IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 November 2023, accepted 20 December 2023, date of publication 29 December 2023,
date of current version 5 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3348342

RESEARCH ARTICLE

BERT-Based Model for Aspect-Based Sentiment Analysis for Analyzing Arabic Open-Ended Survey Responses: A Case Study

KHLOUD A. ALSHAIKH[1], **OMAIMA A. ALMATRAFI**[1], **AND YOOSEF B. ABUSHARK**[2]
[1]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Corresponding author: Khloud A. Alshaikh (khassanalshaikh@stu.kau.edu.sa)

ABSTRACT Educational institutions typically gather feedback from beneficiaries through formal surveys. Offering open-ended questions allows students to express their opinions about matters that may not have been measured directly in closed-ended questions. However, responses to open-ended questions are typically overlooked due to the time and effort required. Aspect-based sentiment analysis is used to automate the process of extracting fine-grained information from texts. This study aims to 1) examine the performance of different BERT-based models for aspect term extraction for Arabic text sourced from educational institution surveys; 2) develop a system that automates the ABSA process in a way that will automatically label survey responses. An end-to-end system was developed as a case study to extract aspect terms, identify their polarity, map extracted aspects to their respective categories, and aggregate category polarity. To accomplish this, the models were evaluated using an in-house dataset. The result showed that FAST-LCF-ATEPC, a multilingual checkpoint, outperformed other models including AraBERT, MARBERT, and QARiB, in the aspect-term extraction task, with an F1 score of 0.58. Hence, it was used for aspect-term polarity classification, showing an F1 score of 0.86. Mapping aspects to their respective categories using a predefined list yielded an average F1 score of 0.98. Furthermore, the polarities of the categories were aggregated to summarize the overall polarity for each category. The developed system can support Arabic educational institutions in harnessing valuable information in responses to open-ended survey questions, allowing decision-makers to better allocate resources, and improve facilities, services, and students' learning experiences.

INDEX TERMS Arabic ABSA, aspect extraction, aspect-based sentiment analysis, BERT-based model, education, polarity classification.

I. INTRODUCTION

Universities and higher education institutions worldwide allocate significant financial resources to enhance their services to maintain existing students and attract new ones [1]. Student satisfaction and opinions about the university's service quality are very important because they have a direct impact on student impressions and the institution's reputation [2]. Students can express their thoughts through official surveys published at the institutional level. Usually, these surveys include closed- and open-ended questions. Close-ended questions are specific and easy to analyze. In contrast, open-ended questions give students the opportunity to express their opinions and sentiments [1]. This type of question is valuable because it encourages them to express their minds and feelings and provide useful information on personal experiences [3], [4]. However, these textual responses require more effort in the analysis process to extract helpful information and obtain sentiments from it. Such analyses also consume considerable human time, especially when the number of responses is large and the questions cover more than one aspect [3], [4]. Students' responses are typically related to university aspects, such as services, professors, and buildings, and their feelings (positive, negative, and neutral) toward

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

these aspects. Extracting useful information from textual responses calls for an automated system that can analyze the text and detect the sentiments of the elements (aspects) presented in the response.

Sentiment analysis or opinion mining is an active area of the natural language processing [5]. The main task of sentiment analysis is to classify expressed opinions in the text [6]. The extracted opinion is typically classified according to its polarity as positive, negative, or neutral [7]. There are three levels of classification in sentiment analysis: document, sentence, and aspect [8]. Although document- and sentence-level analyses are useful for some applications, they are not sufficient for others to search for fine-grained information about a particular aspect. In such cases, aspect-based sentiment analysis (ABSA) is used. Principally, ABSA systems receive a set of texts (product reviews, comments, forum discussions, etc.) that discuss a specific entity. The system attempts to obtain the main aspects of the entity and detect the sentiments expressed toward each aspect [7]. The results of ABSA provide detailed sentiment information that can be highly valuable in various domains. Despite this benefit, ABSA has not been extensively applied in the educational domain. In addition, the majority of prior work on ABSA has focused on English, with a limited number of studies targeting ABSA in Arabic and other languages [9]. Arabic is the primary spoken language for approximately 422 million speakers worldwide [10]. It is a rich language with a large number of vocabulary words with different sentence structures and multiple meanings. It has approximately 10,000 roots and more than 900 forms of nouns and verbs based on their morphology [11]. This results in a variety of derivational morphologies and structural forms, which increase the sparsity of morphemes and words [9] as well as the complexity of the analysis.

An advanced system is needed to analyze students' survey responses offered in Arabic, categorize them based on various aspects of the university, and identify students' sentiments toward these aspects. Such a system will support the integration of student feedback into decision-making processes and aid university leaders in allocating resources and improving the quality of the services provided. Thus, there is a need to examine the literature and identify potential approaches that can improve ABSA for the Arabic language as well as its effectiveness in analyzing educational data.

This study aims to fill this gap by examining a transfer learning approach to assess ABSA in the Arabic educational context and to evaluate its performance. Different bidirectional encoder representations from transformer BERT-based models were evaluated for aspect extraction using an in-house dataset of open-ended survey responses at King Abdulaziz University (KAU). The best-performing model was used to classify the polarity of each aspect. The aspects are then mapped to their category, and the results are summarized by category by simply counting the polarities of the aspects for each category. To the best of our knowledge, this is the first study to assess ABSA using Arabic surveys in the educational domain.

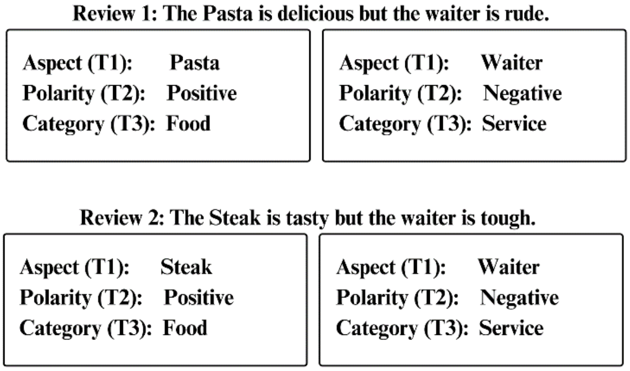

FIGURE 1. ABSA tasks.

II. BACKGROUND

A. ASPECT-BASED SENTIMENT ANALYSIS

ABSA produces finely detailed sentiment information. This information is useful for many applications in various domains. The ABSA consists of four tasks: aspect term extraction (T1), aspect polarity classification (T2), aspect category mapping (T3), and category polarity (T4). T1 extracts all the words or aspects that need to specify their polarity sentiment. This aspect can be implicit or explicit. The task was executed using supervised and unsupervised methods [12]. T2 assigns the polarity of the sentiment analysis to the extracted aspect [13]. T3 identifies the category using a multilabel classifier that classifies each entity into multiple labels, where the label consists of entities and aspects. T4 assigns the polarity of sentiment analysis to the identified categories. Figure 1 shows an example of these tasks.

Assuming that there are only two reviews for a restaurant, tasks T1, T2, and T3 are assigned, as shown in Figure 1. Task T4 for the overall polarity of the category in this example is positive for food and negative for service since pasta and steak are rated as positive to yield an overall positive category polarity for food, whereas the waiter is rated negative, yielding an overall negative category polarity for service.

B. DEEP LEARNING

Deep learning is a rising technique in machine learning that uses a hierarchy of layers to progressively extract higher-level features. During training, the high layers exploit the complex compositional nonlinear functions of the lower layers. This means that the layers in a higher hierarchy have more abstract or divided representations than the lower ones. Consequently, each layer receives input to analyze and classify it to provide the output that feeds the input of the next layer [14], [15]. A variety of algorithms, such as deep neural networks, convolutional neural networks, recurrent neural networks (RNN), and recursive neural networks, help in the analysis of many fields, especially in fine-grained processes for processors with a large number of layers [14]. Additionally, word embedding, long short-term memory (LSTM), and bi-directional LSTM are concepts related to deep learning that allow dealing with various types of data such as text, images, and videos [16].

C. TRANSFORMER-BASED TRANSFER LEARNING

Transfer learning is an emerging machine-learning technique that uses existing knowledge to solve different domain problems and produces state-of-the-art prediction results [17]. Transfer learning methods perform extensively in computer vision tasks such as anomalous activity detection, object classification, and image captioning. Moreover, transfer-learning-based methods, such as BERT, have been successful in several natural language processing (NLP) tasks [18] and in the field of sentiment analysis [17]. BERT is a pre-trained language model developed by Google in 2018. It uses deep neural network architecture with an attention component. It is designed to process sequential data such as text and learn the contextual relationships between words [19].

III. LITERATURE REVIEW

To decide which is the most appropriate approach, a comprehensive literature review on ABSA in the education domain and Arabic ABSA approaches was done. The“ABSA in educational domain” section presents all existing empirical studies in the educational domain to review the source of used data, approaches, and ABSA tasks through a methodical and exhaustive literature review using search queries consisting of the keywords (“aspect-based sentiment analysis” OR“ABSA”) AND“education”. After that, there is still a need for more research on approaches used for ABSA tasks in Arabic datasets in other domains that are covered in the“Arabic ABSA approaches” by exploring the literature review using the keywords (“aspect-based sentiment analysis” OR“ABSA”) AND“Arabic” presented in the“Arabic ABSA Approaches” section. We came up with three subsections: unsupervised learning, supervised learning, and deep learning. The associated studies of these approaches were presented in detail.

A. ABSA IN EDUCATIONAL DOMAIN

Most ABSA studies in the educational field have been conducted on English-language datasets. The aim of these studies was to assist academic institutions in identifying and addressing student issues through feedback analysis. The data for these studies was primarily gathered from social media platforms like Twitter and Facebook, as mentioned in [20], [21], [22], and [23]. Other studies utilized data collected from the institution, such as MOOC platforms or traditional institution surveys, as seen in [24], [25], and [26]. The methods used in these studies included semantic relatedness and sentiment polarity categorization. The researchers employed various classical machine learning algorithms such as k-means clustering, naive Bayes, linear regression, and support vector machine (SVM) with only two studies employed deep neural networks, LSTM [24], [26]. We also found that all studies focused on aspect extraction and polarity classification tasks, with the exception of a study that used a combination of machine learning- and lexicon-based approaches to accomplish all four tasks [23].

ABSA has also found application beyond the English language in education, albeit in a limited capacity. In Serbia, [1] employed ML algorithms to achieve T1 and T2. They examined student reviews on the“Oceni profesora” (“Rate my professors”) website to gain insights into the teaching faculty, courses, and programs offered by the Faculty of Technical Sciences. Another study in Indonesia used an unsupervised lexicon-based method for both tasks [27]. They used recent online learning graduates feedback from BINUS (Bina Nusantara University). Moreover, [28] proposed a hybrid features selection method to address T1 and T2 in Arabic tweets related to Qassim University. It extracts aspects related to the education domain such as teaching quality, services, activities, etc. The purpose of this study is to enhance the SVM classifier in ABSA by decreasing used features. The results showed that the hybrid method successfully improves SVM classifier performance with (F1: 0.70) for T1 and (F1: 0.71) for T2. Table 1 provides a summary of prior work on ABSA in the educational domain showing the year of publication, the targeted language, the data source, the approach used, and the tasks covered by each paper. As shown, there is a lack of ABSA research on the Arabic language in the educational domain. This research aims to contribute to this direction benefitting researchers and practitioners.

B. ARABIC ABSA APPROACHES

1) UNSUPERVISED APPROACHES

A comparative study was conducted to test and assess various lexicon-based approaches for ABSA tasks T3 and T4 based on 63,000 book reviews annotated by humans [29]. This was later extended using enhanced lexicon-based approaches on the same book review dataset to achieve results that exceeded those of the previous study, particularly for T4 (accuracy:0.88) and T3 (F1 score:0.24) [30]. Several studies have combined two approaches or models to produce superior models. Reference [31] combined corpus- and lexicon-based approaches to address tasks T2 and T4 using a large-scale Arabic book review dataset. Furthermore, [32] proposed a hybrid approach to address T1 and T2 from reviews in Arabic government applications. This approach combined lexicons with rule-based models. The authors aimed to develop rules, techniques, and lexicons to address the challenges of sentiment analysis. The results showed an increase in accuracy when compared to the baseline models.

2) SUPERVISED APPROACHES

Supervised approaches depend on the training process using labeled data to train the machine in predicting the output for the new input. Various studies have used the Arabic-language

TABLE 1. A summary of prior work on ABSA in the educational field.

Year	Ref.	Language	Dataset	Approach Type\ Name	ABSA Tasks
2017	[20]	English	Tweets of online student feedback	k-mean clustering and naïve Bayes classification	T1T2
2017	[22]		Review sites and social media networks	OpenNLP POS Standford NLP library	T1 T2
2019	[24]		Last five years students feedback of Sukkur IBA University	Supervised two-layered LSTM model	T1 T2
2019	[23]		Student’s comments from social media	Machine learning and lexicon based	T1 T2 T3 T4
2020	[26]		Students’ reviews on MOOCs from Coursera and students’ feedback in traditional classroom settings	Word-Embedded (fastText- GloVe-Word2Vec- MOOC) With CNN and LSTM	T2 T3
2022	[21]		UAE COVID-19 education related data	Logistic Regression Linear SVC Multinomial NB Random Forest	T1 T2
2022	[25]		Coursera dataset from Kaggle	unsupervised and semi-supervised LDA	T1 T2
2020	[1]	Serbian	Student reviews of Faculty of Technical Sciences in Serbia and a corpus of online reviews from (“Rate my professors”) website	Rule-based and dictionary-based components	T1 T2
2022	[27]	Indonesian	Students’ feedback from Bina Nusantara University	Unsupervised lexicon-based method	T1 T2
2020	[28]	Arabic	Tweets related to Qassim University in KSA	Hybrid feature selection method to Enhance SVM	T1 T2

hotel review dataset as a benchmark to evaluate their proposed approaches or models. The authors in [9] proposed a framework for applying ABSA to Arabic. They suggested the use of a SVM approach for tasks T1, T2, and T3. Reference [33] considered morphological, syntactic, and semantic features to address task T2, in addition to T1 and T3. The authors examined multiple classification methods such as naïve Bayes, Bayes networks, decision trees, k-nearest neighbor (K-NN), and SVM. The results showed that models developed by the supervised learning approach performed better than combined lexicons with rule-based models, whereas SVM performed the best compared with the other classifiers for all tasks in the study. Moreover, [12] evaluated various classifier techniques for T1, and the results showed that the adaptive boosting (AdaBoost) classifier achieved the best results compared with previous methods in terms of precision (97%) and recall (96.9%).

3) DEEP LEARNING APPROACHES

A study by [34] compared two pretrained word-embedding models for ABSA. These models are fastText Arabic Wikipedia and AraVec Web. An SVM classifier was used to train the model for tasks T1 and T2 in a dataset of 5000 Arabic tweets related to airline services that were manually labeled for ABSA. The study showed an enhancement in the SVM classifier performance when extracting features using word embedding. The result was slightly better when fastText Arabic Wikipedia word embedding was used compared with AraVec-Web, indicating the usefulness of word embedding for sentiment analysis.

Other studies used the Arabic-language hotel review dataset to evaluate the proposed approach or model. Reference [35] applied a deep RNN and SVM to hotel reviews to address tasks T1, T2, and T3. The results showed that the SVM exceeded the deep RNN. However, the authors suggested enhancing the proposed deep learning approach by assessing different LSTM networks and using word embedding, such as fastText. Reference [36] applied the suggestions of a previous study by utilizing LSTM neural networks for T1 and T2. The results showed that the method used exceeded the baseline (SVM trained with N-gram features) for both the T1 and T2 tasks. Furthermore, [37] applied two deep learning models: the convolutional independent LSTM model (C-IndyLSTM) for T1, and the memory-based recurrent attention model (MBRA) for T3. The C-IndyLSTM model is based on a convolutional neural network and stacked independent long-short-term memory, whereas the MBRA model is based on stacked bidirectional independent LSTM, a position-weighting mechanism, and multiple attention mechanism layers. Moreover, [38] applied two deep-learning models based on GRU neural networks. The first model, BGRU-CNN-CRF, combines a bidirectional GRU, CNN, and CRF for T1. The second model, IAN-BGRU, is an interactive attention network used for T2.

Recently, increased attention has been paid to the use of large pre-trained language models, such as BERT and its variations, as it achieves superior results for a variety of NLP tasks. Reference [39] proposed a BERT with a simple linear classification layer to accomplish T2 only. Experiments on three Arabic datasets, hotel reviews, book reviews, and Arabic news, showed that the proposed model accuracies were 89.51%, 73.23%, and 85.73%, respectively. The researchers aim to accomplish T1 and T3 in future work. Reference [40] proposed a transfer learning method using the AraBERT pre-trained language model to accomplish tasks T1 and T3.

Most previous studies individually or sequentially handled the T1 and T2 tasks, where independent models were designed for each task. However, T1 and T2 are performed jointly in multi-task learning by other studies. Reference [41] developed a lightweight ABSA framework called Python aspect-based sentiment analysis (PyABSA), which can be used for T1 and T2. The models were trained on various datasets, including restaurants, laptops, MOOCs, Twitter, and other domains in eight languages (one of them was the Arabic language dataset SemEval-2016 Task 5). The Arabic dataset was used to evaluate the BERT-ATESC, Fast LCF-ASESC, and LCF-ATESC models. Performance evaluation showed that the BERT-ATESC model achieved the best results, with an F1 score of 71.18% for T1 and T2.

Furthermore, [42] tested a transfer-learning approach using Arabic-BERT-CRF for tasks T1 and T2 on a human-annotated Arabic dataset for ABSA. The experimental results demonstrated that the model exceeded the baseline model, which relied on conditional random fields (CRF) with features extracted using named entity recognition (NER), POS tagging, parsing, semantic analysis, and other recently proposed models such as AraBERT, MarBERT, and CamelBERT-MSA. Reference [43] proposed a multi-task learning approach called local context focus-aspect term extraction and polarity classification (LCF-ATEPC) and AraBERT as a shared layer for Arabic contextual text representation to accomplish T1 and T2 simultaneously. The reference hotel and product review datasets were used. In addition, the authors proposed a data augmentation technique for T2 that involves generating synthetic data using back-translation and synonym replacement. The results showed that the proposed model outperformed the baseline models on both datasets for both single- and multitask approaches, achieving state-of-the-art performance. Table 2 provides a comparison of the different Arabic language ABSA literature reviews that were summarized above. The comparison is across the year of publication, the data source, the used approach, and the result of the covered tasks by each research.

Overall, Arabic ABSA has evolved significantly over the years, transitioning from lexicon-based approaches to deep learning techniques. Lexicon-based approaches were simple but suffered from scalability constraints and the inability to adapt to context-dependent nuances in sentiment analysis. Supervised learning methods improved scalability but required substantial amounts of labeled data and involved feature engineering. However, in recent years, deep learning has emerged as a dominant approach in ABSA, largely due to the Transformer-based BERT model. BERT has demonstrated remarkable effectiveness in understanding contextual information, capturing complex language patterns, and addressing prior limitations. Moreover, BERT has introduced the concept of transfer learning in NLP, enabling it to learn general language representations through pre-training on vast text corpora. Subsequently, fine-tuning BERT on task-specific data significantly reduces the need for extensive labeled data, making it an ideal choice for this research.

As a part of our research, we have carefully selected the latest and top-performing BERT-based models from the literature - LCF and AraBERT. FAST-LCF-ATEPC model stood out as it efficiently performs aspect term extraction and aspect polarity classification simultaneously. AraBERT, on the other hand, was specifically designed and trained on Arabic data, making it a promising model. While AraBERT has shown significant improvement over baseline approaches in various Arabic NLP tasks, it has been outperformed by MARBERT [44]. QARiB also performed well in Arabic NLP tasks like SA and NER, but its performance for Arabic ABSA has not yet been evaluated. Therefore, it is essential to experiment with the most promising BERT-based models for Arabic ABSA and evaluate their performance on related data to be able to develop an effective ABSA system for educational institutions. Table 3 provides a comprehensive overview of the BERT models used, highlighting their respective areas of focus, advantages, and limitations. Moreover, each model is explained separately in the methodology section.

IV. RESEARCH CONTRIBUTION

This study is unlike prior works, as it delves into the examination and application of ABSA methods in a domain that has received limited attention - Arabic language text obtained from the educational sector. This sector has not been extensively studied, and the effectiveness of pre-trained models, such as BERT, which performed well in various NLP tasks remains unexplored in the intersection of Arabic ABSA and the educational sector. It is essential to acknowledge that models trained for one domain may not perform as well in another, emphasizing the need for rigorous evaluation of different ABSA models on Arabic text derived from educational data. The main contribution of this research is 1) to examine the performance of different BERT-based models for aspect term extraction for Arabic text sourced from educational institution surveys; 2) to develop a system that automates the ABSA process in a way that will automatically label survey responses. This research has significant implications including improving the quality of education and enhancing user satisfaction. Moreover, the automatic identification of areas of concern or success, which, in turn, can inform policymakers and aid in the allocation of resources to meet the evolving needs of students and educators. The benefits of this research are not limited to educational institutions,

TABLE 2. A summary of prior work on ABSA in arabic.

Ref. Year	Dataset	Approach Type\ Name	ABSA Tasks\ Results
		Un-supervised Learning	
[31] 2020	Book reviews (LABR)	Corpus-based and lexical-based approach	T2: Acc.(80.5%) T4: Acc.(78%)
[32] 2020	UAE government mobile application reviews	Lexicon with rule-based	T1 : F1(92.50%) T2 : Acc.(95.81%)
		Supervised Learning	
[45] 2015	News posts on social media	CRF vs. J48	Best: T1: J48 F1(0.82) T2: CRF Acc.(0.87)
[9] 2017	Hotels' reviews (SemEval 2015)	SVM	T1 : F1(30.978) T2: Acc.(76.421) T3: F1(40.336) T4: F1(18.806)
[35] 2018	Hotels' reviews (SemEval 2016)	RNN vs. SVM trained along with lexical, word, syntactic, morphological, and semantic features.	Best: T1: SVM F1(89.8%) T2: SVM Acc.(95.4%) T3: SVM F1(89.8%)
[33] 2019	Hotels' reviews (SemEval 2016)	SVM, K-Nearest Neighbor, Decision Tree, Bayesian Networks, and Naïve Bayes	Best: T1: SVM F1(93.4) T2: SVM F1(89.9) T3: SVM Acc.(95.4)
[12] 2021	Hotels' reviews (SemEval 2016)	ADAL system (Adaboost classifier: rule based with machine learning methods)	T1: P.(97) T1: R.(97)
		Deep Learning	
[34] 2019	Airline services tweets	fastText vs. AraVec with SVM classifier	Best : T1 : fastText F1(79) T2 : fastText Acc.(89)
[36] 2019	Hotels' reviews (SemEval 2016)	T1: Bi-LSTM-CRF (fastText) T2: INSIGHT-1 (CNN)	T1: F1(69.98) T2: Acc.(82.7)
[37] 2021	Hotels' reviews (SemEval 2016)	T2: MBRA T3: C-IndyLSTM	T2: F1(58.05) T3: Acc.(87.31)
[38] 2021	Hotels' reviews (SemEval 2016)	T1: CNN-BGRU-CRF (fastText) T2: IAN-BGRU	T1: F1(70.67) T2: Acc.(83.98)
[40] 2022	News posts on social media	BiLSTM + CRF BERT + linear classification layer BERT + CRF BERT + BiLSTM + CRF BERT + BiGRU + CRF *BERT refers to AraBERT	Best: T1: BERT + BiGRU + CRF F1(88.1)
[39] 2022	Hotels' reviews (SemEval 2016) ABSA book reviews (HAAD) News posts on social media	BERT with a simple linear classification layer	Hotels' T2: Acc.(89.51) Book reviews T2: Acc.(73.23) News posts T2: Acc.(85.73)
[41] 2022	Hotels' reviews (SemEval 2016)	BERT-ATESC Fast-LCF-ASESC LCF-ATESC	Best: BERT-ATESC T1: F1(71.18) T2: F1(71.18)
[42] 2023	ABSA book reviews (HAAD)	AraBERT and text classification by using CRF	T1: F1(47.63) T2: Acc.(95.23)
[43] 2023	Hotels' reviews (SemEval 2016)	LCF-ATEPC model + AraBERT AR-LCF-ATEPC_Fusion AR-LCF-ATEPC_CWD AR-LCF-ATEPC_CMD	Best: AR-LCF-ATEPC_Fusion T1: F1(75.94) T2: Acc.(91.5)

TABLE 3. Overview of the used pre-trained BERT models.

Model	Aspect	Content
FAST-LCF-ATEPC [41], [46]	**Focus**	▪ Joint task of aspect term extraction and aspect polarity classification. ▪ Applies self-attention and local context focus techniques to aspect word extraction task.
	Related Work	▪ English [46], [41], [47] ▪ Chinese [46], [41] ▪ Dutch, Spanish, French, Turkish, Russian, and Arabic (Hotel's reviews) [41]
	Advantages	▪ Extracts aspect term and infers aspect term polarity synchronously. ▪ Integrates the pre-trained BERT model, to leverage its strengths of separate layers for local and global context modeling. ▪ Achieved new state-of-the-art performance, especially the F1 score of T1 task. ▪ Achieved state-of-the-art performance on seven ABSA datasets.
	Limitation	▪ performs better on some datasets when dealing with a single task such as T1 or T2 than ABSA multi-tasking based on experimental results.
AraBERT [48]	**Focus**	▪ Specifically designed for the Arabic language. ▪ Captures the linguistic characteristics and nuances of the Arabic language.
	Related Work	▪ Hotel's reviews: ○ (T1): [49] ○ (T3): [50], [51] ○ (T1, T2): [52] ○ (T2, T4): [53] ▪ Book reviews (HAAD) ○ (T2, T4): [53] ▪ Social media: ○ News posts about the 2014 Gaza Attacks ▪ (T1, T3): [40] ▪ (T2, T4): [53] ○ Tweets about food delivery service reviews (T1, T2, T3): [54] ○ Tweets about sports, politics, and economics (T2): [55]
	Advantages	▪ Better performance on Arabic NLP tasks compared to generic multilingual models. ▪ Provides better language-specific understanding. ▪ Achieved state-of-the-art performance on most tested Arabic NLP tasks. ▪ Evaluated on NLP tasks: SA and NER. ▪ Designed specifically for the Arabic language, which makes it a strong candidate for ABSA tasks in Arabic.
	Limitation	▪ Performance may be suboptimal for certain Arabic dialects or domains. ▪ It may not generalize well to low-resource Arabic varieties.
MARBERT [44]	**Focus**	▪ Multilingual Bert-based model that handles code-switching scenarios and allows cross-lingual transfer learning.
	Related Work	▪ Twitter about sports, politics, and economics (T2): [55]
	Advantages	▪ Enables better performance on Arabic and related languages. ▪ Provides benefits of cross-lingual transfer learning. ▪ Outperformed AraBERT.
	Limitation	▪ Performance may be weaker than language-specific models like AraBERT for Arabic-specific tasks. ▪ Less effective for languages less similar to Arabic.
QARiB [56]	**Focus**	▪ Designed specifically for Arabic, addressing dialectal variations within the Arabic language.
	Related Work	▪ It has not been yet evaluated for ABSA tasks.
	Advantages	▪ Captures linguistic features specific to Arabic, improving performance on NLP tasks. ▪ Achieved state-of-the-art results on emotion and NER tasks.
	Limitation	▪ Performance may not generalize well to other Arabic dialects or languages and may be less effective for tasks involving standard Arabic or other dialects. ▪ Explore dialect-specific models like QARiB for other Arabic dialects and investigate approaches to handle dialectal variations effectively.

which can expedite the analysis through the automation of the four steps of ABSA but also extend to the advancement of natural language processing research, particularly for the Arabic language.

V. METHODOLOGY

This section introduces the used datasets and models. After that, we described the approach used to develop the aspect-based sentiment analysis system. Lastly, we defined the performance measures used to evaluate the different tasks in this research.

A. DATA

This section describes the steps involved in building a reliable annotated dataset from an educational context for testing and evaluation. Data were collected from the KAU service evaluation survey. The responses were usually written in formal Arabic, with a maximum of 200 characters. Students dis-

TABLE 4. Examples of responses to the open-ended question.

Arabic Response	Translation
شكرا جامعتي وفرتي لي المسكن والمأكل والعلم والبيئة الافضل كونها جامعة حكومية	Thank you, my university, for providing me with housing, food, education, and the best environment, being a public university
توفير مكتبة خاصة بكل كلية، توفير مواقف سيارات للطالبات وأماكن استراحة للسائقين	Providing a library for each college, providing car parking for female students and resting places for drivers.
التخصصات المتوفرة لاتتناسب احتياج سوق العمل	The available specializations do not fit the needs of the labor market

closed their feelings transparently regarding the university's main aspects. The total number of collected responses was 1815 responses, 91 of which were written in English, and 218 were garbled data, such as spaces, numbers, and symbols. The remaining 1506 responses were analyzed. Some sample responses to the open-ended questions "Any other additions that were not mentioned in the questionnaire that you would like to mention?" are shown in Table 4.

In the second step, the collected dataset underwent annotation to label the responses manually. The annotation was performed by three KAU employees: A, B, and C. Detailed guidelines were provided to the annotators to help them extract aspects and identify their polarity and categories. After receiving the annotated data from the annotators, the data were explored and cleaned. Table 5 provides three examples of responses to this question, along with human annotations, showing a sample of the in-house built dataset.

In the third step, the labeled responses were assessed and evaluated by calculating Cohen's kappa, which measures the agreement between two annotators to ensure the quality of the annotation process [57]. In our case, Cohen's kappa was used to check the agreement for pairs of annotators (A and B, A and C, and B and C) separately for aspect, polarity, and category. As shown in Table 6, the best agreement for the aspect, polarity, and category was between annotators B and C. Cohen's Kappa showed a **substantial agreement**, with a kappa value of 0.70 for the aspect. Further, B and C gave the largest number of aspects compared to A and B or A and C. Moreover, the polarity and category showed an **almost perfect agreement**, with kappa values of 0.92 and 0.87, respectively.

According to the Cohen kappa results, annotators B and C were selected to construct the golden dataset. When there was a discrepancy between B and C, the annotation of A was consulted. Only matched aspect, polarity, and category were sustained, which resulted in the retention of 448 responses that included 639 aspects related to 13 categories. The polarity of these aspects is skewed toward negative (512 aspects, ~80%), which is expected because people tend to recall and report negative experiences or thoughts more than positive ones, which is also known as a negativity bias [58], [59].

B. MODELS

1) FAST-LCF-ATEPC MODEL

FAST-LCF-ATEPC, proposed in 2021, is a multitask learning model based on self-attention and local context focus (LCF) mechanisms that integrate the pretrained BERT model. Unlike other models, it extracts aspect terms and synchronously infers polarity [46]. It employs two separate BERT layers to capture the global and local context, respectively. To enable simultaneous multi-task training, the input sequences are divided into separate tokens, and each token is assigned two labels. The first label determines whether the token is part of an aspect, while the second label denotes the polarity of the token associated with the aspect.

PyABSA, which is an open framework, has different versions of FAST-LCF-ATEPC trained on the SemEval 2016 Arabic dataset [41]. The checkpoints used in this study were multilingual, multilingual-256, and multilingual-256-2. The main difference between these three models is the number of languages and the size of the embedding layers used in the model as shown in Table 7. For the multilingual checkpoint, the model is trained on a multilingual dataset with 5 languages (English, French, German, Italian, and Spanish) and the embedding layer size used is 768. For the multilingual-256 checkpoint, the model is also trained on a multilingual dataset with 5 languages but uses a smaller embedding layer size of 256. This reduces the memory footprint of the model and can improve training speed on smaller datasets. For the multilingual-256-2, the model is trained on a larger multilingual dataset with 15 languages and uses a smaller embedding layer size of 256. This allows the model to generalize better across languages and reduces the likelihood of overfitting on any particular language [41], [46]. Therefore, it is essential to conduct an empirical evaluation on all three models to determine which one would yield better results considering the differences in the size of the embedding layer and the diversity of languages used in training.

2) ARABERT MODEL

AraBERT was developed in 2021 as a pretrained BERT model specifically for the Arabic language to achieve the same success as BERT for the English language. In addition to BERT base configuration, AraBERT employs two tasks: Masked Language Modeling (MLM) task to improve pre-training tasks by forcing the model to predict the whole word instead of getting hints from parts of the word, and Next Sentence Prediction (NSP) task to helps the model understand the relationship between two sentences, which can be useful for many language understanding tasks such as Question Answering. It was trained on a large-scale Arabic corpus extracted from news articles on the Arabic media. This corpus contained modern standard Arabic (MSA) data.

TABLE 5. Examples of labeled responses.

Response	Aspects	Polarity	Category
Thank you, my university, for providing me with housing, food, education, and the best environment, being a public university شكرا جامعتي وفرتي لي المسكن والمأكل والعلم والبيئة الافضل كونها جامعة حكومية	my university جامعتي	Positive إيجابي	Impression of the University and personal skills الصورة الذهنية للجامعة والمهارات الشخصية
	housing المسكن		University infrastructure and public services البنية التحتية للجامعة والخدمات العامة
	food المأكل		
	environment البيئة		
	education العلم		Educational process العملية التعليمية
Providing a library for each college, providing car parking for female students and resting places for drivers توفير مكتبة خاصة بكل كلية ، توفير مواقف سيارات للطالبات وأماكن استراحة للسائقين	library مكتبة	Negative سلبي	University infrastructure and public services البنية التحتية للجامعة والخدمات العامة
	car parking مواقف سيارات		
	resting places for drivers أماكن استراحة للسائقين		
The available specializations do not fit the needs of the labor market التخصصات المتوفرة لاتناسب احتياج سوق العمل	specializations التخصصات	Negative سلبي	Educational process العملية التعليمية

TABLE 6. Annotators Cohen's kappa for aspect, polarity, and category.

Annotators	Aspect	Polarity	Category
A and B	0.72	0.82	0.86
A and C	0.74	0.89	0.82
B and C	**0.70**	**0.92**	**0.87**

TABLE 7. FAST-LCF-ATEPC checkpoints.

	FAST-LCF-ATEPC		
Checkpoint	multilingual	multilingual-256	multilingual-256-2
Dataset language	5 languages	5 languages	15 languages
Embedding layer size	768	256	256

It includes 70 million sentences and 3 billion words. The authors evaluated the model on three NLP downstream tasks: SA, question answering, and NER. The performance of AraBERT was compared with that of multilingual BERT from Google and other state-of-the-art approaches. The results showed that the newly developed AraBERT achieved state-of-the-art performance on most tested Arabic NLP tasks [48].

3) MARBERT MODEL

MARBERT is an Arabic-focused transformer language model developed in 2021. Unlike AraBERT, MARBERT is trained using data from the Twitter platform (one billion Arabic tweets), which includes both MSA and diverse Arabic dialects. MARBERT uses the same network architecture as the BERT model, but excludes the next sentence prediction objective because of the word count limit in tweets. MARBERT was evaluated using six NLP tasks: sentiment analysis, topic classification, dialect identification, question answering, NER, and social meaning. According to [44], the results of these six tasks showed that MARBERT was significantly better than AraBERT.

4) QARIB MODEL

QARiB is a pretrained model developed in 2021 [56]. The authors trained five BERT models on different sizes of training sets, different linguistic preprocessing, and different text dialects: MSA formal and informal Arabic dialects. The MSA texts include data extracted from newswire sources, online Arabic newspaper websites, and movie and TV subtitles, whereas the dialect text includes Twitter data. The corpus contained 180 M sentences and 440 M tweets composed of 2.7 B words. According to [56], QARiB achieved state-of-the-art results on several tasks such as emotion, NER, and offensive aspects.

C. ABSA TASKS

Our research objective was to develop an Aspect-Based Sentiment Analysis system for KAU that facilitates the analysis of Arabic survey responses. To achieve this, we conducted experiments to determine the most suitable model for our task. Using the PyABSA framework [41], we evaluated the performance of three models: FAST-LCF-ATEPC (multilingual), FAST-LCF-ATEPC (multilingual-256), and FAST-LCF-ATEPC (multilingual-256-2). These models are designed to perform both T1, which involves identifying the aspect term, and T2, which involves assigning its polarity, simultaneously. In addition, we fine-tuned three pre-trained language models designed for Arabic language NLP tasks,

TABLE 8. Snapshot of the output results.

Response	Aspect	Polarity	Category
أشعر بالسعادة والفخر كوني أحد منتسبي هذه الجامعة	الجامعة	Positive	Impression of the University and personal skills
عدم نظافة الاكل في الكفتريات	الاكل	Negative	University infrastructure and public services

namely AraBERT [48], MARBERT [44], and QARiB [56]. The base architecture of these three models remains the same, including the tokenization of input text into subwords or tokens. To fine-tune the models for the aspect extraction task (T1), a new task-specific token classification head called named entity recognition (NER) is added. NER is a technique used in NLP to automatically find and categorize names, words, or phrases in text that refer to real objects such as people, groups, places, dates, amounts, etc. This additional layer is responsible for predicting the NER label for each token in the input. Labeled annotated datasets are required for fine-tuning, where each word in the text is labeled with its corresponding label, such as 'ASP' for aspect and 'O' for others. We utilized the reference multilingual ABSA dataset (SemEval2016-ABSA for Task 5) with 9620 examples [60] for both training and validation purposes. During the fine-tuning process, the cross-entropy loss function was used to measure the dissimilarity between the predicted probabilities of each token's label and the actual labels. This loss is then backpropagated through the network to update the token classification head weights. To avoid overfitting, we monitor the model's performance on the validation set and adjust parameters accordingly. Default hyperparameters were used for all models with an embedding size (100), batch size (32), epochs (8) for optimal performance, and learning rate (5e-5). After completing the fine-tuning process, the model becomes equipped to perform the aspect extraction tasks. Figure 2 illustrates the four tasks we performed to develop an end-to-end ABSA system in this study. The input was Arabic survey responses. The first task (T1) involved extracting aspects from each response. We conducted six experiments to examine the performance of these models to accomplish this task. Then, the best-performing model, the one that has the highest F1-score for extracting aspects from the responses, was used in (T2), which involved identifying the polarity that is associated with each aspect. Following that, we executed the third task (T3), which involved mapping each extracted aspect to a category. We used a predefined list of categories and their associated aspects curated from the golden dataset to accomplish this task. Once we completed task 3 for all responses, we presented the extracted aspects, polarity, and category for each response as shown in Table 8. In the final task (T4), we aggregated the results for each category by counting the polarities of their related aspects. This allowed us to assign an overall polarity for each category.

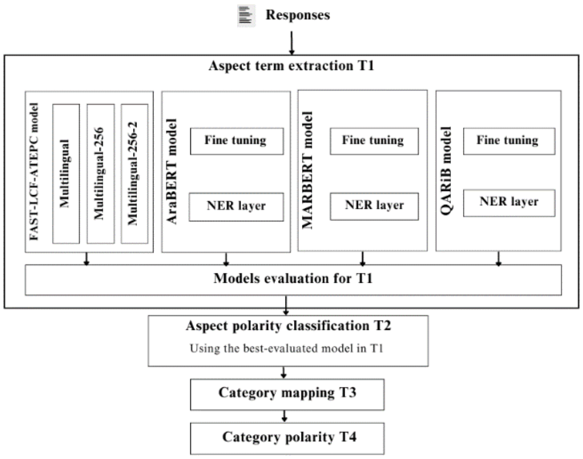

FIGURE 2. An end-to-end ABSA framework used in the study.

D. PERFORMANCE MEASURES

In this study, various evaluation metrics were used. For T1, because aspects were extracted directly from the responses and not from a predefined list, message understanding conference (MUC) metrics were used [61] to obtain detailed results. MUC represents one of the earliest and longest-running efforts to evaluate language-understanding technologies. It is particularly useful for text processing problems such as sentiment analysis and information extraction [62]. MUC considers different categories of errors: correct (COR), incorrect (INC), partial (PAR), missing (MIS), and spurious (SPU). These metrics were defined by comparing the responses of a model against golden annotation. We used COR, INC, MIS, and SPU metrics and eliminated the PAR metric because we considered PAR to be COR in our case. For example, the aspect "university" was considered COR as long as it is part of the actual aspect "KAU university" in the golden dataset and does not need to be identical. Recall (R), precision (P), and F score (F1) were calculated as secondary metrics from MUC-5, as these metrics are commonly used for comparison of models, as shown in (1)–(3):

$$Recall(R) = \frac{COR}{possible} \tag{1}$$

$$Precision\ (P) = \frac{COR}{actual} \tag{2}$$

$$Fscore(F1) = \frac{2 * (PR)}{(P + R)} \tag{3}$$

where:

$possible = COR + INC + MIS$

$actual = COR + INC + SPU$

For the polarity classification (T2) and category mapping (T3), a confusion matrix was used to report the detailed performance of the classification tasks. From the confusion matrix, four commonly used classification metrics were computed: P, R, F1, and accuracy (Acc) [63]. The overall category polarity (T4) is a summation of the polarities that belong to

TABLE 9. Summary of T1 experiment results.

	FAST-LCF-ATEPC					
	Multilingual	**Multilingual-256**	**Multilingual-256-2**	**AraBERT**	**MARBERT**	**QARiB**
Responses (#)	314	297	341	**368**	319	328
COR (#)	**205**	172	201	179	145	146
INC (#)	55	66	71	**46**	58	66
MIS (#)	134	151	107	**80**	129	120
SPU (#)	**54**	59	69	143	116	116
R (%)	0.52	0.44	0.53	**0.59**	0.44	0.44
P (%)	**0.65**	0.58	0.59	0.49	0.45	0.45
F1 (%)	**0.58**	0.50	0.55	0.54	0.44	0.44

the same category, which allows for an overall result representation.

VI. EXPERIMENTAL RESULTS

The experiments were performed on an educational dataset with 448 responses to evaluate T1. The first three experiments used the FAST–LCF–ATEPC model. Each of these experiments used different checkpoints: multilingual, multilingual-256, and multilingual-256-2. The remaining three experiments used AraBERT, MARBERT, and QARiB, respectively using default hyperparameters. The experiments were implemented in Python. PyTorch was used as the deep learning framework. A snapshot of the output results is shown in Table 8.

A. ASPECT EXTRACTION RESULTS

The results of the six experiments are presented in Table 9. Based on the experiments, we've found that the FAST-LCF-ATEPC (multilingual) model has shown promising results for T1 with an F1 score of 0.58, precision of 0.65, and recall of 0.52. The reason behind this could be the large embedding size layer that allows for more expressive representations because it provides a higher-dimensional space in which tokens can be represented. This higher dimensionality enables the model to capture more nuanced relationships and semantic information between words. While the model was successful in extracting aspects from 314 out of 448 responses, there were 134 responses from which no aspects could be extracted.

Upon evaluating the MUC-5 metrics, we found that the model accurately extracted aspects from 205 responses, while 55 contained incorrect aspects and 54 contained spurious aspects that were not in the golden dataset. We believe that with fine-tuning and data augmentation, the model can be further improved to extract aspects from the remaining responses to achieve better results.

AraBERT, on the other hand, was able to extract aspects for the largest number of responses, which could be due to its tailored training for the Arabic language and its ability to capture the unique nuances of the language. However, it also extracted a high number of spurious aspects, leading to a lower precision and F1 score. MARBERT and QARiB had lower performance, which could be due to the original dataset used in the pre-trained models, which included various dialects in addition to formal Arabic. Overall, these observations highlight opportunities for further improvement in aspect extraction for the Arabic language in the educational domain.

In our case, the domain we are working on is relatively unexplored and as mentioned in [64], no technique can guarantee good performance in all domains. For that, we have opted not to compare with existing work in different domains to avoid any potential inaccuracies. Regarding comparison with the educational domain existing work, there was only one study that used Arabic language data collected from Twitter and applied the SVM method, which is completely different from the method used in this study. Nonetheless, we compared various BERT-based methods on our dataset, which consists of Arabic text derived from the educational domain to help us determine the best method for Arabic ABSA related to education.

B. POLARITY CLASSIFICATION RESULTS

The polarity classification task determined the polarity of each extracted aspect. For this task, the FAST-LCF-ATEPC (multilingual) model was used because it achieved the best results for T1. The model results are as follows. Table 10 shows that 29% of the aspects extracted by the model were positive, 70% had a negative polarity, and 1% had a neutral polarity. Compared with the 300 matched aspects in the golden dataset, 14% of the aspects had a positive polarity, and 86% had a negative polarity. Since there was no

TABLE 10. Polarity classification statistics.

Aspects Polarity	FAST-LCF-ATEPC Model	Golden Annotated Data
positive	88	42
negative	209	258
neutral (removed)	3	0
total number of correctly extracted aspects	300	300

TABLE 11. Precision, recall, F1, accuracy, and weighted average of polarity classification.

	Precision	Recall	F1 score	Accuracy	Weighted avg
Negative	1.00	0.82	0.90	0.84	0.86
Positive	0.47	0.98	0.63		

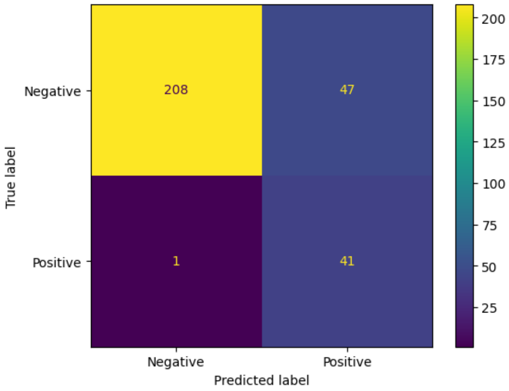

FIGURE 3. Confusion matrix of model polarity classification results.

neutral polarity in the golden dataset, the neutral aspects were removed from the results.

The model was then reevaluated using a confusion matrix, P, R, Acc, and F1, as shown in Figure 2 and Table 11. In the confusion matrix, rows represent the actual number of aspects with negative polarity and those with positive polarity in the golden dataset, whereas columns represent the polarity of the aspects predicted by the FAST-LCF-ATEPC (multilingual) model. As shown, the data were unbalanced, with 255 negative aspects and 42 positive aspects. As per the model prediction, only one aspect was incorrectly classified as negative, and 47 aspects were classified incorrectly as positive.

As shown in Table 11, the accuracy of the model is 84%. For negative polarity, the precision was 100%, recall was 82%, and the F1 score was 90%. For positive polarity, the precision was 47%, recall was 98%, and the F1 score was 63%.

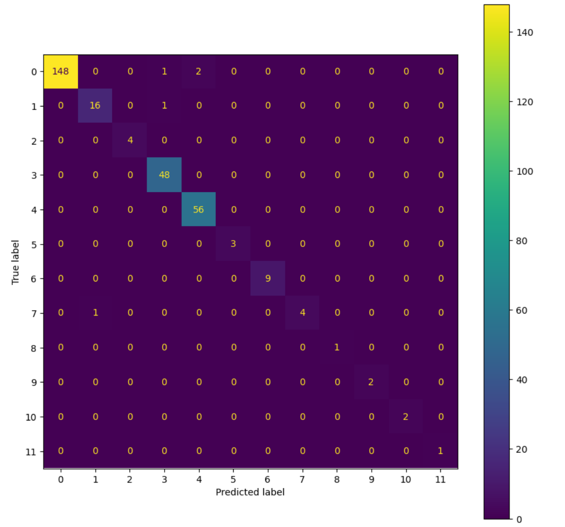

FIGURE 4. Confusion matrix of model category mapping result.

TABLE 12. Summarization of the overall polarity for each category.

Category ID	#Positive	#Negative	Overall Polarity
1	0	5	100% negative
2	0	1	100% negative
3	13	42	76% negative
4	33	15	69% positive
5	0	2	100% negative
6	0	1	100% negative
7	0	2	100% negative
8	5	12	71% negative
9	4	5	56% negative
10	0	4	100% negative
11	0	3	100% negative
12	32	117	78% negative
13	0	4	100% negative

C. CATEGORY MAPPING RESULTS

In this task, each aspect extracted by the model was mapped to a category. There were 13 categories, including university infrastructure and public services, medical administration and its services, and libraries and their services. To achieve this, a predefined list of categories was constructed from the golden dataset. The assigned category was evaluated against the human-assigned category for each sample.

In the confusion matrix, Figure 3, the rows represent the actual categories in the golden dataset, and the columns represent the same categories assigned using the predefined list. As shown, the data were unbalanced. The confusion matrix result shows that for category (0), 148 aspects were labeled correctly, and three aspects were incorrectly classified. The overall accuracy for assigning a category for the extracted aspects showed an overall accuracy of 0.98 and a weighted average F1 score of 0.98.

D. CATEGORY POLARITY RESULTS

In this section, the results are summarized to provide the overall polarity for each category. Table 12 summarizes the number of positive and negative aspects extracted by the model and the overall polarity of each category. Table 12 can inform decision makers about the services that need to be improved, as people tend to leave written feedback when they want to complain. In this summary, all short responses with fewer than three words were removed from the analysis for two reasons. First, they did not have an explicit aspect, and second, they were typically positive sentiments, such as "Thank you" or "Nothing," which are not valuable to decision makers.

VII. CONCLUSION AND FUTURE WORK

In this study, we evaluate different BERT-based models for Arabic ABSA in the educational domain: FAST-LCF-ATEPC (multilingual), FAST-LCF-ATEPC (multilingual-256), FAST-LCF-ATEPC (multilingual-256-2), AraBERT, MARBERT, and QARiB. These models were fine-tuned using a reference multilingual ABSA dataset (SemEval2016-ABSA for Task 5). Six experiments were performed to determine the best method for extracting the aspect terms. The best result was achieved using the FAST-LCF-ATEPC (multilingual) model. This model performs T1 and T2 simultaneously by extracting aspect terms and classifying their polarities, which is better than pipeline solutions that design different models for each task, in which the output from the T1 model is used as the input for the T2 model, thus potentially propagating errors from one step to another. The end-to-end ABSA system achieved good results for all the four tasks. Future research should explore new methods to improve the aspect extraction task as there is still room for improvement. Other methods for optimizing T2 should be investigated. This study contributes to the body of knowledge by enriching research in the Arabic language as well as the educational field. The system can be used by educational institutions to analyze open-ended Arabic responses more efficiently and improve their services and institutions.

REFERENCES

[1] N. Nikolić, O. Grljević, and A. Kovačević, "Aspect-based sentiment analysis of reviews in the domain of higher education," *Electron. Library*, vol. 38, no. 1, pp. 44–64, Feb. 2020, doi: 10.1108/EL-06-2019-0140.

[2] P. A. Rauschnabel, N. Krey, B. J. Babin, and B. S. Ivens, "Brand management in higher education: The university brand personality scale," *J. Bus. Res.*, vol. 69, no. 8, pp. 3077–3086, Aug. 2016, doi: 10.1016/j.jbusres.2016.01.023.

[3] V. Baburajan, J. d. A. Silva, and F. C. Pereira, "Open-ended versus closed-ended responses: A comparison study using topic modeling and factor analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 4, pp. 2123–2132, Apr. 2021, doi: 10.1109/TITS.2020.3040904.

[4] M. Saarela, J. Lahtonen, M. Ruoranen, A. Mäkeläinen, T. Antikainen, and T. Kärkkäinen, "Automatic profiling of open-ended survey data on medical workplace teaching," *Int. J. Emerg. Technol. Learn. (iJET)*, vol. 14, no. 5, p. 97, Mar. 2019, doi: 10.3991/ijet.v14i05.9639.

[5] B. Liu, "Sentiment analysis and opinion mining," *Synth. Lectures Human Lang. Technol.*, vol. 5, no. 1, pp. 1–167, May 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.

[6] H. Liu, I. Chatterjee, M. Zhou, X. S. Lu, and A. Abusorrah, "Aspect-based sentiment analysis: A survey of deep learning methods," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 6, pp. 1358–1375, Dec. 2020, doi: 10.1109/TCSS.2020.3033302.

[7] K. Vivekanandan et al., "Aspect based sentiment analysis survey," *Int. J. Comput. Appl.*, vol. 106, no. 3, 2016. [Online]. Available: https://www.iosrjournals.org/iosr-jce/papers/Vol18-issue2/Version-1/D018212428.pdf, doi: 10.9790/0661-18212428.

[8] K. Schouten and F. Frasincar, "Survey on aspect-level sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 813–830, Mar. 2016, doi: 10.1109/TKDE.2015.2485209.

[9] M. AL-Smadi, O. Qwasmeh, B. Talafha, M. Al-Ayyoub, Y. Jararweh, and E. Benkhelifa, "An enhanced framework for aspect-based sentiment analysis of Hotels' reviews: Arabic reviews case study," in *Proc. 11th Int. Conf. Internet Technol. Secured Trans. (ICITST)*, Dec. 2016, pp. 98–103, doi: 10.1109/ICITST.2016.7856675.

[10] E. Alsharhan and A. Ramsay, "Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition," *Lang. Resour. Eval.*, vol. 54, no. 4, pp. 975–998, Oct. 2020, doi: 10.1007/s10579-020-09505-5.

[11] K. Shaalan, S. Siddiqui, M. Alkhatib, and A. Abdel Monem, "Challenges in Arabic natural language processing," in *Computational Linguistics, Speech and Image Processing for Arabic Language* (Series on Language Processing, Pattern Recognition, and Intelligent Systems), vol. 4. Singapore: World Scientific, 2017, pp. 59–83, doi: 10.1142/9789813229396_0003.

[12] S. Trigui, I. Boujelben, S. Jamoussi, and Y. B. Ayed, "ADAL system: Aspect detection for Arabic language," in *Advances in Intelligent Systems and Computing*. Cham, Switzerland: Springer, 2021, pp. 31–40, doi: 10.1007/978-3-030-49336-3_4.

[13] A. Sabeeh and R. K. Dewang, "Comparison, classification and survey of aspect based sentiment analysis," in *Communications in Computer and Information Science*. Singapore: Springer, Jul. 2019, pp. 612–629, doi: 10.1007/978-981-13-3140-4_55.

[14] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: A comparative review," *Expert Syst. Appl.*, vol. 118, pp. 272–299, Mar. 2019, doi: 10.1016/j.eswa.2018.10.003.

[15] L. Deng, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, Jun. 2014, doi: 10.1561/2000000039.

[16] A. Kumar and A. Sharan, "Deep learning-based frameworks for aspect-based sentiment analysis," Tech. Rep., 2020, pp. 139–158, doi: 10.1007/978-981-15-1216-2_6.

[17] R. Liu, Y. Shi, C. Ji, and M. Jia, "Special section on advanced optical imaging for extreme environments a survey of sentiment analysis based on transfer learning," Tech. Rep., 2019, doi: 10.1109/ACCESS.2019.2925059.

[18] H. Gandhi and V. Attar, "Transfer learning for aspect term polarity determination," *Solid State Technol.*, vol. 63, pp. 956–968, Oct. 2020.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL HLT Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. Conf.*, vol. 1, Oct. 2018, pp. 4171–4186.

[20] M. Sivakumar and U. S. Reddy. *Aspect Based Sentiment Analysis of Students Opinion Using Machine Learning Techniques*. Accessed: Sep. 25, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8365231

[21] H. Ismail, A. Khalil, N. Hussein, and R. Elabyad, "Triggers and tweets: Implicit aspect-based sentiment and emotion analysis of community chatter relevant to education Post-COVID-19," *Big Data Cognit. Comput.*, vol. 6, no. 3, p. 99, Sep. 2022, doi: 10.3390/bdcc6030099.

[22] L. Balachandran and A. Kirupananda. *Online Reviews Evaluation System for Higher Education Institution: An Aspect Based Sentiment Analysis Tool*. Accessed: Sep. 25, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8294118

[23] G. S. Chauhan, P. Agrawal, and Y. K. Meena, "Aspect-based sentiment analysis of students' FeedBack to improve teaching-learning process," in *Information and Communication Technology for Intelligent Systems* (Smart Innovation, Systems and Technologies), S. C. Satapathy and A. Joshi, Eds. Singapore: Springer, 2019, pp. 259–266, doi: 10.1007/978-981-13-1747-7_25.

[24] I. Sindhu, S. M. Daudpota, K. Badar, M. Bakhtyar, J. Baber, and M. Nurunnabi, *Aspect-Based Opinion Mining on Student's FeedBack for Faculty Teaching Performance Evaluation*. Accessed: Sep. 25, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8763969

[25] J. M. Rosalind and S. Suguna, "Predicting students' satisfaction towards online courses using aspect-based sentiment analysis," in *Computer, Communication, and Signal Processing*, E. J. Neuhold, X. Fernando, J. Lu, S. Piramuthu, and A. Chandrabose, Eds. Cham, Switzerland: Springer, 2022, pp. 20–35, doi: 10.1007/978-3-031-11633-9_3.

[26] Z. Kastrati, A. S. Imran, and A. Kurti, *Weakly Supervised Framework for Aspect-Based Sentiment Analysis on Students' Reviews of MOOCs*. Accessed: Sep. 25, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9110884

[27] Y. Heryadi, B. D. Wijanarko, D. F. Murad, C. Tho, and K. Hashimoto, *Aspect-Based Sentiment Analysis for Improving Online Learning Program Based on Student FeedBack*. Accessed: Sep. 25, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9865450

[28] M. Alassaf and A. M. Qamar, *Aspect-Based Sentiment Analysis of Arabic Tweets in the Education Sector Using a Hybrid Feature Selection Method*. Accessed: Sep. 25, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9299026

[29] I. Obaidat, R. Mohawesh, M. Al-Ayyoub, M. AL-Smadi, and Y. Jararweh, "Enhancing the determination of aspect categories and their polarities in Arabic reviews using lexicon-based approaches," in *Proc. IEEE Jordan Conf. Appl. Electr. Eng. Comput. Technol. (AEECT)*, Nov. 2015, pp. 1–6, doi: 10.1109/AEECT.2015.7360595.

[30] M. A. Smadi, I. Obaidat, M. Al-Ayyoub, R. Mohawesh, and Y. Jararweh, "Using enhanced lexicon-based approaches for the determination of aspect categories and their polarities in Arabic reviews," *Int. J. Inf. Technol. Web Eng.*, vol. 11, no. 3, pp. 15–31, Jul. 2016, doi: 10.4018/IJITWE.2016070102.

[31] R. Masadeh, "A hybrid approach of lexicon-based and corpus-based techniques for Arabic book aspect and review polarity detection," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 4336–4340, Aug. 2020, doi: 10.30534/ijatcse/2020/24942020.

[32] S. Areed, O. Alqaryouti, B. Siyam, and K. Shaalan, "Aspect-based sentiment analysis for Arabic government reviews," in *Studies in Computational Intelligence*, vol. 874. Cham, Switzerland: Springer, 2020, pp. 143–162, doi: 10.1007/978-3-030-34614-0_8.

[33] M. Al-Smadi, M. Al-Ayyoub, Y. Jararweh, and O. Qawasmeh, "Enhancing aspect-based sentiment analysis of Arabic Hotels' reviews using morphological, syntactic and semantic features," *Inf. Process. Manage.*, vol. 56, no. 2, pp. 308–319, Mar. 2019, doi: 10.1016/j.ipm.2018.01.006.

[34] M. M. Ashi, M. A. Siddiqui, and F. Nadeem, "Pre-trained word embeddings for Arabic aspect-based sentiment analysis of airline tweets," in *Advances in Intelligent Systems and Computing*. Cham, Switzerland: Springer, Sep. 2019, pp. 241–251, doi: 10.1007/978-3-319-99010-1_22.

[35] 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, Jul. 2018, doi: 10.1016/j.jocs.2017.11.006.

[36] 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, Aug. 2019, doi: 10.1007/s13042-018-0799-4.

[37] S. Al-Dabet, S. Tedmori, and M. AL-Smadi, "Enhancing Arabic aspect-based sentiment analysis using deep learning models," *Comput. Speech Lang.*, vol. 69, Sep. 2021, Art. no. 101224, doi: 10.1016/j.csl.2021.101224.

[38] 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.*, vol. 34, no. 9, pp. 6652–6662, Oct. 2022, doi: 10.1016/j.jksuci.2021.08.030.

[39] M. M. Abdelgwad, T. H. A. Soliman, and A. I. Taloba, "Arabic aspect sentiment polarity classification using BERT," *J. Big Data*, vol. 9, no. 1, p. 115, Dec. 2022, doi: 10.1186/s40537-022-00656-6.

[40] 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, doi: 10.1007/s13278-021-00794-4.

[41] H. Yang and K. Li, "PyABSA: Open Framework for Aspect-based Sentiment Analysis," 2022, *arXiv:2208.01368*.

[42] H. Chouikhi, M. Alsuhaibani, and F. Jarray, "BERT-based joint model for aspect term extraction and aspect polarity detection in Arabic text," *Electronics*, vol. 12, no. 3, p. 515, Jan. 2023, doi: 10.3390/electronics12030515.

[43] A. S. Fadel, O. A. Abulnaja, and M. E. Saleh, "Multi-task learning model with data augmentation for Arabic aspect-based sentiment analysis," *Comput., Mater. Continua*, vol. 75, no. 2, pp. 4419–4444, 2023, doi: 10.32604/cmc.2023.037112.

[44] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," Jun. 2021, *arXiv:2101.01785*.

[45] 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, doi: 10.1109/UCC.2015.78.

[46] H. Yang, B. Zeng, J. Yang, Y. Song, and R. Xu, "A multi-task learning model for Chinese-oriented aspect polarity classification and aspect term extraction," 2019, *arXiv:1912.07976*.

[47] A. Boumhidi, A. Benlahbib, and E. H. Nfaoui, "Cross-platform reputation generation system based on aspect-based sentiment analysis," *IEEE Access*, vol. 10, pp. 2515–2531, 2022, doi: 10.1109/ACCESS.2021.3139956.

[48] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," 2020, *arXiv:2003.00104*.

[49] A. S. Fadel, M. E. Saleh, and O. A. Abulnaja, "Arabic aspect extraction based on stacked contextualized embedding with deep learning," *IEEE Access*, vol. 10, pp. 30526–30535, 2022, doi: 10.1109/ACCESS.2022.3159252.

[50] M. A. Almasre. *Enhance the Aspect Category Detection in Arabic Language using AraBERT and Text Augmentation | IEEE Conference Publication | IEEE Xplore*. Accessed: Oct. 3, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10067648

[51] R. Bensoltane and T. Zaki, "Comparing word embedding models for Arabic aspect category detection using a deep learning-based approach," in *Proc. E3S Web Conf.*, vol. 297, 2021, Art. no. 01072, doi: 10.1051/e3sconf/202129701072.

[52] H. Chouikhi, F. Jarray, and M. Alsuhaibani, "A sequence-to-sequence neural network for joint aspect term extraction and aspect term sentiment classification tasks," Tech. Rep., 2023, p. 123, doi: 10.5220/0011620500003393.

[53] A. S. Dahou, "Aspect-based sentiment classification model employing dialect normalization and deep learning," M.S. thesis, Dept. Math. Comput. Sci., Univ. Ahmed DRAIA Adrar, Adrar, Algeria, 2022. Accessed: Oct. 04, 2023. [Online]. Available: https://dspace.univ-adrar.edu.dz/jspui/handle/123456789/7924

[54] I. Al-Jarrah, A. M. Mustafa, and H. Najadat, "Aspect-based sentiment analysis for Arabic food delivery reviews," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 7, pp. 1–18, Jul. 2023, doi: 10.1145/3605146.

[55] A. Israeli, A. Naaman, Y. Nahum, R. Assi, S. Fine, and K. Bar, "Love me, love me not: Human-directed sentiment analysis in Arabic," in *Proc. 3rd Int. Workshop NLP Solutions Under Resourced Lang. (NSURL) Co-Located With ICNLSP*, Trento, Italy: Association for Computational Linguistics, Dec. 2022, pp. 22–30. Accessed: Oct. 03, 2023. [Online]. Available: https://aclanthology.org/2022.nsurl-1.4

[56] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-training BERT on Arabic tweets: Practical considerations," 2021, *arXiv:2102.10684*.

[57] R. G. Pontius and M. Millones, "Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment," *Int. J. Remote Sens.*, vol. 32, no. 15, pp. 4407–4429, Aug. 2011, doi: 10.1080/01431161.2011.552923.

[58] P. Rozin and E. B. Royzman, "Negativity bias, negativity dominance, and contagion," *Personality Social Psychol. Rev.*, vol. 5, no. 4, pp. 296–320, Nov. 2001, doi: 10.1207/S15327957PSPR0504_2.

[59] R. F. Baumeister, E. Bratslavsky, C. Finkenauer, and K. D. Vohs, "Bad is stronger than good," *Rev. Gen. Psychol.*, vol. 5, no. 4, pp. 323–370, Dec. 2001, doi: 10.1037/1089-2680.5.4.323.

[60] M. Pontiki, "SemEval-2016 task 5: Aspect based sentiment analysis," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 19–30. Accessed: Jun. 3, 2023. [Online]. Available: http://hdl.handle.net/1854/LU-8131987

[61] N. Chinchor and B. Sundheim, "MUC-5 evaluation metrics," in *Proc. 5th Message Understand. Conf. (MUC-5) Conf. Held* Baltimore, MD, USA, Aug. 1993. Accessed: Jan. 31, 2023. [Online]. Available: https://aclanthology.org/M93-1007

[62] L. Hirschman, "The evolution of evaluation: Lessons from the message understanding conferences," *Comput. Speech Lang.*, vol. 12, no. 4, pp. 281–305, Oct. 1998, doi: 10.1006/csla.1998.0102.

[63] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "Multi-label classifier performance evaluation with confusion matrix," in *Computer Science & Information Technology*. Chennai, India: AIRCC Publishing Corporation, Jun. 2020, pp. 1–14, doi: 10.5121/csit.2020.100801.

[64] R. Hajrizi and K. P. Nuçi, "Aspect-based sentiment analysis in education domain," 2020, *arXiv:2010.01429*.

KHLOUD A. ALSHAIKH received the bachelor's degree in information technology, in 2018. She is currently pursuing the master's degree in information systems with the Faculty of Computing and Information Technology, King Abdulaziz University (KAU). Her research interests include machine learning and deep learning.

OMAIMA A. ALMATRAFI received the B.S. degree in computer science from King Abdulaziz University (KAU), Jeddah, Saudi Arabia, in 2008, and the M.S. degree in information systems and the Ph.D. degree in information technology from George Mason University, USA, in 2013 and 2018, respectively. She is currently an Assistant Professor with the Department of Information Systems, KAU. Her research has been published in several conferences and academic journals. Her research interests include computer-supported collaborative learning, learning analytics, and educational data mining.

YOOSEF B. ABUSHARK is currently an Associate Professor with the Computer Science Department, King Abdulaziz University (KAU). He has been publishing several research outcomes in leading venues. His research interests include software engineering with a focus on engineering intelligent systems and building agent-based simulations.