Social Network Analysis and Mining (2022) 12:7
https://doi.org/10.1007/s13278-021-00794-4

ORIGINAL ARTICLE

Towards Arabic aspect-based sentiment analysis: a transfer learning-based approach

Rajae Bensoltane[1] · **Taher Zaki**[1]

Received: 7 May 2021 / Revised: 11 July 2021 / Accepted: 19 August 2021 / Published online: 15 November 2021

Abstract

Aspect-based sentiment analysis (ABSA) is a fine-grained sentiment analysis task that aims to extract the discussed aspects and identify their corresponding sentiment polarities from a given text. Most of the existing Arabic ABSA methods rely heavily on tedious preprocessing and feature-engineering tasks in addition to the use of external resources (e.g., lexicons). Hence, this paper tries to overcome these shortcomings by proposing a transfer learning method using pre-trained language models to perform two ABSA tasks in Arabic, namely aspect term extraction (ATE) and aspect category detection (ACD). The proposed models are built based on the Arabic version (AraBERT) of the BERT model. Different implementations of BERT are compared, including fine-tuning and feature-based methods. The main findings of this paper are: (1) Fine-tuning is more suitable in low-resource settings. (2) Designing customized downstream layers enhances the default fine-tuned BERT model results. The experiments were conducted on a reference ABSA dataset on Arabic news posts. The results show that our models outperform the baseline methods and the related work with an overall enhancement of more than 6% for the ATE task and more than 19% for the ACD task.

Keywords Deep learning · Transfer learning · Natural language processing · Aspect-based sentiment analysis · Arabic language · BERT

1 Introduction

With the emergence of web 2.0, different tools have been developed (e.g., social networks, forums, blogs, e-commerce websites) and enabled users to share their opinions and thoughts about various subjects such as products, services, and events. Analyzing this unstructured shared data automatically to extract valuable information about users' opinions and experiments is of great importance for many individuals or groups such as companies, governments, and customers. Therefore, many research fields have become very active such as sentiment analysis and big data.

Sentiment analysis is a sub-field of natural language processing (NLP) that enables the extraction of sentiments expressed in users' reviews. However, identifying the polarity of the whole document or sentence is not always useful. A review can express different opinions towards different aspects, e.g., "*The pizza was delicious, but the ambiance was very bad*". In this example, positive sentiment is expressed toward the aspect term *pizza*, whereas the polarity of *ambiance* is negative. Therefore, a more challenging task of aspect-based sentiment analysis (hereafter SA) was introduced to handle this kind of text.

The Arabic language is the official language of 22 countries, and it is spoken by more than 400 million people in the world (Guellil et al. 2019). It is classified in the 4th position as the most used language on the internet by more than 230 million users, according to the Internet World Stats.[1] In the last decade, SA in Arabic text has attracted the interest of many researchers. However, the number of the released works is still limited compared to the English language, especially for the ABSA tasks (Al-Dabet et al. 2020). This can be justified by the fact that the Arabic language has a rich and complicated morphology and a large number of dialects. Additionally, the lack of resources and reliable NLP tools makes the task of sentiment analysis more difficult.

The majority of the proposed Arabic sentiment analysis systems were built based on traditional machine learning

✉ Rajae Bensoltane
r.bensoltane@uiz.ac.ma

[1] IRF-SIC Laboratory, Faculty of Science, Ibn Zohr University, FP-Agadir, Morocco

[1] http://www.internetworldstats.com/stats7.htm.

classifiers and rule-based approaches. In addition to the use of external resources (e.g., lexicons and NLP tools), especially for morphologically rich languages such as Arabic, these methods generally require tedious and time-consuming preprocessing and feature-engineering tasks. Previously, deep learning models have achieved state-of-the-art results in many domains, including NLP tasks (Goldberg 2016), without relying on hand-crafted features. However, these models also require a huge amount of data to be trained accurately. These large datasets are not always available, especially in low-resource languages (e.g., Arabic) and domains. Furthermore, building high-quality annotated corpora is an expensive and time-consuming task.

Recently, transfer learning based on pre-trained language models has achieved state-of-the-art results in NLP. These models are already pre-trained on a large unlabeled corpus and require much less labeled data to be fine-tuned to downstream NLP tasks. Therefore, the aim of this study is to perform ABSA tasks, namely aspect term extraction and aspect category detection, on Arabic reviews using transfer learning techniques based on pre-trained language models.

The main contributions of this paper can be summarized as follows:

- To the best of our knowledge, this is the first time to apply a transfer learning method based on the BERT model for handling ABSA tasks on an Arabic dataset.
- In contrast to most of the existing Arabic ABSA methods that rely heavily on hand-crafted features and preprocessing steps, we propose end-to-end models for handling ABSA tasks without the need for external resources or feature engineering tasks.
- Implementing modified methods of fine-tuning the BERT model to improve the default fine-tuned BERT model results for both tasks, aspect term extraction, and aspect category detection.

The structure of this paper thus consists of the following sections: A literature review of Arabic ABSA and transfer learning is provided in Sect. 2. The research objectives of our study are presented in Sect. 3. In Sect. 4, the research methodology is discussed. The dataset and baselines are detailed in Sect. 5. Section 6 illustrates the results. Findings, error analysis, and implications of this study are discussed in Sect. 7. Section 8 concludes the paper and provides future directions for this work.

2 Literature review

This section is divided into two sub-sections. The first of which overviews the related work to the ABSA in Arabic (AABSA), whereas the second subsection sheds light on the transfer learning technique using pre-trained languages in general, and the work performed the ABSA based on the BERT model in particular. Because of the lack of Arabic ABSA work using BERT, work on the English language is overviewed.

2.1 Aspect-based sentiment analysis in the Arabic language

The ABSA is a fine-grained SA task that aims to extract the aspects and their related polarities from users' reviews. The ABSA on Arabic reviews is recent. No work handled this task before 2015 (Al-Twairesh et al. 2014), yet the first AABSA work was that of Al-Smadi et al. (2015). The authors have provided an annotated ABSA corpus based on book reviews following the annotation scheme of SemEval 2014 task 4 (Pontiki et al. 2014). They have also offered a set of baseline models for each ABSA task. The obtained results were an $F1$-score of 23.4% and 15.2% for aspect term extraction (ATE) and aspect category detection (ACD), respectively, and an accuracy (Accu) of 29.7% and 42.6% for aspect term polarity (ATEP) and aspect category polarity (ACP), respectively.

Another benchmark dataset was introduced in Al-Ayyoub et al. (2017). The authors have attempted to evaluate the effect of Arabic news posts on readers using an ABSA approach. They have also annotated their corpus following the scheme of SemEval 2014 task 4. The baseline models achieved an $F1$-score of 39.4% for ATE, an accuracy of 65.9% for ATEP, an $F1$-score of 64.9% for ACD, and an accuracy score of 74% for ACP. Moreover, a lexicon-based approach was proposed in the same paper to improve the baseline results for the ATE and ATEP tasks. This method obtained an $F1$-score of 51.4% for the ATE task and an accuracy score of 67.3% for ATEP. The authors have justified this improvement by the fact that the lexicon-based method used extra features (POS tags, N-gram, position, and pruning) that were not implemented in the baseline methods.

Mohammad et al. (2016) have sought to enhance the previous results on aspect term extraction and aspect term polarity detection. The proposed approach used multiple features, including N-grams, named entity recognition (NER), and part-of-speech (POS) tagging. Additionally, multiple classifiers were evaluated. These include decision tree (J48: WEKA implementation), conditional random fields (CRF), Naïve Bayes (NB), and K-nearest neighbor (Ibk: WEKA implementation). Experimental results have shown that all the considered classifiers outperformed the baseline models. The best result for ATE was obtained using J48 ($F1$-score = 81.7%). For the ATPE task, the best result was achieved using CRF (Accu = 86.5%). We notice that we conducted our experiments using the same dataset as used in these two previous studies.

An attempt to enhance the results of SemEval 2016 task 5 was presented in Al-Smadi et al. (2019a). The authors have trained several classifiers, namely K-nearest neighbor (KNN), Naïve Bayes, support-vector machine (SVM), and decision tree, with morphological, syntactic, and semantic features. The SVM classifier outperformed the other classifiers in the three tasks. It achieved $F1$ scores of 93.4% and 89.8% for aspect category detection and opinion target expression, respectively. For sentiment polarity identification, an accuracy score of 95.4% was achieved. The authors have claimed that NER and POS tags features positively impact the results, which affirmed the findings of Mohammad et al. (2015). Additionally, employing morphological features has yielded improved results, which, according to the authors, has confirmed the findings of Abdul-Mageed et al. (2011), Duwairi and El-Orfali (2014), Refaee and Rieser (2014).

Areed et al. (2020) have also introduced a new dataset that consisted of Arabic reviews of government smart applications in Dubai. They proposed an ABSA method to provide insights into the expectations and needs of clients. They combined rule-based with lexicon-based methods in order to extract aspects and classify their related sentiments. The experimental results showed improvements in f-measure and accuracy of 17% and 6%, respectively, compared to the baseline model. They have achieved an $F1$-score of 92.5% for aspect extraction and an accuracy score of 95.8% for sentiment classification.

On the other hand, the use of DL techniques in Arabic SA is in its infancy (Oueslati et al. 2020), especially in the ABSA field, which progresses slowly compared to the English language (Al-Dabet et al. 2020). Al-Smadi et al. (2019b) have implemented two models based on a long short-term memory network (LSTM) to enhance the results on the Hotel datasets (Al-Smadi et al. 2016; Pontiki et al. 2016) in slot 2 and slot 3. A bidirectional long short-term memory (BiLSTM) with CRF classifier (BiLSTM-CRF) was implemented for the sub-task of opinion target expression (OTE) following the work of Lample et al. (2016). They have also investigated two word embedding techniques, a word-level based on Word2vec and a character-level using FastText. For sentiment polarity classification, the aspect-OTEs were considered as attention expressions in an LSTM-based model to help identify the aspect's sentiment polarity. The best result for slot 2 was obtained using BiLSTM-CRF (FastText) by achieving an improvement of 39% over the baseline results ($F1$-score = 69.9% vs. 30.9%). Whereas for slot 3, a comparable result to the best model in SemEval 2016 task 5 (Pontiki et al. 2016) was achieved (Accu = 82.6%).

In Al-Dabet et al. (2020), the previous model for extracting the opinion target expression was enhanced by adding two other layers: a CNN layer for extracting the character-level and concatenate it with word-level vectors and an attention layer to capture the key parts of the sentence. The model was evaluated on the same Hotel reviews dataset. Different pre-trained word embeddings models were also investigated. Experimental results showed that the proposed model outperformed the baseline and the related research work, including the previous model. It has achieved an $F1$-score of 72.8% using a CBOW model trained on Wikipedia datasets with a dimensional size of 100. Different experiments were also conducted with and without using the CNN model and have shown that the character-level vectors extracted by CNN positively impacted the model's performance.

Unlike the existing ABSA work on Arabic reviews, the proposed method uses transfer learning based on pre-trained language models to enhance the ATE and ACD tasks in Arabic. This work tries to overcome the main shortcomings of the previous methods that rely mainly on laborious preprocessing steps and feature extraction tasks. Moreover, this method tries to address the challenge of the limited size of annotated Arabic ABSA datasets by exploiting the recent trends of transfer learning techniques in NLP.

2.2 Transfer learning

Transfer learning (Pratt 1992) is a machine learning technique that stores the knowledge learned while solving one problem and exploiting it to handle different but related problems. Recently, transfer learning using pre-trained language models has contributed to the state-of-the-art of many NLP tasks, including SA. These models are already pre-trained on a large unlabeled text and can be fine-tuned to downstream NLP tasks by using relatively small labeled data. Thus, this can overcome the shortcoming of the limited annotated dataset in some domains and languages and save time and computational resources needed for building a new model from scratch. Many pre-trained language models were proposed in the last few years, such as OpenAI GPT (Radford et al. 2019), XLNET (Yang et al. 2019), and BERT (Devlin et al. 2019).

BERT is a deeply bidirectional and unsupervised language representation model. Unlike word2vec and glove word embedding models that are context-independent, BERT is a contextualized model that can generate different word embeddings for the same word based on the context it occurs. BERT has achieved state-of-the-art results in many NLP tasks, including SA.

For the English language, many papers that have applied BERT to ABSA tasks can be found. The authors in Sun et al. (2019) have investigated different methods of constructing an auxiliary sentence to transform ABSA from a single sentence classification task to a sentence pair classification task. They fine-tuned the pre-trained model from BERT on the sentence pair classification and obtained state-of-the-art

results on SentiHood ($F1$ score = 87.9% for aspect detection and Accu = 97% for sentiment classification) and SemEval task 4 datasets ($F1$ score = 92.3% for ACD and Accu = 95.6% for ACP).

In Li et al. (2019), the authors jointly have solved the aspect detection and aspect sentiment classification tasks as a sequence labeling problem. They investigated different neural models for implementing a layer (called E2E-ABSA) on top of the BERT embeddings layer, including linear layer, CRF layer, recurrent neural networks, and self-attention networks (SAN). The proposed approach outperformed state-of-the-art results and achieved an F-1 score of 61.1% (using BERT + GRU) in the laptop domain and an $f1$-score of 74.7% in the restaurant domain (using BERT + SAN).

In Hoang et al. (2019), the authors fine-tuned BERT on pair sentence classification to handle ABSA tasks. They implemented three models, one for sentiment classification, a second for the aspect category detection, and a third one for jointly detecting the aspect category and its sentiment polarity. The models were evaluated on text-level, sentence-level, in-domain, and out-of-domain using SemEval 2016 task 5 datasets in the laptop, restaurant, and hotel domains. The achieved results showed that the combined model outperformed state-of-the-art results for aspect-based sentiment classification.

The authors in Rietzler et al. (2020) handled the aspect-target sentiment classification by fine-tuning BERT on domain-specific data. They implemented a two-stage procedure: the first step is a self-supervised domain-specific BERT fine-tuning, followed by a supervised task-specific fine-tuning. They reported a state-of-the-art result on the SemEval 2014 Task 4 restaurants dataset (Accu = 87.1%). They also showed that the cross-domain adapted BERT model performed better than strong baseline models like vanilla BERT-base and XLNet-base.

In Meškelė and Frasincar (2020), a hybrid model, called ALDONAr, was proposed for aspect-based sentiment analysis. The authors combined BERT for producing word embeddings, a sentiment domain ontology for detecting domain information, and two CNN layers to enhance sentiment classification. The proposed model achieved accuracy scores of 83.8% and 87.1% on SemEval 2015 task 12 (Pontiki et al. 2015) and SemEval 2016 task 5 (Pontiki et al. 2016) restaurants datasets, respectively.

Abas et al. (2020) pre-trained the BERT model on three domain-specific corpora (laptops, restaurant, and Twitter) for domain adaption and local and global context features extraction. They then used a continuous dynamic masking technique for distinguishing aspects' local and global semantic features based on a predefined relative distance measure. Finally, several multi-head attention (MHA) mechanisms and a convolutional operation are used to capture hidden semantic interaction and predict the targeted aspect's sentiment polarity. The experimental results showed that the proposed model outperformed the compared state-of-the-art models by achieving accuracy scores of 91.6%, 85.2%, and 80.6% for restaurant, laptop, and Twitter datasets, respectively.

For the Arabic language, there exist two types of pre-trained language models. The first type is multilingual models such as mBERT (Devlin et al. 2019), which provides sentence representations for 104 languages, including Arabic. The other type is monolingual models such as Hulmona (ElJundi et al. 2019), AraBERT (Baly and Hajj 2020), Qarib (Chowdhury et al. 2020), and MamBERT (Abdul-Mageed et al. 2101). Hulmona used the ULMfit structure (Howard and Ruder 2018) and was trained on a large general-domain Arabic corpus, whereas the other models were developed based on Google's BERT architecture. These models have achieved enhanced results in many Arabic NLP downstream tasks, including sentiment analysis, named entity recognition, and question answering.

The authors in Chowdhury et al. (2020) tried to enhance the text categorization task in the Arabic dataset using a monolingual language model (called Qarib) trained on formal and informal Arabic content. They also provided two new Arabic text categorization datasets. The first data (called ASND) are a set of MSA posts collected from the official Aljazeera news channel accounts on Facebook, YouTube, and Twitter, whereas the second data are a set of dialectical tweets collected from Arab influencers' accounts on Twitter. The experimental results showed the effectiveness of training a BERT model using a mixture of formal and informal datasets compared to BERT models trained using formal data only ($F1$: 81% using Qarib vs. 60% using AraBERT on ASND dataset). Moreover, the empirical studies showed that diversifying the training dataset for the specific task yielded improved results.

Abuzayed and Al-Khalifa (2021) introduced their systems submitted to the shared task of sarcasm and sentiment detection in Arabic (Farha et al. 2021). They fine-tuned seven BERT-based models, including MamBERT and Qarib. They claimed that the MamBERT model achieved promising results compared to the other models ($F1$-score: 65% vs. 60% using Qarib on Sarcasm detection). They also evaluated the MamBERT-based model with data augmentation methods to handle the imbalanced data problem. Experimental results showed the effectiveness of the data augmentation methods by enhancing the results of the MamBERT model in both tasks ($F1$-score: 80% for sarcasm detection and F-PN: 86% for sentiment analysis). Furthermore, a conducted error analysis showed that the main reasons for misclassifying some tweets were related to the nature of the dataset besides the errors in the human annotation.

In Boudjellal et al. (2021), a BERT-based model was implemented to identify biomedical entities in the Arabic

text. The authors developed the ABioNER model by further pre-trained the AraBERT model on a small-scale biomedical dataset. Experiments were conducted on "Disease or Syndrome" and "Therapeutic Or Preventive Procedure" entities only. Results showed that the proposed model outperformed two state-of-the-art BERT models (AraBERT and mBERT) by achieving an $F1$-score of 85%. The authors claimed that pre-training a monolingual BERT model on small-scale biomedical data can enhance the biomedical domain text model understanding.

The previous studies used pre-trained language models to perform other NLP tasks than ABSA in Arabic. Hence, this work tries to fill this gap by investigating the use of BERT-based models in handling two ABSA tasks (ATE and ACD) on an Arabic reference dataset.

3 Research objectives

The aim of this study is to investigate the effectiveness of a transfer learning approach based on the BERT model in Arabic ABSA and implement efficient end-to-end models for ATE and ACD tasks. The following research objectives are considered in this paper:

- Examine the effectiveness of the BERT model in handling the complexity and ambiguity of the Arabic language without the need for tedious preprocessing and feature-engineering tasks.
- Investigate the use of a transfer learning method to overcome the challenge of data scarcity in Arabic ABSA.
- Examine the effectiveness of fine-tuning the BERT model against well-known deep neural networks in a low-resource setting.
- Investigate the impact of adding more powerful layers on top of the BERT model in handling the ATE task.
- Examine the use of aspect terms information in enhancing the knowledge learned by BERT for detecting the aspect categories.

4 Methodology

4.1 Tasks description

In the SemEval 2014 task 4 (Pontiki et al. 2014), the aspect-based sentiment analysis task was divided into four sub-tasks:

- *Aspect term extraction* This task aims to extract aspects or features of products, services, or news that have been evaluated in a given sentence, e.g., "*The camera of this phone is very powerful*". In this sentence, the reviewer evaluated the *camera* of the phone.
- *Aspect term polarity* This task aims to identify the semantic orientation (e.g., positive, negative, neutral, or conflict) of each aspect that has been evaluated within a sentence, e.g., "*The camera of this phone is amazing*". This review expresses a positive opinion on the *camera*.
- *Aspect category detection* The objective of this task is to identify the aspect category discussed in a given sentence based on a predefined list of aspect categories, e.g., "*The pizza was very delicious*". The aspect category is *FOOD*. It is considered as the hypernym of the aspect term *pizza*.
- *Aspect category polarity* This task aims to detect the sentiment polarity of the discussed aspect categories in a given sentence, e.g., "*The dishes were delicious, but the music was horrible*". The reviewer in this sentence expressed positive sentiment towards the category *FOOD* but a negative opinion about the *AMBIENCE* category.

In this paper, only two of these tasks are investigated, namely the aspect term extraction and the aspect category detection.

4.2 Models overview

4.2.1 Bert architecture

BERT is implemented based on a transformer, which is an attention mechanism that uses an encoder to read the text input and a decoder that generates a prediction for the task. BERT employs only the encoder part for generating a language representation model. It uses an input of a single sequence of tokens that can represent a single text sentence as well as a pair of sentences. Before feeding to BERT, two extra tokens are added at the beginning ([CLS]) and the end ([SEP]) of the tokenized sentence. The first token is used as an input representation for classification tasks, while the second one separates a pair of input texts. BERT can be fine-tuned to a downstream NLP task by adding additional layer/s on top of BERT and train all layers together.

AraBERT is an Arabic language representation model that was developed based on the BERT model. It used the base configuration of BERT with 12 transformer blocks, 768 hidden dimensions, 12 attention heads, a maximum sequence length of 512 tokens, and a total of ~ 110 M parameters. To avoid the lexical sparsity of Arabic, the words are first segmented into stems, prefixes, and suffixes. Then, a SentencePiece tokenizer (Kudo 2018) is trained in unigram mode on the segmented pre-training dataset (total size of vocabulary ~ 60 K tokens). Another version of AraBERT (AraBERTv0.1) is created by training SentencePiece on a non-segmented text and has a vocabulary size of 64 k tokens.

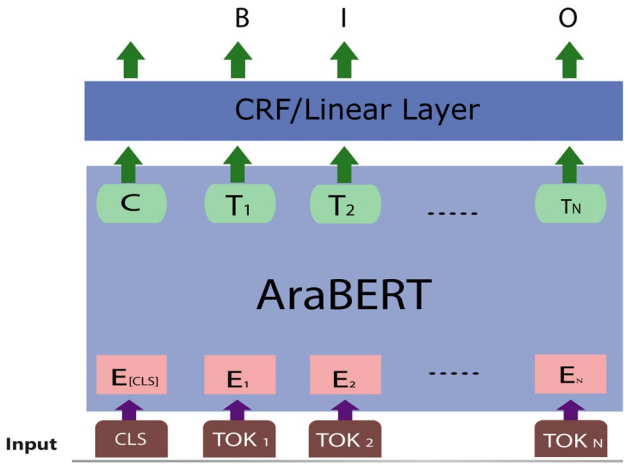

Fig. 1 Architecture of BERT + Linear and BERT + CRF models

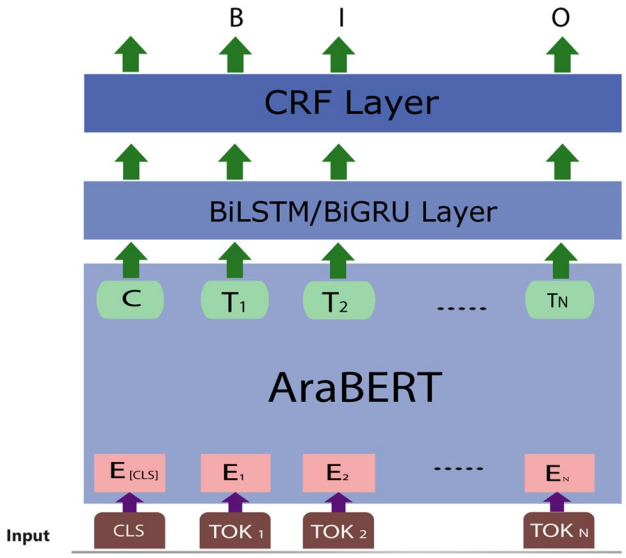

Fig. 2 Architecture of BERT + BiLSTM + CRF and BERT + BiGRU + CRF models

4.2.2 Aspect term extraction

This task can be treated as a sequence labeling problem (Al-Smadi et al. 2019b; Xu et al. 2018). We annotated the sentences following the BIO scheme, where the "B" tag designates the beginning of the aspect term, "I" corresponds to the rest of the words of the same aspect, and the "O" tag indicates that the tagged word is a non-aspect term.

The performance of the pre-trained BERT model with different layers was evaluated. We notice that during fine-tuning, the entire BERT parameters and the additional layers were trained on our specific task.

- The first model (BERT + Linear) fine-tuned BERT by adding a linear layer with a SoftMax activation function on top of BERT, which took as input the last hidden state of every input token and outputted a prediction for each token. Figure 1 illustrates the model used.
- Instead of adding a linear layer on top of BERT, this model (BERT + CRF) used a CRF layer for final sequence tagging prediction. CRF is known for its efficiency in sequence labeling tasks (Yu et al. 2010). It predicts the labels not only based on the current token but also based on the previously predicted label. Figure 1 illustrates the model used.
- In addition to CRF, we also tried to evaluate the impact of adding a BiLSTM layer, which has achieved significant results in sequence labeling tasks (Steingrímsson et al. 2019). The proposed model (BERT + BiLSTM + CRF) incorporated BiLSTM into the BERT model and used the CRF layer to label the entire sequence. Figure 2 represents the model employed.
- The last model is the same as the previous model. Still, instead of using a BiLSTM layer, we used a bidirectional gated recurrent unit (BiGRU) layer (BERT + BiGRU + CRF) to evaluate its impact on the whole system. Especially that GRU has shown its superiority over basic RNN and LSTM models in previous studies, e.g., the work of Chung et al. (2014) and Jozefowicz et al. (2015). Figure 2 represents the model employed.

Since AraBERT uses a tokenizer that might split a single word into several sub-words, we assigned the tag of the token to its first piece, and the special tag 'U' was assigned to the rest of the pieces. U tokens were ignored in training and testing predictions.

4.2.3 Aspect category detection

For this task, we compared two approaches using the BERT model, namely feature-based approach and fine-tuning approach.

For the fine-tuning method, two architectures of BERT were explored, namely a single sentence classification (S-BERT) and a sentence pair classification (P-BERT). During fine-tuning, the entire layers were trained, and the BERT parameters were updated.

- S-BERT: The dataset we used in our experiments assigned only one category to each sentence. Thus, the first model fine-tuned BERT for a multi-class classification problem. It used the sentence review as a single input to the BERT model and. Then, a fully connected layer with a SoftMax activation function added on top of BERT took as input the [CLS] token representation in the final layer of BERT and outputted the probabili-

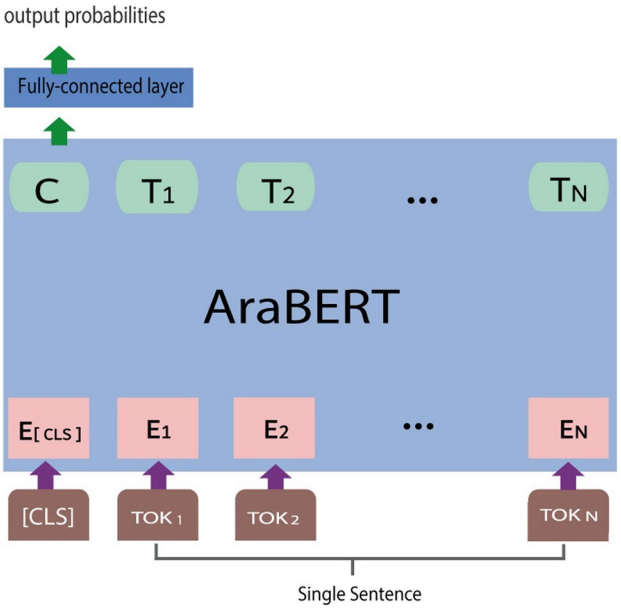

Fig. 3 Architecture of S-BERT model

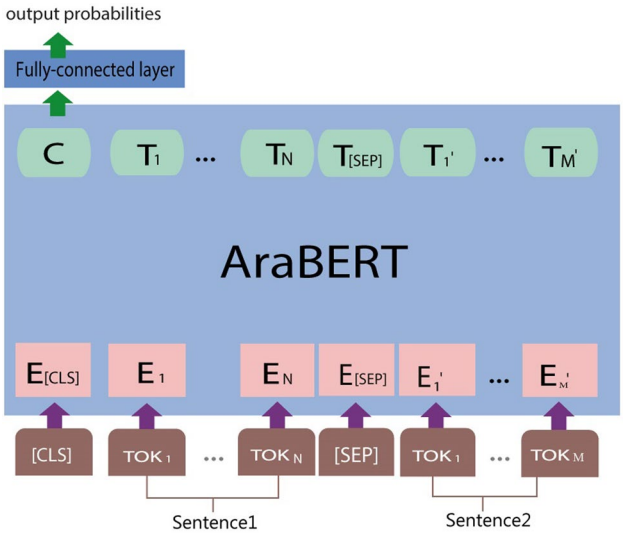

Fig. 4 Architecture of P-BERT model

ties over each label category. Figure 3 illustrates the model employed.

- P-BERT: The second model exploited the relation between aspect categories and aspect terms, as the categories can be considered as the generalization of aspect terms (Xue et al. 2017). Thus, we used BERT's pair sentence classification architecture, which took the sentence review as the first input and aspect terms as the second inputs. The last hidden state of the first token [CLS] was then inputted to the classification layer with a SoftMax function to compute the label probabilities. Figure 4 shows the model employed.

For the feature-based method, the BERT model was used to produce word embeddings to train a neural network model. We evaluated two strategies of extracting the embeddings without fine-tuning BERT parameters. The first of which was to extract the representations from the last hidden layer (BERT-Last + CNN), and the second was to concatenate the representations from the last four hidden layers (BERT-4LAST + CNN).

The neural model we used was based on the CNN model proposed by Kim (2014), as it has shown to be effective for the text classification tasks (Guo et al. 2019). The model was implemented using a single convolutional layer with multiple filter sizes and feature maps, followed by a max-pooling layer. A dropout layer was added to prevent overfitting and co-adaptation of hidden units. A final SoftMax layer was then used for computing probabilities over each class.

Table 1 Distribution of aspect terms over category classes

Category	Number of aspects
Plans	3183
Results	4083
Peace	1815
Parties	520

5 Experiments

5.1 Dataset description

The dataset we used in our experiments was introduced in Al-Ayyoub et al. (2017). The authors collected news posts and comments from Facebook about the 2014 Gaza Attacks to evaluate the effect of the news on readers. They manually annotated the dataset following the SemEval 2014 task 4 annotation guidelines and using an extensible markup language (XML) schema. They annotated the aspect terms, aspect categories, and the corresponding sentiment polarity of each term and category for each post.

The dataset consists of 2265 posts written in modern standard Arabic (MSA). Four predefined category classes are considered (الخطط/Plans, النتائج/Results, الهدنة/Peace, الأطراف/Parties). The distribution of aspect terms over the category classes is presented in Table 1, while Table 2 illustrates the number of posts per aspect category. Moreover, an example of annotated sentences from the used dataset is illustrated in Fig. 5.

Table 2 Number of posts per aspect category

Category	Number of posts
Plans	813
Results	961
Peace	350
Parties	141

5.2 Preprocessing

We have tested our models with different preprocessing steps, and we have remarked that:

- Removing punctuations, symbols, URLs, non-Arabic words, diacritics, numbers, and elongation has no benefit to our models.
- Applying normalization, stemming, or removing stop words has decreased the performance of our models. This can be justified by the fact that these methods have affected the contextual meaning learned by BERT.

Therefore, we decided not to apply any of these steps to the dataset. Meanwhile, the text must be tokenized before feeding to the BERT model. No segmentation was needed as we used the AraBERT v01 model. Additionally, the special token [CLS] was added at the first position of each single input sequence, whereas the [SEP] token was used to separate the tokens of two sentences in case of the sentence pair classification.

5.3 Evaluation metrics

To compare our results to the baselines and related work, we used the same metrics. Precision (Eq. 1), recall (Eq. 2), and $F1$ score (Eq. 3) were calculated for both ATE and ACD tasks.

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$F1 = \frac{2PR}{P + R} \tag{3}$$

TP is the set of true positives, which are the relevant aspect terms (T1), or aspect categories (T2) that were retrieved from the test dataset.
FP is the set of false positives, which means the irrelevant aspect terms or categories identified in the same dataset.
FN is the set of relevant terms or categories that have not been extracted.

Fig. 5 Example of sentences annotated based on the SemEval-2014 task 4 annotation guidelines

```
<sentence id="Event1Post_0007">
        <text>اوباما يعبر عن مواساته لمقتل المستوطنين الثلاثة ويدين العنف ويدعو الاطراف للامتناع عن الاعمال التي تزعزع الاستقرار</text>
         <aspectTerms>
                <aspectTerm term="اوباما" from="0" to="6" polarity="neutral"/>
                <aspectTerm term="مقتل" from="24" to="28" polarity="positive"/>
                <aspectTerm term="المستوطنين" from="29" to="39" polarity="positive"/>
                <aspectTerm term="العنف" from="54" to="59" polarity="positive"/>
                <aspectTerm term="الاطراف" from="66" to="73" polarity="positive"/>
                <aspectTerm term="الاعمال" from="86" to="93" polarity="negative"/>
                <aspectTerm term="الاستقرار" from="105" to="114" polarity="negative"/>
        </aspectTerms>
        <aspectCategories>
                <aspectCategory polarity="neutral" category="الاطراف"/>
        </aspectCategories>
</sentence>

<sentence id="Event6Post_0244">
        <text>سكاي نيوز ،، الشرطة الاسرائيلية تفرق تظاهرة في تل ابيب ضد العملية العسكرية على قطاع غزة</text>
         <aspectTerms>
                <aspectTerm term="الشرطة الاسرائيلية" from="14" to="32" polarity="negative"/>
                <aspectTerm term="تظاهرة" from="38" to="44" polarity="negative"/>
                <aspectTerm term="العملية العسكرية" from="59" to="75" polarity="positive"/>
        </aspectTerms>
        <aspectCategories>
                <aspectCategory polarity="negative" category="النتائج"/>
        </aspectCategories>
</sentence>
```

Table 3 Results of aspect term extraction

Model	Precision (%)	Recall (%)	*F*1 score (%)
Baseline (Al-Ayyoub et al. 2017)	37.9	41	39.4
Lexicon-based (Al-Ayyoub et al. 2017)	48.5	54.6	51.4
BiLSTM + CRF	75.4	79.6	77.4
J48 (Mohammad et al. 2016)	81.3	82.5	81.7
BERT + Linear	85.8	88.5	87.1
BERT + CRF	86.7	88.1	87.4
BERT + BiLSTM + CRF	87.1	88.3	87.7
BERT + BiGRU + CRF	**87.7**	**88.5**	**88.1**

Bold values indicate the best results

5.4 Baseline and related research results

5.4.1 Aspect term extraction

In order to evaluate the performance of our models, we compare them with several baseline methods. We list them as follows:

- Baseline (Al-Ayyoub et al. 2017): following the SemEval 2014 task 4 baseline model. All tokens in the testing dataset were tagged as an aspect if they matched with a dictionary of aspect terms extracted from the training dataset.
- Lexicon-based approach (Al-Ayyoub et al. 2017): the method consisted of three steps. Firstly, a lexicon of aspect terms that appeared in the training set was built. Then, multiple preprocessing methods were applied. Many features were used in the last step, including POS tags, *N*-gram, position, and pruning, to finally capture the aspect terms in the testing set.
- J48 (Mohammad et al. 2016): different classifiers (NB, J48, CRF, Ibk) were trained using multiple features, including part-of-speech tagging, NER, *N*-grams, and aspect terms position. The J48 model achieved the best result in this task. Thus, we compared our results against it.
- BiLSTM-CRF: this model has shown its efficiency in handling sequence labeling tasks such as named entity recognition (NER), part-of-speech (POS) tagging. To convert words into their corresponding word embeddings, we used the AraVec (wiki skip-gram 300) (Soliman et al. 2017) word embedding model, which was built for the Arabic language following Word2vec (Mikolov et al. 2013) approach.

5.4.2 Aspect category detection

In order to further examine and validate the performance of our models, we list two baseline methods:

- Baseline (Al-Ayyoub et al. 2017): based on the SemEval 2014 task 4 baseline model. The test sentence t was assigned to the most frequent category of the k similar training sentences to t. The Dice similarity coefficient was used to calculate the distance between two sentences.
- AraVec-CNN: we evaluated the same CNN model described in Sect. 4.2.3 but using AraVec (wiki skip-gram 300) embeddings instead of BERT embeddings.

6 Results

6.1 Aspect term extraction

The experimental results for this task are illustrated in Table 3. They show that our models outperform the baseline model significantly and surpass the results of the related work on the same dataset. The evaluation results can be summarized as follows:

- The reported results for the baseline and related work are: *F*1-score = 39.4%, *F*1-score = 51.4% for the lexicon-based approach, and *F*-1 score = 81.7% for J48 classifier.
- The default fine-tuned BERT model outperforms the well-known BiLSTM + CRF model with AraVec embeddings (*F*1 score = 87.1% vs. 77.4%)
- Adding different layers on the top of BERT enhances the results of the default fine-tuned model. The best result (*F*1-score: 88.1%) in this task is achieved using the BERT + BiGRU + CRF model.

6.2 Aspect category detection

The evaluation results show that our models outperform the baseline model significantly. However, no related work can be found for this task on the same dataset. As presented

Table 4 Results of aspect category detection

Model	Precision (%)	Recall (%)	*F*1 score (%)
Baseline (Al-Ayyoub et al. 2017)	64.9	64.9	64.9
AraVec + CNN	72.4	72.4	72.4
BERT-LAST + CNN	74.8	74.8	74.8
BERT-4LAST + CNN	77.5	77.5	77.5
S-BERT	82.8	82.8	82.8
P-BERT	**84.1**	**84.1**	**84.1**

Bold values indicate the best results

in Table 4, the experimental results can be summarized as follows:

- The baseline model achieves an *F*1-score of 64.9%, which is lower than that of our best model by ~ 19%.
- The BERT feature-based method (BERT-LAST + CNN) outperforms the CNN model using AraVec embeddings (*F*1 score: 77.5% vs. 74.8%)
- BERT-4LAST + CNN achieves better results than the BERT-LAST + CNN model (*F*1-score: 77.5% vs. 74.8%)
- The BERT fine-tuning methods outperform the BERT feature-based methods, and the best result in this task is obtained using P-BERT (*F*1-score: 84.1%).

7 Discussion

7.1 Aspect term extraction

The evaluation results show that the proposed models outperform the baseline and related work on the same dataset. Our Arabic BERT-based models have achieved better results than a machine learning-based method trained on semantic and syntactic features (*F*1-score = 87.1% using BERT + linear vs. 81.7% using J48 (Mohammad et al. 2016)). This indicates that the AraBERT model can handle the morphological complexity and the structural ambiguity in Arabic without the need for hand-crafted features or tedious preprocessing steps.

Additionally, the default fine-tuned BERT model outperforms the well-known BiLSTM + CRF model using AraVec embeddings. This can be due to the fact that deep neural network models require a large amount of dataset to be trained from scratch. Therefore, this shows that fine-tuning is more suitable in case of having a limited size of task-specific labeled datasets. Moreover, BERT models word meaning and context information better than a context-independent word embeddings model. Besides, the BERT model breaks down unseen words into multiple sub words, which is better in handling the out of vocabulary (OOV) issue that constitutes a significant problem for NLP tasks, especially in morphologically rich languages like Arabic.

Furthermore, adding more powerful layers on the top of BERT has enhanced the default fine-tuned model results, which shows that it is efficient to design customized downstream models for a specific task. The BERT + BiGRU + CRF model has achieved the best results in this task, which shows that in our case, the BiGRU layer has improved the context information extracted by BERT better than the BiLSTM layer.

7.2 Aspect category detection

The experimental results show that the proposed models outperform the baseline model significantly. Besides, the CNN model with both strategies of extracting BERT embeddings outperforms the AraVec-CNN model. This can be explained by the fact that the BERT model extracts the semantic representations better than a context-free word embeddings model, especially that BERT learns information from both directions during the training phase. Additionally, as mentioned before, BERT can handle the OOV problem better than a word embeddings model.

Moreover, the features extracted by concatenating the last four hidden layers yield better results than extracting the last hidden layer (77.5% vs. 74.8%), which is compatible with the findings in the original BERT paper (Devlin et al. 2019). However, the fine-tuning method proves to be more accurate than the feature-based method using the BERT model by boosting the *F*1-score by more than 6%. This can be justified by the fact that the pre-trained model has weights that require smaller data to be fine-tuned compared to a task-specific deep neural network. Hence, this proves that further fine-tuning on the specific task is necessary to release BERT's true power, especially in low-resource settings. Furthermore, P-BERT has achieved the best results and has outperformed the S-BERT model (*F*1: 84.1% vs. 82.8%), which shows that integrating the aspect terms information has enhanced the knowledge learned by BERT for detecting the aspect categories.

Table 5 Case study of different models on the used dataset

Sentence	Predicted label using the proposed approach	Predicted label using AraVec-BiLSTM-CRF	True label
Case 1			
Ar: عباس: دولة قطر اسهمت في اتفاق وقف اطلاق النار وكذلك وزير الخارجية الأمريكي	B	B	B
En: Abbas: The State of Qatar contributed to the ceasefire agreement, as did the US Secretary of State			
Ar: المواقع الاخبارية العبرية لم تورد اي وقف اطلاق نار نبا حول التوصل الى اتفاق في غزة حتى هذه اللحظة	B	B	B
En: The Hebrew news websites have not reported any news about reaching a cease-fire agreement in Gaza until this moment			
Ar: عزام الاحمد: بعد اتفاقنا مع وزير خارجية قطر فوجئنا بتصريحات خالد مشعل التي تقول لا داعي للمفاوضات	B	O	B
En: Azzam Al-Ahmad: After our agreement with the Minister of Foreign Affairs of Qatar, we were surprised by Khaled Meshaal's statements that there is no need for negotiations			
Case 2			
Ar: في اول رحلة صيد بحرية لمسافة 6ميل صيادي غزة يصطادون سمكة قرش صغيرة	I	I	I
En: On their first 6-mile fishing trip, Gaza fishermen caught a small shark			
Ar: افيخاي ادرعي ،،، من يعتقد ان بامكانه اطلاق صواريخه العبثية بحرية على مواطنينا فهو واهم	O	B	O
En: Avichai Adraee: Whoever thinks that he can fire his absurd missiles freely at our citizens is deluded			
Case 3			
Ar: المقاومة تتمكن برغم القصف المستمر على القطاع واستمرارية تحليق كافة انواع المحاذية الطائرات، من قصف المستوطنات لقطاع غزة بعدد من الصواريخ والاحتلال يعترف بسقوطها بالنقب الغربي	B B	I B	B B
En: Despite the continuous bombardment on the Strip and the continued flight of all kinds of aircraft, the resistance was able to bombard the settlements adjacent to the Gaza Strip with a number of missiles, and the occupation acknowledges their fall on the western Negev			
Ar: الصحة: وصول اصابتين خطيرة في استهداف معبر رفح جنوب القطاع الي مستشفى ابو يوسف النجار	B B	I B	B B
En: Health: Two serious injuries in the targeted Rafah crossing in the southern Gaza Strip arrived at Abu Yousef Al-Najjar Hospital			

7.3 Case study

The evaluation results have shown the effectiveness of our models compared to feature-based machine learning methods and deep learning models using context-free word embeddings. To better understand the proposed method's efficacy in handling the ambiguity and the morphological complexity in the Arabic language, we illustrate frequently

Table 6 Example of errors for the ATE task

Sentence	Error type
Ar: المعبر سيدار باشراف القوات التابعة لحرس الرئيس **En**: The crossing will be managed under the supervision of the forces of the Presidential Guards. Predicted labels: I B O B True labels: I I I B	Error due to long and complex aspect terms
Ar: الخلية الثانية اطلقت صاروخ مضاد للدروع باتجاه سيارة عسكرية **En**: The second cell fired an anti-tank missile at a military vehicle Predicted labels: B I B True labels: I I B	

confirmed correct cases of our model compared to the AraVec-BiLSTM-CRF in Table 5.

Arabic is an agglutinative language that attaches affixes to roots to generate different words from the same stem. Given examples of case 1, adding suffixes to the word اتفاق has generated a different word اتفاقنا (Our agreement) based on the context in which it appears. However, this word has the same surface (agreement) in the English translation of all of these sentences. Our BERT-based model correctly predicted the label of these words in the three examples without any preprocessing (e.g., stemming or lemmatization), whereas the BiLSTM-CRF model failed to detect the correct label in the last example. This can be justified by the fact that AraBERT splits unknown words into known WordPieces (e.g., اتفاقنا is split into اتفاق ##ن ##ا) unlike AraVec that generates different word embeddings for each of these words.

Taking examples of case 2, the same word بحرية has two different meanings in these sentences. The first of which is naval (sentence 1), and it was widely used as an adjective of the first aspect term (multi-word aspect terms, e.g., وحدة بحرية زودن اموك/Marine commando unit) in the training set. The second meaning is freely (sentence 2) where the word بحرية is segmented into ب/with and حرية/freedom. This lexical ambiguity presents one of the main challenges of Arabic, especially that diacritics are omitted in the majority of existing Arabic text (particularly in MSA). However, our model correctly labeled this word in sentence 2, unlike the AraVec-BiLSTM-CRF model that wrongly tagged this word as an aspect term. This can be justified by the fact that, unlike AraVec, the BERT model provides different vectors for the same word depending on the context in which it occurs.

The third case shows successive words that are tagged as independent aspect terms in the test set. BERT correctly predicted the labels of these words, while AraVec-BiLSTM-CRF wrongly tagged these words as multi-word aspect terms. This shows the effectiveness of the semantic and context information captured by the Arabic BERT model.

7.4 Error analysis

We conducted an error analysis to analyze the limitations of the proposed models and identify the possible improvements for future work. Tables 5 and 6 illustrate examples of errors incurred by our best models (BERT + BiGRU + CRF and P-BERT) for the ATE and ACD tasks, respectively.

For the ATE task, the main error is related to the long and complex structure of some aspect terms. The proposed model failed to detect the entire aspect term. For example, it wrongly tagged some words as new aspect terms, e.g., حرس/Guards in القوات التابعة لحرس الرئيسي/The forces of the Presidential Guards. This can be due to the fact that this type of term is specific to the war and security domains, which might not be detected by the AraBERT model that was pre-trained on general corpora. Thus, a future improvement includes further pre-training the AraBERT model on a related domain dataset to examine its effect on the proposed model's performance. Examples of misclassified aspect terms are illustrated in Table 6.

For the ACD task, the first error is due to wrongly annotated sentences. For example, we believe that the two first sentences in Table 7 discuss *Results*, but annotators assigned them the *Plans* category. Another error is related to the confusion between two categories: *Parties* and *Results*. Sentences 3 and 4 both talk about *Results* (يرفضون التعليق للجزيرة على مجازر اسرائيل في غزة / T h e y

Table 7 Example of errors for the ACD task

Sentence	Predicted label	Test set annotated label	Error type
Ar: كتائب القسام تستهدف دبابة صهيونية في موقع الارسال شمال بيت لاهيا بصاروخ كورنيت وتدمرها تدميرا كاملا	Results	Plans	Error in human annotation
En: Al-Qassam Brigades have targeted a Zionist tank at the Irsal site north of Beit Lahia with a Kornet missile and have completely destroyed it			
Ar: روتر العبري: انفجاران كبيران هزا منطقة رحوفوت وانباء عن سقوط عدة صواريخ	Results	Plans	
En: Reuters: Two large explosions have rocked the Rehovot area, and the landing of several rockets has been reported			
Ar: العربية ،،، السلطات المصرية تفتح معبر رفح غدا لاستقبال الجرحى	Parties	Results	Confusion between two categories: parties and results
En: Al-Arabiya: The Egyptian authorities will open the Rafah crossing tomorrow to receive the wounded victims			
Ar: الجزيرة ،،، مسؤولون كبار في الجامعة العربية يرفضون التعليق للجزيرة على مجازر اسرائيل في غزة	Parties	Results	
En: Al-Jazeera: Senior officials in the Arab League have refused to comment to Al-Jazeera on Israel's massacres in Gaza			

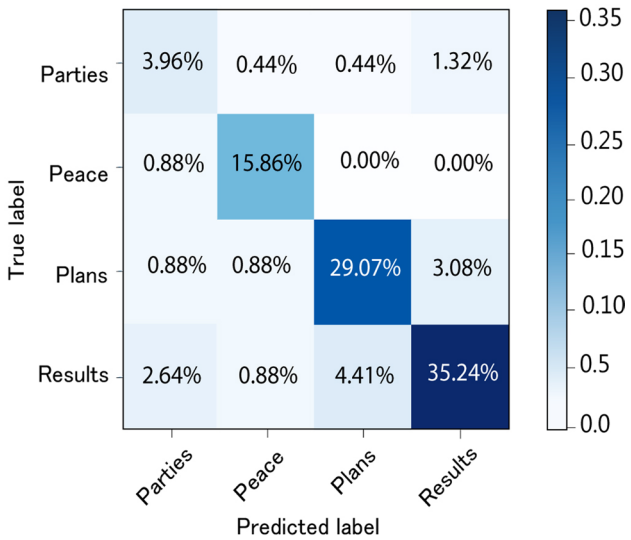

Fig. 6 The confusion matrix of the P-BERT model

refuse to comment to Al Jazeera on Israel's massacres in Gaza; تفتح معبر رفح غدا/will open the Rafah crossing tomorrow) but in relation with *Parties* (مسؤولون كبار في الجامعة العربية/Senior officials in the Arab League; السلطات المصرية/The Egyptian authorities) and they were both assigned to *Results*. However, other similar sentences in the test set were assigned to *Parties* (e.g., وفدا تضامنيا يضم مصريين وماليزيين واطباء من دخول غزة / Al-Jazeera,,, السلطات المصرية برفح تمنع الجزيرة ،،، The Egyptian authorities in Rafah prevent a solidarity delegation that includes Egyptians, Malaysians, and doctors from entering Gaza; لا يتوجب على اي دولة ان تقبل باطلاق صواريخ تجاه مدنييها بشكل عشوائي الرئيس الامريكي اوباما يدافع عن العملية الاسرائيلية ،،،/US President Obama defends the Israeli operation,,, No country should accept firing rockets at its civilians indiscriminately). Hence, the proposed model failed in detecting the correct label between these two categories in many cases. Future improvements include fixing these issues in the annotated dataset to enhance the achieved results. Another issue is related to the class imbalance problem. Table 1 shows that the minority label in the used dataset is *Parties*, which was the least detected category given the confusion matrix illustrated in Fig. 6. Future work includes investigating the effect of data augmentation methods on the proposed model's performance.

7.5 Theoretical and practical implications

For theoretical implications, this paper makes certain contributions to the existing field of Arabic ABSA. As we can conclude from the literature review section, most of the existing AABSA studies rely on laborious preprocessing and feature extraction tasks. Additionally, the deep learning methods were implemented based on neural network models. Our approach, instead, investigates the use of transfer learning based on pre-trained language models to perform ABSA tasks. Additionally, the AraBERT model shows its effectiveness in handling the ambiguity and the morphological complexity in Arabic without the need for manually engineered features or external resources, which can be very promising for other morphologically rich languages. Moreover, our findings can also have an important implication for other researchers by providing a new perspective on the ABSA tasks in low-resource languages that suffer from data scarcity and lack resources such as lexicons and NLP tools.

On the other hand, the online shared reviews and thoughts are precious for many entities. Analyzing these data can help companies evaluate their product or services and improve them to satisfy their clients. Governments can get closer to their citizens by analyzing their opinions and thoughts faced to public issues. Customers can search for other users' opinions before making some purchase decisions. The reviewers generally discuss an entity (product, service, etc.) in a fine-grained way, making the aspect-level more suitable for real-life applications. The first and essential step in implementing any ABSA system is to identify the discussed aspect terms and aspect categories in each review, which are the tasks we handled in this work. The next step is to identify the sentiment polarity of each aspect term or category, which will be targeted in future work.

8 Conclusion and future work

In this paper, we investigate the effectiveness of transfer learning based on a pre-trained language model on two ABSA tasks, namely aspect term extraction (ATE) and aspect category detection (ACD), on a reference Arabic news dataset. Experimental results demonstrated that a simple fine-tuned BERT model outperformed the baselines and related work methods that require a laborious task of preprocessing and feature engineering. Moreover, fine-tuning the BERT model has shown to achieve better results on a relatively small labeled dataset compared to well-known neural networks using free-context word embeddings models. For the ATE task, the default fine-tuned BERT model results were enhanced by incorporating more powerful layers into the BERT model (CRF, BiLSTM + CRF, and BiGRU + CRF). The best result was achieved using the BERT + BiGRU + CRF model. For the ACD task, the BERT fine-tuning approach achieved better results than the BERT feature-based method. Furthermore, the default fine-tuned BERT model's performance was boosted by exploiting the aspect terms information in detecting the aspect categories.

Future work includes evaluating this approach on the other ABSA tasks, namely aspect term polarity and aspect category polarity. Additionally, these methods' effectiveness can be evaluated on dialectical Arabic reviews and other domains such as hotels and restaurants. Moreover, we intend to examine the impact of designing more complex downstream models in enhancing the achieved results. Additionally, we plan to augment the size of the used dataset by collecting and annotating new posts to evaluate the effect of using larger data on the proposed models.

Funding The authors did not receive support from any organization for the submitted work.

Declaration

Conflict of interest The authors declare that they have no conflict of interests.

References

Abas AR, El-Henawy I, Mohamed H, Abdellatif A (2020) Deep learning model for fine-grained aspect-based opinion mining. IEEE Access 8:128845–128855

Abdul-Mageed M, Diab M, Korayem M (2011) Subjectivity and sentiment analysis of modern standard Arabic. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies

Abdul-Mageed M, Elmadany A, Nagoudi EMB (2020) ARBERT and MARBERT: deep bidirectional transformers for Arabic. arXiv preprint arXiv:2101.01785

Abuzayed A, Al-Khalifa H (2021) Sarcasm and sentiment detection In Arabic tweets using BERT-based models and data augmentation. In: Proceedings of the sixth Arabic natural language processing workshop

Al-Ayyoub M, Al-Sarhan H, Al-So'ud M, Al-Smadi M, Jararweh Y (2017) Framework for affective news analysis of Arabic news: 2014 Gaza attacks case study. J UCS 23(3):327–352

Al-Dabet S, Tedmori S, Al-Smadi M (2020) Extracting opinion targets using attention-based neural model. SN Comput Sci 1:10

Al-Smadi M, Qawasmeh O, Talafha B, Quwaider M (2015) Human annotated Arabic dataset of book reviews for aspect based sentiment analysis. In: 2015 3rd International conference on future internet of things and cloud. IEEE

Al-Smadi M, Qawasmeh O, Talafha B, Al-Ayyoub M, Jararweh Y, Benkhelifa E (2016) An enhanced framework for aspect-based sentiment analysis of Hotels' reviews: Arabic reviews case study. In: 2016 11th International conference for internet technology and secured transactions (ICITST), pp 98–103

Al-Smadi M, Al-Ayyoub M, Jararweh Y, Qawasmeh O (2019a) Enhancing aspect-based sentiment analysis of Arabic hotels'

reviews using morphological, syntactic and semantic features. Inf Process Manag 56(2):308–319
Al-Smadi M, Talafha B, Al-Ayyoub M, Jararweh Y (2019b) Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. Int J Mach Learn Cybern 10(8):2163–2175
Al-Twairesh N, Al-Khalifa H, Al-Salman A (2014) Subjectivity and sentiment analysis of Arabic: trends and challenges. In: 2014 IEEE/ACS 11th International conference on computer systems and applications (AICCSA). IEEE
Areed S, Alqaryouti O, Siyam B, Shaalan K (2020) Aspect-based sentiment analysis for Arabic Government reviews. In: Abd-Elaziz M et al (eds) Recent advances in NLP: the case of Arabic language. Springer International Publishing, Cham, pp 143–162
Baly F, Hajj H (2020) AraBERT: transformer-based model for Arabic language understanding. In: Proceedings of the 4th workshop on open-source Arabic Corpora and processing tools, with a shared task on offensive language detection
Boudjellal N, Zhang H, Khan A, Ahmad A, Naseem R, Shang J, Dai L (2021) ABioNER: a BERT-based model for Arabic biomedical named-entity recognition. Complexity 2021:1–6
Chowdhury SA, Abdelali A, Darwish K, Soon-Gyo J, Salminen J, Jansen BJ (2020) Improving Arabic text categorization using transformer training diversification. In: Proceedings of the fifth Arabic natural language processing workshop
Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555
Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers). Association for Computational Linguistics
Duwairi R, El-Orfali M (2014) A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. J Inf Sci 40(4):501–513
ElJundi O, Antoun W, El Droubi N, Hajj H, El-Hajj W, Shaban K (2019) hulmona: the universal language model in Arabic. In: Proceedings of the fourth Arabic natural language processing workshop
Farha IA, Zaghouani W, Magdy W (2021) Overview of the wanlp 2021 shared task on sarcasm and sentiment detection in Arabic. In: Proceedings of the sixth Arabic natural language processing workshop
Goldberg Y (2016) A primer on neural network models for natural language processing. J Artif Intell Res 57:345–420
Guellil I, Saâdane H, Azouaou F, Gueni B, Nouvel D (2019) Arabic natural language processing: an overview. J King Saud Univ Comput Inf Sci 33:497–507
Guo B, Zhang C, Liu J, Ma X (2019) Improving text classification with weighted word embeddings via a multi-channel TextCNN model. Neurocomputing 363:366–374
Hoang M, Bihorac OA, Rouces J (2019) Aspect-based sentiment analysis using bert. In: Proceedings of the 22nd nordic conference on computational linguistics
Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. In: Proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 1: long papers)
Jozefowicz R, Zaremba W, Sutskever I (2015) An empirical exploration of recurrent network architectures. In: International conference on machine learning. PMLR
Kim Y (2014) Convolutional neural networks for sentence classification. In: Proceedings of the 2014 conference on empirical methods in natural language processing
Kudo T (2018) Subword regularization: improving neural network translation models with multiple subword candidates. In: Proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 1: long papers)
Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C (2016) Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: human language technologies
Li X, Bing L, Zhang W, Lam W (2019) Exploiting BERT for end-to-end aspect-based sentiment analysis. In: Proceedings of the 5th workshop on noisy user-generated text (W-NUT 2019)
Meškelė D, Frasincar F (2020) ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model. Inf Process Manag 57(3):102211
Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 26:3111–3119
Mohammad A-S, Al-Ayyoub M, Al-Sarhan H, Jararweh Y (2015) Using aspect-based sentiment analysis to evaluate Arabic news affect on readers. In: 2015 IEEE/ACM 8th International conference on utility and cloud computing (UCC). IEEE
Mohammad AS, Al-Ayyoub M, Al-Sarhan HN, Jararweh Y (2016) An aspect-based sentiment analysis approach to evaluating Arabic news affect on readers. J Univ Comput Sci 22(5):630–649
Oueslati O, Cambria E, HajHmida MB, Ounelli H (2020) A review of sentiment analysis research in Arabic language. Futur Gener Comput Syst 112:408–430
Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S (2014) SemEval-2014 Task 4: aspect based sentiment analysis. In: COLING 2014
Pontiki M, Galanis D, Papageorgiou H, Manandhar S, Androutsopoulos I (2015) Semeval-2015 task 12: aspect based sentiment analysis. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)
Pontiki M, Galanis D, Papageorgiou H, Androutsopoulos I, Manandhar S, Al-Smadi M, Al-Ayyoub M, Zhao Y, Qin B, De Clercq O (2016) Semeval-2016 task 5: aspect based sentiment analysis. In: International workshop on semantic evaluation
Pratt LY (1992) Discriminability-based transfer between neural networks. Adv Neural Inf Process Syst 5:204–211
Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9
Refaee E, Rieser V (2014) An Arabic twitter corpus for subjectivity and sentiment analysis. In: LREC
Rietzler A, Stabinger S, Opitz P, Engl S (2020) Adapt or get left behind: domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In: Proceedings of the 12th language resources and evaluation conference
Soliman AB, Eissa K, El-Beltagy SR (2017) AraVec: a set of Arabic word embedding models for use in Arabic NLP. Procedia Comput Sci 117:256–265
Steingrímsson S, Kárason Ö, Loftsson H (2019) Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In: Proceedings of the international conference on recent advances in natural language processing (RANLP 2019)
Sun C, Huang L, Qiu X (2019) Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)
Xu H, Liu B, Shu L, Philip SY (2018) Double embeddings and CNN-based sequence labeling for aspect extraction. In: Proceedings of

the 56th annual meeting of the Association for Computational Linguistics (volume 2: short papers)

Xue W, Zhou W, Li T, Wang Q (2017) MTNA: a neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews. In: Proceedings of the eighth international joint conference on natural language processing (volume 2: short papers)

Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) XLNet: generalized autoregressive pretraining for language understanding. Adv Neural Inf Process Syst 32:5753–5763

Yu D, Wang S, Deng L (2010) Sequential labeling using deep-structured conditional random fields. IEEE J Sel Top Signal Process 4(6):965–973