IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 13, 2021, accepted November 7, 2021, date of publication November 10, 2021,
date of current version November 18, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3127140

Arabic Aspect-Based Sentiment Analysis: A Systematic Literature Review

RUBA OBIEDAT[1], DUHA AL-DARRAS[2], ESRA ALZAGHOUL[1], AND OSAMA HARFOUSHI[1]
[1]King Abdullah II School for Information Technology, The University of Jordan, Amman 11942, Jordan
[2]Software Engineering Department, Bethlehem University, Bethlehem 11407, Palestine

Corresponding author: Ruba Obiedat (r.obiedat@ju.edu.jo)

ABSTRACT Recently sentiment analysis in Arabic has attracted much attention from researchers. A modest number of studies have been conducted on Arabic sentiment analysis. However, due to the vast increase in users' comments and reviews on social media and e-commerce websites, the necessity to detect sentence-level and aspect-level sentiments has also increased. The aspect-based sentiment analysis has emerged to detect sentiments at the aspect level. Few studies have attempted to perform aspect-based sentiment analysis on Arabic texts because Arabic natural language processing is a challenging task and because of the lack of available Arabic annotated corpora. In this paper, we conducted a systematic review of the methods, techniques, and datasets employed in aspect-based sentiment analysis on Arabic texts. A total of 21 articles published between 2015-2021 were included in this review. After analysing these articles, we found a lack of annotated datasets that can be used by researchers. In addition, the used datasets were limited to few fields. This review will serve as a foundation for researchers interested in Aspect-Based Sentiment Analysis, it will assist them in developing new models and techniques to tackle this task in the future.

INDEX TERMS Arabic sentiment analysis, aspect-based sentiment analysis, feature-based sentiment analysis, multi-aspect sentiment analysis, sentiment analysis.

I. INTRODUCTION

In the current century, a massive amount of data is created and added to the web each day. This data contains users' reviews, ratings, and opinions about an issue, place, service, product, and so on. Analysing and understanding this kind of data plays an important role in decision-making in multiple fields [1]. For example, companies are interested in identifying users feedback about their products. Sometimes, a single review may contain multiple aspects (i.e. product colour, product quality, product price) with different polarities (positive, neutral, and negative), as the user might like the quality of a product but dislike its design. However, knowing general feedback about the product doesn't provide enough information to the company to recognise the weaknesses and strengths of its product. Moreover, some of these aspects could be more important and valuable to the company than others in its decisions regarding the product. Considering the previous example, there is a necessity to be able to identify the user's impression for each aspect mentioned in the review.

Sentiment analysis (SA) techniques have been developed to analyse these reviews and understand what impressions they carry. SA can be applied at three different levels; each level has its importance and applications. The three levels are the document level [2], sentence level [3], and aspect level [4]. At the document level, the sentiment polarity is detected for the whole document, while at the sentence level, the polarity is extracted for each sentence in the document. At the aspect level, the polarity is detected for each aspect in the sentence [5]. The first two levels have gained significant attention from researchers, and the majority of studies have focused on these levels. The aspect level is largely ignored because it is associated with more sentiment analysis difficulties and challenges than the other levels [6].

The three main SA approaches that were found in the literature for both English and Arabic languages are the machine learning approach, lexicon-based approach, and hybrid approach as shown in FIGURE 1 [7]. In the machine learning approach, supervised, unsupervised or

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

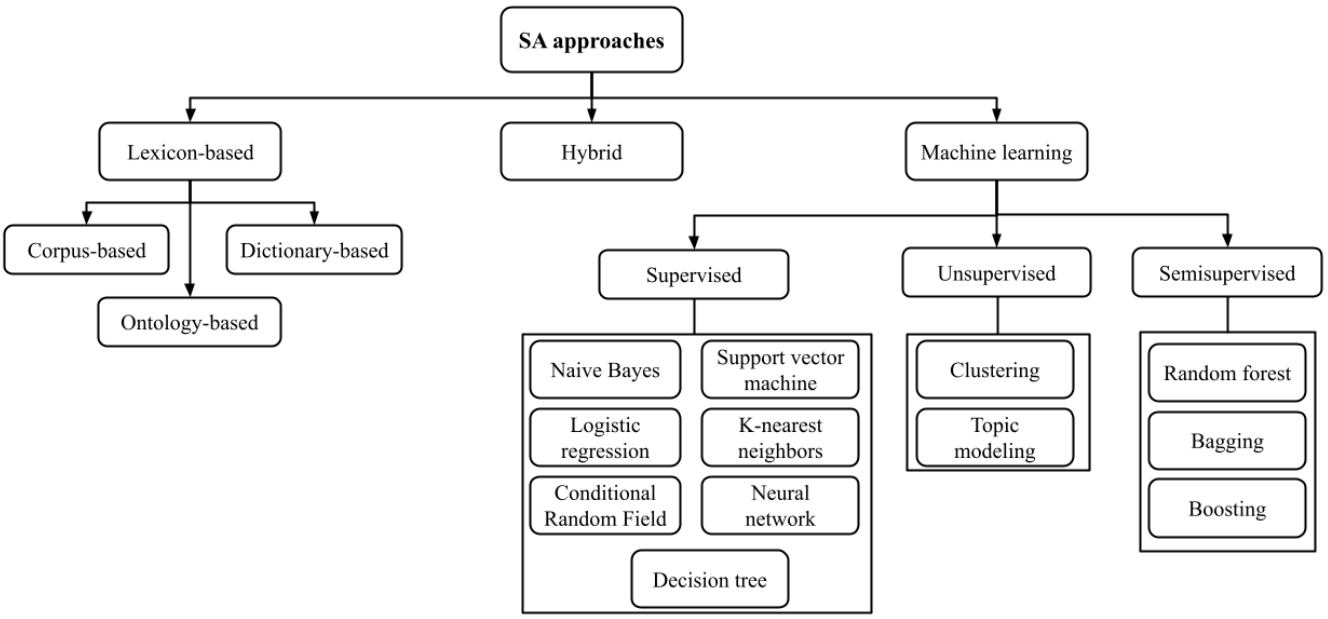

FIGURE 1. Sentiment analysis methods.

semi-supervised classification algorithms can be used to identify the polarity of a given text. Whereas the supervised algorithms require an annotated dataset, unsupervised algorithms don't. The semi-supervised method is a combination of the supervised and unsupervised methods since it deals with labelled and unlabelled data. Most studies that have utilised this approach used supervised algorithms, mainly SVM and NB, to develop their classification models [8]. In the lexicon-based approach, a lexicon is created that contains a set of sentiment words and their polarity values [9]. The two main types of created lexicons are dictionary-based lexicons, which are general-purpose lexicons, and corpus-based lexicons, which are domain-specific [10]. Another type of lexicon is ontology-based lexicons [11], [12]. However, this type of lexicon is still not as widely used as the previous ones because building and utilising an ontology-based corpus is difficult and time-consuming. Finally, the hybrid approach combines machine learning approach techniques and lexicon-based approach techniques [13]. A survey of approaches and techniques used in Arabic SA was done in [14].

According to Feldman [5], the multi-aspect or aspect-based sentiment analysis (ABSA) is defined as follows: ''The research problem that focuses on the recognition of all sentiment expressions within a given document and the aspects to which they refer''. The aspects to be studied could be identified by the researcher based on the dataset domain, for example, in restaurant reviews, the most important aspects are food quality, price, and service quality. The aspects could also be automatically extracted from the data using methods like topic modelling. The main stages in ABSA are: aspect term (expression) extraction and aspect sentiment classification.

In aspect term extraction, the terms related to each aspect are extracted from the text. Meanwhile, in the aspect sentiment classification stage the polarity of each aspect is determined [15]. In the aspect sentiment classification stage, the methods and approaches that are used for SA are also applied. FIGURE 1 summarises these methods. In the aspect term extraction stage, four main approaches could be used to extract the aspect-related terms. The first one is extraction based on frequent nouns and noun phrases, where a part-of-speech (POS) tagger is used to recognise these nouns; after that, only frequent ones are kept. The second approach is extraction by exploiting opinion and target relations. Since sentiment terms are usually used to describe an aspect, the nearest noun or noun phrase to each sentiment word is considered an aspect, and, thus, that sentiment word is a term for the specified aspect. Another approach utilises supervised machine learning algorithms such as hidden Markov models (HMMs) and conditional random fields (CRFs) for terms extraction. The last approach is using topic modelling algorithms, such as probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA), to model both sentiment words (aspect terms) and topics (aspects) at the same time [15], [16].

A modest number of studies have been conducted on Arabic sentiment analysis [17], but few have been done on Arabic ABSA. This is because Arabic ABSA is a more challenging task than Arabic SA. One of these challenges is the need to identify the terminologies for each aspect in the sentence. Another challenge is the lack of standard available annotated corpora, which means researchers need to make an extra effort when working on Arabic ABSA [3], [18]. In addition to these challenges the complexity of the Arabic language itself makes Arabic text processing and feature extraction a

difficult task. This difficulty comes first from Arabic language morphology, as Arabic is a high derivational language, meaning that multiple words with different meanings have the same root or stem. Another reason for the complexity is the use of short vowels, which affects the phonetics and the meanings of words. Also, the Arabic language's richness (e.g. its many synonyms, where different words have the same or similar meanings) contributes in the difficulty [19]. Lastly, the vast difference between Modern Standard Arabic (MSA) and Dialectical Arabic (DA) presents a significant challenge. MSA is used in formal situations - for instance, in education, in books, and on the news. MSA is a standardised version that is the same in all Arabic spoken countries. On the other hand, DA is primarily used by Arabs in daily life, in songs, on TV shows, on social media, and on microblogging channels. DA is dependent on the region, as each region has a unique dialect. Some examples of dialects are Egyptian Arabic, Gulf Arabic, Iraqi Arabic, and Levantine (including the dialects of Lebanon, Syria, Jordan, and Palestine). Dialects differ from each other in their vocabularies, pronunciations, and even grammatical rules [20].

The purpose of this paper is to survey all studies that have been published on Arabic ABSA. In this survey, various techniques used in ABSA are identified and categorised with brief details. Also, the datasets used in ABSA are discussed in terms of the annotation process, language type, and the dataset domain, which presents the target field of the study and the source from where the reviews dataset was collected such as hotels, restaurants, airlines, telecommunication, books and other. This survey reveals the lack of ABSA studies in a variety of important domains. Moreover, the survey provides a basis for future studies on Arabic ABSA.

The remainder of this paper is structured as follows. In the Section 2, the methodology used to conduct this review is explained. Then, Section 3 summarises and explains the selected articles. Section 4 provides answers to the research questions. This is followed by the discussion in Section 5. Finally Section 6 presents the study's conclusions.

II. SYSTEMATIC LITERATURE REVIEW METHODOLOGY

In this research, a systematic review will be conducted based on the guidelines proposed in [22]. These guidelines consist of three main stages: planning, execution and reporting of the review.

A. REVIEW PLANNING

The planning stage consists of four main components: the research question, keywords, query string, and the selection of search sources.

1) Research questions: The aim of this systematic review is to answer the following questions.
Question1: What are the main ABSA datasets' domains, i.e. what areas do the collected reviews cover (hotels, books, news, education, etc.)?.
Question2:What approaches and algorithms are used in Arabic ABSA?
Question3: What frequent features are used in Arabic ABSA?
Question4: What are the evaluation criteria of techniques used in Arabic ABSA?
2) Keywords: A set of English keywords will be considered in this research, with their singular (S) and plural (P) forms, if applicable. These words are multi-aspect (S/P), sentiment (S/P), aspect-based (s), and Arabic (S).
3) Query string: All keywords mentioned above will be used to formulate the search query, which will be used to search through the search engines of the selected sources. The search string will be as follows: (''multi-aspect'' OR ''multi-aspects'' OR ''aspect-based'' OR ''feature-based'') AND (''sentiment'' OR ''sentiments'' OR ''opinion'' OR ''opinions'') AND ''analysis'' AND ''Arabic''.
4) Selection of search sources: The selection criteria of search sources was dependent on the databases, the search engines of which were accessible through the available subscriptions. The available subscriptions cover full text access. These sources were Web of Science, Scopus, IEEE Xplore, and ACM Digital Library. All the articles present in these sources that are related to Arabic ABSA are written in English, thus the review will only consider articles written in English language.

B. STUDY SELECTION CRITERIA (EXECUTION)

After carrying out the search query, the study selection criteria were needed to identify articles that are related to the search question. The selection criteria consisted of the inclusion and exclusion criteria that are explained below. However, duplicate articles that appear in the search result of more than one source will be included only once. FIGURE 2 illustrates the sequential application of the inclusion and exclusion criteria in the article selection process, these criteria were adopted from [21].

The inclusion criteria (IC) for selecting related articles were applied sequentially as listed below:

- IC 1: Articles whose title maintains a relationship with some or all the keywords used in the search.
- IC 2: Articles whose keywords are a subset of the keywords used in the search.
- IC 3: Articles whose abstracts describe multi-aspect sentiment analysis (or its synonym, ABSA).
- IC 4: Articles that propose a new model or techniques for multi-aspect analysis or that modify an existing one.
- IC 5: Articles that employed already-existing multi-aspect techniques in their experiments.

The exclusion criteria (EC) for excluding the unrelated articles were as follows:

- EC 1: Articles that do not contain experiments on the Arabic language.
- EC 2: Articles that only proposed or used techniques for basic sentiment analysis.
- EC 3: Articles that do not comply with any inclusion criteria.

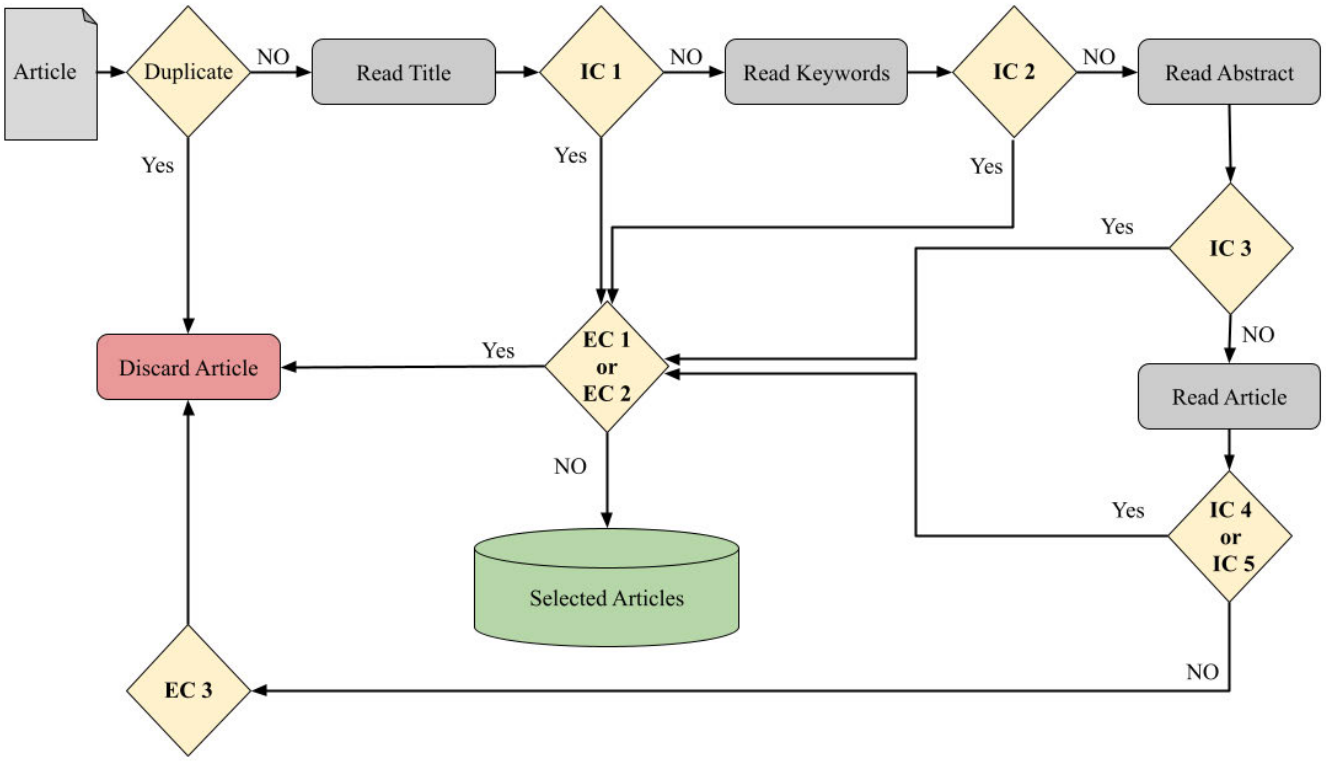

FIGURE 2. Articles selection process flowchart adopted from [21].

TABLE 1. Number of retrieved and selected articles by source.

Search source	Number of articles retrieved		
	Based on search query terms	After Removing duplicate	For final analysis
IEEE Xplore	9	9	9
Scopus	38	23	10
Web of science	21	2	1
ACM Digital Library	53	52	1
Total	**121**	**86**	**21**

C. BIBLIOGRAPHY MANAGEMENT AND DOCUMENT RETRIEVAL

We used Zotero 5.0.96.2 to organise and manage all bibliographic information and citations.Then, all studies were scanned by their title, keywords, and abstract. Any studies that needed to be checked for IC 4 and IC 5 were downloaded via the aforementioned subscriptions to complete the selection process. In the end, all the selected articles were also downloaded to be used in the reporting stage of the review. See TABLE 1 for a summary of the numbers of discovered and selected articles for each of the search resources. For more details about the number of selected studies at each step, refer to the PRISMA flowchart provided in FIGURE 3.

III. SUMMARY OF SELECTED RESEARCH STUDIES

The 21 selected articles were analysed in terms of dataset collection and annotation, dataset domain, extracted features, ABSA utilised approach, and results. A summary of the goal and results of these articles (sorted from earliest to most recent) is provided in TABLE 2. All selected articles were published from 2015-2021, justifying the small number of selected articles and reaffirming that ABSA is still an emerging research area, especially in the Arabic language. Moreover, there has been noticeable interest from researchers in Arabic ABSA in the last year (see FIGURE 4).

The research of Al-Smadi *et al.* [24] was the first study on Arabic ABSA. This research includes a dataset that contains 1,513 book reviews written in MSA. These reviews were selected from the large-scale Arabic book review (LABR) dataset [25]. Then, the selected review were annotated manually to be used as a baseline for later studies on Arabic ABSA. The annotation process consisted of four parts: aspect identification, aspect polarity, aspect term identification, and aspect term polarity. The annotated data were provided in XML format. A simple evaluation method based on the Dice coefficient similarity was used. This method's results were as follows: For the aspect term extraction task, the F-1 measure was 0.23, and for the aspect sentiment classification task, the achieved accuracy was 42.57%. This dataset was used later only in another two studies (S2 and S17). In S2 the researchers used the same reviews selected in S1 but they re-annotated the reviews. In S17, the researchers selected and annotated 1,000 review from the same dataset: LABR. The research of Al-Smadi *et al.* [24] did not succeed in providing a baseline for other studies. Therefore, the same researchers conducted a new research work on ABSA [26].

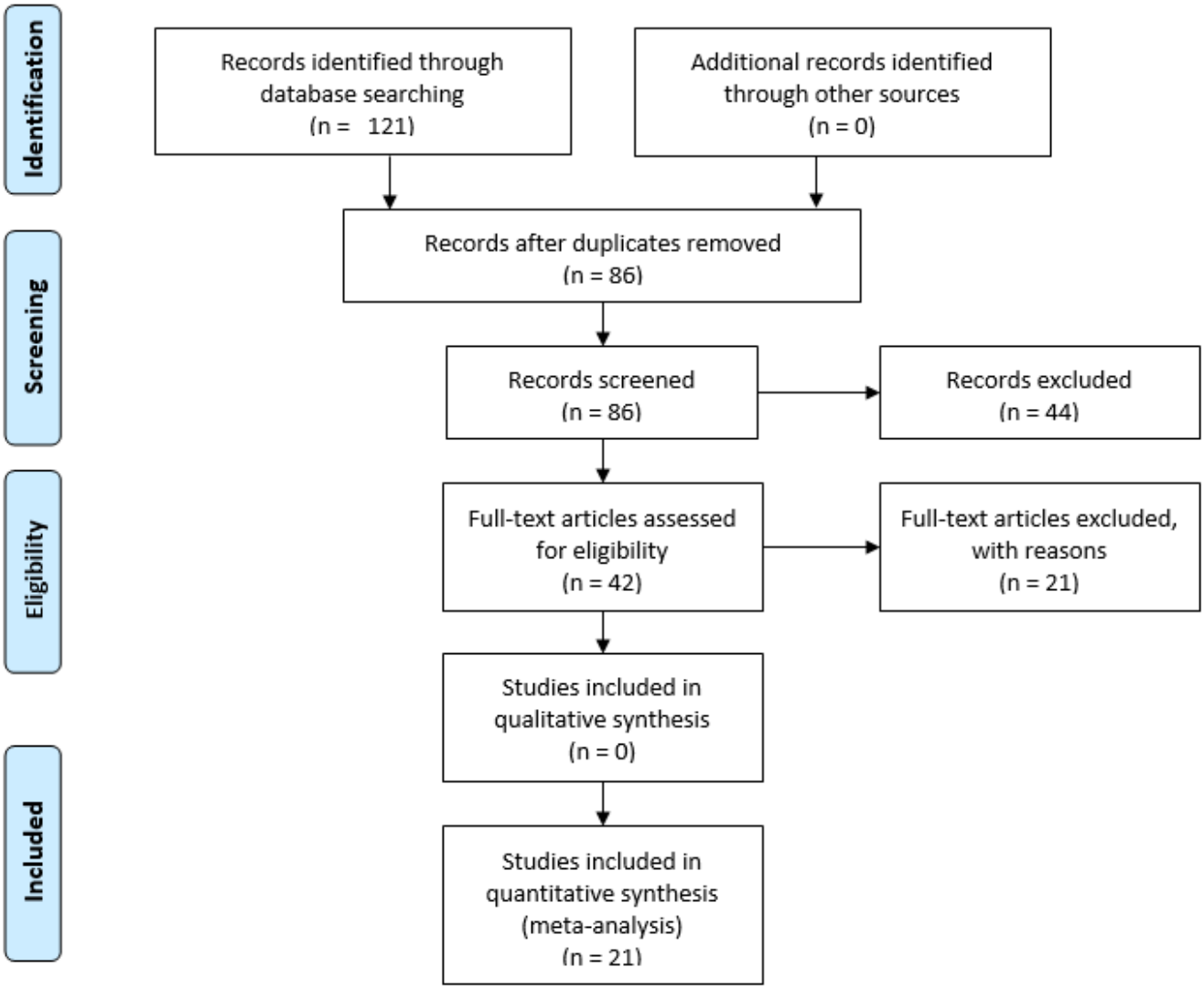

FIGURE 3. PRISMA flowchart adopted from [23].

In the latter research [26], a dataset for Arabic ABSA was prepared via the Semantic Evaluation Workshop 2016 (SemEval 2016) [41]. The dataset contains around 15,562 hotel reviews collected from well-known hotels' booking websites, such as TripAdvisor.com and Booking.com [53]. The reviews were written in both MSA and multiple DAs. A subset of 2,291 reviews was selected and annotated based on the annotation guidelines in [54]. From the selected reviews, 1,839 reviews were used for the training task and 452 were used for the testing task. The annotated dataset was provided in XML format. As a baseline evaluation method the SVM classifier with a linear kernel was trained using N-Unigrams extracted from the training reviews. Then, for the classification decision during the testing phase threshold value of 0.2 was used. For the aspect term extraction task, the model achieved an F-1 value of 0.40; the model's accuracy for the aspect sentiment classification task was 73.2%. The results of this study showed a noticeable improvement over the previous one [24]. This research succeeded in providing a baseline for other researches, and the same annotated dataset was used in the six later studies (S7, S11, S12, S13, S14, and S21). Of these, S11 and S13 achieved the best F-1 measure (0.93) for aspect aspect detection, and the best accuracy (95.4%) for the aspect sentiment classification task.

IV. RESULTS

The 21 selected Arabic ABSA studies consist of 11 journal articles and 10 conference papers (see FIGURE 5). These studies were analysed to extract information related to the research questions. This section is organised to answer the four research questions that were proposed in the review planning section.

TABLE 2. Summary of methodologies and findings of the selected research studies (n = 21).

Study	Methodology / Finding	Study ID
Human Annotated Arabic Dataset of Book Reviews for Aspect Based Sentiment Analysis [24]	This research was the first to study Arabic ABSA. It also provided a benchmark human-annotated Arabic book review dataset, which was manually annotated with aspect terms and their polarities. A total of 1,513 reviews were selected from the LABR dataset [25] and annotated to be used in the training and testing of the model. A simple evaluation method was adopted based on the Dice coefficient similarity measure to compute the distance between sentences in the training and testing datasets.	S1
	The result of this study was a baseline for Arabic ABSA, the F-1 result for the aspect term extraction task was 23%, and the accuracy for aspect sentiment classification was 42.57%.	
Enhancing the Determination of Aspect Categories and Their Polarities in Arabic Reviews Using Lexicon-Based Approaches [27]	This study focused on the lexicon-based approach for both tasks of ABSA; the used lexicons were generated manually. Then, these lexicons were used for aspect category determination and for the classification task. The model was evaluated based on 1,513 book reviews that were selected from the LABR dataset [25] (the same dataset as S1). The selected reviews were reannotated to be used in the ABSA task.	S2
	The best-achieved accuracies were 42.6% for aspect category determination and 71% for aspect category polarity determination. The model developed in this research suffers from drawbacks like the need to use an automatic technique for lexicon creation, in addition to improving the weighting scheme used for the decision-making task.	
A Generic Approach for Extracting Aspects and Opinions of Arabic Reviews [28]	A domain-independent technique was developed by adding opinion tags (polarity and score) to an existing Arabic root based lemmatiser lexicon, both at the root and pattern levels. The lexicon consists of 3,829 roots and 69 patterns. This technique was based on the concept that each opinion has a target aspect or entity. Thus, when an opinion word is detected, the target aspect also exists in the sentence. N-Grams and POS tagger were used to identify the aspects.	S3
	The proposed system was evaluated based on three different datasets containing reviews on multiple domains like restaurants, movies, hotels, products, and novels. The system achieved a recall of 80.8%, precision of 77.5%, and F-Measure of 79.1%. The system showed that a generic sentiment annotated lexicon can be used to detect opinions in the sentence.	
Feature-Based Sentiment Analysis in Online Arabic Reviews [29]	In this research, an automatic model for extracting sentiments and features were developed based on POS tagging features and an online Arabic lexicon. Five rules were proposed for extracting feature-sentiment pairs, and the lexicon was used to extract weights for a classification task.	S4
	The proposed model achieved high accuracy (92.15%) for aspect-term extraction and aspect sentiment classification. However, the proposed model does not consider negation cases or the composite words feature. Moreover, the model was evaluated based on only 200 reviews extracted from forums, Facebook, YouTube, and a Google search.	
An Aspect-Based Sentiment Analysis Approach to Evaluating Arabic News Affect on Readers [30]	The aim of this study was to evaluate news posts' effects on readers by utilising an ABSA approach. A total of 2,265 Arabic news posts related to the Israel-Gaza conflict in 2014 were collected. After preprocessing the collected posts, features like POS, NER, and N-Grams were extracted. The extracted features were used to train the following supervised classifiers: conditional random field (CRF), decision tree (J48), Naïve Bayes, and K-nearest neighbour.	S5
	The results of the proposed model showed that the J48 achieved the best F-1 measure (82%) for the aspect terms extraction task. Meanwhile, CRF and Naïve Bayes performed better in the aspect term polarity identification task with accuracies of 86.5% and 86.47%, respectively. The results also showed that features like POS and NER play a significant role in the evaluation.	

TABLE 2. *(Continued.)* **Summary of methodologies and findings of the selected research studies (n = 21).**

An Enhanced Framework for Aspect-Based Sentiment Analysis of Hotels' Reviews: Arabic Reviews Case Study [26]	An improved version of a model proposed in [24] was developed. For improvement, the SVM classifier with a linear kernel was trained using n-Unigrams extracted from training Hotel's Arabic reviews. For classifier training, the 1,000 most frequent Unigrams were used. However, for the classification decision during the testing phase, a threshold value of 0.2 was used.	S6
	The model's enhancements improved the results compared with the previous related study of Al-Smadi et al. [24]. For the aspect term extraction task, the F-1 score was 40%. On the other hand, the model exhibited an accuracy of 73.2% for the aspect sentiment classification task.	
Semantic Feature Based Arabic Opinion Mining Using Ontology [31]	This study utilised the semantics of ontologies and lexicons for feature term extraction and feature sentiment classification. An ontology for hotels' related features was developed, in addition to reusing and creating a lexicon for terms polarity identification. The ontology was used for feature extraction, and three different configurable N-Grams methods were proposed for feature polarity identification. Also, the proposed model was used for sentence level opinion identification.	S7
	The proposed model achieved good opinion mining results for the whole review, but the results for aspects opinion mining were not as good. The best accuracy achieved for aspect sentiment classification was 67.5%, whereas the best accuracy for opinion mining for the whole review was 95.5%.	
Framework for Affective News Analysis of Arabic News: 2014 Gaza Attacks Case Study [32]	The aim of this study was the same as in the study of Readers [30]. However, while the previous study utilised supervised classifiers, this study used a lexicon-based approach. An aspect term lexicon was manually created for each aspect before being used for the evaluation of both ABSA tasks.	S8
	The results were not as satisfying as those of the previous study. Specifically, for the aspect term extraction task, the achieved F-1 score was 39%; for the aspect sentiment classification task, an accuracy of 74% was achieved.	
Feature-Based Opinion Summarization for Arabic Reviews [33]	A feature-based opinion summarisation model for Arabic hotel reviews was proposed using NLP techniques and sentiment lexicons. The model has three tasks: aspect extraction, opinion identification for each aspect, and generating sentences for each aspect as a summary. Aspects were extracted by identifying frequent and relevant nouns while removing unrelated nouns. Then a domain-specific sentiment lexicon was used.	S9
	The model was the first to extract a feature-based opinion summary from Arabic reviews. Even though the model's results on hotels reviews were promising and achieved an accuracy of 81.48% for feature extraction and 71.2% for sentiment classification.	
Pre-trained Word Embeddings for Arabic Aspect-Based Sentiment Analysis of Airline Tweets [34]	Two-word embedding models were utilised for the ABSA of Arabic Airline tweets. The WE models were used for the aspect detection task, and the SVM classifier was used for the sentiment classification task. The used embeddings models include fastText Arabic Wikipedia [35] and AraVec-Web pre-trained models [36]. A total of 5,000 tweets were collected and annotated manually.	S10
	The fastText Arabic Wikipedia word embeddings model performed slightly better than the AraVec-Web model, with accuracies of 70% for aspect detection and 89% for sentiment polarity detection.	
Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews [37]	Two state-of-the-art approaches based on supervised machine learning were developed to handle the tasks of ABSA. The first approach was based on a deep recurrent neural network (RNN), and the other one was based on a support vector machine (SVM). The approaches were trained on 1,839 Arabic hotel reviews and tested on 452 reviews from the same dataset. Lexical, syntactic, morphological, and semantic features were extracted to train the models. These features included POS, NER, WE, and N-Grams.	S11

TABLE 2. *(Continued.)* **Summary of methodologies and findings of the selected research studies (n = 21).**

	Results showed that SVM outperforms RNN classifiers (F-1 = 93.4% in the aspect category identification task and an accuracy of 95.4% for the sentiment polarity classification task).	
Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews [38]	Two models of deep long short-term memory (LSTM) neural networks were developed for aspect-based sentiment analysis of Arabic hotel reviews. The two models were based on LSTM neural networks. The first one was named bi-directional LSTM and was used along with a conditional random field classifier (Bi-LSTM-CRF) at the character and word levels. This model was developed to extract aspect terms. The second model was aspect-based LSTM for sentiment polarity classification (AB-LSTM-PC). It was developed to handle the sentiment classification task.	S12
	Experimental results showed that the proposed model outperformed the baseline, as the Bi-LSTM-CRF (based on word2vec word embeddings) attained F-1 scores of 66% and 69% based on faestText character embeddings for the aspect term extraction task.	
Enhancing Aspect-Based Sentiment Analysis of Arabic Hotels' reviews using morphological, syntactic, and semantic features [39]	An enhanced approach for ABSA of Arabic hotel reviews was developed based on supervised machine learning. A set of classifiers was used in model development. These classifiers were trained with morphological, syntactic, and semantic features extracted from the reviews. The extracted features include N-Grams, POS, NER, TF, and TF-IDF. The model was evaluated using the Semantic Evaluation 2016 workshop dataset for Arabic hotel reviews.	S13
	The results were satisfactory. The SVM classifier outperformed all other classifiers (F-1= 93% for the aspect category identification task and an accuracy of 95.4% for aspect sentiment classification).	
ADAL System: Aspect Detection for Arabic Language [40]	A supervised system for the ABSA aspect detection task was proposed. The model is based on extracting multiple features from the SemEval 2016 corpus containing hotel reviews [41]. The extracted features were lexical (POS), semantic (NER), syntactic, and numeric features. These features were extracted from 4,802 training reviews. The extracted features are then fed into the following classifiers: Naïve Bayes, decision tree, RepTree, and Adaboost to build a training model.	S14
	The results of the model were promising. Among all classifiers, Adaboost techniques achieved the best results in terms of both precision (97%) and recall (96.9%).	
Aspect-based Sentiment Analysis for Arabic Content in Social Media [42]	An experiment was conducted to compare machine learning and deep learning sentiment analysis approaches on Arabic tweets, in addition to evaluating the impact of using POS tagger and word embedding on the performance of deep learning techniques. The used dataset consisted of 1,098 Arabic tweets about Saudi telecommunication companies.	S15
	The findings showed that the deep learning technique with word embedding method achieved promising results (F-1 = 81%). Also, the results showed that using the Unigram language model is significantly better than using the bigram language model.	
Aspect-Based Sentiment Analysis of Arabic Tweets in the Education Sector Using a Hybrid Feature Selection Method [43]	The researcher analysed real-world tweets about educational institutions to provide a picture of these institutions' strengths and weaknesses. The 7,934 tweets collected for this research were about Qassim University in Saudi Arabia. The proposed model used a hybrid feature selection method to reduce the number of features used during classification.	S16
	The best F1 measure achieved was 70% (for aspect detection), while an F1 score of 87% was achieved for the classification task. The results showed that the hybrid method enhanced the SVM performance both for tasks aspect detection and aspect opinion classification.	

TABLE 2. *(Continued.)* **Summary of methodologies and findings of the selected research studies (n = 21).**

A Hybrid Approach of Lexicon-based and Corpus-based Techniques for Arabic Book Aspect and Review Polarity Detection [44]	A hybrid approach combining a corpus-based approach and a lexical-based approach was developed for the tasks of aspect polarity and review polarity detection. The dataset was created by selecting 1,000 reviews from the LABR dataset [25]. The selected reviews were manually annotated by researchers to be used to test and train the model.	S17
	The proposed approach's experimental results were acceptable, with accuracies of 80.5% for aspect polarity and 78% for review polarity prediction.	
Aspect-Based Sentiment Analysis for Arabic Government Reviews [45]	A lexicon-based and rule-based approaches were adopted for aspect extraction and sentiment classification. A total of 2,071 Arabic government's apps reviews were selected and annotated manually.	S18
	The experimental results showed that for aspect extraction, the best accuracy was 96.6%, with an F-1 score of 92%; for the sentiment classification task, the proposed model achieved an accuracy of 95.8% and an F-1 of 90%.	
Feature-Based Sentiment Analysis for Arabic Language [46]	A lexicon-rule-based approach was developed to analyse phone-related reviews. Three lexicons were extracted: the first one was for phone-related features, the second one was for sentiment words with their polarities, and the last one was for entities. For the aspect sentiment classification task, the polarities were computed. Then, based on a threshold, each aspect class was identified.	S19
	The model was tested on a total of 1,594 phone comments collected from Facebook phone advertisement posts. The result of the model achieved an F-1 score of 88%.	
Towards Semantic Aspect-Based Sentiment Analysis for Arabic Reviews [47]	This paper proposed a model to improve aspect-based sentiment analysis for Arabic reviews in a specific domain. The proposed approach suggested employing an ontology to identify entities, aspects, and opinions. Then, a set of semantic of description logics and linguistic was employed for the identification of opinion targets and their polarity. Linguistic rules are suggested to be used with Arabic sentiment lexicon (ArSenL) [48] for the extraction of opinion targets and the determination of the sentiment polarity for each aspect.	S20
	No evaluation measures were available for this study since it simply proposed the approach. The researchers recommended evaluating the model on the books domain using the Human Annotated Arabic Dataset [24].	
Enhancing Arabic aspect-based sentiment analysis using deep learning models [49]	This research proposed two deep learning models to solve ABSA tasks: aspect-category identification and aspect-sentiment classification. For the first task, a model for aspect-category identification was proposed by combining a convolutional neural network (CNN) [50] with independent long-short term memory (Indy-LSTM) [51] networks. For the aspect sentiment classification task, a classification model was proposed by combining a stacked bidirectional-Indy-LSTM (Bi-Indy- LSTM) network, a position-weighting mechanism, and multiple attention layers [52]. The proposed models were evaluated using the ArabicSemEval-2016 dataset for hotel reviews [41].	S21
	This study was the first to use a stacked Bi-Indy-LSTM network for an aspect sentiment classification task. This network was used to overcome the weaknesses of the standard RNN network. The proposed models outperformed the baseline for both tasks. For the first task, the C-Indy-LSTM model achieved an F-1 score of 58%; for the second task, the Bi-Indy-LSTM model achieved an accuracy of 87.31%.	

A. DATASETS DOMAINS, EXTRACTION, AND ANNOTATION Techniques (QUESTION 1)

After analysing the selected articles, it was found that researchers used multiple datasets in various domains. These domains include hotels, books, products, restaurants, political conflicts, airlines, educational institutes, telecommunication companies, and government applications. For each of these domains (except hotels and books), a single piece of research was conducted. Nine studies used hotel reviews to develop and evaluate their models, and another four used book reviews (see FIGURE 6).

The datasets were either collected and annotated by the researchers or extracted from available corpora. The main resource for the researchers' collected reviews was Twitter

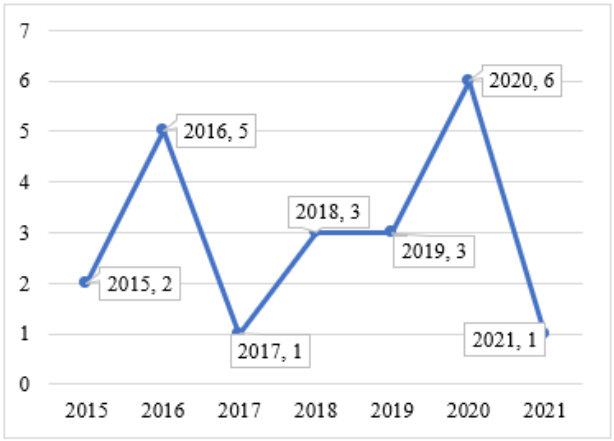

FIGURE 4. Distribution of the Arabic ABSA studies per publication year.

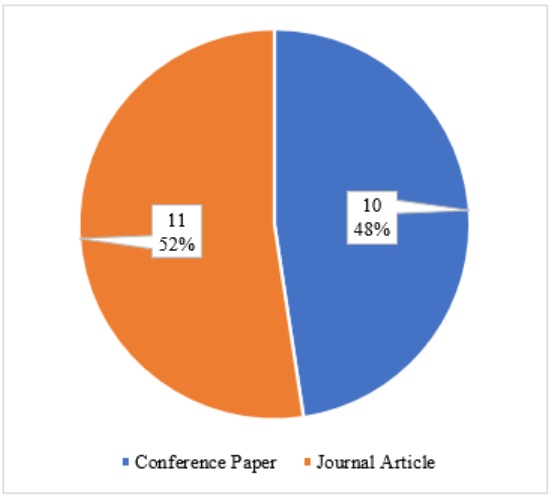

FIGURE 5. Type of selected studies.

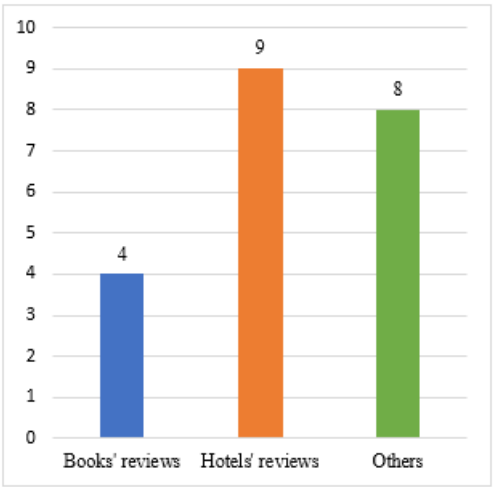

FIGURE 6. Most common datasets domains.

as in S10, S15, and S16. The second resource was Facebook as in S4 and S 9. Other considered resources included forums,

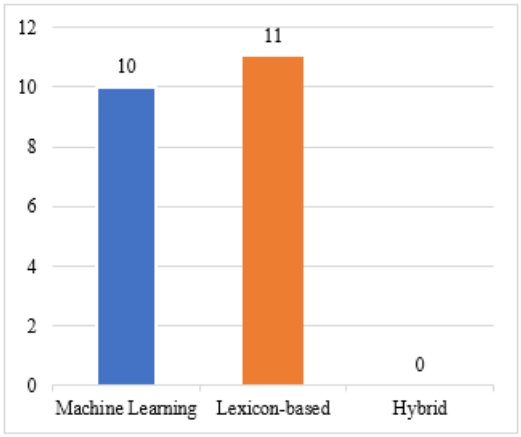

FIGURE 7. Number of studies addressing each approach.

Youtube, and websites. However, most of these datasets are not publicly available.

On the other hand, the hotel reviews dataset provided by Semantic Evaluation Workshop 2016 (task 5) [41] was the resource of reviews dataset for seven studies. This dataset contains a total of 2,291 hotel review; 1,839 reviews were prepared for the training task and 452 were prepared for the testing task. In S6, this dataset was annotated for ABSA tasks; the same annotated dataset was used in S7, S11, S12, S13, S14, and S21. Also, the LABR dataset [25] was a resource for reviews in S1, S2, and S17. This dataset contains more than 63,000 book reviews in Arabic. The book reviews were collected from the Goodreads website during March 2013.

TABLE 3 below provides a summary of the datasets used in the studies in terms of their domain, language type, source, and the number of reviews.

B. TECHNIQUES FOR ARABIC ABSA (QUESTION 2)

As explained in the introduction, several approaches have been used for ABSA. These approaches were classified into three groups lexicon-based, machine learning, and hybrid. In this section, the approaches and algorithms used for Arabic ABSA will be explained. All of the studies applied either machine learning or lexicon-based approach, while none have applied a hybrid approach. As shown in FIGURE 7, 10 studies adopted a machine learning approach, whereas 11 studies adopted a lexicon-based approach.

FIGURE 8 displays the popularity of used machine learning algorithms in the selected articles. The frequencies in the figure stand for the number of studies that applied each algorithm. It can be seen that the SVM was the most used machine learning algorithm, as it was utilised in seven studies. This is followed by the neural network, with its two classes recurrent neural network (RNN) and convolutional neural network (CNN). Other classifiers like Naive Bayes, decision tree, and K-nearest neighbour (K-NN) were also used.

The lexicons used for the lexicon-based approach can be classified into two groups: manually constructed lexicons

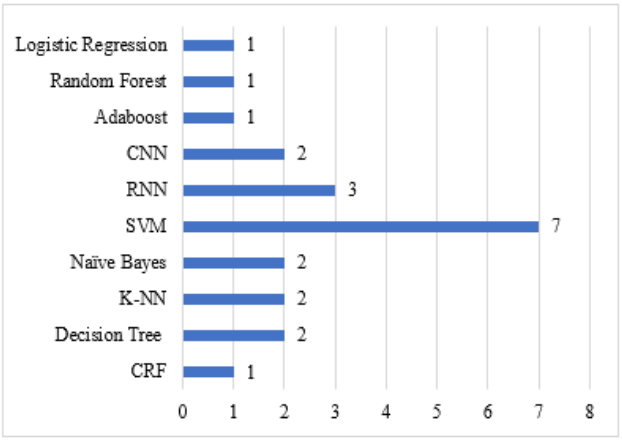

FIGURE 8. Used machine learning algorithms.

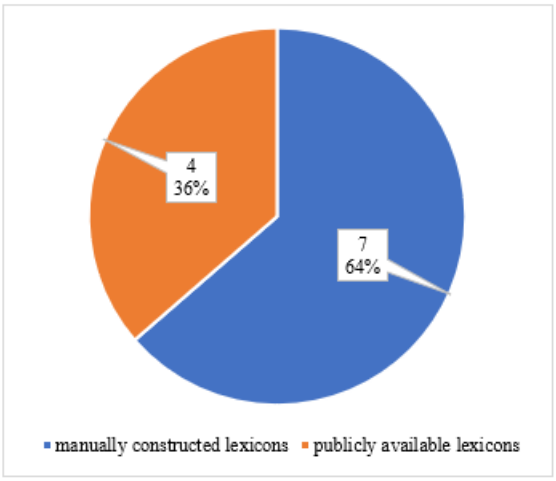

FIGURE 9. Percentages of the used lexicons.

and publicly available lexicons. Only in four studies were previously available lexicons utilised; In seven studies, the researchers built their own lexicons (see FIGURE 9). The three publicly available lexicons that were utilised are the SentiWordNet lexicon [55] (S2), ARBL lexicon [56] (S3), and Arabic Sentiment Lexicon (ArSenL) [48] (S9 and S20).

The SentiWordNet lexicon is a publicly available lexical resource developed for opinion mining and sentiment classification tasks. Each word in this lexicon is assigned three numerical sentiment scores (positivity, negativity, and objectivity), such that the sum of the three scores is 1.0 [55]. These scores are used by researchers for the sentiment classification task. The second used lexicon was the Arabic root based lemmatiser(ARBL) lexicon. It consists of 38,29 roots, 69 patterns, and a closed set of 346 Arabic words. These roots are categorised into 16 groups, some of which are nouns, verbs, numerals, prepositions, and conjunctions. The roots and patterns are assigned with tags that specify whether they are opinion-bearing words or not [56]. Researchers used this lexicon to identify sentiment bearing words. A score is then assigned to each identified word. The last lexicon is the Arabic Sentiment Lexicon (ArSenL). This lexicon was developed based on WordNet, SentiWordNet, and an Arabic morphological analyser. It consists of 28,760 lemmas with their sentiment scores (positivity, negativity, and objectivity) [48]. These scores are used by researchers for sentiment classification.

C. THE FREQUENT EXTRACTED FEATURES USED ARABIC ABSA (QUESTION 3)

Multiple features were extracted from the datasets, such as part of speech (POS) tags, N-Grams, named entity recognition (NER), word embedding (WE), term frequencies (TF), and term frequency-inverse document frequency (TF-IDF). The most dominant features were POS and N-Grams, and the least dominant one was TF. FIGURE 10 shows the frequencies of the extracted features among the 21 selected research works.

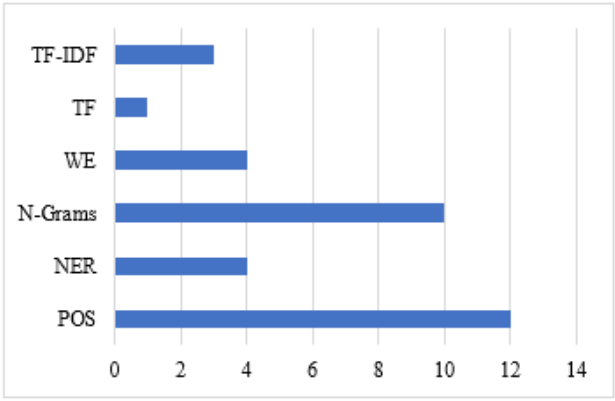

FIGURE 10. Frequent extracted features.

From TABLE 4, it can be concluded that for the lexicon-based approach, POS and N-Grams are the only extracted features, while for the machine learning approach, all six features mentioned above were extracted and used. Multiple tools were used to extract these features. For POS, three different tools were used: the Stanford POS tagger [57] was used in S2, S5, and S9. S4 used ATKS tools for POS tagging [58]. Meanwhile, S11, S13, and S14 used MADAMIRA [59] to extract both POS and NER features. MADAMIRA is a toolkit that provides services for POS tagging, NER, tokenization, diacritic analysis, and lemmatization of Arabic text. Another tool Polyglot-NER was utilised as a web service for NER extraction in S5 [60]. For WE, three tools were used: Word2Vec [61], fastText [35], and AraVec-Web [36]. Word2Vec was utilised in S10 and S11, whereas fastText and AraVec-Web were used in S12. The fastText is an extension of the Word2Vec skip-gram model trained on Arabic Wikipedia with a dimension of 300 and a vocabulary size of 610,977. On the other hand, AraVec-Web is a pre-trained Word2Vec skip-gram on World Wide Web

TABLE 3. Summary of datasets used for evaluation of Arabic MASA methods.

Author/Year	Dataset Domain	Dataset Description	Language	Document Type	ID
Al-Smadi et al. (2015)	Books' reviews	1513 Arabic book reviews were selected and annotated from LABR dataset [25]	MSA	Conference paper	S1
Obaidat et al. (2015)	Books' reviews	Same dataset as study S1	MSA	Conference paper	S2
Ismail et al. (2016)	Reviews about hotels, novels, products, restaurants, and events	Three datasets were used: the first one contains 500 movie reviews collected from different web pages and blogs in Arabic. The second one contains 1000 Arabic reviews about restaurants. The last one contains 500 Arabic reviews in different domains, collected from different websites	MSA, DA	Conference paper	S3
Abd-Elhamid et al. (2016)	User reviews in Multiple domains	200 reviews collected by researchers from Forums, Facebook, YouTube, and google search	DA	Conference paper	S4
Al-Smadi, Al-Ayyoub, et al. (2016)	Arabic news posts related to Israel-Gaza conflict in 2014	A total of 2,265 news posts from well-known Arabic news networks like Al-Jazeera and Al-Arabiya. Collected by researcher	MSA	Journal Article	S5
Al-Smadi, Qwasmeh, et al. (2016)	Hotels' Reviews	A total of 1839 training and 452 testing review were selected from the dataset provided by Semantic Evaluation workshop 2016 task 5 [41]	MSA, DA	Conference paper	S6
Alkadri & ElKorany (2016)	Hotels' Reviews	Same dataset as study S6	MSA, DA	Journal Article	S7
Al-Ayyoub et al. (2017)	Arabic news posts related to Israel-Gaza conflict in 2014	Same dataset as study S5	MSA	Journal Article	S8
El-Halees & Salah (2018)	Hotels' Reviews	Consists of 2860 hotels' related reviews collected by the author	MSA, DA	Conference paper	S9
Ashi et al. (2018)	Saudi Airline service-related tweets	A total of 5000 tweets were collected by author using Twitter API	MSA, DA (Saudi dialect)	Conference paper	S10
Al-Smadi et al. (2018)	Hotels' reviews	Same dataset as study S6	MSA, DA	Journal Article	S11
Al-Smadi, Talafha, et al. (2019)	Hotels' reviews	Same dataset as study S6	MSA, DA	Journal Article	S12

TABLE 3. *(Continued.)* **Summary of datasets used for evaluation of Arabic MASA methods.**

Al-Smadi, Al-Ayyoub (2019)	Hotels' reviews	Same dataset as study S6	MSA, DA	Journal Article	S13
Trigui et al. (2019)	Hotels' reviews	Same dataset as study S6	MSA, DA	Conference paper	S14
Alshammari & AlMansour (2020a)	Arabic tweets about Saudi telecommunication companies	Collected by authors, consists of 1098 tweets	DA	Conference paper	S15
Alshammari & AlMansour, (2020b)	Tweets related to educational institutions	Collected by authors consist of 7,934 tweets about Qassim university in Saudi Arabia	DA	Conference paper	S16
Masadeh & Sa'ad Al-Azzam (2020)	Books' reviews	1000 Arabic book reviews were selected and annotated from LABR dataset [25]	MSA	Journal Article	S17
Areed et al. (2020)	Government mobile app Reviews	A total of 2071 Arabic reviews were selected from Apple Store and Google Play which are related to 60 different mobile apps for governments in United Arab Emirates (UAE)	MSA, DA	Journal Article	S18
Alhamad & Kurdy (2020)	Mobile phones reviews	A total of 85 posts of mobile phones were collected, these posts also include 1024 comments, which include 570 replies obtained from mobile pages like souq.com and 3mobihall.com pages on Facebook	MSA, DA	Journal Article	S19
Behdenna et al. (2020)	Books' reviews (suggested to be used)	Same dataset as study S1	MSA	Journal Article	S20
Al-Dabet et al. (2021)	Hotels' reviews	Same dataset as study S6	MSA, DA	Journal Article	S21

pages' Arabic content with a dimension of 300 and a vocabulary size of 145,428.

In addition to features-extraction tools, a set of preprocessing techniques were utilised by the researchers. Tokenisation and normalisation were the most often-used preprocessing techniques, as these techniques were almost used by all studies. Also, seven studies applied stop words removal before extracting the features. Moreover, word stemming was used in six studies (refer to TABLE 4 for details). The AraNLP tool [62] was utilised in S2, S5, S11, and S13 to handle the required preprocessing steps. The AraNLP tool provides the following preprocessing steps: punctuation, numbers, and non-Arabic word removal, tokenization, tatweel removal, and diacritic removal. Another tool developed by Althobaiti et.al [63] was used in S7 for tokenization, normalization, and stemming.

D. EVALUATION CRITERIA USED FOR ARABIC ABSA (QUESTION 4)

The evaluation criteria used for Arabic ABSA are accuracy, recall, precision, and F1. Accuracy is mainly used for the evaluation of the aspect sentiment classification step, whereas

TABLE 4. Summary of approaches, algorithms, extracted features, and pre-processing steps.

ID	Approach	Algorithms	Features	Pre-processing
S1	Lexicon-based	Manual lexicon construction	N/A	N/A
S2	Lexicon-based	SentiWordNet lexicon [55]	POS	Tokenization, Normalization
S3	Lexicon-based	ARBL lexicon [56]	POS, N-Grams	Lemma generation, Stop words removal
S4	Lexicon-based	Manual lexicon construction	POS	Normalization, Stemming, Stop words removal
S5	Machine Learning	Supervised Conditional Random Fields (CRF), Decision Tree (DT), Naive Bayes (NB), and K-Nearest Neighbour (K-NN)	POS, NER, N-Grams	Tokenization, Normalization
S6	Machine Learning	Support Vector Machine (SVM)	N-Grams	Tokenization, Stop words removal
S7	Lexicon-based	Manual lexicon construction	POS, N-Grams	Normalization,Tokenization, Stemming
S8	Lexicon-based	Manual lexicon construction	POS, N-Grams	Tokenization, Normalization
S9	Lexicon-based	Arabic Sentiment Lexicon (ArSenL) [48]	POS	Tokenization,Normalization, Stemming
S10	Machine Learning	SVM	Word Embeddings (WE), N-Grams, cosine similarity measure	Text pre-processing (no details available)
S11	Machine Learning	SVM, Deep Recurrent neural network (RNN)	N-Grams, POS, NER, WE	Tokenization, Normalization
S12	Machine Learning	Long Short-Term Memory deep neural networks (LSTM)	N-Grams, WE	Text pre-processing (no details available)
S13	Machine Learning	NB, Bayes Networks , DT, K-NN, SVM	N-Grams, POS, NER, TF, TF-IDF	Tokenization, Normalization
S14	Machine Learning	NB, DT, RepTree, Adaboost	POS, NER	N/A

TABLE 4. *(Continued.)* **Summary of approaches, algorithms, extracted features, and pre-processing steps.**

S15	Machine Learning	Logistic Regression (LR), SVM, Random Forest (RF), and Convolutional Neural Network (CNN)	N-Grams, POS, WE, TF-IDF	Removing non-Arabic characters, and numbers, Stop words removal
S16	Machine Learning	SVM	TF-IDF	Tokenization,Normalization, Stemming, Stop words removal
S17	Lexicon-based	Manual lexicon construction	N/A	Normalization, Stop Words removal
S18	Lexicon-based	Manual lexicon construction	N/A	Tokenization,Normalization, Stemming, Stop words removal
S19	Lexicon-based	Manual lexicon construction	N/A	Tokenization, Normalization, Stemming
S20	Lexicon-based	(ArSenL) lexicon	POS	Tokenization, Normalization
S21	Machine Learning	CNN, LSTM	N/A	N/A

recall, precision, and F1 are used for evaluating the aspect term extraction step. F1, which is the harmonic mean of precision and recall, is the most popular evaluation criterion – 19 studies out of the 21 investigated studies used it. The second-most used evaluation criterion is accuracy, which was computed in 16 studies. Accuracy is used for the polarity identification of each aspect and is computed as the total number of correctly identified aspects polarities to the overall number of available aspects polarities tuples. The least popular evaluation criteria are precision and recall, which were adopted in 13 studies. Precision is computed by dividing the correctly identified/extracted terms by the total terms, and recall is computed by dividing the correctly identified/extracted terms by the total identified/extracted terms. FIGURE 11 shows the popularities of various evaluation criteria.

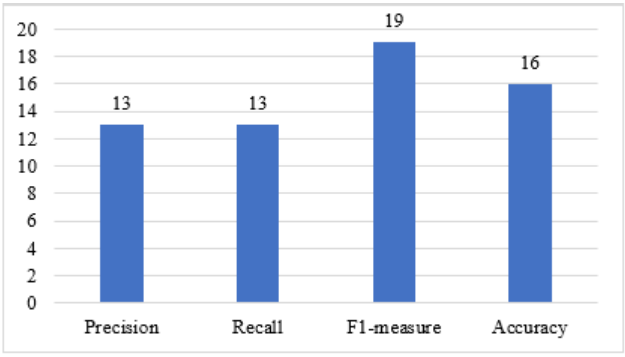

FIGURE 11. Popularity of various evaluation criteria.

V. DISCUSSION

Almost all the studies on Arabic ABSA were designed to solve ABSA for a specific domain. This is because the available datasets prepared for ABSA were domain-specific (e.g. hotel and book review datasets). Moreover, building an Arabic ABSA corpus is time-consuming and requires customised preprocessing tools for DA, as most of the content found on Arabic forms and blogs is written in many forms of DA, such as Egyptian, Levantine, and Gulf Arabic. As mentioned in the introduction, these dialects differ in syntax and vocabulary, which means that each dialect has its own lexicon. The availability of these lexicons is still limited, and 64% of the studies that were lexicon-based have developed their own lexicons. A generic Arabic Lexicon is needed to assist in the task of ABSA. In addition to the sentiment lexicons, aspects-related lexicons are required for ABSA. These lexicons contain all terms related to a specific aspect. After analysing the reviewed studies, it was found that these lexicons were not available for researchers; thus, researchers have to build their own lexicons if their approach requires one.

A variety of features were extracted, with POS and N-Grams the most commonly extracted features among the studies. However, it can be noticed that feature-like word embedding has recently gained attention from researchers. The results of studies in terms of accuracy and F1 measure were satisfying for most studies (see TABLE 5 for details).

TABLE 5. The best achieved accuracy and F1-measure for each study.

Study	Approach	Accuracy	F1-measure
S1	Lexicon-based	42.57	23
S2	Lexicon-based	71	22.8
S3	Lexicon-based	-	79.1
S4	Lexicon-based	92.15	98
S5	Machine Learning	86.5	82
S6	Machine Learning	73.2	40
S7	Lexicon-based	67.5	-
S8	Lexicon-based	74	39
S9	Lexicon-based	71.22	70.08
S10	Machine Learning	89	79.9
S11	Machine Learning	95.4	93.4
S12	Machine Learning	82.6	69.98
S13	Machine Learning	95.4	93.4
S14	Machine Learning	-	96.9
S15	Machine Learning	81	81
S16	Machine Learning	-	87
S17	Lexicon-based	80.5	78
S18	Lexicon-based	96.6	92.5
S19	Lexicon-based	-	-
S20	Lexicon-based	-	88
S21	Machine Learning	87.31	58

The highest accuracy (96.6%) was achieved by S18, and the best F1 value (98%) was achieved by S4. Most of the accuracy values ranged between 70% and 90%. From table TABLE 5, it can be concluded that studies that utilised a machine learning approach achieved better accuracy than studies that utilised a lexicon-based approach. This is because the employed machine learning methods are designed to solve classification tasks. On the other hand, lexicon-based approach studies have achieved better F1 scores.

The hotel reviews dataset was the most-used dataset in the assessed Arabic ABSA studies. TABLE 6 shows the features extracted in these studies and the results achieved by each. It can be found that for this dataset, the machine learning approach was the predominantly adopted approach. Moreover, the results achieved by machine learning methods were promising. The highest accuracy was achieved by S11 and S13, whereas the best F1-measure was achieved by S14.

Al-Smadi *et al.* have published 4 articles on hotel reviews dataset these articles are S6, S11, S12, and S13. Where in each one of these articles an algorithm or set of algorithms were utilized with a different set of extracted features. In S6 SVM classifier with a linear kernel was trained using unigrams extracted features. The achieved accuracy was 73.2% and F1-measure was 40. The next conducted study was S11, in this study a comparison conducted between SVM and RNN classifiers, both classifiers were trained using n-grams, POS, NER, and WE extracted features. The results showed that SVM outperforms RNN classifiers with F1-measure of 93.4 in aspect category identification task, and with accuracy of 95.4% for sentiment polarity classification task. Moreover, combining n-grams with other features has improved the accuracy of SVM classifier compared with S6.In the third study S12 two models of deep LSTM neural networks were developed for Arabic ABSA, these models were trained with n-gram and WE features. The results of these models were not as good as S11, with an accuracy of 82.6% and F1-measure of 70. In the last related study S13 a comparison between a set of machine learning algorithms (NB, Bayes Networks, DT, K-NN, and SVM) was conducted. These algorithms were trained using n-grams, POS, NER, TF, TF-IDF features. Also, in this study the SVM classifier outperforms all other classifiers with results similar to S11.

TABLE 6. Hotels' reviews dataset related studies.

Study	Extracted Features	Accuracy	F1-measure
S6	N-Grams	73.2	40
S7	N-Grams, POS	67.5	-
S11	N-Grams, POS, NER, WE	95.4	93.4
S12	N-Grams, WE	82.6	69.98
S13	N-Grams, POS, NER, TF, TF-IDF	95.4	93
S14	POS, NER	-	96.9
S21	-	87.31	58

VI. CONCLUSION

Most of the research efforts in Arabic sentiment analysis have been directed towards sentence-level and text-level sentiment analysis, while few studies have explored aspect-level sentiment analysis. This is due to the lack of annotated datasets that researchers can use to train an ABSA model, thus requiring extra from researchers working in this field. In addition, the differences between MSA and the multiple variations of DA decrease the accuracy of ABSA models.

The purpose of this systematic review was to view trends in Arabic ABSA over the past years to help researchers address the Arabic ABSA task. This is significant because ABSA plays an important role in decision-making in multiple fields. This review highlighted the main ABSA challenges, resources, and techniques that have been developed for Arabic reviews. It is clear that few studies have been done to solve

the Arabic ABSA task. Moreover, the number of available datasets is small and restricted to a couple of domains like hotels and books reviews. Moreover, there is no available gold-standard dataset that covers multiple domains. Future work could aim to construct such a dataset.

More research and datasets are needed to gain benefits from Arabic ABSA. There are several recommendations for researchers to follow in the coming years. First of all, developing multiple domain-annotated datasets for Arabic ABSA is a worthwhile task. Also, different methods for learning can be developed that take into account the differences between MSA and DA in terms of structure, syntax, and vocabulary. Another recommendation is to develop and use domain-independent models that can learn features from multiple-domain datasets. In addition to domain-independent models, a domain-independent lexicon is also needed. Also, ontologies and word embedding techniques can be utilised to improve the quality of extracted features and their polarities. Finally, lexicon-based and machine learning approaches could be combined into hybrid approaches.

REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, pp. 1–135, 2008.
[2] S. Behdenna, F. Barigou, and G. Belalem, "Document level sentiment analysis: A survey," *EAI Endorsed Trans. Syst. Appl.*, vol. 4, no. 13, Mar. 2018, Art. no. 154339.
[3] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.
[4] K. Yadav, "A comprehensive survey on aspect based sentiment analysis," 2020, *arXiv:2006.04611*.
[5] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, no. 4, pp. 82–89, 2013.
[6] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," in *A Practical Guide to Sentiment Analysis*. Cham, Switzerland: Springer, 2017, pp. 1–10.
[7] H. Thakkar and D. Patel, "Approaches for sentiment analysis on Twitter: A state-of-art study," 2015, *arXiv:1512.01043*.
[8] H. Sankar and V. Subramaniyaswamy, "Investigating sentiment analysis using machine learning approach," in *Proc. Int. Conf. Intell. Sustain. Syst. (ICISS)*, Dec. 2017, pp. 87–92.
[9] A. Jurek, M. D. Mulvenna, and Y. Bi, "Improved lexicon-based sentiment analysis for social media analytics," *Secur. Informat.*, vol. 4, no. 1, pp. 1–13, Dec. 2015.
[10] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in *Proc. IEEE Jordan Conf. Appl. Electr. Eng. Comput. Technol. (AEECT)*, Dec. 2013, pp. 1–6.
[11] E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades, "Ontology-based sentiment analysis of Twitter posts," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4065–4074, Aug. 2013.
[12] S. Bandari and V. V. Bulusu, "Survey on ontology-based sentiment analysis of customer reviews for products and services," in *Data Engineering and Communication Technology*. Singapore: Springer, 2020, pp. 91–101.
[13] O. Appel, F. Chiclana, J. Carter, and H. Fujita, "A hybrid approach to the sentiment analysis problem at the sentence level," *Knowl.-Based Syst.*, vol. 108, pp. 110–124, Sep. 2016.
[14] M. Alrefai, H. Faris, and I. Aljarah, "Sentiment analysis for Arabic language: A brief survey of approaches and techniques," 2018, *arXiv:1809.02782*.
[15] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*. Boston, MA, USA: Springer, 2012, pp. 415–463.
[16] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.
[17] A. Ghallab, A. Mohsen, and Y. Ali, "Arabic sentiment analysis: A systematic literature review," *Appl. Comput. Intell. Soft Comput.*, vol. 2020, pp. 1–21, Jan. 2020.
[18] S. Tedmori and A. Awajan, "Sentiment analysis main tasks and applications: A survey," *JIPS*, vol. 15, pp. 500–519, Oct. 2019.
[19] M. K. Saad and W. M. Ashour, "OSAC: Open source Arabic corpora," in *Proc. 6th ArchEng Int. Symp.*, Oct. 2010, pp. 1–6.
[20] N. Y. Habash, "Introduction to Arabic natural language processing," *Synth. Lectures Hum. Lang. Technol.*, vol. 3, no. 1, pp. 1–187, Jan. 2010.
[21] L. Gutiérrez and B. Keith, "A systematic literature review on word embeddings," in *Proc. Int. Conf. Softw. Process. Improvement*. Cham, Switzerland: Springer, 2018, pp. 132–141.
[22] B. Kitchenham, "Procedures for performing systematic reviews," Dept. Comput. Sci., Keele Univ., Keele, U.K., Tech. Rep. 0400011T.1, 2004, pp. 1–26, vol. 33.
[23] D. Moher, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *PLoS Med.*, vol. 6, no. 7, 2009, Art. no. e1000097.
[24] 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.
[25] M. Aly and A. Atiya, "LABR: A large scale Arabic book reviews dataset," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2013, pp. 494–498.
[26] 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. 11st Int. Conf. Internet Technol. Secured Trans. (ICITST)*, Dec. 2016, pp. 98–103.
[27] 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.
[28] S. Ismail, A. Alsammak, and T. Elshishtawy, "A generic approach for extracting aspects and opinions of Arabic reviews," in *Proc. 10th Int. Conf. Informat. Syst.*, 2016, pp. 173–179.
[29] L. Abd-Elhamid, D. Elzanfaly, and A. S. Eldin, "Feature-based sentiment analysis in online Arabic reviews," in *Proc. 11st Int. Conf. Comput. Eng. Syst. (ICCES)*, Dec. 2016, pp. 260–265.
[30] A.-S. Mohammad, M. Al-Ayyoub, H. N. Al-Sarhan, and Y. Jararweh, "An aspect-based sentiment analysis approach to evaluating Arabic news affect on readers," *J. Universal Comput. Sci.*, vol. 22, no. 5, pp. 630–649, 2016.
[31] A. M. Alkadri and A. M. ElKorany, "Semantic feature based Arabic opinion mining using ontology," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 5, pp. 577–583, 2016.
[32] M. Al-Ayyoub, H. Al-Sarhan, M. Al-So'ud, M. Al-Smadi, and Y. Jararweh, "Framework for affective news analysis of Arabic news: 2014 Gaza attacks case study," *J. Univers. Comput. Sci.*, vol. 23, no. 3, pp. 327–352, 2017.
[33] A. M. El-Halees and D. Salah, "Feature-based opinion summarization for Arabic reviews," in *Proc. Int. Arab Conf. Inf. Technol. (ACIT)*, Nov. 2018, pp. 1–5.
[34] M. M. Ashi, M. A. Siddiqui, and F. Nadeem, "Pre-trained word embeddings for Arabic aspect-based sentiment analysis of airline tweets," in *Proc. Int. Conf. Adv. Intell. Syst. Inform.* Cham, Switzerland: Springer, 2018, pp. 241–251.
[35] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.
[36] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "AraVec: A set of Arabic word embedding models for use in Arabic NLP," *Proc. Comput. Sci.*, vol. 117, pp. 256–265, Jan. 2017.
[37] 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.
[38] 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, 2018.
[39] 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.

[40] S. Trigui, I. Boujelben, S. Jamoussi, and Y. B. Ayed, "Adal system: Aspect detection for Arabic language," in *Proc. Int. Conf. Hybrid Intell. Syst.* Cham, Switzerland: Springer, 2019, pp. 31–40.

[41] M. Pontiki, "Semeval-2016 task 5: Aspect based sentiment analysis," in *Proc. Int. workshop Semantic Eval.*, 2016, pp. 19–30.

[42] N. F. Alshammari and A. A. AlMansour, "Aspect-based sentiment analysis for Arabic content in social media," in *Proc. Int. Conf. Electr., Commun., Comput. Eng. (ICECCE)*, Jun. 2020, pp. 1–6.

[43] M. Alassaf and A. M. Qamar, "Aspect-based sentiment analysis of Arabic tweets in the education sector using a hybrid feature selection method," in *Proc. 14th Int. Conf. Innov. Inf. Technol. (IIT)*, Nov. 2020, pp. 178–185.

[44] 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.

[45] 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.

[46] G. Alhamad and M.-B. Kurdy, "Feature-based sentiment analysis for Arabic language," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, pp. 455–462, 2020.

[47] S. Behdenna, F. Barigou, and G. Belalem, "Towards semantic aspect-based sentiment analysis for Arabic reviews," *Int. J. Inf. Syst. Service Sector*, vol. 12, no. 4, pp. 1–13, Oct. 2020.

[48] G. Badaro, R. Baly, H. Hajj, N. Habash, and W. El-Hajj, "A large scale Arabic sentiment lexicon for Arabic opinion mining," in *Proc. EMNLP Workshop Arabic Natural Lang. Process. (ANLP)*, 2014, pp. 165–173.

[49] 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.

[50] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2014, pp. 1746–1751. [Online]. Available: https://aclanthology.org/D14-1181

[51] P. Gonnet and T. Deselaers, "Indylstms: Independently recurrent LSTMS," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3352–3356.

[52] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[53] H. ElSahar and S. R. El-Beltagy, "Building large Arabic multi-domain resources for sentiment analysis," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*. Cham, Switzerland: Springer, 2015, pp. 23–34.

[54] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "SemEval-2015 task 12: Aspect based sentiment analysis," in *Proc. 9th Int. Workshop Semantic Eval. (SemEval)*, 2015, pp. 486–495.

[55] F. Sebastiani and A. Esuli, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proc. 5th Int. Conf. Lang. Resour. Eval.*, 2006, pp. 417–422.

[56] T. El-Shishtawy and F. El-Ghannam, "An accurate Arabic root-based lemmatizer for information retrieval purposes," 2012, *arXiv:1203.3584*.

[57] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, 2003, pp. 252–259.

[58] *Arabic Toolkit Service (ATKS)*. Accessed: Aug. 26, 2021. [Online]. Available: https://www.microsoft.com/en-us/research/project/arabic-toolkit-service-atks/

[59] A. Pasha, M. Al-Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. M. Roth, "MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic," in *Proc. LREC*, vol. 14, 2014, pp. 1094–1101.

[60] R. Al-Rfou, V. Kulkarni, B. Perozzi, and S. Skiena, "POLYGLOT-NER: Massive multilingual named entity recognition," in *Proc. SIAM Int. Conf. Data Mining*, Jun. 2015, pp. 586–594.

[61] Y. Goldberg and O. Levy, "Word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method," 2014, *arXiv:1402.3722*.

[62] M. Althobaiti, U. Kruschwitz, and M. Poesio, "AraNLP: A Java-based library for the processing of Arabic text," in *Proc. 9th Int. Conf. Lang. Resour. Eval.*, 2014, pp. 4134–4138.

[63] M. Althobaiti, U. Kruschwitz, and M. Poesio, "Automatic creation of Arabic named entity annotated corpus using Wikipedia," in *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2014, pp. 106–115.

RUBA OBIEDAT received the B.Sc. degree in computer science from The University of Jordan, in 2003, the M.Sc. degree in information system from DePaul University, in 2007, and the Ph.D. degree in E-business from University del Salento, Lecce, Italy, in 2010. She has been an Associate Professor with the Department of Information Technology, The University of Jordan, since 2014. Her research interests include data mining, machine learning, business intelligence, sentiment analysis, and E-business. She was awarded a full-time, competition-based Ph.D. scholarship from the Italian Ministry of Education and Research to pursue her Ph.D. degree.

DUHA AL-DARRAS received the B.Sc. degree in computer information system from Bethlehem University, Palestine, in 2015, and the M.Sc. degree in web intelligence from The University of Jordan, Jordan, in 2019. She is currently a Lecturer with the Department of Software Engineering, Faculty of Science, Bethlehem University. Her research interests include area of natural language processing and computational linguistics in Arabic.

ESRA ALZAGHOUL received the Ph.D. degree in software engineering from the University of Birmingham. She is currently an Assistant Professor with The University of Jordan. She was the Assistant Director of the Accreditation and Quality Assurance Center. In addition, she was the Director of the Website Department, The University of Jordan. Earlier, she was a Senior Web Developer (J2EE) at ATS. The unique academic experience along with her administrative positions has inspired her research that currently focuses on managing technical debt in cloud-based service selection and composition. She is a member of the Program Committee of the International Conference on Technical Debt and the International Workshop on Managing Technical Debt.

OSAMA HARFOUSHI received the B.Sc. degree in CIS from the Jordan University of Science and Technology, Jordan, in 2003, the M.Sc. degree in E-business from the University of Huddersfield, U.K., and the Ph.D. degree in mobile learning from the University of Bradford, U.K. He is currently a Full Professor with the Information Technology Department, King Abdullah II School of Information Technology, The University of Jordan. His research interests include cloud computing, E-business, and business data mining.