IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received October 24, 2018, accepted November 12, 2018, date of publication December 28, 2018, date of current version January 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2890076

Arabic Natural Language Processing and Machine Learning-Based Systems

SOUAD LARABI MARIE-SAINTE[1], NADA ALALYANI[2], SIHAAM ALOTAIBI[3], SANAA GHOUZALI[2], (Member, IEEE), AND IBRAHIM ABUNADI[1], (Member, IEEE)

[1]Computer Science Department, College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia
[2]Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh 11495, Saudi Arabia
[3]Computer Science and Information Technology Department, College of Community, Princess Noura bint Abdulrahman University, Riyadh 84428, Saudi Arabia

Corresponding author: Sanaa Ghouzali (sghouzali@ksu.edu.sa)

This work was supported by the Research Center of the Female Scientific and Medical Colleges, Deanship of Scientific Research, King Saud University.

ABSTRACT Arabic natural language processing (ANLP) consists of developing techniques and tools that can utilize and analyze the Arabic language in both written and spoken contexts. ANLP makes an important contribution to many existing developed systems. It provides Arabic and non-Arabic speakers with helpful and convenient tools that can be used in different domains. Modern ANLP tools are developed using machine learning (ML) techniques. ML algorithms are widely used in NLP because of their high accuracy rate regardless of the robustness of the data that is used and because of the ease with which they can be implemented. On the other hand, the methodology of ANLP applications based on ML involves several distinct phases. It is, therefore, crucial to recognize and understand these phases in detail as well as the most widely used ML algorithms. This survey discusses this concept in detail, shows the involvement of ML techniques in developing such tools, and identifies well-known techniques used in ANLP. Moreover, this survey discusses the characteristics and complexity of the Arabic language in addition to the importance and needs of ANLP.

INDEX TERMS Arabic natural language processing, classification, feature selection, machine learning.

I. INTRODUCTION

Natural language processing (NLP) is a field of computer science that seeks to create concepts, find methods, and construct software that is able to comprehend, study, and produce natural human languages to enable human interaction with computers through writing and speech. In other words, NLP helps computers identify the ways in which humans use language.

Arabic natural language processing (ANLP) has attracted many researchers after significant research has been carried out on English NLP and that of other languages. Many ANLP laboratories have been established. Recently, ANLP has received more attention, and several applications have been developed including text categorization, web page spam detection, and sentiment analysis [1]–[3]. However, developing ANLP tools requires additional efforts due to two principal difficulties: the combination of letters in the Arabic language and the removal of diacritics that represent the vowels [4].

Recently, ANLP tools are developed using machine learning algorithms. Machine learning (ML) falls within the domain of artificial intelligence. The goal of such technologies is to enable computers to learn without explicit programming. ML has been successfully applied in many difficult and complex computing tasks (such as ANLP) without designing and programming explicit algorithms. In addition, the range of ML algorithms that yield satisfactory results has led ML to become vastly involved in ANLP and NLP in general.

Different surveys in the literature describe and discuss the difficulty and complexity of Arabic Language Processing (for example, [5]). This article provides a comprehensive review of ANLP-based ML systems. This survey focuses on the following: 1) presenting the characteristics, challenges, and complexity of the Arabic language to provide new researchers in the field with brief background knowledge about ANLP; 2) stating the predominant domains that utilize ANLP applications; and 3) addressing the concept of supervised ML techniques used to develop modern applications/tools based on ANLP and identifying well-known techniques.

This article provides a brief background of the Arabic language, its characteristics, and its complexity in sections II, III, and IV, respectively. An overview of Arabic

Natural Language Processing, along with the necessity of and needs in ANLP are presented in section V. The involvement of ML algorithms in ANLP and specifically in ANLP applications is shown in section VI and VII, respectively. Finally, a discussion of the challenges encountered in ANLP applications and a suggestion for some solutions conclude the last section.

II. ARABIC LANGUAGE

Arabic is an Afro-Asiatic language that developed in the Middle East. More than 250 million individuals speak the Arabic language across the world [1]. The Arabic language was widely disseminated after the emergence of Islam, although it existed centuries before the religion. As a universal religion, Islam has delivered the Arabic language to its followers, estimated to be 1.5 billion people [1].

Historically speaking, Arabic is rooted in Classical Arabic (CA), which has been used as the Arab peoples' native language since 600 AD. It is associated with Islam and the Quran. However, over the centuries, the language has evolved and been simplified to create what is known as Modern Standard Arabic (MSA). The terminology and the linguistic features of MSA differ from those of CA, but the structure of words and sentences have remained. In addition to CA and MSA, each region has a dialect of Arabic spoken in the community (between friends and family) [5].

III. ARABIC LANGUAGE CHARACTERISTICS

The Arabic language consists of grammar, spelling, punctuation marks, slang as informal language, idioms, and pronunciation. Many characteristics make the Arabic language distinctive [6]: 1) reading and writing in Arabic moves from right to left. 2) The language consists of 28 characters. 3) In Arabic, upper and lower cases are not distinguished, like Chinese, Japanese, and Korean. 4) Numbers are divided into plural, dual, and singular, with two genders–feminine and masculine. 5) The language comprises several words formed from roots, and several root words are composed of three letters. 6) Verbs in the past tense are identified by suffixes, and verbs in the present or future tenses are designated by prefixes; for example, "dahabat" means "she went," but "tadhabu" means "she goes." 7) Sentences start with verbs, followed by subjects, and are finished with objects for the predicate@perio 8) Arabic tolerates the deletion of subject pronouns (pro-drop language) like Italian and Chinese [5].

IV. ARABIC LANGUAGE COMPLEXITY

Arabic can be considered more complex than English. Arabic writing does not possess vowels; rather, diacritics are placed above or below letters. Modern writers have abandoned these diacritics; readers are expected to understand the lost diacritics based on their knowledge of the language [5]. This characteristic induces both structural and lexical ambiguity in Arabic texts because various diacritics may lead to different meanings [3], [8]. In SYSTRAN, an Arabic-to-English transfer machine translation system, the ambiguity for a token in MSA achieves approximately 19.2, while it reaches 2.3 in most languages [5]. This ambiguity renders challenging the Arabic Text translation.

Another complexity consists of dots, which are frequently used in Arabic. The structure of many letters is similar or even identical, so letters are differentiated by the number of dots and their locations, such that the letters (n-ن, b-ب, t-ت) all have the same structure but with different dot locations and numbers.

In addition, some letters possess diverse forms that rely on their location in the word. Of the 28 letters in the language, 22 take four different shapes each (at the beginning/middle/end of the word, and at a non-linked letter). The letter / ع ain/ can be used as an example here, as it is characterized by the following shapes ("عـ", "ـعـ", "ـع", "ع"). The remaining letters hold two forms each (at the end of the word and when it is not connected to another letter). Moreover, nouns and adjectives in Arabic can be masculine or feminine [6].

Furthermore, vocabulary of the Arabic language can have different meanings. For example, the word darkness has 52 synonyms, the word rain has 34, the word moon has 16, the word light has 21, the word short has 164, the word long has 91, and there are 50 synonyms for the word cloud [6].

Moreover, capital letters do not exist in Arabic, which is a major characteristic in recognizing nouns in NLP contexts [3], [7], [8], especially Named Entity Recognition discussed in Section VII-A

Arabic also uses specific inflections; usually, a term may be stated as a mix of prefix (es) (which can be articles, prepositions, or conjunctions), lemma, and suffix (es) (which are objects or personal/possessive anaphora) [3], [7]. Figure 1 is an example of Arabic inflection.

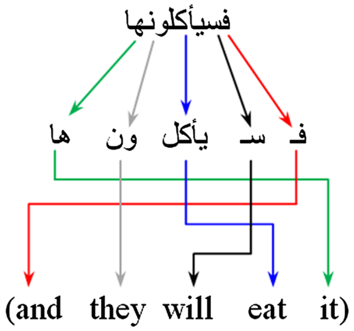

FIGURE 1. Example of Arabic inflection [9].

These complexities affect the translation task which cannot reach a high performance compared to the translation of English language to other languages. Two common machine translation software (Google Translate and Babylon) are still far from having a high accuracy in translating Arabic to other languages [7]. Henceforth, the Arabic language requires superior treatment. However, standard language processing systems that are intended for other languages are incapable of dealing with the Arabic language. The researchers are making great efforts in preprocessing the Arabic language while considering these previous characteristics, which make the preprocessing of Arabic texts very different from the preprocessing of the other languages' text. The tokenization and the morphological Analysis are mainly based on the identification of words and root of words. Nevertheless, the Arabic inflection makes stemming (one step in the preprocessing) difficult to perform. Even though many efforts have been done in this regard, some improvements are still needed. The preprocessing phase is discussed in detail in Section VIII-B.

V. ARABIC NATURAL LANGUAGE PROCESSING

A. NATURAL LANGUAGE PROCESSING

NLP is an area of investigation that aims to enable computers to analyze and use original text or human speech as it is spoken to perform valuable applications. NLP researchers seek to identify how humans manipulate language, which helps them develop techniques and tools that allow computer systems to handle natural languages and perform necessary tasks. Several disciplines have formed the basics of NLP, such as artificial intelligence, computer and information sciences, linguistics, electronic and electrical engineering, robotics, mathematics, and psychology. NLP has many applications in different domains, such as summarizing natural language texts, machine translation, and language information retrieval, among others. Many NLP applications are now available in many languages such as Chinese, Arabic, and English [2], [7], [9].

B. ARABIC NATURAL LANGUAGE PROCESSING

ANLP has become an exciting research domain. It involves the development of techniques and tools using the Arabic language. Numerous existing systems have been created for different applications such as machine translation, information retrieval and extraction, localization, and multilingual information retrieval systems [11], [12]. These applications encounter numerous intricate problems related to the structure and nature of the Arabic language.

C. NECESSITY AND NEEDS IN ANLP

Most developed ANLP systems are dedicated to allowing non-Arabic speakers in the Western world to understand Arabic texts. For example, sentiment analysis, machine translation, and Arabic named entity recognition are the most commonly used ANLP tools [5].

ANLP is highly demanded in several sectors. For example, the task of correctly identifying Arabic names is challenging for non-Arabic government institutions. ANLP tools that scan and recognize names, places, and dates are becoming necessary and beneficial because they save time spent waiting for language experts to carry out the task.

On the other hand, Google Translate and Babylon are always looking for translating English sentences into Arabic (or vice versa) by giving the meaning of each word without considering the meaning of the whole sentence. Machine translation intends to translate a sentence from one language (e.g. Arabic), providing the closest meaning in an alternative language(s) [13]. ANLP application can handle a given text at the sentence level; to identify the order of the words, grammar, and sense of the whole phrase, which is very useful in machine translation tasks.

In the Arab world, several tools have been developed to help understand the sentiments and impressions of users of social networks. Marketing researchers and business owners, in general, can take advantage of such tools. These tools usually give an indication of user's impressions about a specific product or place. An example of these tools is Salesforce [http://www.salesforce.com/]. ANLP applications are useful in accomplishing the task of sentiment analysis; they can be used worldwide to identify the sense of words in addition to morphological structure. They are also used at the sentence level to recognize the sense of whole phrases.

VI. ANLP AND MACHINE LEARNING

Machine learning involves statistical techniques and algorithms that permit the generation of structures in the learning stage based on training data in order to predict results for new data in the testing stage. The learning stage involves optimizing a numerical measure to compute the parameters that characterize a given algorithm's underlying model. Existing machine learning techniques are expressed as supervised or unsupervised [14].

In supervised learning, training examples are annotated with pre-specified class labels. Among supervised learning techniques are Support Vector Machines, Maximum Entropy, Neural Networks, and Bayesian Networks. On the other hand, unsupervised learning consists of searching for similarities between training examples, in the absence of pre-specified class labels, to group data into clusters/classes. The clustering approach is a common unsupervised learning technique [14].

Machine learning techniques have evolved over the last decade and have been useful in different domains including ANLP [15], giving rise to several language-processing software that rely on a variety of supervised and unsupervised models [16]. In this article, we mainly focused on the supervised ML techniques as discussed in Section VIII-D.

Contrary to prior applications of language processing tasks that encompassed manual coding of large sets of rules [17], [18], machine learning techniques used in modern NLP algorithms in different languages such as Arabic offer many advantages:

- ML algorithms are used to automatically focus on common cases, whereas in the manual coding of rules, it is not clear where the effort must be directed.
- ML algorithms can produce models for unfamiliar data.
- ML can be accurate by merely increasing the input data, whereas systems based on the manual coding of the rules can be effective only if the complexity of the rules is increased, which is a much more challenging task [19].

VII. ANLP APPLICATIONS-BASED MACHINE LEARNING

Several applications of ANLP-based machine learning that have been proposed in the literature are surveyed below, with a focus on supervised techniques.

A. NAMED ENTITY RECOGNITION (NER)

Named entity recognition (NER) refers to the identification and classification of specific proper nouns into predefined target entity categories such as persons, locations, and organizations. NER is usually integrated as a useful preprocessing component of various ANLP applications [20], including Machine Translation [7], Search Results clustering [8], and Question Answering [21].

Contrary to English language, NER resources based Arabic language are either not up to date or costly. Thus, researchers have to build their own resources for research which require human effort for annotation and validation. Consequently, most of the available NER resources based Arabic language are annotated manually rendering their use time consuming. Moreover, the research studies in NER systems based Arabic language have usually considered few entities (e.g. person, location, and organization) from source corpora while other entity types are being ignored. The reasons have not been mentioned or investigated [20].

B. READABILITY ASSESSMENT

Text readability is defined as ''the ease of understanding a text due to its writing style'' [22]. Text readability is used to provide different reader groups with the appropriate texts. It is widely used in education to select appropriate textbooks for students, in healthcare to provide medical instructions understood by the average patient, and in the industry to measure the readability of user manuals of the products. ANLP tools are applied to retrieve information or features that represent text readability levels such as lexical, morphological, and semantic features [23], [24].

English and other foreign languages have a long history in the field of readability. Nowadays, Arabic readability has drawn the attention of the Arab community. Most of the research has focused merely on the use of shallow text features such as features extracted after segmentation of the words by simply counting words and characters. There are more important features that reflect the readability of text such as the attributes associated with the difficulty or unfamiliarity of vocabulary and the attributes measuring cohesion and coherence between text elements. In addition, researchers have mainly focused on the readability of MSA and Pedagogical text. Focus should be directed towards informal text as well (e.g. twitter tweets). Also, researchers may have to examine the possibility of applying the methods of readability over text coming from other domains. Moreover, providing such tools online and with friendly interfaces is very important for all kind of users (e.g. the teachers).

C. WEB SPAM DETECTION

Web spamming refers to actions that mislead search engines into placing some pages in a rank that is higher than what they deserve [3]. Web spamming can be very dangerous because it spreads malware, which can affect users' privacy by obtaining sensitive information. Web spam detection using ANLP tools are used to determine information and features included in the content of web pages to ensure that only web pages that present useful content are retrieved [18], [25]–[28]. This allows threats posed by suspicious web pages to be mitigated [3].

The previous works on Web spam detection in Arabic language have focused mostly on the use of features such as link, content, or hybrid features [18], [26], [27]. However, more sophisticated features (e.g. Language models) have not yet been used. Although, studies in other languages (for example English) have proved that, the application of language models besides other features has improved the performance of Web spam detection tasks.

D. SENTIMENT ANALYSIS

Sentiment analysis, or opinion mining, attempts to analyze people's attitudes, opinions, and emotions regarding entities like services, products, and organizations. It has immense importance in social media monitoring, as it can provide an overview of broader public opinions on specific topics. Social media monitoring tools make the process quicker and easier than ever before. Much research has been concerned with studying sentiment analysis in Arabic using machine learning approaches [29]. ANLP-based machine learning tools were used in sentiment analysis to select the features from Arabic public comments on Twitter [30], [31] and Facebook [32] to correctly classify them.

Even though the preprocessing and feature selection steps are extremely significant to improve the classification accuracy,some existing sentiment analysis studies did not include these steps, for example, authors in [32] classified the input texts directly without any preprocessing. Other sentiment analysis applications (e.g. [30] and [31]) used tools, such as RapidMiner and WEKA, in order to filter the Arabic text tweets. Unfortunately, these tools have limited Arabic preprocessing services. Some of them either do not include feature selection process [31] or there are no sufficient methods for selecting features from tweets [30].

Moreover, these studies depend on ''in-house'' developed dataset of tweets/comments. So, it is more than important to build efficient feature selection algorithms for any Arabic text.

E. ARABIC TEXT CATEGORIZATION

Information generation and sharing expose the community to a flood of content, which demands automatic text categorization algorithms. Text categorization automatically assigns a category to a document. The Arabic automatic categorization documents system has gained high significance in education, health, and information sciences. ANLP tools were used in text categorization to extract the top features from documents and correctly classify them. Many investigations have studied Arabic document categorization employing supervised ML techniques [33]–[36], [63]. The Hadith classifier is one example in this domain, which categorizes the talks of the Prophet Mohammad (Peace be upon him, PBUH).

However, there is a lack of publicly available preprocessing and feature selection tools and reusable libraries for Arabic text documents. Therefore, the researchers are required to build the preprocessing stage from the scratch, which consumes lots of time and effort until a high accuracy level is achieved. Most of the existing studies used different swarm optimization techniques, including Bee Swarm Optimization (BSO) [33], Particle Swarm Optimization (PSO) [34], [35] and Firefly algorithm [60]. These methods, in contrast to others such as TF/IDF (term frequency / inverse document frequency) [37], are very effective in selecting the main features in Arabic text document.

F. WEB DOCUMENT CLASSIFICATION

Web mining is an important technology due to the huge amount of text information stored online. Web mining involves using tools and techniques to rapidly retrieve desired information from the web. Web document classification is a critical technique in web mining and it refers to the process of assigning an explicit document, based on its content, to one or more predefined categories, which makes it easier to manage and sort through the huge database of information on the web [37], [68]. Several studies have examined web document classification based on ANLP tools [38]–[40], [68].

For web Arabic document classification systems, the preprocessing and the feature selection steps are very important to enhance the classification process. However, in [37] and [38] there is no mention to the filtering process. In [39], the preprocessing stage is implemented but with low impact on the documents. The feature selection employed in [37]–[39] was based on TF/IDF; this feature selection method has low impact on the classification process compared to other sophisticated feature selection methods.

It is worthwile to build a standard tool that gather all essential preprocessing and feature selection methods for the Arabic data mining. There is no doubt that such tool will mitigate the significant efforts, cost, and time of building Arabic applications that require data mining and Machine Learning processes. If we look at the other languages, such as English language, there are many tools (e.g. WEKA) that collect all efficient algorithms for data mining.

VIII. HOW ANLP IS EVOLVED IN MACHINE LEARNING

The ANLP applications we have described earlier use machine learning approaches (mainly supervised) and have recently attracted the focus of researchers. The development of these applications often includes many phases, as shown in Figure 2 and presented in detail in the following subsections.

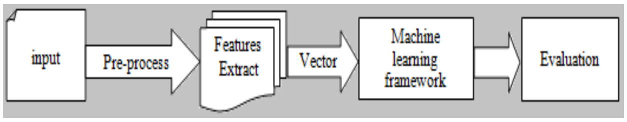

FIGURE 2. The typical phases of ANLP applications-based machine learning.

A. CORPORA

The first step in ANLP-based machine learning studies is data collection. These data are text samples that are suitable for the concerned subject area. Few freely available collected and classified Arabic corpora exist, such as Al-Nahar, Al-Jazeera, Al-Ahram, Al-hayat, and Al- Dostor newspapers [3], Hadith corpus [13], Akhbar-Alkhaleej corpus [8], Arabic NEWSWIRE [33], Quranic Arabic Corpus [13], corpus Watan-2004 [35], KACST Arabic corpus [42], BBC Corpus [43], CCN Corpus [44] and Open Source Arabic Corpora (OSAC) [45], NADA corpus [59]. These corpora have been used in Arabic processing studies, especially in text categorization. For named entity recognition studies, benchmark data exist such as ACE 2003 [46] and ANERcorp [41]. On the other hand, for the other applications like spam detection, sentiment analysis, readability, and web document classification, benchmark corpora are still missing.

B. PREPROCESSING

Text preprocessing is a core natural language processing task. It applies operations to create another form from the inputted text. In ANLP, MADAMIRA and RapidMiner offer natural language processing operations. MADAMIRA provides the study of the structure of words and part of words (root words, prefixes, and suffixes) in the Arabic language [47]. It includes the critical characteristics of the well-known Arabic processing systems MADA [48]–[50] and AMIRA [51]. RapidMiner consists of many preprocessing operations including stemming, cleaning, and visualization. It is implemented using Java and can be operated with any operating system [52]. Data Cleaning, normalization, tokenization, and stemming are common text preprocessing operations in most ANLP applications. Table 1 summarizes these preprocessing operations.

- Data Cleaning: comprises the removal and/or correction of incorrect records from a data set [24], [31], [32], [34], [35].
- Normalization: focused on the removal of inconsistent variations of Arabic text [31]–[33].
- Tokenization: aims to detect and separate individual words by eliminating additional components such as punctuation marks, white space, and unique characters [21], [30], [31], [53].
- Stemming: is used to reduce multiple forms of the word to one form producing root or stem [30], [31], [33].

TABLE 1. Summary of the preprocessing operations.

Operation	*Example [Before applying the operation]*	*Example [After applying the operation]*	*Studies*
Data Cleaning	ذَهـبَ أحمدُ إلى المدرسةِ (Ahmed went to school)	ذهب أحمد المدرسة	([24], [31], [32], [35] [34]
Normalization	ذهب أحمد إلى المدرسة	ذهب احمد الى المدرسه	[31], [32], [33]
Tokenization	ذهب أحمد إلى المدرسة	[ذهب, أحمد, إلى, المدرسة]	[21], [30], [31], [53]
Stemming	(المكتبة, الكاتب, الكتاب) (The library), (the writer), (the book)	كتب (Wrote)	[30], [31], [33]

Benajiba *et al.* [21] presented an Arabic NER method. In the preprocessing stage, they used AMIRA toolkit for tokenization. They stressed that, contrary to English language, Arabic tokenization is significant and challenging because of the Arabic's morphological structure. So, it is primordial to avoid any error in this stage in order to obtain accurate results. They also prevented the removal of the suffixes to decrease ambiguity of the text. Duwairi *et al.* [31] handled the sentiment analysis of tweets. They started by splitting the tokens using colon, semicolon, comma, space. Then, they applied the normalization on the tweets. After that, they removed the Arabic stop words using a new dictionary integrated to RapidMiner. This is done to overcome the deletion of the negation that is considered as stop words in RapidMiner while it is capital in the analysis of sentiments. Finally, they used light stemming and stemming. The authors stressed that the integration of the dictionary increased the performance of their approach. Zahran and Kanaan [34] proposed a new approach for Arabic text categorization. In the preprocessing phase, they started by removing the non-Arabic letters. Then, the normalization is applied. After that, the stop words are deleted. The authors did not apply the stemming to avoid any conflation with the same form of root.

C. FEATURE SELECTION

Dimensionality reduction or feature selection is a central stage in pattern recognition, especially in ANLP applications. FS aims to increase the efficiency and accuracy of ANLP applications by selecting relevant words (sufficient features) from a text document.

Not all the features (the words of the text document) are useful for the classification stage because the large dimensionality of features affects the performance of classification [1]. To overcome these challenges, many feature selection methods are used in ANLP research such as term frequency/inverse document frequency (TF/IDF) [54]. Moreover, Chi-Squared statistics (X2), information gain (IG), Mutual information (MI) [18], and document frequency and information gain [44] have also been used in Feature Selection. In addition, Latent Dirichlet Allocation (LDA) [42], Particle Swarm Optimization (PSO) [30], [55], [57], Ant Colony Optimization (ACO) [43], Bee Swarm Optimization (BSO) [14], Genetic Algorithm (GA) [19], [36], [45], Singular Value Decomposition (SVD) [56], and Firefly Algorithm (FFS) [58] have shown to be promising. Figure 3 shows the state-of-the-art feature selection methods applied in ANLP from 2004 to 2018.

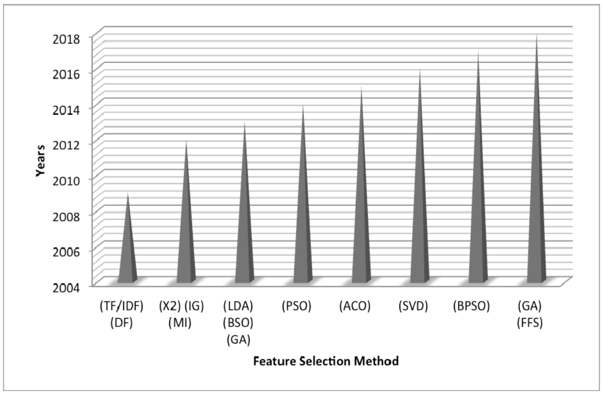

FIGURE 3. Classification of surveyed feature selection methods.

D. SUPERVISED MACHINE LEARNING TECHNIQUES

Supervised learning approaches are related to a labeled training data. Several types of supervised classifiers exist. In the following section, we will present a review of the most frequently used classifiers in ANLP applications.

1) SUPPORT VECTOR MACHINES (SVMs)

are well-known supervised machine learning techniques [53]. SVMs have been effectively employed in many problems related to pattern recognition in fields such as bioinformatics and biometrics. Regarding text processing, SVMs have achieved the best results in text categorization and are used extensively in NLP-related problems in different languages such as Arabic for methods like readability prediction and sentiment analysis [17], [24], [30]–[32].

2) NAIVE BAYES CLASSIFICATION (NB)

is the most straightforward and the second-most-used classifier. The Naive Bayes classifier is a prevalent technique for text categorization that assigns documents to associated categories such as spam or legitimate [28], [31], [32], [36] and positive, negative, or neutral [30]–[32], [60].

3) DECISION TREES

have been used in many NLP-related classification problems [61]. In addition to the Naive Bayes classifier, Decision Tree classifiers provide excellent results for spam detection [18], [25], [27], [28]. The decision tree classifier is a favored machine learning technique because the model is easy to understand. Abdallah *et al.* [62] showed that their hybrid approach with the rule-based approach using decision tree performed better than the existing classifiers for Arabic Named Entity Recognition applications.

4) k-NEAREST NEIGHBOR (k-NN) ALGORITHM

has been successfully applied to several problems related to ANLP due to its simplicity (e.g., Extraction of Semantic Relations between Concepts, [16]). The k-NN consists of instance-based learning, or lazy learning, where the learning is delayed until classification is conducted. k-NN is well-known in sentiment analysis applications [30]–[32]. It is also utilized in other ANLP applications for classification tasks such as spam detection [28] and web document classification [5], [40].

Wahsheh and Al-Kabi [28] applied three different classification algorithms to detect Arabic spammed Web pages using content based analysis. The obtained results using KNIME software showed that the k-NN (k = 1) achieved better accuracy than the Naive Bayes and decision tree. The main problem faced in this study is the lack of large number of Arabic Web pages with valuable information. Duwairi and Qarqaz [30] presented an approach for Arabic sentiment analysis. They first generated an in-house dataset by collecting and labeling tweets and Facebook comments from the Internet using crowdsourcing. RapidMiner is then used for the preprocessing purpose. Finally, Naïve Bayes, SVM, and k-NN have been used for the classification task. Results showed that SVM gives the highest Precision while k-NN (k = 10) gives the highest Recall. The reason for these results is the fact that the dataset is not balanced as the number of negative reviews is larger than the positive reviews.

Several other classifiers have been used in different ANLP applications such as Conditional Random Fields (CRF) classifier [15], Neural Network ANN [33], Association rules [37], LogitBoost [27]. However, these classifiers are not commonly used in ANLP literature as the classifiers discussed earlier in this section, as shown in Figure 4.

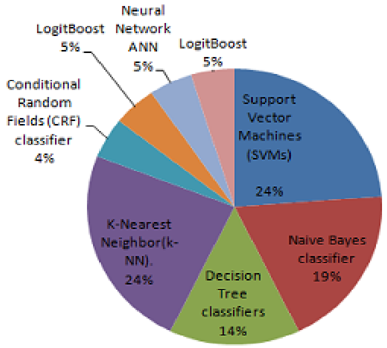

FIGURE 4. The supervised machine learning algorithms used in the literature.

E. EVALUATION PROCESS

In ANLP applications, experiments are conducted to evaluate these applications. The effectiveness measures are needed to assess the performance of applications. These measures represent the ability to take the correct decision and classify an Arabic text to predefined class. Many measures are commonly used to evaluate the effectiveness of the classifiers such as the accuracy, the precision, the recall, the F-measure, etc.

IX. CONCLUSION

High-quality corpora are representative and balanced considering the genre or language that they represent. However, one of the leading challenges in most ANLP applications is having standard gold corpora where all text is tagged correctly according to the most recent guidelines [64]. This challenge is critical for machine learning that relies on data to have the more accurate classification process. There is a lack of specialized tools to aid in the process of data collection and specifically in the Arabic language, considering all its types. Data include corpora, gazetteers, etc.

One of the most important aspects that affects the performance of text categorization is the existence of an adequate number of classes. Furthermore, there is a need to have representational and more accurate features from the text that indicate the text representational class and capture as many Arabic characteristics as possible. However, researchers can take advantage of language-independent features, the features that correlate across languages even though there is dissimilarity between each language [17]. In sentiment analysis applications, it is difficult to identify the direction of Arabic sentiment terms in tweets. This difficulty may be due to the complexity of the Arabic language itself and to the use of Twitter as an informal channel of communication. The readability measurement of Arabic text is still in the improvement stage. It is difficult to identify features that affect the readability of a given text. Choosing representative features affects classification performance. Spam detection researchers are still trying to identify the highest number of spam types. However, as the number of new spam types increases every year, this field should be up to date to eliminate the negative impacts of these new types of spam. By considering the features of a web page language, spam detection tools would be highly effective at capturing spam content [25].

There is a need for tools and data collection that are tailored to the Arabic language. These tools can be combined with a standard and formal specification to annotate an application representative text to create a gold-standard corpus. These specifications can be consulted with experts in the field. An initial attempt to create such a tool is MADAD [65], which is a general-purpose annotation tool for Arabic text with a focus on readability annotation. The same is also applied to ANLP preprocessing tools. Arabic language characteristics are intrinsically challenging for Arabic language processing developers and researchers. The most notable characteristics are the high inflection and the absence of capitalization and punctuation rules. A few basic tools have been established by ANLP researchers to process Arabic text such as sentence splitters, tokenizers, and light stemmers. These tools are used to prepare data for high-level processing stages. As a consequence of heterogeneity and interoperability issues, researchers are continuously developing these basic tools from scratch to be used in their projects. There is an urgent need to have such tools that contain similar and interoperable entities that can be used in one single project [66].

Moreover, we need to identify features that have an effect on text readability either on the word level, sentence level, or document level. An example of word level text readability is the NER application, which can help tag common entities in text and provide an indication of its readability. NER applications are also useful in other applications such as the process of identifying sentiment terms. This application will have a high impact on the sentiment analysis field. NER also has an impact on text categorization, where it helps remove named entities that do not have an impact on classification processes, such as people's names and locations. Since NER applications are essential in many fields, they must use features and models appropriate to the nature of the language used to have superior performance [67]. Readability measurements are also used in applications like spam detection. As almost all spam types represented as text are unreadable, the use of an agent that measures the readability of dubious text will be able to determine whether the text is spam.

This survey provided a comprehensive review of the available literature on ANLP and machine learning-based systems. Although some papers only reviewed the English language while others only reviewed ANLP machine learning techniques, this paper provided a comprehensive view on ANLP, covering this gap in the literature. This survey can serve as an important theoretical foundation for researchers who are interested in this topic. Additionally, we show different ANLP applications and their significance in such machine learning-based systems. Also, we present several systems that utilize ANLP-based machine learning techniques, their methodologies, challenges, and solutions. As this would lay a foundation for researchers and technology practitioners in the field providing them essentials about this topic. In the future work, we will cover more systems that utilize ANLP-based machine learning with semi-supervised and unsupervised approaches. We will also focus on the Deep Learning which is a new and interesting research direction.

ACKNOWLEDGMENT

The authors would like to thank the Deanship of Scientific Research and RSSU at King Saud University for their technical support.

REFERENCES

[1] A. A. Al-Ajlan, H. S. Al-Khalifa, and A. S. Al-Salman, "Towards the development of an automatic readability measurements for Arabic language," in *Proc. 3rd Int. Conf. Digit. Media*, Nov. 2008, pp. 506–511.

[2] P.-C. Chang, M. Galley, and C. D. Manning, "Optimizing Chinese word segmentation for machine translation performance," in *Proc. 3rd Workshop Stat. Mach. Transl.*, 2008, pp. 224–232.

[3] S. Sahu, B. Dongre, and R. Vadhwani, "Web spam detection using different features," *Int. J. Soft Comput. Eng.*, vol. 1, no. 3, pp. 70–73, 2011.

[4] N. Boukhatem, "The Arabic natural language processing: Introduction and challenges," *Int. J. English Lang. Transl. Stud.*, vol. 2, no. 3, pp. 106–112, 2014.

[5] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Trans. Asian Lang. Inf. Process.*, vol. 8, no. 4, p. 14, 2009.

[6] H. Hasanuzzaman, "Arabic language: characteristics and importance. The echo," *J. Humanities Social Sci.*, vol. 1, no. 3, pp. 11–16, 2013.

[7] B. Babych and A. Hartley, "Improving machine translation quality with automatic named entity recognition," in *Proc. EACL-EAMT*, 2003, pp. 1–8.

[8] H. Toda and R. Kataoka, "A search result clustering method using informatively named entities," in *Proc. 7th ACM Int. Workshop Web Inf. Data Manage.*, 2005, pp. 81–86.

[9] H. Abdelbaki, M. Shaheen, and O. Badawy, "ARQA high performance Arabic question answering system," in *Proc. Arabic Lang. Technol. Int. Conf. (ALTIC)*, 2011, pp. 4541–4564.

[10] R. Florian, A. Ittycheriah, H. Jing, and T. Zhang, "Named entity recognition through classifier combination," in *Proc. 7th Conf. Natural Lang. Learn. (HLT-NAACL)*, vol. 4, 2003, pp. 168–171.

[11] L. S. Larkey, L. Ballesteros, and M. E. Connell, "Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis," in *Proc. ACM 25th Annu. Int. Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2002, pp. 275–282.

[12] N. Habash, "Arabic morphological representations for machine translation," in *Arabic Computational Morphology*. Dordrecht, The Netherlands: Springer, 2007, pp. 263–285.

[13] M. N. Al-Kabi, T. M. Hailat, E. M. Al-Shawakfa, and I. M. Alsmadi, "Evaluating english to Arabic machine translation using BLEU," in *Proc. Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 1, pp. 1–8, 2013.

[14] M. Hijjawi, Z. Bandar, and K. Crockett, "User's utterance classification using machine learning for Arabic conversational agents," in *Proc. 5th Int. Conf. Comput. Sci. Inf. Technol. (CSIT)*, Mar. 2013, pp. 223–232.

[15] S. AbdelRahman, M. Elarnaoty, M. Magdy, and M. Fahmy, "Integrated machine learning techniques for Arabic named entity recognition," *Int. J. Comput. Sci. Issues*, vol. 7, no. 4, pp. 27–36, 2010.

[16] V. K. Verma and N. Khanna, "Indian language identification using K-means clustering and support vector machine (SVM)," in *Proc. Students Conf. Eng. Syst.*, Apr. 2013, pp. 1–5.

[17] W. Shen, J. Williams, T. Marius, and E. Salesk, "A language-independent approach to automatic text difficulty assessment for second-language learners," in *Proc. 2nd Workshop Predicting Improving Text Readability Target Reader Populations*, 2010, pp. 1–9.

[18] H. A. Wahsheh, M. N. Al-Kabi, and I. M. Alsmadi, "Spam detection methods for Arabic Web pages," in *Proc. 1st Taibah Univ. Int. Conf. Comput. Inf. Technol.-Inf. Syst. (ICCIT)*, 2012, pp. 486–490.

[19] M. C. Surabhi, "Natural language processing future," in *Proc. Int. Conf. Opt. Imag. Sensor Secur. (ICOSS)*, Jul. 2013, pp. 1–3.

[20] K. Shaalan, "A survey of Arabic named entity recognition and classification," *J. Comput. Linguistics*, vol. 40, no. 2, pp. 469–510, 2014.

[21] Y. Benajiba, M. Diab, and P. Rosso, "Arabic named entity recognition: An SVM-based approach," in *Proc. Arab Int. Conf. Inf. Technol. (ACIT)*, 2008, pp. 16–18.

[22] G. R. Klare, "The measurement of readability: Useful information for communicators," *ACM J. Comput. Document.*, vol. 24, no. 3, pp. 107–121, 2000.

[23] M. El-Haj and P. Rayson, "OSMAN—A novel Arabic readability metric," in *Proc. Lang. Resour. Eval. Conf.*, Portorož, Slovenia, 2016, pp. 250–255.

[24] H. S. Al-Khalifa and A. A. Al-Ajlan, "Automatic readability measurements of the Arabic text: An exploratory study," *Arabian J. Sci. Eng.*, vol. 35, no. 2C, pp. 103–124, 2010.

[25] M. Alsaleh and A. Alarifi, "Analysis of Web spam for non-english content: Toward more effective language-based classifiers," *PLoS ONE*, vol. 11, no. 11, p. e0164383, 2016.

[26] A. A. Hammad and A. El-Halees, "An approach for detecting spam in Arabic opinion reviews," *Int. Arab J. Inf. Technol.*, vol. 12, no. 1, pp. 9–16, 2015.

[27] R. Jaramh, T. Saleh, S. Khattab, and I. Farag, "Detecting Arabic spam Web pages using content analysis," *Int. J. Rev. Comput.*, vol. 6, p. 18, Jul. 2011.

[28] H. A. Wahsheh and M. N. Al-Kabi, "Detecting Arabic Web spam," in *Proc. 5th Int. Conf. Inf. Technol. (ICIT)*, 2011, pp. 1–8.

[29] N. Boudad, R. Faizi, R. O. H. Thami, and R. Chiheb, "Sentiment analysis in Arabic: A review of the literature," *Ain Shams Eng. J.*, vol. 9, no. 4, pp. 2479–2490, 2017, doi: 10.1016/j.asej.2017.04.007.

[30] R. M. Duwairi and I. Qarqaz, "Arabic sentiment analysis using supervised classification," in *Proc. Int. Conf. Future Internet Things Cloud (FiCloud)*, Aug. 2014, pp. 579–583.

[31] R. M. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat, "Sentiment analysis in Arabic tweets," in *Proc. 5th Int. Conf. Inf. Commun. Syst. (ICICS)*, Apr. 2014, pp. 1–6.

[32] R. T. Khasawneh, H. A. Wahsheh, M. N. Al-Kabi, and I. M. Alsmadi, "Sentiment analysis of Arabic social media content: A comparative study," in *Proc. 8th Int. Conf. Internet Technol. Secured Trans. (ICITST)*, Dec. 2013, pp. 101–106.

[33] R. Belkebir and A. Guessoum, "A hybrid BSO-Chi2-SVM approach to Arabic text categorization," in *Proc. ACS Int. Conf. Comput. Syst. Appl. (AICCSA)*, May 2013, pp. 1–7.

[34] B. M. Zahran and G. Kanaan, "Text feature selection using particle swarm optimization algorithm," *World Appl. Sci. J.*, vol. 7, pp. 69–74, Jan. 2009.

[35] H. K. Chantar and D. W. Corne, "Feature subset selection for Arabic document categorization using BPSO-KNN," in *Proc. 3rd World Congr. Nature Biol. Inspired Comput. (NaBIC)*, Oct. 2011, pp. 546–551.

[36] A. S. Ghareb, A. A. Bakara, Q. A. Al-Radaideh, and A. R. Hamdan, "Enhanced filter feature selection methods for arabic text categorization," *Int. J. Inf. Retr. Res.*, vol. 8, no. 2, pp. 10–24, 2018, doi: 10.4018/IJIRR.2018040101.

[37] A. T. Al-Taani and N. A. K. Al-Awad, "A comparative study of Webpages classification methods using fuzzy operators applied to Arabic Webpages," *Int. J. Comput. Inf. Eng.*, vol. 1, no. 7, 2007.

[38] J. Z. Liang, "SVM multi-classifier and Web document classification," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 3, Aug. 2004, pp. 1347–1351.

[39] A. Shdaifat and M. Alian, "Arabic WebPages classification based on fuzzy association," *Int. J. Comput. Sci. Issues*, vol. 11, no. 2, pp. 110–119, 2014.

[40] S. A. Salloum, A. Q. AlHamad, M. Al-Emran, and K. Shaalan, "A survey of arabic text mining," in *Intelligent Natural Language Processing: Trends and Applications* (Studies in Computational Intelligence), vol. 740. Cham, Switzerland: Springer, 2018.

[41] Y. Benajiba, P. Rosso, and J. M. BenedíRuiz, "ANERsys: An Arabic named entity recognition system based on maximum entropy," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*. Berlin, Germany: Springer, Feb. 2007, pp. 143–153.

[42] B. Hawashin, A. Mansour, and S. Aljawarneh, "An efficient feature selection method for Arabic text classification," *Int. J. Comput. Appl.*, vol. 83, no. 17, pp. 1–6, 2013.

[43] A. M. Mesleh and G. Kanaan, "Support vector machine text classification system: Using ant colony optimization based feature subset selection," in *Proc. Int. Conf. Comput. Eng. Syst. (ICCES)*, Nov. 2008, pp. 143–148.

[44] F. Harrag, E. El-Qawasmeh, and P. Pichappan, "Improving Arabic text categorization using decision trees," in *Proc. 1st Int. Conf. Netw. Digit. Technol. (NDT)*, Jul. 2009, pp. 110–115.

[45] W. Zhao, Y. Wang, and D. Li, "A new feature selection algorithm in text categorization," in *Proc. Int. Symp. Comput. Commun. Control Autom. (CA)*, vol. 1, May 2010, pp. 146–149.

[46] A. Mitchell *et al.*, *TIDES Extraction (ACE) 2003 Multilingual Training Data*. Philadelphia, PA, USA: Linguistic Data Consortium, 2003.

[47] A. Pasha *et al.*, "MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic," in *Proc. LREC*, vol. 14, May 2014, pp. 1094–1101.

[48] N. Habash and O. Rambow, "Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop," in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics*, Jun. 2005, pp. 573–580.

[49] N. Habash, O. Rambow, and R. Roth, "MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization," in *Proc. 2nd Int. Conf. Arabic Lang. Resour. Tools (MEDAR)*, Cairo, Egypt, vol. 41, Apr. 2009, p. 62.

[50] N. Habash, R. Roth, O. Rambow, R. Eskander, and N. Tomeh, "Morphological analysis and disambiguation for dialectal Arabic," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2013, pp. 426–432.

[51] M. Diab, K. Hacioglu, and D. Jurafsky, "Automated methods for processing Arabic text: From tokenization to base phrase chunking," in *Arabic Computational Morphology: Knowledge-Based and Empirical Methods*. Norwell, MA, USA: Kluwer, 2007.

[52] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: Rapid prototyping for complex data mining tasks," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2006, pp. 935–940.

[53] G. Grefenstette and P. Tapanainen, "What is a word, what is a sentence?: Problems of tokenisation," Rank Xerox Res. Centre, Tech. Rep., 1994, pp. 79–87.

[54] A.-S. Riyad, K. Ghassan, and G. Manaf, "Arabic text categorization using KNN algorithm," in *Proc. 4th Int. Multiconf. Comput. Sci. Inf. Technol.*, vol. 4, 2006, pp. 5–7.

[55] A. M. Al-Zahrani, H. Mathkour, and H. Abdalla, "PSO-based feature selection for Arabic text summarization," *J. Universal Comput. Sci.*, vol. 21, no. 11, pp. 1454–1469, 2015.

[56] F. S. Al-Anzi and D. AbuZeina, "Toward an enhanced Arabic text classification using cosine similarity and latent semantic indexing," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 29, no. 2, pp. 189–195, 2017, doi: 10.1016/j.jksuci.2016.04.001.

[57] H. A. Naji, W. M. Ashour, and M. A. Alhanjouri, "A new model in Arabic text classification using BPSO/REP-tree," *J. Eng. Res. Technol.*, vol. 4, no. 1, pp. 28–42, 2017.

[58] S. L. Marie-Sainte and N. Alalyani, "Firefly algorithm based feature selection for Arabic text classification," *J. King Saud Univ.-Comput. Inf. Sci.*, to be published, doi: 10.1016/j.jksuci.2018.06.004.

[59] N. Alalyani and S. L. Marie-Sainte, "NADA: New Arabic dataset for text classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 9, pp. 206–212, 2018, doi: 10.14569/IJACSA.2018.090928.

[60] H. M. Noaman, S. Elmougy, A. Ghoneim, and T. Hamza, "Naive Bayes classifier based Arabic document categorization," in *Proc. 7th Int. Conf. Inform. Syst. (INFOS)*, Mar. 2010, pp. 1–5.

[61] A. Selamat, N. C. Ching, and Y. Mikami, "Arabic script Web documents language identification using decision tree-ARTMAP model," in *Proc. Int. Conf. Converg. Inf. Technol.*, Nov. 2007, pp. 721–726.

[62] S. Abdallah, K. Shaalan, and M. Shoaib, "Integrating rule-based system with classification for Arabic named entity recognition," in *Computational Linguistics and Intelligent Text Processing*. Berlin, Germany: Springer, 2012, pp. 311–322.

[63] K. Hamouda, "New techniques for Arabic document classification," M.S. thesis, Dept. Math. Comput. Sci., Heriot-Watt Univ., Edinburgh, U.K., 2013.

[64] J. Pustejovsky and A. Stubbs, *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. Chicago, IL, USA: O'Reilly Media, 2012.

[65] N. N. Al-Twairesh *et al.*, "MADAD: A readability annotation tool for Arabic text," in *Proc. LREC*, 2016, pp. 4093–4097.

[66] Y. Jaafar and K. Bouzoubaa, "A survey and comparative study of Arabic NLP architectures," in *Intelligent Natural Language Processing: Trends and Applications*. Cham, Switzerland: Springer, 2018, pp. 585–610.

[67] R. E. Salah and L. Q. B. Zakaria, "Comparative review of machine learning for Arabic named entity recognition," *Int. J. Adv. Sci., Eng. Inf. Technol.*, vol. 7, no. 2, pp. 511–518, 2017.

[68] J. Zhang, Y. Niu, and H. Nie, "Web document classification based on fuzzy k-NN algorithm," in *Proc. Int. Conf. Comput. Intell. Secur. (CIS)*, vol. 1, Dec. 2009, pp. 193–196.

SOUAD LARABI MARIE-SAINTE graduated in operational research from the University of Sciences and Technology Houari Boumediene, Algeria. She received the dual M.Sc. degrees in mathematics, decision and organization from Paris Dauphine University, France, and in computing and applications from Sorbonne Paris1 University, France, and the Ph.D. degree in bio-inspired algorithms (artificial intelligence) from the Computer Science Department, Toulouse1 University Capitole, France, in 2011. She was an Assistant Professor with the College of Computer and Information Sciences, King Saud University. She was also an Associate Researcher with the Department of Computer Science, Toulouse1 Capitole University, France. She also co-supervised bachelor's and graduate students from the Department of Computer Science. She is currently an Assistant Professor with the Department of Computer Science, Prince Sultan University. She has written several articles and has attended various specialized international conferences. She has also participated in several research projects. Her research interests include combinatorial optimization, heuristics, metaheuristics artificial intelligent methods, and especially bio-inspired algorithms applied to multidisciplinary domains, algorithms analysis and design, bioinformatics, data mining, biometric identification, and natural language processing. She is a Vice-Chair of the ACM Professional Chapter with Prince Sultan University, Riyadh.

NADA ALALYANI received the B.Sc. and M.Sc. degrees in information technology from King Saud University, Riyadh, in 2014 and 2016, respectively, where she is currently pursuing the Ph.D. degree in computer science. Her research interests include natural language processing, machine learning, data mining, human–computer interface, software engineering, and algorithm development.

SIHAAM ALOTAIBI received the B.Sc. and M.Sc. degrees in information technology from King Saud University, Riyadh, in 2014 and 2016, respectively. In 2017, she joined the Deanship of Community Service and Continuing Education with Princess Nourah bint Abdul Rahman University, where she is currently a Collaborator Lecturer with the Community College. Her current research interests include machine learning, natural language processing, and HCI.

SANAA GHOUZALI (M'09) received the master's and Ph.D. degrees in computer science and telecommunications from University Mohamed V-Agdal, Rabat, Morocco, in 2004 and 2009, respectively. From 2009 to 2011, she was an Assistant Professor with the National school of Applied Sciences, Abdelmalek Essaâdi University, Tétouan, Morocco. In 2012, she joined King Saud University, where she is currently an Associate Professor with the College of Computer and Information Sciences. She has supervised over 10 master's and Ph.D. theses. She is involved in many research projects as a principal investigator and a co-principal investigator. She has authored or co-authored over 50 publications, including the IEEE, Springer, and Elsevier journals, and flagship conference papers. Her research interests include image processing, statistical pattern detection and recognition, information security, biometrics, and biometric template protection. She is a member of the IEEE Signal Processing Society–Morocco Section. In 2005, she received a Fulbright Grant for a Joint-Supervision Program with the Visual and Communication Laboratory, Cornell University, Ithaca, NY, USA.

IBRAHIM ABUNADI received the Ph.D. degree in information systems from the School Information Communication Technology, Griffith University, Australia. He taught many courses including human–computer interaction, business process management, enterprise architecture, technology innovations, business analysis, and computer databases and computer applications for business. He was an IT Analyst for Computer Associates and a Strategic Consultant for the Saudi Computer Association. He is currently an Assistant Professor with the College of Computer and Information Sciences and a fellow of the British Higher Education Academy. He has numerous publications in the field of information systems and software engineering and continuously conduct reviews for many conferences and journals in the same fields. His research interests include data mining, technology adoption, e-government, and human–computer interaction. He is a member of Saudi and Australian Computer Societies, ACM, and the Association of Information Systems.