IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 18, 2022, accepted March 4, 2022, date of publication March 14, 2022, date of current version March 23, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3159252

Arabic Aspect Extraction Based on Stacked Contextualized Embedding With Deep Learning

ARWA SAIF FADEL[1,2], **MOSTAFA ELSAYED SALEH**[1], **AND OSAMA AHMED ABULNAJA**[1]
[1]Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University (KAU), Jeddah 21589, Saudi Arabia
[2]Computer Science Department, Faculty of Computer Sciences and Engineering, Hodeidah University, Hodeidah, Yemen

Corresponding author: Arwa Saif Fadel (arwa.cs.2006@gmail.com)

ABSTRACT The exponential growth of the internet and a multi-fold increase in social media users in the last decade have resulted in a massive growth of unstructured data. Aspect-Based Sentiment Analysis (ABSA) is challenging because it performs a fine-grain analysis; it is a text analysis technique where the opinions group is based on the aspect. The Aspect Extraction (AE) task is one of the core subtasks of ABSA; it helps to identify aspect terms in the text, comments, or reviews. The challenge of the Arabic AE task increases due to the complexity of the Arabic language. This work aims to develop the Arabic AE task by proposing transfer learning using state-of-art pre-trained contextual language models. We concatenate the Bidirectional Encoder Representation from Transformers (BERT) language model and contextualize string embeddings (Flair embedding) as a stacked embeddings layer for better word representation for Arabic language. Then, we extend it with different deep learning network architectures. For Arabic AE, the model is developed by concatenating the Arabic contextual language model, AraBERT, and Flair embedding as a contextual stacked embeddings layer with an extended layer, BiLSTM-CRF or BiGRU-CRF, for sequence labeling. Our proposed models are called BF-BiLSTM-CRF and BF-BiGRU-CRF. The proposed model is evaluated using the Arabic Hotel's reviews dataset. For performance evaluation, we used the F1 score. The experimental results show that the proposed BF-BiLSTM-CRF configuration outperformed the baseline and other models by achieving an F1score of 79.7%.

INDEX TERMS Arabic aspect extraction, AraBERT, BERT, flair embedding, BiLSTM, BiGRU, CRF.

I. INTRODUCTION

Sentiment Analysis (SA), often known as opinion mining, is a popular study topic in Natural Language Processing (NLP). In recent years, it has attracted the research community's attention due to the explosion of social media data and the abundance of online reviews on services and products [1]–[3]. However, it is problematic to identify the opinion or sentiment expressed about services and products in text.

SA is classified into document, sentence, and aspect levels. The document and sentence levels are considered coarse-grained analysis, where the opinion is about the whole given text (document or sentence), and is not sufficient, in many cases, to indicate an opinion about the specific aspects given in the text [4]. On the other hand, aspect level analysis, or ABSA, is a fine-grained analysis where the sentiment is predicted based on the entities (products, services, organizations, or events) in a particular domain that SA cannot cover [5]. ABSA helps, for example, organizations, governments, and decision-makers to know which features are considered attractive for customers and which are not favored by users to avoid or enhance them in the future. For example, a sentence about a hotel, "The buffet was very delicious, but it is a little price," gives positive and negative remarks about a hotel on two different aspects, "food" and "price," respectively. Another example is a review in Arabic about a new mobile phone "وعمر البطارية قصيرسعر الجهاز معقول، ولكن عيبه في الكاميرا الأمامية," which means "The price of the device is reasonable, but its disadvantage is the front camera and the battery life is short." The main aspects in this review are price, battery, and camera; the cost of the mobile phone has a positive polarity while negative polarities are indicated about the

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

battery and camera, which means a negative polarity about its properties.

ABSA comprises four main subtasks: aspect term extraction, aspect term sentiment analysis, aspect category classification, and aspect–category sentiment analysis [6]. Every task can be performed independently or combined with others. Aspect term extraction constitutes the basis of ABSA and governs the accuracy of ABSA results [5], [7], [8]. We concentrate on the Aspect Extraction (AE) task for the Arabic language [6].

An AE task is treated as a sequence labeling problem, similar to Named Entity Recognition (NER). Well-known methods to extract aspects are the Conditional Random Field (CRF) and the Hidden Markova Model (HMM)) [9], [10], which rely heavily on handcrafted features, for instance, bi-gram and part of speech. Aspect category classification is solved by a support vector machine, logistic regression with several feature representations, and extraction methods such as frequency features, n-gram, a bag of words, term frequency-inverse document frequency (tf-idf), and word embedding [11]. The accuracy of applying these techniques to AE depends on quality handcrafted features, and it needs feature engineering at a high level, which is time-consuming. With the increase in the number of datasets and developed computation resources, the researchers of this paper utilized a deep neural network for ABSA tasks to give rich feature representations. Deep learning consists of cascaded layers with nonlinear processing units for feature extractions. Applying deep learning in NLP problems shows excellent potential, and it can train complex models on big datasets [12]. Deep learning reports more accuracy through several algorithms and techniques, or a hybrid of them, such as [13], Convolutional Neural Networks (CNNs) [14], Recurrent Neural Networks (RNNs), and RNN extensions, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) [15]. In addition, using Attention mechanisms for AE improves the performance of ABSA [16]. Arabic is the fifth most widely spoken language in the world. More than 422 million people speak it, and it is the official language of 22 countries. Arabic has some unique characteristics, such as containing 28 letters, its orientation from right to left, and lack of capitalization. Unlike other languages such as English, Arabic is a Semitic language; it has rich morphology and words that could have different meanings within a given context. Moreover, diacritics in Arabic serve the same purpose as short vowels in English; they determine how letters are pronounced. Arabic is a derivative language where All Arabic words are derived from a root composed of constants. These are usually three or four letters [17].

The Arabic language has a complex inflectional and derivative morphology, leading to the rise of sentiment analysis and Arabic ABSA (AABSA) challenges. Some of these challenges are [17], [18]:

With the lack of reliable NLP resources that deal with DA, there is a lack of opinion written in classic and MSA Arabic forms. The users typically use dialects forms to express opinions on social media, for example, Twitter, forums, and online websites.

Additionally, large AABSA-annotated corpora are not available for learning accurate models and no annotated datasets for social media and different dialects.

Moreover, many challenges occur during the preprocessing phase [18], such as:

- A given term can take on several meanings depending on the context, such as (ذهب), which may mean “gold” or “went.”
- Capital letters are a problem in AE.
- Transliteration and misspelling in microblogs such as Twitter and hashtags result in noisy and dirty datasets.
- With many word forms and removing diacritics, the sentence holds different meanings; e.g., (علم) can be “Flag,” “taught,” “understood,” “science,” and “knew.”

In comparison to a large number of studies on ABSA in English, there are few works on AABSA for the following reasons: (1) the difficulty and complexity of the Arabic language; (2) the lack of reliable NLP tools for the Arabic language; and (3) less availability of large and annotated datasets.

Training deep learning models is data-intensive. Thus, deep-learning-based NLP tasks need more data for training available in limited resource languages such as Arabic; this problem could be solved by transfer learning [19]. Transfer learning enables transferring the knowledge from pre-trained models such as XLNeT [20], ELMO [21], Flair [22], and Bidirectional Encoder Representations from Transformers (BERT) to other limited domains [23]—for example, product reviews and healthcare. In other words, it allows using the pre-trained models on a vast text corpus and using them in other downstream tasks with small datasets [24].

Moreover, deep learning methods such as LSTM cannot parallelize and capture semantics over longer sequences than transformers. In addition, deep learning methods based on word embedding vectors such as Wor2vec [11] and Glove [25] cannot consider the polysemies of the word in different contexts where one vector can be generated for each word in the vocabulary regardless its surrounding context.

Recently, transfer learning has been shown a significant advantage in the semantic representation of text through a pre-trained contextual model. BERT and Flair have a strong semantic text representation; they achieve state-of-the-art results in several NLP tasks. BERT and Flair can deal with polysemy, Out-Of-Vocabulary OOV and misspelling [23], [26].

In our work for better word representation and to develop Arabic AE, we find that the concatenation BERT with contextual string embedding Flair (as stacked embedding layer) strengthens word representation, enriches word representations with additional semantic and syntactic information, and their use is the state-of-art in sequence labeling. Thus, it can improve the performant Arabic AE task.

The contribution of our work can be summarized as follows:

- We propose using transfer learning based on stacked contextualized embedding for Arabic AE. To our knowledge, this is the first work using a transfer learning based on a combination of fine-tuned AraBERT and contextual Character-level embedding (Flair embedding) for better word representation to solve the AE task on an Arabic dataset.
- We propose combining LSTM (BiLSTM) or Bidirectional GRU (BiGRU) and CRF on top of the contextual embedding layer for sequence labeling.
- We use the Arabic Hotel's reviews dataset to train and evaluate our proposed model. Extensive experimental results demonstrate that our proposed fine-tuning BERT BF-BiLSTM-CRF model outperforms baseline and state-of-the-art works on the Arabic AE task.

The rest of this paper is organized as follows. Section II discusses the related work. The proposed models are provided in Section III. The details of the experiments and the evaluation results are presented in Section IV. Finally, the conclusion of this work is presented in Section V.

II. RELATED WORK

A. ENGLISH AND OTHER LANGUAGES

There are three categories of AE Approaches: early methodologies, rule-based or based on a dictionary; traditional approaches based on machine learning; and modern approaches rely on deep learning and transformers [5]. Additionally, the approaches can be classified into supervised methods and unsupervised methods. Unsupervised techniques to extract aspects are rule-based, frequency-based, and statistical methods [27], [28]. Rule-based approaches are based on predefined rules to extract aspect terms manually [29], [30] or automatically [31], [32]. With limited grammatical information, frequency-based algorithms extract aspects based on the more frequent features, such as nouns or noun phrases [33].

On the other hand, the supervised approach considers the AE task a sequence-labeling problem. Several machine learning-based approaches have been performed on AE such as Support Vector Machine (SVM) [8]. Also, topic modeling is widely used for AE and aspect grouping [34]. Many hybrid approaches combine more than one of these methods. The limitations of machine learning methods lie in handcrafted features that usually need many experts in the domains and human labor.

In recent years, researchers have adopted deep learning rather than traditional techniques for AE; deep learning models enhance performance by automatically learning the semantic and syntactic features. One work used a CNN for aspect extraction [14], which had seven layers, one input layer, two conventional layers, two max-pooling layers, and fully connected layers with Softmax output. For features, the researchers combined word embedding and parts of speech. The results showed a significant performance improvement. In [7], the authors proposed two embedding layers combined with a CNN network, where the two embedding layers were a general embedding layer and an in-domain layer. For sequence labeling, they used four CNN layers. Their results showed that the two embedding mechanisms with CNN enhanced the performance of the model.

Tang *et al.* [35] extended LSTM into Target-Dependent LSTM (TD-LSTM) and Target-Connection LSTM (TC-LSTM) to consider aspect target. In TD-LSM, they used two LSTM layers—LSTML to represent the left context from the aspect and LSTMR to represent the right context of the aspect plus the aspect for learning. After that, they concatenated the last hidden vector of LSTML and LSTMR and pass them to the last layer to predict sentiment polarity related to the aspect. To capture the interaction between aspect word and its context. Some works combined RNN with CRF methods to improve the AE task to identify the aspect boundary [5], [36]. In addition, several studies used an attention mechanism to assist the model learn representation more effectively, highlighting aspect-related words while de-emphasizing irrelative aspect words [16], [37]–[41].

Most deep learning models are based on word embedding. Word embedding is a neural method for text representation that translates text into vectors (vector representation for a particular word). It is used as input for deep learning. The vector space model represents words in a continuous vector, where the semantically and syntactically similar words are mapped to nearby points and embedded near each other. Wor2vec [11] and Glove [25] are the most common examples of word embedding. The limitation of word embedding is context-free, and it cannot consider polysemy words, limiting the performance of models that rely on word embedding.

Recently, pre-trained models, such as ELMO [21], XLNeT [20], and BERT [23] accomplished state-of-the-art results in NLP tasks because they adjusted the word vector based on the context. Utilizing BERT in ABSA, Gao *et al.* [42] converted the ABSA problem into a sentence-pair classification task. At the same time, TD-BERT concentrated on using the positioned output at the target word as the classification input rather than the first (CLS) tag; the features from these two base models were concatenated. Li *et al.* [43] also proposed the BERT-CRF model for End-to-End ABSA. In [44], they presented the FAGOM model for aspect level opinion mining; they used the BERT model for context embedding and a multi-head attention mechanism. Finally, a pooling layer is added to extract local and global features.

MA *et al.* [45] proposed combining LDA with lexicon for aspect extraction from Chinese reviews for the Chinese language. Yu *et al.* [46] proposed Fine-tuned BERT to extract implicit aspect terms from online clothing reviews. ABSA was studied on some low-resource Language such as Urdu [47]; the authors annotated data set for ABSA task for Roman Urdu and validated it by using several machine learning models.

To investigate the effect of transfer learning in a low resource language, Winatmoko *et al.* [48] used the multilingual version of BERT with the auxiliary label and CRF as an output layer to extract the aspect term and opining tasks on hotel reviews in Bahasa Indonesian. The results showed improvement in the F1 score. Lopes *et al.* [49] also used BERT for an AE task on a Portuguese dataset.

B. ARABIC LANGUAGE

In this paper, our primary focus is on AABSA research. Several works on Arabic SA and some earlier works used traditional methods for SA [50]–[52]. Lately, most works are based on modern techniques such as deep learning and transformers [53]–[55]. Compared to English, there was a limited number of research works targeting AABSA. There was no work on AABSA data before 2015. The first presented research for AABSA was in 2015.

The first Arabic dataset supporting ABSA is HAAD (Human Annotated Arabic Dataset of Book Review); it contains Arabic books reviews [56]. The following work used the HAAD dataset to enhance ABSA tasks aspect category extraction and aspect polarity classification [57]. Recently, In [58], the author extracted the explicit aspects from HAAD by using description logic to describe terminological knowledge. Then, they combine linguistic Rule and Description Logic to extract opinion targets. Areed *et al.* [59] present dataset for government reviews, the combined rule-based and lexicon models for AE.

In [60], ABSA was used to study and analyze the effect of Arabic news on readers. Another dataset for Arabic laptop reviews was prepared to support ABSA. Concerning using deep learning in AABSA, few studies are found in the literature. The first one is INSIGHT-1; the authors used CNN for aspect category and sentiment polarity detection for multilingual ABSA (11 languages) [61]. For Arabic, they used the review Hotel' dataset, and the results showed an enhancement in performance over the baseline result by 11% for aspect category detection and 6% for sentiment polarity classification. Another research used the LSTM approach for eight languages, including Arabic [62]. For Arabic, their results did not perform well compared to the baseline result; they achieved (F1 = 47.3%), with an enhancement of around 7%. Al-Smadi *et al.* [63] used the Arabic Hotel's reviews dataset for SemEval-2016 Task 5. The authors compared RNN with SVM for ABSA tasks related to the dataset: aspect category identification, aspect opening target expression extracting, and aspect polarity classification. They extracted lexical, syntactic, semantic, and morphological features for training SVM classifiers. Then, they compared a trained RNN with SVM, but it had a long execution time in the training and testing phases.

Two approaches were developed to handle the AABSA in [64]. The first one used Bi-LSTM and CRF on the word and character level to extract aspects from the review. For the sentiment polarity classification, LSTM was used. The result showed enhancement over baseline research on both tasks (39% for task 1 and 6% for task 2). In [65], the authors used an attention mechanism for AE, and the performance improved with a 72.8 F-score. In [66], a BiGRU was used. For AE, the result was close to [65].

A few works used fine-tuning BERT with linear classification for Arabic aspect polarity classification [67]. Bensoltan *et al.* [68], proposed Bert-BiLSTM-CRF model for AE from the News dataset that outperformed the previous works on this dataset.

To the best of our knowledge, no prior work used a combination of BERT and string embedding for Arabic Aspect Extraction.

III. PROPOSED MODEL ARCHITECTURE

This section introduces our proposed models; we propose two models for the Arabic AE task: the BF-BiLSTM-CRF model and the BF-BiGRU-CRF model. The proposed models integrate two contextual pre-trained models, namely BERT language model and contextual string embedding (Flair) as stacked embeddings. The architectures of the models consist of three main layers: (1) the combination of the Arabic BERT (AraBERT) and Flair embeddings is used as the input layer, (2) BiLSTM and BiGRU represent the encoder layer, and (3) CRF serves as the decoding layer.

As shown in Figure 1, the proposed model consists of three main layers: embedding layer, BiLSTM or BiGRU layer, and CRF layer. For the embedding layer or input representation, we use Arabic BERT (AraBERTv02) and Flair embeddings, in which every word of the sentence is mapped to a contextual vector of concatenated embedding. BiLSTM/BiGRU-CRF is used for sequence labeling. BiLSTM/BiGRU encodes the contextual information for each word in the input sequence; it is used for semantic encoding and obtaining global sequence features. The embedding vectors from the embedding layer are used as input to BiLSTM or BiGRU; they generate scores representing the probability for the tags; for example, the highest score is selected as the final prediction for a particular token. However, some predicted labels are invalid, so the output from BiLSTM is passed to the CRF Layer for correction.

The CRF is a decoding layer that predicts the final sequence labels, considers the dependency relationship between adjacent tags, and selects the best sequence tagging.

A. EMBEDDING LAYER

The embedding layer receives a sequence of N words (w_1, $w_2, \ldots, w_N$) and generates an embedding vector for each word ($e_1, e_2, \ldots, e_N$). The final embedding vector can have stacked embeddings, where a Flair embedding concatenated with BERT forms the final embedding vector for a word w in position i, and it is given by:

$$e_i = \begin{bmatrix} e_i^{Flair} \\ e_i^{BERT} \end{bmatrix} \tag{1}$$

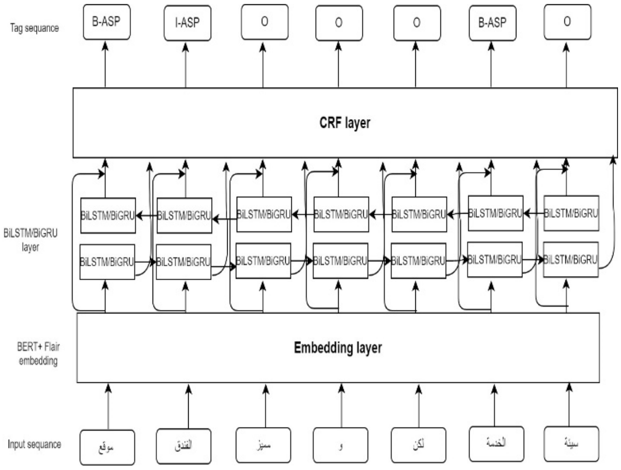

FIGURE 1. **The proposed model architecture.**

Here, e_i^{Flair} and e_i^{BERT} represent the contextual Flair and BERT embedding vectors, respectively. The concatenation combines the best of these pre-trained models to enhance the performance; it is beneficial to enrich word representations with additional syntactic and semantic information.

Moreover, the performance of sequence labeling tasks, such as NER, is enhanced by using BiLSTM-CRF architecture on the top embedding layer. As we use a combination of pre-trained embedding, we can assume that the performance of the Arabic AE task will be improved.

1) BERT

BERT uses a transformer encoder in a bidirectional way to encode a word, representing the semantic of the word in the context based on its semantic relationship with the other words [23]. Thus, the output is a contextual embedding vector for each word. It is implemented using a multi-layer transformer encoder, proposed and developed by [69]. The transformer's core is a self-attention mechanism that obtains the relationship between words in the context by calculating the attention between each word in the input sequence. As shown in (2), the attention is calculated using three input word vector matrixes: Query Vector Q, Value Vector V, and Key Vector (K). d_k is the input vector dimension, while QK^t is used to calculate the relationship between input words. The weights representations are obtained by softmax normalization, and the final output is the sum of the weights of all input vectors. In addition, the transformer uses multi-head attention, which computes the attention h at different times from different points of view with different weight matrices, and concatenates them together.

$$Attention\,(Q, K, V) = softmax\left(\frac{QK^t}{\sqrt{d_k}}\right)V \tag{2}$$

The BERT input is composed of three types of embeddings. The first is token embedding to encode the word by adding special tokens to the input [CLS] and [SEP] at the beginning and end of each sentence, respectively. Second, segment embedding is used to encode the sentence position. Third, position embedding encodes the word's position in the sequence. BERT is pre-trained on a massive corpus, enabling it to obtain a much better feature representation.

BERT is pre-trained on two tasks to understand the language: A Masked Language Model (MLM) and Next Sentence Prediction (NSP). In MLM, 15% of the input words are masked randomly to predicted by BERT based on the context information. In NSP, BERT predicts whether two sentences occurred consecutively in a given text or not by predicting if the second sentence continues the previous sentence.

In this paper, we apply the AraBERTv02 [70] pre-trained on the Arabic Wikipedia and a large Arabic news corpus containing 8.5 articles, 70 million sentences, and 2.5B tokens; the size of the dataset used is 24 GB. It covers different topics from different regions and covers a wide range of Arabic language in the Arabic world.

2) FLAIR EMBEDDING

Flair embedding is pre-trained contextual string embedding [22]. It is based on a character-level language model (CharLM); each letter of the words is sent into the character language model. The words and characters are a probability distribution of words, and characters are a probability distribution. Every new character or word depends on the character or words that come before it. The input representation is taken from the Backward and Forward language models. The final embedding from both hidden states is concatenated after the last character in the word.

B. BiLSTM LAYER

The Recurrent Neural Network (RNN) is an artificial neural network which processes sequential data. RNN has internal memory that enables it to process a data sequence by applying the same function and set of parameters to every sequence element (in each layer). Each output depends on all previous information; this enables inferring the following word or character in the sequence. Theoretically, RNNs can use the information for a long sequence, but practically, back-propagated gradient growth (becoming extremely high) or shrinkage (closing on zero) explodes gradients or causes gradient problems to vanish after many steps.

To solve RNN problems, extension networks were developed from RNNs, such as LSTM [15]. LSTM does not differ from RNN; however, it has different computations in a hidden state. It has memory cells and three nonlinear gates that control the information that should be kept (positive values), and the information should be forgotten. The forget gate controls the gradient passing through it. It allows for explicit memory deletions and updates; the input gate determines the vital information in the current state, and the output gate is used to determine the next hidden layer state. The LSTM network structure is shown in the following equations.

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{3}$$

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \quad (4)$$
$$\tilde{C}_t = tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \quad (5)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (6)$$
$$O_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_0\right) \quad (7)$$
$$h_t = o_t * tanh(C_t) \quad (8)$$

where f_t i_t, and $\mathrm{O_t}$ represent the forget gate, input gate, an output gate, respectively. $\mathrm{C_t}$ and $\mathrm{h_t}$ represent the cell state and hidden state at time t. *b* is the bias vector, and W is the weight matrix. σ is the sigmoid function, which is the hyperbolic tangent activation function. $\tilde{\mathrm{C}}_\mathrm{t}$ represents the candidate value created by the layer, it added to the output of the input gate to obtain a new state cell $\mathrm{C_t}$ at time t.

This work uses the Bi-LSTM as a layer in this architecture to capture both forward and backward long dependencies between sequence words. Two hidden units are represented starting from the first word to the last and separately in reverse order. The two hidden units are concatenated simultaneously as the final output.

C. BiGRU LAYER

GRU is a simple version of LSTM, where it controls the flow of information through two gates, reset and update gates, like LSTM, but without a memory unit. The update gate determines the information that should be passed from the previous state to the current state; the reset gate determines the information that should be forgotten of the prior time step. The hidden unit's calculations are shown in the following equations:

$$u_t = \sigma\left(W_u\left[h_{t-1}, x_t\right] + b_u\right) \quad (9)$$
$$r_t = \sigma\left(W_r\left[h_{t-1}, x_t\right] + b_r\right) \quad (10)$$
$$\tilde{h}_t = tanh\left(W * \left[h_{t-1} * r_t, x_t\right] + b\right) \quad (11)$$
$$h_t = \left(1 - u_t\right) * h_{t-1} + u_t * \tilde{h}_t \quad (12)$$

where $\mathrm{r_t}$, $\mathrm{u_{t_t}}$, and $\tilde{h}_t$ represent the reset gate, update gate, and hidden state, respectively. $\mathrm{x_t}$ is the input vector at time t, $\mathrm{h_t}$ is the hidden state and the output vector. W_u, W_r, and W represent the weights for the reset, cell, and update states, respectively. b_r, b, *and* b_u represent bias parameters for the rest, cell, and update states, respective BiGRU is a bidirectional variant of GRU. Like BiLSTM, the input sequence is read in a bidirectional way from left to right (from the first word to the last) by a forwarding layer and in reverse from right to left (starting from the last word to the first word) by a backward layer. The final hidden layers from the forward direction and the final hidden state from the backward direction are concatenated to produce the last hidden state.

D. CRF LAYER

In sequence-labeling tasks, such as aspect extraction, the strong dependency relationships between labels should be considered. BiLSTM or BiGRU can consider the long-term context information, but they cannot consider the tag dependency for output results. CRF can solve these problems [71]. The CRF layer is used with highly interdependent output labels. Instead of modeling labeling decisions independently, they are jointly modeled with a CRF layer, which aims to produce the optimal global sequence of labels given a sequence of input [72]. The main advantage of CRF is learning some restrictions of the output labels that follow the BIO labeling scheme to ensure the validity of the predicted sequence labels. These restrictions are learned automatically during the learning process. Some examples of these restrictions in the case of our AE task are:

The first prediction label can start with "B-ASP" or "O" but not "I-ASP."

The "O I-ASP" pattern is not valid because "I-ASP" should be preceded by "B-ASP."

For sequence input $X = (x_1, x_2, \ldots, x_N)$, X is the input to the model for training. is $Y = (y_1, y_2, \ldots, y_N)$ is tag sequence. CRF determines the final score of the prediction sequence label from two types of scores; emission scores are the probability of the output from the BiLSTM layer, where I and y_i are the indices of word and label, respectively. N, X, and K represent the size of the output matrix P, where N is the number of words and K is the number of tags. Additionally, the transition score, A, represents the transition matrix, and the transition probability from one tag to another is represented by $A_{yi,yi+1}$. The final total score is:

$$score\left(x, y\right) = \sum\nolimits_{i=1}^{n} A_{yi,yi+1} + \sum_{i=1}^{n} p_{i,yi} \quad (13)$$

The Softmax function, the overall possible tag sequences, is used to obtain the score of the probability of sequence y.

$$p = \frac{e^{s(X,y)}}{\sum_{\tilde{y} \in Y_x} e^{s(X,\tilde{y})}} \quad (14)$$

Then using the logarithm to maximize the correct tag sequence:

$$log\left(p\left(\frac{y}{X}\right)\right) = s\left(X, y\right) - log\sum_{\tilde{y} \in Y_x} e^{s(X,\tilde{y})} \quad (15)$$

Finally, the output sequence of the maximum score is given by:

$$y^* = aragmax_{\tilde{y} \in Y_x} s\left(X, \tilde{y}\right) \quad (16)$$

IV. EXPERIMENT AND RESULT

A. DATASET

We test the proposed models on the Arabic Hotel's reviews dataset. The dataset was part of the SemEval2016 competition task 5 for ABSA analysis [6]. Semeval2016 is a multilingual task that covers customer reviews for seven domains and eight different languages. It is considered the benchmark dataset for AABSA tasks.

The dataset consists of several sentences, and each sentence is divided into a list of tuples. The dataset was annotated on text level with 2029 reviews, 1839 pieces for training, and 425 testing. Moreover, it was annotated on the sentence level with 6029 sentences: 4082 for training and 1227 for testing. It consisted of 24,028 annotated tuples split into 19,226 for training and 4802 for testing. We adopted the BIO annotation strategy for labeling the dataset. The aspect term contains one word or phrase; there are three types of labels, in which B-ASP indicates the first word of the aspect term, I-ASP indicates the inside the aspect term (but not the first word) O indicates not aspect words. For example, the input sentence ‘‘لكن الخدمة سيئةموقع الفندق جيد’’ (which means ‘‘the hotel location is good but the service is bad’’) is annotated as follows:

O	B-ASP	O	O	I-ASP	B-ASP
سيئة	الخدمة	لكن	جيد	الفندق	موقع

B. EVALUATION

For the Arabic AE task, the F1-score was used for performance evaluation. F1 is a standard measure for sequence labeling problems; it combines precision and recall rates. Precision indicates the number of correct predicted aspect entities to all detected aspect entities, and recall indicates the number of correctly predicted aspect entities to the number of entities in the standard result. F1 is the harmonic average of Precision and recall. The calculation method for Precision, recall, and F1 are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

$$F1 = \frac{2 \cdot Precision \cdot Recal}{Precision + Recal} \tag{19}$$

C. EXPERIMENT SETTING

We used a Flair Framework to implement the proposed models in this experiment [26]. The whole experiment was run on Google Colaboratory with Tesla T4 GPU. During the training, the Hyperparameters are updated using Stochastic Gradient Descent algorithms (SGD). The model's hyperparameters are shown in Table 1. Both models were trained with one and two layers to assess whether they increase or decrease the quality of overall performance. The pre-trained AraBERTv0.2 was utilized as a feature extractor in two ways: a feature-based method where the weights are a frozen and fine-tuned-based method where all model parameters are fine-tuned, including BERT. Due to limited computational resources, the generic pre-trained AraBERT model (AraBERTv02-base) was utilized rather than the large version (AraBERT-large).

V. RESULT AND DISCUSSION

We compare the proposed models with baseline [6] and previous models that used traditional deep learning, RNN [63], BiLSTM-CRF with Word2vec/fastText as word embedding [64], and an attention-based neural model [65].

TABLE 1. Experimental hyperparameters setting.

Parameter	Values
Optimizer	SGD
Mini-Batch size	32
Max epochs	50
Learning rate	0.1
Anneal factor	0.5
Recurrent units	256
BiLSTM/BiGRU layers	1, 2

TABLE 2. Comparing proposed model based on fine-tuned BERT performance with the previous model.

Model	Precision	Recall	F1
Baseline (SVM) [6]			30.97
OTE Extraction Bi-LSTM-CRF (word2vec) [63]	-	-	66.32
Bi-LSTM-CRF (fastText) [64]	-	-	69.9
Attention-Based Neural Model [65]	-	-	72.8
BERT only	71.75	73.57	72.6
BERT-CRF	72.06	73.74	72.8
BERT-BiLSTM-CRF (1 layer)	72.30	0.7428	73.1
BERT-BiLSTM-CRF (2 layers)	71.29	78.98	74.9
BERT-BiGRU-CRF (1 layer)	72.05	75.08	73.5
BERT-BiGRU-CRF (2 layers)	70.41	77.66	73.8
BF-BiGRU-CRF	79.33	79.99	79.6
BF-BiLSTM-CRF	82.31	77.33	**79.7**

In addition, we conducted a series of experiments to verify the proposed BF-BiLSTM-CRF and BF-BiGRU-CRF models on the Arabic AE task. First, the fine-tuned BERT with a linear layer and the BERT-CRF model was run. Then, the BERT-BiLSTM-CRF model was run with one BiLSTM layer and then with stacked two layers. Then, the BERT-BiGRU-CRF model was run with the same configuration. In the end, in our proposed models BF-BiLSTM-CRF and BF-BiGRU-CRF, we used BF as an abbreviation for ‘‘BERT + Flair.’’ All the models were tested on the same training and testing datasets.

We used two training methods for BERT: a feature-based method with fixed parameters and a fine-tuning method training the whole model.

As shown from Table 2, BERT-BiLSTM-CRF and BERT-BiGRU-CRF outperformed BERT with a linear layer and BERT-CRF, reflecting the effectiveness of BiLSTM-CRF and Bi-GRU-CRF on top of the BERT layer for sequence labeling. That proves the positive effect of using CRF to consider the dependency between adjacent labels.

In addition, the results demonstrate that using two layers of BiLSTM/BiGRU enhances the model's performance in all cases. For that, the proposed models BF-BiLSTM-CRF and BF-BiGRU-CRF were tested with stacked BiLSTM and BiGRU layers (two layers).

As for the overall models' performance, our proposed models, the fine-tuned BF-BiLSTM-CRF, outperformed the

TABLE 3. The results with models that used BERT as feature extractor (feature-based).

Model	Precision	Recall	F1
BERT-CRF	75.30	63.24	68.7
BERT-BiLSTM-CRF	72.74	76.05	74.3
BERT-BiGRU-CRF	73.33	74.66	73.9
BF-BiLSTM-CRF	79.22	75.14	**77.1**

baseline, previous related works, and other models based on BERT by achieving a 79.7% F1-score. That demonstrates the effectiveness of using a combination of stacked contextual word embedding to improve the performance of the Arabic AE task. That proves that using stacked contextual embeddings (BERT and Flair) enhances the semantic representation of words and semantic relationships of the words in the text.

Moreover, we were used BERT as a feature-based model with different model configurations BERT-CRF and BERT-BiLSTM-CRF, BERT-BiGRU-CRF, and BF-BiLSTM-CRF. As shown in Table 2 and Table 3, the proposed model, BF-BiLSTM-CRF (fine-tuning BERT) outperformed the BF-BiLSTM-CRF model based on a feature-based method by 2.5% F1 score points, which shows the effectiveness of the fine-tuning method rather than feature-based methods.

VI. CONCLUSION AND FUTURE WORK

Compared to the English language, limited works target Arabic AE because the Arabic language has richer inflectional and derivative morphology and Lack of Available NLP tools and resources. In this paper, we integrated the BERT language model and contextualized string embedding (Flair) for better word representation to enhance Arabic AE. It extended with a variant of neural network architecture and CRF.

The model's performance was investigated using Arabic pre-trained AraBERTv02 and Flair embedding as stacked embedding layers for contextual representation, and then it was combined with BiLSTM/BiGRU and CRF. The proposed models are called BF-BiLSTM-CRF and BF- BiGRU-CRF.

We experimented with two kinds of BERT training methods: a fine-tuning-based method and a feature-based method. Our results showed that the integration of BERT and Flair and BiLSTM-CRF, in addition, improves the outcomes because it combines the advantage of pre-trained contextual embedding, the BiLSTM network, and the CRF model. In all evaluation experiments, fine-tuning BF-BiLSTM-CRF with two BiLSTM stacked layers can outperform all the other models and previous works on the AE task on the same dataset with a 79.7% F1 score.

For future work, we can use contextual embeddings to improve the results of other Arabic ABSA tasks, such as aspect polarity classification and aspect category detection. This area has many challenges that the research community must exploit in the future.

REFERENCES

[1] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowl.-Based Syst.*, vol. 89, pp. 14–46, Nov. 2015, doi: 10.1016/j.knosys.2015.06.015.

[2] K. Zhang, Y. Geng, J. Zhao, J. Liu, and W. Li, "Sentiment analysis of social media via multimodal feature fusion," *Symmetry*, vol. 12, no. 12, pp. 1–14, 2020, doi: 10.3390/sym12122010.

[3] K. Dashtipour, M. Gogate, A. Adeel, H. Larijani, and A. Hussain, "Sentiment analysis of Persian movie reviews using deep learning," *Entropy*, vol. 23, no. 5, pp. 1–16, 2021, doi: 10.3390/e23050596.

[4] K. Schouten and F. Frasincar, "Survey on aspect-level sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 813–830, Mar. 2015.

[5] Ł. Augustyniak, T. Kajdanowicz, and P. Kazienko, "Comprehensive analysis of aspect term extraction methods using various text embeddings," *Comput. Speech Lang.*, vol. 69, Sep. 2021, Art. no. 101217, doi: 10.1016/j.csl.2021.101217.

[6] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra, and G. Eryiğit, "SemEval-2016 task 5: Aspect based sentiment analysis," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 19–30, doi: 10.18653/v1/s16-1002.

[7] H. Xu, B. Liu, L. Shu, and P. S. Yu, "Double embeddings and CNN-based sequence labeling for aspect extraction," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 592–598.

[8] A. S. Manek, P. D. Shenoy, M. C. Mohan, and K. R. Venugopal, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier," *World Wide Web*, vol. 20, no. 2, pp. 135–154, Mar. 2017, doi: 10.1007/s11280-015-0381-x.

[9] N. Jakob and I. Gurevych, "Extracting opinion targets in a single- and cross-domain setting with conditional random fields," in *Proc. Conf. Empir. Methods Nat. Lang. Process. Conf. (EMNLP)*, Oct. 2010, pp. 1035–1045.

[10] W. Jin and H. H. Ho, "A novel lexicalized HMM-based learning framework for web opinion mining," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 1–8.

[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013, *arXiv:1310.4546*.

[12] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: A comprehensive review," *ACM Comput. Surv.*, vol. 54, no. 3 pp. 1–40, Apr. 2021.

[13] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Sci. China Technol. Sci.*, 2020.

[14] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowl.-Based Syst.*, vol. 108, pp. 42–49, Sep. 2016, doi: 10.1016/j.knosys.2016.06.009.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[16] K. Srividya and A. Mary Sowjanya, "NA-DLSTM—A neural attention based model for context aware aspect-based sentiment analysis," *Mater. Today, Proc.*, Feb. 2021, doi: 10.1016/j.matpr.2021.01.782.

[17] M. El-Masri, N. Altrabsheh, and H. Mansour, "Successes and challenges of Arabic sentiment analysis research: A literature review," *Social Netw. Anal. Mining*, vol. 7, no. 1, pp. 1–22, Dec. 2017, doi: 10.1007/s13278-017-0474-x.

[18] N. Al-Twairesh, H. Al-Khalifa, and A. Al-Salman, "Subjectivity and sentiment analysis of Arabic: Trends and challenges," in *Proc. IEEE/ACS 11th Int. Conf. Comput. Syst. Appl. (AICCSA)*. Doha, Qatar: IEEE, 2014.

[19] S. Panigrahi, A. Nanda, and T. Swarnkar, "A survey on transfer learning," *Smart Innov. Syst. Technol.*, vol. 194, pp. 781–789, Oct. 2020, doi: 10.1007/978-981-15-5971-6_83.

[20] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[21] M. E. Peters, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2018, pp. 2227–2237, doi: 10.18653/v1/N18-1202.

[22] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *Proc. 27th Int. Conf. Comput. linguistics*, 2018, pp. 1638–1649.

[23] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[24] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *Proc. Conf. North*, 2019, pp. 15–18.

[25] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation Jeffrey," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[26] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An easy-to-use framework for state-of-the-art NLP," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Demonstrations*, 2019, pp. 54–59.

[27] J. Zhu, H. Wang, M. Zhu, B. K. Tsou, and M. Ma, "Aspect-based opinion polling from customer reviews," *IEEE Trans. Affect. Comput.*, vol. 2, no. 1, pp. 37–49, Jan. 2011, doi: 10.1109/T-AFFC.2011.2.

[28] S. M. Al-Ghuribi, S. A. M. Noah, and S. Tiun, "Unsupervised semantic approach of aspect-based sentiment analysis for large-scale user reviews," *IEEE Access*, vol. 8, pp. 218592–218613, 2020, doi: 10.1109/ACCESS.2020.3042312.

[29] S. Poria, E. Cambria, L.-W. Ku, C. Gui, and A. Gelbukh, "A rule-based approach to aspect extraction from product reviews," Assoc. Comput. Linguistics, Dublin City Univ., Dublin, Ireland, Tech. Rep., 2015, pp. 28–37, doi: 10.3115/v1/w14-5905.

[30] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *Proc. 19th Nat. Conf. Artif. Intell.*, 2004, pp. 755–760.

[31] Q. Liu, Z. Gao, B. Liu, and Y. Zhang, "Automated rule selection for aspect extraction in opinion mining," in *Proc. 24th Int. Conf. on Artif. Intell.*, 2015, pp. 1291–1297.

[32] T. A. Rana and Y.-N. Cheah, "A two-fold rule-based model for aspect extraction," *Expert Syst. Appl.*, vol. 89, pp. 273–285, Dec. 2017, doi: 10.1016/j.eswa.2017.07.047.

[33] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, "Red opal: Product-feature scoring from reviews," in *Proc. 8th ACM Conf. Electron. Commerce (EC)*, 2007, pp. 182–191.

[34] X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid," in *Proc. Conf. Empirical MeAutomated Rule Selection Aspect Extraction Opinion Mininthods Natural Lang. Process.*, Cambridge, MA, USA: MIT, Oct. 2010, pp. 56–65.

[35] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," in *Proc. 26th Int. Conf. Comput. Linguistics*, 2016, pp. 3298–3307.

[36] H. Gandhi and V. Attar, "Extracting aspect terms using CRF and bi-LSTM models," *Proc. Comput. Sci.*, vol. 167, pp. 2486–2495, Jan. 2020, doi: 10.1016/j.procs.2020.03.301.

[37] C.-H. Lai, D.-R. Liu, and K.-S. Lien, "A hybrid of XGBoost and aspect-based review mining with attention neural network for user preference prediction," *Int. J. Mach. Learn. Cybern.*, vol. 12, no. 5, pp. 1203–1217, May 2021, doi: 10.1007/s13042-020-01229-w.

[38] K. Gao, H. Xu, C. Gao, X. Sun, J. Deng, and X. Zhang, "Two-stage attention network for aspect-level sentiment classification," in *Neural Information Processing* (Lecture Notes in Computer Science) vol. 11304. Cham, Switzerland: Springer, 2018.

[39] L. Li, L. Yang, and Y. Zeng, "Improving sentiment classification of restaurant reviews with attention-based bi-GRU neural network," *Symmetry*, vol. 13, no. 8, p. 1517, Aug. 2021, doi: 10.3390/sym13081517.

[40] 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.*, 2017, pp. 452–461.

[41] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Coupled multi-layer attentions for co-extraction of aspect and opinion terms," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 3316–3322.

[42] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with BERT," *IEEE Access*, vol. 7, pp. 154290–154299, 2019, doi: 10.1109/ACCESS.2019.2946594.

[43] X. Li, L. Bing, W. Zhang, and W. Lam, "Exploiting BERT for end-to-end aspect-based sentiment analysis?" Assoc. Comput. Linguistics, Hong Kong, Tech. Rep., 2019, doi: 10.18653/v1/d19-5505.

[44] A. R. Abas, I. El-Henawy, H. Mohamed, and A. Abdellatif, "Deep learning model for fine-grained aspect-based opinion mining," *IEEE Access*, vol. 8, pp. 128845–128855, 2020, doi: 10.1109/ACCESS.2020.3008824.

[45] B. Ma, D. Zhang, Z. Yan, and T. Kim, "An LDA and synonym lexicon based approach to product feature extraction from online consumer product reviews," *J. Electron. Commer. Res.*, vol. 14, no. 4, p. 304, 2013.

[46] L. Yu and X. Bai, "Implicit aspect extraction from online clothing reviews with fine-tuning BERT algorithm," *J. Phys., Conf.*, vol. 1995, no. 1, p. 12040, 2021.

[47] R. Zahid, M. O. Idrees, H. Mujtaba, and M. O. Beg, "Roman Urdu reviews dataset for aspect based opinion mining," in *Proc. 35th IEEE/ACM Int. Conf. Automated Softw. Eng. Workshops*, Sep. 2020, pp. 138–143.

[48] Y. A. Winatmoko, A. A. Septiandri, and A. P. Sutiono, "Aspect and opinion term extraction for hotel reviews using transfer learning and auxiliary labels," 2019, *arXiv:1909.11879*.

[49] E. Lopes, U. Correa, and L. Freitas, "Exploring BERT for aspect extraction in Portuguese language," *Proc. Int. FLAIRS Conf.*, vol. 34, 2021, pp. 1–4.

[50] S. Alowaidi, M. Saleh, and O. Abulnaja, "Semantic sentiment analysis of Arabic texts," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 2, pp. 256–262, 2017, doi: 10.14569/ijacsa.2017.080234.

[51] S. R. El-Beltagy, T. Khalil, A. Halaby, and M. Hammad, "Combining lexical features and a supervised learning approach for Arabic sentiment analysis," in *Computational Linguistics and Intelligent Text Processing* (Lecture Notes in Computer Science), vol. 9624. Cham, Switzerland: Springer, 2018.

[52] D. Gamal, M. Alfonse, E.-S.-M. El-Horbaty, and A.-B.-M. Salem, "Implementation of machine learning algorithms in Arabic sentiment analysis using N-gram features," *Proc. Comput. Sci.*, vol. 154, pp. 332–340, Jan. 2019, doi: 10.1016/j.procs.2019.06.048.

[53] A. Mohammed and R. Kora, "Deep learning approaches for Arabic sentiment analysis," *Social Netw. Anal. Mining*, vol. 9, no. 1, pp. 1–12, Dec. 2019, doi: 10.1007/s13278-019-0596-4.

[54] B. Nouhaila, A. Habib, A. Abdellah, and I. El Farouk Abdelhamid, "Arabic sentiment analysis based on 1-D convolutional neural network," in *Innovations in Smart Cities Applications*, vol. 183. Cham, Switzerland: Springer, 2021.

[55] A. H. Ombabi, W. Ouarda, and A. M. Alimi, "Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks," *Social Netw. Anal. Mining*, vol. 10, no. 1, pp. 1–13, Dec. 2020, doi: 10.1007/s13278-020-00668-1.

[56] M. Al-Smadi, O. Qawasmeh, B. Talafha, and M. Quwaider, "Human annotated Arabic dataset of book reviews for aspect based sentiment analysis," in *Proc. 3rd Int. Conf. Future Internet Things Cloud*, Aug. 2015, pp. 726–730, doi: 10.1109/FiCloud.2015.62.

[57] M. Al-Ayyoub, A. Gigieh, A. Al-Qwaqenah, M. N. Al-Kabi, B. Talafhah, and I. Alsmadi, "Aspect-based sentiment analysis of Arabic laptop reviews," in *Proc. Int. Arab Conf. Inf. Technol.*, Dec. 2017. p. 7. [Online]. Available: http://files/1172/Al-Ayyoub - 2017 - Aspect-Based Sentiment Analysis of Arabic Laptop R.pdf%0Ahttp://files/950/Al-Ayyoub - 2017 - Aspect-Based Sentiment Analysis of Arabic Laptop R.pdf%0Ahttps://www.researchgate.net/publication/329557349_Asp.

[58] S. Behdenna, G. Belalem, and F. Barigou, "An ontology-based approach to enhance explicit aspect extraction in standard Arabic reviews," *Int. J. Comput. Digit. Syst.*, vol. 11, no. 1, pp. 277–287, Jan. 2022.

[59] S. Areed, O. Alqaryouti, B. Siyam, and K. Shaalan, "Aspect-based sentiment analysis for Arabic government reviews," in *Recent Advances in NLP: The Case of Arabic Language*. Cham, Switzerland: Springer, 2020, pp. 143–162.

[60] M. AL-Smadi, M. Al-Ayyoub, H. Al-Sarhan, and Y. Jararweh, "Using aspect-based sentiment analysis to evaluate Arabic news affect on readers," in *Proc. IEEE/ACM 8th Int. Conf. Utility Cloud Comput. (UCC)*, Dec. 2015, pp. 436–441.

[61] S. Ruder, P. Ghaffari, and J. G. Breslin, "INSIGHT-1 at SemEval-2016 task 5: Deep learning for multilingual aspect-based sentiment analysis," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 330–336, doi: 10.18653/v1/s16-1053.

[62] A. Tamchyna and K. Veselovská, "UFAL at SemEval-2016 task 5: Recurrent neural networks for sentence classification," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 367–371, doi: 10.18653/v1/s16-1059.

[63] M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, "Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews," *J. Comput. Sci.*, vol. 27, pp. 386–393, 2018, doi: 10.1016/j.jocs.2017.11.006.

[64] M. Al-smadi, B. Talafha, M. Al-Ayyoub, and Y. Jararweh, "Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 8, pp. 2163–2175, 2019, doi: 10.1007/s13042-018-0799-4.

[65] S. Al-Dabet, S. Tedmori, and M. Al-Smadi, "Extracting opinion targets using attention-based neural model," *Social Netw. Comput. Sci.*, vol. 1, no. 5, pp. 1–10, Sep. 2020, doi: 10.1007/s42979-020-00270-4.

[66] M. M. Abdelgwad, T. H. A Soliman, A. I. Taloba, and M. F. Farghaly, "Arabic aspect based sentiment analysis using bidirectional GRU based models," *J. King Saud Univ. Comput. Inf. Sci.*, Sep. 2021, doi: 10.1016/j.jksuci.2021.08.030.

[67] M. M. Abdelgwad, "Arabic aspect based sentiment classification using BERT," 2021, *arXiv:2107.13290*.

[68] R. Bensoltane and T. Zaki, "Towards Arabic aspect-based sentiment analysis: A transfer learning-based approach," *Social Netw. Anal. Mining*, vol. 12, no. 1, pp. 1–16, Dec. 2022.

[69] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5999–6009.

[70] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," in *Proc. 4th Work. Open-Source Arab. Corpora Process. Tools, Shar. Task Offensive Lang. Detect.*, May 2020., pp. 9–15.

[71] C. Sutton, K. Rohanimanesh, and A. McCallum, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, pp. 282–289.

[72] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 260–270.

ARWA SAIF FADEL received the B.Sc. degree from Hodeidah University, Yemen, in 2006, and the M.Sc. degree in computer science from King Abdulaziz University, Saudi Arabia, in 2015, where she is currently pursuing the Ph.D. degree with the Department of Computer Science. Her research interests include natural language processing, sentiment analysis, machine learning, and deep learning.

MOSTAFA ELSAYED SALEH received the B.S. degree in computer engineering and the M.S. and Ph.D. degrees in computer engineering from Mansoura University, Egypt, in 1992, 1995, and 2000, respectively. He is currently an Associate Professor with King Abdulaziz University, Jeddah, where he is also a member of the HPC Research Group, Department of Computer Science. His research interests include data engineering, semantic web, and high-performance computing.

OSAMA AHMED ABULNAJA received the B.S. degree in computer science from King Abdulaziz University (KAU), Jeddah, Saudi Arabia, in 1986, and the M.S. degree in computer science and the Ph.D. degree in engineering (computer science) from the University of Wisconsin–Milwaukee, Milwaukee, WI, USA, in 1990 and 1996, respectively. He is currently a Professor in computer science with KAU, where he is also a member of the HPC Research Group, Department of Computer Science. His research interests include fault tolerance, high-performance computing (HPC), systems performance, and systems programming.