sentiment classification, and multi-layer attention with recurrent neural network to combine the output of multiple layers.

In the neural network, LSTM with target-level attention is most frequently adopted. However, LSTM has certain limitations, i.e., it is difficult to remember long-term patterns for the complex input sequence. Therefore, few researchers [19], [32] used pre-trained transformer-based models for aspect-based sentiment classification to their success especially BERT [15]. Moreover, Song *et al.* [19] proposed an attention encoder network and performed target word-based attention on context without using recurrent structure. Zeng *et al.* [32] proposed a local context mechanism by dividing context into the local context and global context. Sun *et al.* [33] constructed an auxiliary sentence to convert aspect-based sentiment classification tasks as sentence pair classification tasks.

Although the approaches mentioned above are proposed for mono-lingual aspect-based sentiment classification, most models are built explicitly for mono-lingual data (i.e., English). Therefore, these models cannot process and capture the context information from cross-lingual data in an effective manner.

B. CROSS-LINGUAL ASPECT-BASED SENTIMENT CLASSIFICATION

Tubishat *et al.* reported that most of the work in NLP is limited to mono-lingual data related to English, Chinese, and French [6]. Due to digital and social media advancement, many other languages need to be explored to use data effectively. The significant challenges of dealing with cross-lingual data are: each language has different word resources, and the grammatical structure of one language is different from other languages. To overcome such challenges, the following approaches are proposed for cross-lingual data.

Based on Dashtipour *et al.* [34] research work that summarizes the mono-lingual approaches, Deriu *et al.* [35] leveraged large amounts of weekly supervised data with multi-layer CNN to perform sentiment classification for multiple languages. Moreover, Balahur and Turchi [36] used three different machine translation methods: Bing, Google, and Moses to perform sentiment classification for French, German and Spanish. Similarly, Lambert [37] proposed a machine learning-based model for aspect level sentiment analysis for English and Spanish language. The proposed model is assumed to have parallel annotated data for both source and target languages. If similar data is unavailable, the author used a translation medium to convert source language data into the target language or target language data into the source language. However, this method is highly dependent on parallel data and translation medium.

Zhou *et al.* [7] Proposed a supervised machine learning-based model (CLOpinionMiner) for Opinion target extraction from cross-lingual scenarios for English and Chinese. The proposed approach used Bing translator online services for translating English annotated data into Chinese. Although this technique was effective, it has certain disadvantages, i.e., this technique depends on the translation medium.

Conneau *et al.* [38] proposed an unsupervised method for cross-lingual word embedding. They used an adversarial learning method to map source and target language embeddings into the same vector space. This multi-lingual word method supports up to 30 languages. However, the accuracy of the mapping source and target language into the same vector space decreased gradually from resource-rich to resource-poor language due to the lack of availability of data for resource-poor languages.

Ghadery *et al.* [8] proposed a multi-lingual n-gram based CNN for aspect category detection in online reviews. The author used multi-lingual word embedding to deal with

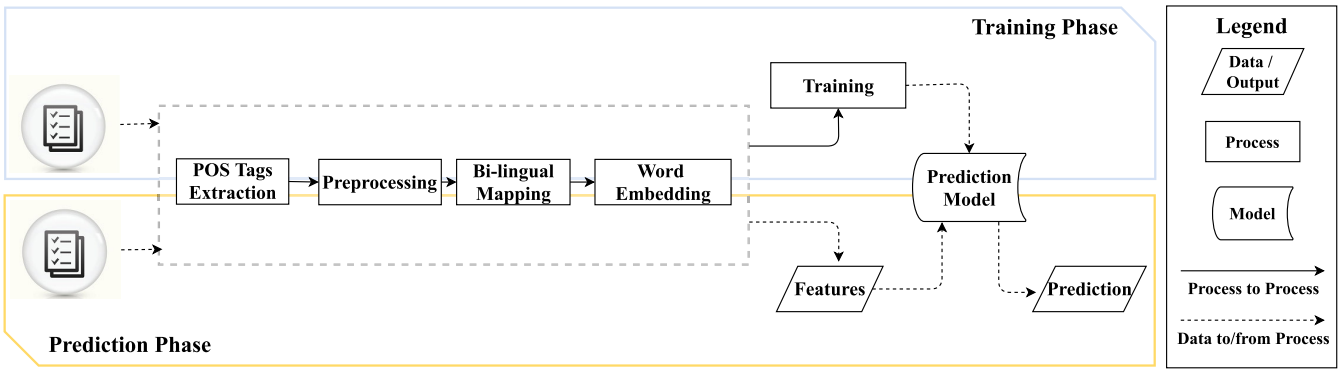

FIGURE 1. Overview of the proposed approach.

multi-lingual data. The author divided aspect category detection into three subtasks: entity detection, attribute detection, and aspect category detection. The significant advantage of this method is that it does not depend on translation techniques. However, this method is only used to get pre-defined aspect categories and is less effective for low-resource languages.

In conclusion, researchers have proposed many approaches for aspect-based sentiment classification. However, only three studies [15], [19], [32] focus on the mono-lingual aspect-based sentiment classification. Our proposed approach differs from the existing approaches in that we apply an attention-based deep learning algorithm for the multi-lingual aspect-based sentiment classification.

III. APPROACH

A. OVERVIEW

Fig. 1 depicts an overview of the proposed approach. The multi-lingual aspect-based identification of sentiments of reviews from 4 different languages is essentially a binary class classification. All submitted multi-language reviews are automatically classified into two classes (i.e., positive or negative) against each aspect based on the identified factors. The proposed approach predicts aspect-based sentiments of reviews as follows:

A brief introduction is presented as follows:

- First, we extract the POS tags from the text of each review.
- Second, we preprocess each review to remove punctuation and stop-words and to split it into tokens.
- Third, we extract the aspects from each review and map them into other languages using ground-truth bilingual dictionaries [38].
- Fourth, we convert each aspect into vectors using multi-lingual BERT.
- Fifth, we pass the POS information and the generated vectors of each aspect and their classification information to an attention-based convolutional neural network as input for training.
- Finally, we extract the POS tagging information and aspect vectors of new reviews and input them to the trained binary-class classifier to predict their labels (i.e., positive or negative) against each aspect.

Each of the essential steps of the proposed approach is presented in the following sections.

B. ANNOTATING EXAMPLES

The following examples are considered to annotate how the proposed approach predicts the aspect-based sentiment of reviews. The example reviews of English and French languages are taken from the *restaurant* domain of *SemEval-2016 Task 5* dataset, respectively. Notably, the dataset is public and contains the annotated reviews.

Example Review 1:
Text = "The food was lousy - too sweet or too salty and the portions tiny".
Language = "English".
Aspect Terms with Polarity = "food (negative) and portions (negative)".
Example Review 2:
Text = "Endroit sympa en plein centre touristique".
Language = "French".
Aspect Terms with Polarity = "Endroit (positive)".

In the above examples, The features ("Text", "Language", and "Aspect Terms and Polarity") represent the text of the reviews, the languages in which the reviews are written, and the associated term/aspects and their polarities/sentiments with the reviews, respectively. The details on how the proposed approach performs for the annotating example are given in the following Section.

C. PROBLEM DEFINITION

A review r from a set of reviews R can be formalized as

$$r = < tr, lr, \sum_{i \epsilon I_r} \{atr_i, pr_i\} > \quad (1)$$

where, tr is the textual information of r, lr is the language of r, atr_i are the terms/aspects of r, and pr_i are the polarities/sentiments of atr_i of r.

TABLE 1. An example of preprocessing.

Original Review	Endroit sympa en plein centre touristique.
After Punctuation Removal	Endroit sympa en plein centre touristique
After Stop-words Removal	Endroit sympa plein centre touristique
After Tokenization	'Endroit', 'sympa', 'en', 'plein', 'centre', 'touristique'

For the example review 2 from the annotating examples presented in Section III-B, we have

$$r_e = < tr_e, lr_e, atr_{e_i}, pr_{e_i} > \tag{2}$$

where, tr_e, lr_e, atr_{e_i}, and pr_{e_i} are "Endroit sympa en plein centre touristique", "French", "Endroit", and "positive", respectively.

The proposed approach predicts the aspect-based sentiments of new reviews as either *positive* (noted as p), or *negative* (noted as n). Consequently, the automatic prediction of aspect-based sentiment of r could be defined a mapping function f:

$$f : r \rightarrow c$$
$$c \in \{p, n\}, \quad r \in R \tag{3}$$

where, c is a suggested sentiment from a polarity set (p, n) against each aspect.

D. POS TAGS EXTRACTION

For each r, we extract the POS tags by exploiting *spaCy*[1] (a Python library). We exploit *spaCy* as it is an open-source library and provides a NLP operation for multi-lingual text. After extracting POS tagging information, a review r can be formalized as

$$r' = < tr, lr, \sum_{i \epsilon I_r} \{atr_i, pr_i\}, tr_{pos} > \tag{4}$$

where, tr_{pos} contains the tagging information of r.

For the example review 2 from the annotating examples presented in Section III-B, we have

$$r'_e = < tr_e, lr_e, atr_{e_i}, pr_{e_i}, tr_{e_{pos}} > \tag{5}$$

where, $tr_{e_{pos}}$ = *Endroit/ NOUN, sympa/ ADJECTIVE, en/ PREPOSITION, plein/ ADJECTIVE, centre/ NOUN, touristique/ ADJECTIVE*.

E. PREPROCESSING

The multi-language reviews contain irrelevant text, e.g., punctuation and stop-words. The input of such text into deep learning algorithms is an overhead in terms of memory and processing time. Therefore, we exploit *spaCy* to preprocess each r to avoid to make the proposed approach cost-effective. Our preprocessing steps remove punctuation and stop-words, and split into tokens. After preprocessing, a review r can be formalized as

$$r'' = < tr, lr, \sum_{i \epsilon I_r} \{atr_i, pr_i\}, tr_{pos}, \sum_{i \epsilon I_r} w_i > \tag{6}$$

where, w_i represents the preprocessed tokens of r.

[1]https://spacy.io/api

For the example review 2 from the annotating examples presented in Section III-B, we have

$$r''_e = < tr_e, lr_e, atr_{e_i}, pr_{e_i}, tr_{e_{pos}}, w_{e_i} > \tag{7}$$

where, w_{e_i} are the preprocessed token as presented in Table 1.

F. DATA MAPPING

An effective data modeling is the most critical step for cross-lingual data mapping from one language (source language) l_s to another language (target language) l_t because word embedding vectors for common words of two different languages may not be similar. To handle such dissimilarity, we propose a data mapping technique by exploiting and combining the ground-truth bi-lingual dictionaries [38].

For the data mapping from l_s to l_t, we exploit bi-lingual dictionaries. Notably, if the source/target word is not listed in bi-lingual dictionary, both words will be considered identical. We map w_i of a preprocessed review r'' into 4 languages. The mapping of w_i can be formalized as

$$w_i \rightarrow \acute{w}_i \tag{8}$$

where, all tokens from w_i of l_s is mapped as $\acute{W}_i$ into l_t.

For the example review 2 from the annotating examples presented in Section III-B, Table 2 shows the mapping of tokens between two languages.

G. WORD EMBEDDING

To convert the tokens of l_s, l_t, mapping from l_s to l_t, mapping from l_t to l_s into vectors, and POS tags, we exploit multi-lingual BERT. Notably, we only consider 'noun', 'verb', 'adjective', and 'adverbs' among POS tags as they contains information about aspects and generate vectors for them.

For the example review 2 from the annotating examples presented in Section III-B, Table 3 presents the generated vectors from multi-lingual BERT.

H. DEEP LEARNING CLASSIFIER

Fig. 2 depicts the composition of the deep neural network-based classifier. Convolutional Neural Network (CNN) is exploited for the prediction of the aspect-based sentiment of R. We use CNN for the composition of the proposed model because of the following reasons: 1) the deep semantic relationships may be learned through CNN layers between input preprocessed words for the aspect-based sentiment classification; 2) CNN has the ability of parallel computation on modern powerful GPUs that reduces its training time [39]; and 3) the proposed model may avoid the exploding gradient problem of recurrent neural network [40], [41] by assigning different filter sizes.

TABLE 2. An example of words mapping.

Settings	Language	Preprocessed Token
Source Language	French	'Endroit', 'sympa', 'en', 'plein', 'centre', 'touristique'
Target Language	English	'place', 'nice', 'full', 'center', 'tourist'

TABLE 3. An example of word embedding.

Settings	Language	Text Type	Output
Source Language	French	Preprocess Tokens	'Endroit', 'sympa', 'en', 'plein', 'centre', 'touristique'
Source Language	French	BERT Vectors	[[101, 142, 102], ..., [101, 173, 102]]
Mapping	French	Mapping Vectors	[[101, 147, 102], ..., [101, 186, 102]]
Target Language	English	Preprocessed Tokens	'place', 'nice', 'full', 'center', 'tourist'
Target Language	English	BERT Vectors	[[101, 147, 102], ..., [101, 186, 102]]

In order to the training of the proposed deep learning model, we first concatenate the embeddings of preprocessed token of each source language review SE_{W_i} (Eq. 6) and mappings of SE_{W_i} into target language $ME_{\hat{W}_i}$ (Eq. 8) and pass into a CNN. Second, we input the POS tags (tr_{pos}) of SE_{W_i} (Eq. 6) and preprocessed tokens TE_{W_i} of target language to separate CNNs. Third, the output of each CNN is passed to separate dense layer to equalize them. We use three layers of CNN with settings: $filter = 128$, $kernel\ size = 1$ and $activation = tanh$. Fourth, the equalized outputs of all dense layers are passed to an divided attention layer. The divided attention is the psychological term to simultaneously paying attention to two or more factors. Notably, we exploit POS words, preprocessed target words, and aspect classes to performed multi-head attention. The divided attention layer also merges the given outputs by the merge layer [42]. Fifth, the output of the divide attention is given to the dense layer that fully connects the 128 *neurons* to those in the next layer. Finally, the output layer 2 *neurons*) map both inputs into a single output (prediction) that predicts the sentiment (p or n) of each aspect of r. We set the loss function for the proposed model as *binary_crossentropy* that computes the performance of a classification model. Notably, we evaluate the proposed deep learning model on different settings, i.e., epoch (5, 10, 15, and 20), batch size (16, 32, 64, and 128) and activation function (tanh, sigmoid, and relu), and find the optimal hyperparameters with $epoch = 10$, $batch\ size = 16$, and $activation = tanh$. We also incorporate the Pooling unit between the merge layer and divided attention layer to reduce the dimensions of the features.

IV. EVALUATION

In this section, the performance of the proposed approach is evaluated by comparing the proposed approach with the state-of-the-art approaches.

A. RESEARCH QUESTIONS

The following questions are investigated for the evaluation of the proposed approach.

- *RQ1:* Does the proposed approach outperform the state-of-the-art approaches? If yes, to what extent?
- *RQ2:* How do the bi-lingual dictionaries influence the proposed approach?
- *RQ3:* How does the divided attention influence the proposed approach?
- *RQ4:* How does the POS tagging information influence the proposed approach?

To answer the research question **RQ1**, the proposed approach is compared with three state-of-the-art approaches: *SPC-BERT* [15], *AEN-BERT* [19], and *LCF-BERT* [32] to check performance improvement. We choose these approaches for comparison because these are recent approaches in aspect-based sentiment analysis.

To answer the research question **RQ2**, the performance results of the proposed approach (without divided attention) are compared with *AEN-BERT* which is best among the state-of-the-art approaches and does not use the bi-lingual mapping and divided attention to check the influence of bi-lingual mapping on the proposed approach.

To answer the research question **RQ3**, the performance results of the proposed approach are compared by excluding the divided attention to check its influence on the proposed approach.

To answer the research question **RQ4**, the performance results of the proposed approach are compared by excluding the POS tagging information to check its influence on the proposed approach.

B. DATASET

We exploit *SemEval-2016 Task 5* dataset[2] created by Pontiki *et al.* [43]. It is a multi-lingual dataset for aspect-based sentiment analysis tasks, available in eight different languages: English, French, Dutch, Spanish, Turkish, Chinese, Arabic, and Russian. Moreover, the dataset contains reviews from seven domains. Although the dataset is available in eight languages and having reviews from seven domains, we only include reviews from "restaurant" in

[2]https://alt.qcri.org/semeval2016/task5/index.php?id=data-and-tools

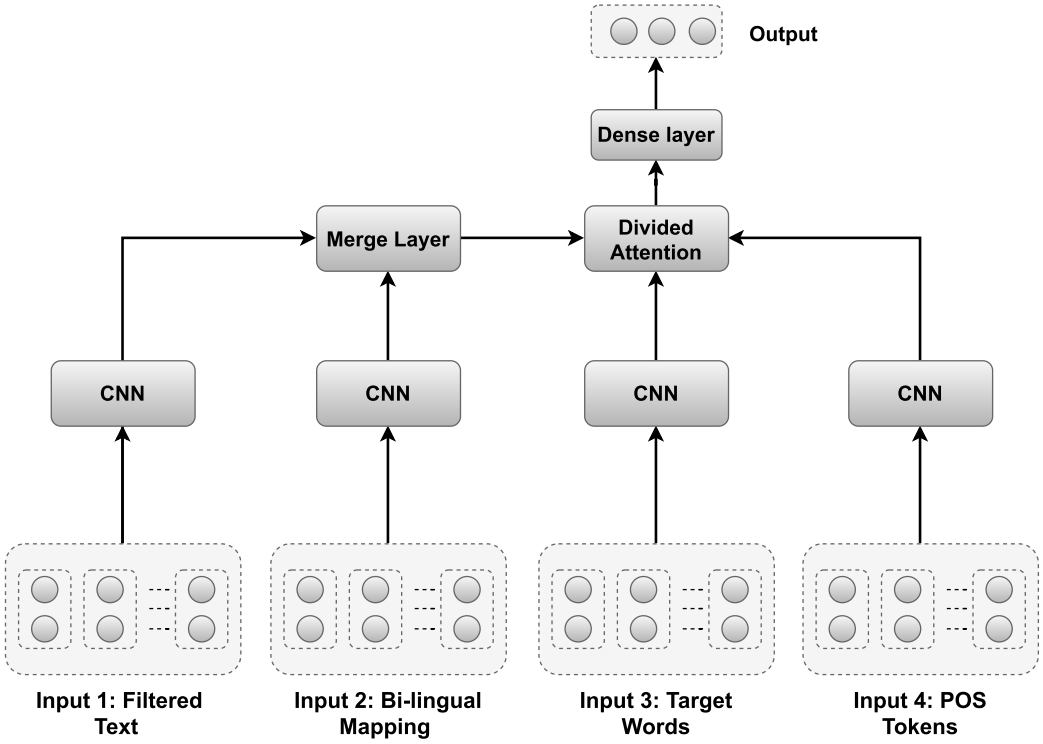

FIGURE 2. The deep learning classifier.

four languages (English, French, Dutch, and Spanish) in our experiments due to the limitations of bilingual dictionaries. The total number of selected reviews is 10351, in which approximately 25.85%, 23.45%, 22.19%, and 28.51% of reviews belong to English, French, Dutch, and Spanish, respectively.

C. PROCESS

We evaluate the proposed approach as follows. First, we exploit the multi-lingual reviews R from an open-source dataset and extract their POS tagging information as discussed in Section III-D. Second, we preprocess each review r from R as discussed in Section III-E. Third, the preprocess reviews (tokens) of l_s are mapped into l_t using bi-lingual dictionaries as discussed in Section III-F. Fourth, given the preprocessed information, we generate the input vectors for the proposed deep learning model using multi-lingual BERT as discussed in Section III-G. Finally, we carry out a cross-language validation on R. We divide R into four sets based on their language notated as $l_i (i = 1 \ldots 4)$. For the i^{th} cross-validation, we consider all reviews except for those in l_i as a training dataset and consider the reviews in l_i as a testing dataset. For the i^{th} cross-validation, the evaluation process as follows:

- First, all the reviews (R_{train}) are selected from training dataset that is a combination of all sets but l_i.

$$R_{train_i} = \sum_{i \in [1,4], j \neq i} l_j \tag{9}$$

- Second, we train SPC-BERT, AEN-BERT, and LCF-BERT with data from R_t against each language.
- Third, we train the proposed approach (CNN-BERT) with data from R_t against each language on different scales, i.e., with and without bi-lingual mapping and multi-lingual BERT, with and without divided attention, and with and without POS tagging information.
- Fourth, for each review R_{test_i} from the testing dataset, we predict its aspect-based sentiment using the trained *SPC-BERT*, *AEN-BERT*, *LCF-BERT*, and *CNN-BERT* to compare its original sentiment.
- Finally, we compute the evaluation metrics for each approach to compare their performances.

D. METRICS

Given the reviews R, we calculate the aspect-based sentiment specific precision Pre, recall Rec and f-measure $F1$ for the evaluation of the proposed approach as these metrics are well-known and have been used in previous studies [44], [45]. The metrics Pre, Rec, and $F1$ can be formalized as

$$Pre = \frac{TP}{TP + FP} \tag{10}$$

$$Rec = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2 * Pre * Rec}{Pre + Rec} \tag{12}$$

where, Pre, Rec and $F1$ present the precision, recall and f-measure of the approaches for aspect-based sentiment prediction of R whose actual aspect-based sentiment is as_i.

TP is the number of *R* that are truly predicted as as_i, *FP* is the number of *R* that are falsely predicted as as_i, and *FN* is the number of *R* that are not predicted as as_i but they are actually as_i.

E. RESULTS

1) RQ1: COMPARISON AGAINST STATE-OF-THE-ART

To answer the research question RQ1, we compare *CNN-BERT* with state-of-the-art approaches (*SPC-BERT*, *AEN-BERT*, and *LCF-BERT*). The results of all approaches are presented in Table 4. The first column represents the languages of the training dataset, the second column represents the languages of the testing dataset for each cross-language validation, and columns 3-5, 6-8, 9-11, and 12-14 represent the performance of *SPC-BERT*, *AEN-BERT*, *LCF-BERT*, and *CNN-BERT*, respectively. The first row represents the evaluation metrics against each approach, rows 2-4, 6-8, 10-12, and 14-16 represent the performance of the approaches against each testing language dataset, and rows 5, 9, 13, and 17 represents the average performance of the approaches against all testing languages dataset. The table presents the best performance for each testing category in bold.

From Table 4, the following observations are made.

- First, *CNN-BERT* significantly improves the state-of-the-art. Compared to *SPC-BERT*, *AEN-BERT*, and *LCF-BERT*, the improvement of *CNN-BERT* in average *Pre*, average *Rec*, and average *F1* is (**23.81%** = (45.92% − 37.09%) / 37.09%, **20.20%** = (45.26% − 37.65%) / 37.65%, and **22.28%** = (45.57% − 37.27%) / 37.27%), (**6.48%** = (45.92% − 43.13%) / 43.13%, **8.89%** = (45.26% − 41.56%) / 41.56%, and **7.74%** = (45.57% − 42.29%) / 42.29%), (**16.76%** = (45.92% − 39.33%) / 39.33%, **14.48%** = (45.26% − 39.53%) / 39.53%, and **15.67%** = (45.57% − 39.39%) / 39.39%), respectively.
- Second, concerning the *F1*, *CNN-BERT* significantly outperforms *SPC-BERT*, *AEN-BERT*, and *LCF-BERT* on every testing category. The improvement in *F1* varies from **1.16%** = (55.89% − 55.15%) / 55.15% to **73.43%** = (61.04% − 35.20%) / 35.20%, **1.21%** = (48.65% − 48.06%) / 48.06% to **20.52%** = (61.04% − 50.65%) / 50.65%, and **0.07%** = (57.15% − 57.11%) / 57.11% to **29.15%** = (56.40% − 43.67%) / 43.67%, respectively.
- Third, *CNN-BERT* has significant improvement in *Rec* and *F1* on all testing categories against all training languages. However, *CNN-BERT* has slight reduction in *Pre* against *SPC-BERT*, *AEN-BERT*, and *LCF-BERT* on one testing category each, respectively (*English* against *Spanish* training language), i.e., the reduction on *English* is **4.92%** = (61.41% − 58.53%) / 58.53%, (*Dutch* against *Spanish* training language), i.e., the reduction on *Dutch* and *English* is **1.97%** = (50.06% − 49.09%) / 49.09%, and (*English* against *French* training language), i.e., the reduction on *English* is **2.76%** = (59.92% − 58.31%) / 58.31%. The reason of such reduction is that the bi-lingual dictionaries (i.e., Spanish-to-English, Spanish-to-Dutch, and French-to-English) are not contextually rich.

The one-way *ANOVA* is applied on *F1* to further check the performance of *CNN-BERT*. It compares the given approaches to computes the performance difference among them. The results of *ANOVA* are presented in Fig. 3. The results suggests that $F > F_{crit}$, i.e., 21.16 > 2.76, and P_{value} is $1.8E$-09 that is less than 0.05 that indicates a significant difference among the *F1* of the given approaches. Note that, we also apply *ANOVA* on *Pre* and *Rec* that confirms the significant improvement of *CNN-BERT*.

Based on the above analysis, we conclude that *CNN-BERT* significantly improves the state-of-the-art in aspect-based sentiment classification of reviews.

2) RQ2: INFLUENCE OF BI-LINGUAL DICTIONARIES

To answer the research question RQ2, we compare *CNN-BERT* (without attention) against the best state-of-the-art approach (*AEN-BERT*). The results of the approaches are presented in Table 5. The first column represents the languages of the training dataset, the second column represents the languages of the testing dataset for each cross-language validation, and columns 3-5 and 6-8 represent the performance of *AEN-BERT* and *CNN-BERT*, respectively. The first row represents the evaluation metrics against each approach, rows 2-4, 6-8, 10-12, and 14-16 describe the performance of the approaches against each testing language dataset, and rows 5, 9, 13, and 17 represents the average performance of the approaches against all testing languages dataset. The table presents the best performance for each testing category in bold.

From Table 5, the following observations are made.

- First, the use of bi-lingual dictionaries (*CNN-BERT* without attention) improves the performance of the proposed approach. Compared to best state-of-the-art approach (*AEN-BERT*) that does not use bi-lingual dictionaries, the improvement of the proposed approach in average *Pre*, average *Rec*, and average *F1* is **2.99%** = (44.02% − 42.74%) / 42.74%, **5.57%** = (43.38% − 41.09%) / 41.09%, and **4.28%** = (43.65% − 41.86%) / 41.86%, respectively.
- Second, concerning the *F1*, the use of *CNN-BERT* (without attention) significantly outperforms *AEN-BERT* on every testing category. The improvement in *F1* varies from **0.21%** = (52.72% − 52.61%) / 52.61% to **11.94%** = (56.70% − 50.65%) / 50.65%.
- Third, the use of *CNN-BERT* (without attention) has significant improvement in *Rec* on all testing categories against all training languages. However, it has slight reduction in *Pre* against *AEN-BERT* on few testing categories. The reason for such reduction is that some bi-lingual dictionaries are not contextually rich.

TABLE 4. Performance of the proposed approach.

Training Language	Testing Language	SPC-BERT			AEN-BERT			LCF-BERT			CNN-BERT		
		Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
English	Dutch	35.53%	34.87%	35.20%	50.78%	50.52%	50.65%	54.07%	49.13%	51.48%	**63.67%**	**58.62%**	**61.04%**
	French	35.59%	35.38%	35.48%	52.11%	50.74%	51.42%	49.03%	47.95%	48.48%	**56.03%**	**54.65%**	**55.33%**
	Spanish	38.67%	34.84%	36.66%	50.61%	51.93%	51.26%	51.49%	53.98%	52.71%	**58.10%**	**57.72%**	**57.91%**
	Average	**36.60%**	**35.03%**	**35.78%**	**51.17%**	**51.06%**	**51.11%**	**51.53%**	**50.35%**	**50.89%**	**59.27%**	**57.00%**	**58.09%**
Spanish	English	**61.41%**	50.05%	55.15%	56.11%	49.52%	52.61%	49.84%	49.69%	49.76%	58.53%	**53.30%**	**55.79%**
	Dutch	39.31%	42.03%	40.63%	**50.06%**	46.22%	48.06%	41.26%	43.89%	42.53%	49.09%	**48.21%**	**48.65%**
	French	38.52%	41.94%	40.16%	55.91%	50.93%	53.30%	43.85%	46.46%	45.12%	**56.01%**	**54.34%**	**55.16%**
	Average	**46.41%**	**44.67%**	**45.31%**	**54.03%**	**48.89%**	**51.33%**	**44.98%**	**46.68%**	**45.81%**	**54.54%**	**51.95%**	**53.20%**
Dutch	English	52.96%	59.71%	56.13%	54.30%	54.90%	54.60%	48.18%	52.09%	50.06%	**57.24%**	**60.87%**	**59.00%**
	Spanish	41.50%	44.52%	42.96%	53.66%	54.21%	53.93%	44.24%	43.12%	43.67%	**56.02%**	**56.79%**	**56.40%**
	French	44.57%	44.08%	44.32%	49.38%	43.91%	46.48%	44.00%	40.79%	42.33%	**54.72%**	**53.88%**	**54.30%**
	Average	**46.34%**	**49.44%**	**47.80%**	**52.45%**	**51.01%**	**51.67%**	**45.47%**	**45.33%**	**45.36%**	**55.99%**	**57.18%**	**56.57%**
French	English	42.35%	48.51%	45.22%	52.43%	50.74%	51.57%	**59.92%**	54.56%	57.11%	58.31%	**56.04%**	**57.15%**
	Spanish	49.35%	50.86%	50.09%	56.75%	59.09%	57.90%	55.56%	59.36%	57.40%	**62.23%**	**63.72%**	**62.97%**
	Dutch	47.69%	43.84%	45.68%	50.55%	50.91%	50.73%	46.47%	46.66%	46.56%	**54.48%**	**53.18%**	**53.82%**
	Average	**46.46%**	**47.74%**	**47.00%**	**53.24%**	**53.58%**	**53.40%**	**53.98%**	**53.53%**	**53.69%**	**58.34%**	**57.65%**	**57.98%**
Average		**37.09%**	**37.65%**	**37.27%**	**43.13%**	**41.56%**	**42.29%**	**39.33%**	**39.53%**	**39.39%**	**45.92%**	**45.26%**	**45.57%**

Based on the above analysis, we conclude that bi-lingual dictionaries improve the aspect-based sentiment classification of reviews.

3) RQ3: INFLUENCE OF DIVIDED ATTENTION

To answer the research question RQ3, we compare the performances of *CNN-BERT* with and without divided attention. The results of the approaches are presented in Table 6. The first column represents the languages of the training dataset, the second column represents the languages of the testing dataset for each cross-language validation, and columns 3-5, and 6-8 represent the performance of *CNN-BERT* (without attention) and *CNN-BERT* (with attention), respectively. The first row represents the evaluation metrics against each approach, rows 2-4, 6-8, 10-12, and 14-16 describe the performance of the approaches against each testing language dataset, and rows 5, 9, 13, and 17 represents the average performance of the approaches against all testing languages dataset. The table presents the best performance for each testing category in bold.

From Table 6, the following observations are made.

- First, the divided attention in the proposed approach improves the performance of the proposed approach. Compared to *CNN-BERT* (without attention), the improvement of the proposed approach (*CNN-BERT* with attention) in average *Pre*, average *Rec*, and average *F1* is $\mathbf{4.72\%} = (46.10\% - 44.02\%) / 44.02\%$, $\mathbf{4.05\%} = (45.13\% - 43.38\%) / 43.38\%$, and $\mathbf{4.43\%} = (45.59\% - 43.65\%) / 43.65\%$, respectively.
- Second, concerning the *F1*, the use of *CNN-BERT* (with attention) significantly outperforms *CNN-BERT* (without attention) on every testing category. The improvement in *F1* varies from $\mathbf{0.42\%} = (48.65\% - 48.44\%) / 48.44\%$ to $\mathbf{10.35\%} = (57.15\% - 51.79\%) / 51.79\%$.

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
CNN-BERT	16	9.033623	0.564601	0.001094
SPC-BERT	16	7.035779	0.439736	0.004247
AEN-BERT	16	8.300237	0.518765	0.000672
LCF-BERT	16	7.829737	0.489359	0.002286

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.131706	3	0.043902	21.16077	1.8E-09	2.758078
Within Groups	0.124481	60	0.002075			
Total	0.256187	63				

FIGURE 3. ANOVA analysis on *F1*.

Based on the above analysis, we conclude that the divided attention significantly influences the proposed approach in aspect-based sentiment classification of reviews.

4) RQ4: INFLUENCE OF POS TAGGING INFORMATION

To answer the research question RQ4, we compare the performances of *CNN-BERT* without POS tagging information (i.e., verbs and adjectives). The results of the approaches are presented in Table 7. The first column represents the languages of the training dataset, the second column represents the languages of the testing dataset for each cross-language validation, and columns 3-5, 6-8, and 9-11 represent the performance of *CNN-BERT* (without verbs), *CNN-BERT* (without adjectives), and *CNN-BERT*, respectively. The first row represents the evaluation metrics against each approach, rows 2-4, 6-8, 10-12, and 14-16 describe the performance of the approaches against each testing language dataset, and rows 5, 9, 13, and 17 represents the average performance of the approaches against all testing languages dataset. The table presents the best performance for each testing category in bold.

TABLE 5. Influence of bi-lingual dictionaries.

Training Language	Testing Language	AEN-BERT Pre	AEN-BERT Rec	AEN-BERT F1	CNN-BERT Without Attention Pre	CNN-BERT Without Attention Rec	CNN-BERT Without Attention F1
English	Dutch	50.78%	50.52%	50.65%	**57.42%**	**55.99%**	**56.70%**
	French	52.11%	50.74%	51.42%	**55.71%**	**52.96%**	**54.30%**
	Spanish	50.61%	51.93%	51.26%	**56.14%**	**52.88%**	**54.46%**
	Average	**51.17%**	**51.06%**	**51.11%**	**56.42%**	**53.94%**	**55.15%**
Spanish	English	56.11%	49.52%	52.61%	**56.22%**	**49.63%**	**52.72%**
	Dutch	**50.06%**	46.22%	48.06%	48.70%	**48.19%**	**48.44%**
	French	**55.91%**	50.93%	53.30%	54.50%	**53.18%**	**53.83%**
	Average	**54.03%**	**48.89%**	**51.33%**	**53.14%**	**50.33%**	**51.66%**
Dutch	English	54.30%	54.90%	54.60%	**55.67%**	**60.62%**	**58.04%**
	Spanish	**53.66%**	54.21%	53.93%	52.29%	**56.15%**	**54.15%**
	French	49.38%	43.91%	46.48%	**51.34%**	**49.72%**	**50.52%**
	Average	**52.45%**	**51.01%**	**51.67%**	**53.10%**	**55.50%**	**54.23%**
French	English	**52.43%**	50.74%	51.57%	52.19%	**51.40%**	**51.79%**
	Spanish	**56.75%**	59.09%	57.90%	55.68%	**60.65%**	**58.06%**
	Dutch	50.55%	50.91%	50.73%	**53.04%**	**52.82%**	**52.93%**
	Average	**53.24%**	**53.58%**	**53.40%**	**53.64%**	**54.96%**	**54.26%**
Average		**42.74%**	**41.09%**	**41.86%**	**44.02%**	**43.38%**	**43.65%**

TABLE 6. Influence of divided attention.

Training Language	Testing Language	CNN-BERT Without Attention Pre	CNN-BERT Without Attention Rec	CNN-BERT Without Attention F1	CNN-BERT With Attention Pre	CNN-BERT With Attention Rec	CNN-BERT With Attention F1
English	Dutch	57.42%	55.99%	56.70%	**63.67%**	**58.62%**	**61.04%**
	French	55.71%	52.96%	54.30%	**56.03%**	**54.65%**	**55.33%**
	Spanish	56.14%	52.88%	54.46%	**58.10%**	**57.72%**	**57.91%**
	Average	**56.42%**	**53.94%**	**55.15%**	**59.27%**	**57.00%**	**58.09%**
Spanish	English	56.22%	49.63%	52.72%	**58.53%**	**53.30%**	**55.79%**
	Dutch	48.70%	48.19%	48.44%	**49.09%**	**48.21%**	**48.65%**
	French	54.50%	53.18%	53.83%	**56.01%**	**54.34%**	**55.16%**
	Average	**53.14%**	**50.33%**	**51.66%**	**54.54%**	**51.95%**	**53.20%**
Dutch	English	55.67%	60.62%	58.04%	**57.24%**	**60.87%**	**59.00%**
	Spanish	52.29%	56.15%	54.15%	**56.02%**	**56.79%**	**56.40%**
	French	51.34%	49.72%	50.52%	**54.72%**	**53.88%**	**54.30%**
	Average	**53.10%**	**55.50%**	**54.23%**	**55.99%**	**57.18%**	**56.57%**
French	English	52.19%	51.40%	51.79%	**58.31%**	**56.04%**	**57.15%**
	Spanish	55.68%	60.65%	58.06%	**62.23%**	**63.72%**	**62.97%**
	Dutch	53.04%	52.82%	52.93%	**54.48%**	**53.18%**	**53.82%**
	Average	**53.64%**	**54.96%**	**54.26%**	**58.34%**	**57.65%**	**57.98%**
Average		**44.02%**	**43.38%**	**43.65%**	**46.10%**	**45.13%**	**45.59%**

From Table 7, the following observations are made.

- First, the POS tagging information in *CNN-BERT* improves the performance of the proposed approach. Compared to *CNN-BERT* (without verbs and adjectives), the improvement of the proposed approach (*CNN-BERT* with POS tagging information) in average *Pre*, average *Rec*, and average *F1* is (**3.97%** = (45.92% − 44.17%) / 44.17% and **3.72%** = (45.26% − 43.63%) / 43.63%), (**3.86%** = (45.58% − 43.89%) / 43.89% and **5.55%** = (45.92% − 43.51%) / 43.51%), and (**4.03%** = (45.26% − 43.51%) / 43.51% and **4.78%** = (45.58% − 43.50%) / 43.50%), respectively. The possible reason of such improvement is that the proposed approach passes the POS tagging information along with each word that helps CNN to distill local and global features of each text. Avoiding POS tagging information may have some effect on the context and the meaning might changed based on token position.

TABLE 7. Influence of POS tagging information.

Training Language	Testing Language	CNN-BERT Without Verbs Pre	Rec	F1	CNN-BERT Without Adjectives Pre	Rec	F1	CNN-BERT Pre	Rec	F1
English	Dutch	59.11%	58.57%	58.84%	58.44%	54.70%	56.51%	**63.67%**	**58.62%**	**61.04%**
	French	55.98%	54.18%	55.07%	53.14%	53.20%	53.17%	**56.03%**	**54.65%**	**55.33%**
	Spanish	57.41%	57.04%	57.22%	54.90%	57.14%	56.00%	**58.10%**	**57.72%**	**57.91%**
	Average	**57.50%**	**56.60%**	**57.04%**	**55.49%**	**55.01%**	**55.25%**	**59.27%**	**57.00%**	**58.11%**
Spanish	English	56.32%	52.40%	54.29%	53.13%	51.59%	52.35%	**58.53%**	**53.30%**	**55.79%**
	Dutch	49.05%	46.83%	47.91%	46.83%	47.18%	47.00%	**49.09%**	**48.21%**	**48.65%**
	French	54.87%	52.61%	53.72%	55.46%	53.09%	54.25%	**56.01%**	**54.34%**	**55.16%**
	Average	**53.41%**	**50.61%**	**51.98%**	**51.81%**	**50.62%**	**51.21%**	**54.54%**	**51.95%**	**53.22%**
Dutch	English	55.76%	60.12%	57.86%	55.35%	57.06%	56.19%	**57.24%**	**60.87%**	**59.00%**
	Spanish	53.77%	55.20%	54.48%	52.18%	56.21%	54.12%	**56.02%**	**56.79%**	**56.40%**
	French	53.93%	53.77%	53.85%	50.51%	50.31%	50.41%	**54.72%**	**53.88%**	**54.30%**
	Average	**54.49%**	**56.36%**	**55.41%**	**52.68%**	**54.53%**	**53.59%**	**55.99%**	**57.18%**	**56.58%**
French	English	52.15%	50.39%	51.25%	55.19%	55.94%	55.56%	**58.31%**	**56.04%**	**57.15%**
	Spanish	58.84%	59.36%	59.10%	57.73%	59.43%	58.57%	**62.23%**	**63.72%**	**62.97%**
	Dutch	52.22%	50.46%	51.32%	54.11%	49.99%	51.97%	**54.48%**	**53.18%**	**53.82%**
	Average	**54.40%**	**53.40%**	**53.90%**	**55.68%**	**55.12%**	**55.40%**	**58.34%**	**57.65%**	**57.99%**
Average		**44.17%**	**43.63%**	**43.89%**	**43.51%**	**43.51%**	**43.50%**	**45.92%**	**45.26%**	**45.58%**

- Second, concerning the *F1*, the use of *CNN-BERT* (with POS tagging Information) significantly outperforms *CNN-BERT* (without verbs and adjectives) on every testing category. The improvement in *F1* varies from **0.48%** = (55.33% − 55.07%) / 55.07% to **11.51%** = (57.15% − 51.25%) / 51.25%, and from **1.68%** = (57.16% − 54.25%) / 54.25% to **8.02%** = (61.04% − 56.51%) / 56.51%, respectively.

Based on the above analysis, we conclude that the POS tagging information significantly influences the proposed approach in aspect-based sentiment classification of reviews.

F. THREATS

1) THREATS TO VALIDITY

A threat to external validity is that only a limited number of reviews from the *restaurant* domain are considered (mentioned in Section IV-B) for the evaluation of the proposed approach. Although the performance of the proposed approach is significant for the selected reviews, the results may not hold for other domains. Notably, bi-lingual dictionaries are either not available for other domains are not contextually rich.

A threat to construct validity is that the aspect-based classification (i.e., labels) in the exploited dataset may be incorrect. Consequently, the results may not hold for false labeling or revised version(s) of the exploited dataset.

Another threat to construct validity is that the proposed approach is not evaluated on the external validation data as the experts from each language are required to validate results. Consequently, the results of such experiments decrease the performance of the proposed approach.

A threat to internal validity is that we replicate *SPC-BERT*, *AEN-BERT*, and *LCF-BERT* for the comparison/evaluation of the proposed approach. There could be some unseen coding issues. However, we verify the implementation and evaluation results to mitigate the threat.

V. CONCLUSION

Aspect-based sentiment classification for cross-lingual data is a challenging task due to the diversity of language structures. To perform effective aspect-based sentiment classification for cross-lingual data, a deep learning-based classifier is proposed. The proposed approach extracts POS tagging information, preprocesses the given reviews and tokenizes them, performs mapping of tokens from one language to another language, and generates vectors for the preprocessed reviews for the training and evaluation of the proposed approach. The results of cross-lingual validation suggest that the proposed approach significantly outperforms the state-of-the-art approaches and improves the precision, recall, and F1 by more than 23%, 20%, and 22%, respectively.

In future, we are intended to validate the proposed approach for multiple domains of cross-lingual data. Moreover, performing co-extraction of aspect-term and sentiment-polarity and finding the co-occurrences and dependency-relationships for cross-lingual data could be the key future directions.

REFERENCES

[1] H. H. Dohaiha, P. W. C. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: A comparative review," *Expert Syst. Appl.*, vol. 118, pp. 272–299, Mar. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417418306456

[2] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and challenges of aspect-based sentiment analysis: A comprehensive survey," *IEEE Trans. Affect. Comput.*, early access, Jan. 30, 2020, doi: 10.1109/TAFFC.2020.2970399.

[3] A. Tripathy, A. Anand, and S. K. Rath, "Document-level sentiment classification using hybrid machine learning approach," *Knowl. Inf. Syst.*, vol. 53, no. 3, pp. 805–831, 2017, doi: 10.1007/s10115-017-1055-z.

[4] B. Liu, "Sentiment analysis and subjectivity," in *Handbook of Natural Language Processing*, vol. 2. London, U.K.: Chapman & Hall, 2010, pp. 627–666.

[5] K. Schouten and F. Frasincar, "Survey on aspect-level sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 813–830, Mar. 2016.

[6] M. Tubishat, N. Idris, and M. A. M. Abushariah, "Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges," *Inf. Process. Manage.*, vol. 54, no. 4, pp. 545–563, 2018.

[7] X. Zhou, X. Wan, and J. Xiao, "CLOpinionMiner: Opinion target extraction in a cross-language scenario," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 4, pp. 619–630, Apr. 2015.
[8] E. Ghadery, S. Movahedi, H. Faili, and A. Shakery, "MNCN: A multilingual Ngram-based convolutional network for aspect category detection in online reviews," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 6441–6448.
[9] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter sentiment classification," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics*, Jun. 2011, pp. 151–160. [Online]. Available: https://www.aclweb.org/anthology/P11-1016
[10] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Nov. 2016, pp. 214–224. [Online]. Available: https://www.aclweb.org/anthology/D16-1021
[11] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2017, pp. 452–461. [Online]. Available: https://www.aclweb.org/anthology/D17-1047
[12] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," Tech. Rep., 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
[13] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018, *arXiv:1801.06146*. [Online]. Available: http://arxiv.org/abs/1801.06146
[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*. [Online]. Available: http://arxiv.org/abs/1802.05365
[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[16] K. Karthikeyan, Z. Wang, S. Mayhew, and D. Roth, "Cross-lingual ability of multilingual bert: An empirical study," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–12.
[17] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Nov. 2016, pp. 606–615. [Online]. Available: https://www.aclweb.org/anthology/D16-1058
[18] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 4068–4074.
[19] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao, "Targeted sentiment classification with attentional encoder network," in *Proc. Int. Conf. Artif. Neural Netw.* Budapest, Hungary: Springer, 2019, pp. 93–103.
[20] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2004, pp. 168–177, doi: 10.1145/1014052.1014073.
[21] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intell. Syst.*, vol. 32, no. 6, pp. 74–80, Nov./Dec. 2017.
[22] H. Tang, S. Tan, and X. Cheng, "A survey on sentiment detection of reviews," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10760–10773, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417409001626
[23] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2090447914000550
[24] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.
[25] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Feb. 2003.
[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781
[27] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Doha, Qatar, Oct. 2014, pp. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162
[28] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proc. Int. Conf. Knowl. Capture (K-CAP)*, 2003, pp. 70–77, doi: 10.1145/945645.945658.
[29] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," 2015, *arXiv:1512.01100*. [Online]. Available: http://arxiv.org/abs/1512.01100
[30] B. Huang and K. M. Carley, "Parameterized convolutional neural networks for aspect level sentiment classification," 2019, *arXiv:1909.06276*. [Online]. Available: http://arxiv.org/abs/1909.06276
[31] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," 2016, *arXiv:1605.08900*. [Online]. Available: http://arxiv.org/abs/1605.08900
[32] B. Zeng, H. Yang, R. Xu, W. Zhou, and X. Han, "LCF: A local context focus mechanism for aspect-based sentiment classification," *Appl. Sci.*, vol. 9, no. 16, p. 3389, Aug. 2019.
[33] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," 2019, *arXiv:1903.09588*. [Online]. Available: http://arxiv.org/abs/1903.09588
[34] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Y. Hawalah, A. Gelbukh, and Q. Zhou, "Multilingual sentiment analysis: State of the art and independent comparison of techniques," *Cognit. Comput.*, vol. 8, no. 4, pp. 757–771, 2016.
[35] J. Deriu, A. Lucchi, V. De Luca, A. Severyn, S. Müller, M. Cieliebak, T. Hofmann, and M. Jaggi, "Leveraging large amounts of weakly supervised data for multi-language sentiment classification," in *Proc. 26th Int. Conf. World Wide Web (WWW)*, 2017, pp. 1045–1052, doi: 10.1145/3038912.3052611.
[36] A. Balahur and M. Turchi, "Multilingual sentiment analysis using machine translation," in *Proc. 3rd Workshop Comput. Approaches Subjectivity Sentiment Anal. (WASSA)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 52–60.
[37] P. Lambert, "Aspect-level cross-lingual sentiment classification with constrained SMT," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process. (Short Papers)*, vol. 2, 2015, pp. 781–787. [Online]. Available: https://www.aclweb.org/anthology/P15-2128
[38] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," 2017, *arXiv:1710.04087*. [Online]. Available: http://arxiv.org/abs/1710.04087
[39] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[40] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012, *arXiv:1207.0580*. [Online]. Available: http://arxiv.org/abs/1207.0580
[41] W. Y. Ramay, Q. Umer, X. C. Yin, C. Zhu, and I. Illahi, "Deep neural network-based severity prediction of bug reports," *IEEE Access*, vol. 7, pp. 46846–46857, 2019.
[42] (Nov. 1, 2018). *Keras. Merge Layer*. [Online]. Available: https://github.com/keras-team/keras/blob/master/keras/layers/merge.py
[43] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clecq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. Jiménez-Zafra, and G. Eryigit, "Semeval-2016 task 5: Aspect based sentiment analysis," in *Proc. Int. Workshop Semantic Eval.*, Jun. 2016, pp. 19–30. [Online]. Available: https://www.aclweb.org/anthology/S16-1002
[44] N. Q. K. Le and T.-T. Huynh, "Identifying SNAREs by incorporating deep learning architecture and amino acid embedding representation," *Frontiers Physiol.*, vol. 10, p. 1501, Dec. 2019, doi: 10.3389/fphys.2019.01501.
[45] N. Q. K. Le, T. N. K. Hung, D. T. Do, L. H. T. Lam, L. H. Dang, and T.-T. Huynh, "Radiomics-based machine learning model for efficiently classifying transcriptome subtypes in glioblastoma patients from MRI," *Comput. Biol. Med.*, vol. 132, May 2021, Art. no. 104320. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521001141

KALIM SATTAR received the B.S. degree in computer science from the COMSATS Institute of Information Technology, Pakistan, in 2014, and the M.S. degree in computer science from the Institute of Southern Punjab, Pakistan, in 2017. He is currently pursuing the Ph.D. degree with Shenzhen Graduate School, Harbin Institute of Technology, China. His research interests include natural language processing, machine learning, and data mining.

QASIM UMER received the B.S. degree in computer science from Punjab University, Pakistan, in 2006, and the M.S. degree in net distributed system development and the M.S. degree in computer science from the University of Hull, U.K., in 2009 and 2013, respectively, and the Ph.D. degree from Beijing Institute of Technology, China. He is currently an Assistant Professor with the Department of Computer Sciences, COMSATS University Islamabad, Vehari Campus, Pakistan. His research interests include machine learning, data mining, and software maintenance. He is also interested in developing practical tools to assist software engineers.

DINARA G. VASBIEVA received the Specialist degree in world economics, the Ph.D. degree in economics, and the Specialist degree in foreign language from the Russian University of Cooperation, in 1996, 2000, and 2003, respectively. She is currently an Associate Professor with the Department of English Language for Professional Communication, Financial University under the Government of the Russian Federation. She is particularly interested in renewable energy, environmental economics, sustainable development, and cross-cultural communication and linguistics. Her current research interest includes the development of practical tools for software engineers in the area of cross-lingual sentiment.

SUNGWOOK CHUNG received the B.S. degree in computer science from Sogang University, South Korea, in 2002, and the M.S. and Ph.D. degrees from the Department of Computer and Information Science and Engineering (CISE), University of Florida, USA, in 2005 and 2010, respectively. From 2010 to 2012, he worked as a Research Engineer with Korea Telecom (KT), developing the IPTV network architectures and IPTV services. He has been an Associate Professor with the Department of Computer Engineering, Changwon National University, South Korea, since 2012. His research interests include the IoT network architectures and services, high-quality real-time content delivery and distribution, and high-performance computing configurations and services.

ZOHAIB LATIF received the B.S. degree in electrical engineering and the M.S. degree in electronics and electrical engineering from the University of Glasgow, U.K., in 2006 and 2008, respectively, and the Ph.D. degree from the School of Computer Science, Beijing Institute of Technology, Beijing, China, in 2020. He was working as an Assistant Professor with the Department of Computing, Riphah International University, Faisalabad Campus, Pakistan. Since 2011, he has been working as a Senior Lecturer with the School of Computer Science, Beijing Institute of Technology. He is currently a Postdoctoral Research Associate with the School of Computer Science, Hanyang University, South Korea. His major research interests include software defined networks (SDN), machine/deep learning, edge computing, and the Internet of Things.

CHOONHWA LEE (Member, IEEE) received the B.S. and M.S. degrees in computer engineering from Seoul National University, South Korea, in 1990 and 1992, respectively, and the Ph.D. degree in computer engineering from the University of Florida, Gainesville, in 2003. He is currently a Professor with the Department of Computer Science, Hanyang University, Seoul, South Korea. His research interests include cloud computing, peer-to-peer and mobile networking and computing, and distributed computing technology.