A Closer Look at Arabic Text Classification

Mohammad A.R. Abdeen[1], Sami AlBouq[2]
Faculty of Computer and Information Systems
Islamic University of Madinah
Madinah, Saudi Arabia

Ahmed Elmahalawy[3]
Faculty of Engineering
Menoufia University
Menoufia, Egypt

Sara Shehata[4]
Faculty of Computer Science
Ain-Shams University
Cairo, Egypt

***Abstract*—The world has witnessed an information explosion in the past two decades. Electronic devices are now available in many varieties such as PCs, Laptops, book readers, mobile devices and with relatively affordable prices. This and the ubiquitous use of software applications such as social media and cloud applications, and the increasing trend towards digitalization, the amount of information on the global cloud has surged to an unprecedented level. Therefore, a dire need exists in order to mine this massively large amount of data and produce meaningful information. Text Classification is one of the known and well established data mining techniques that has been used and reported in the literature. Text classification methods include statistical and machine learning algorithms such as Naive Baysian, Support Vector Machines and others have widely been used. Many works have been reported regarding text classification of various languages including English, Chinese, Russian, and many others. Arabic is the fifth most spoken language in the world. There has been many works in the literature for Arabic text classification. However, and to the best of our knowledge, there is no recent work that presents a good, critical and comprehensive survey of the Arabic text classification for the past two decades. The aim of this paper is to present a concise and yet comprehensive review of the Arabic text classification. We have covered over 50 research papers covering the past two decades (2000 - 2019). The main focus of this paper is to address the following issues: 1) The techniques reported in the literature including. 2) New Techniques. 3) Most claimed efficient technique. 4) Datasets used and which ones are most popular. 5) Which feature selection techniques are used? 6) Popular classes/categories used. 7) Effect of stemming techniques on classification results.**

Keywords*—*Arabic text classification; support vector machines; k-NN; Naive Bayesian; decision trees; C4.5; maximum entropy; feature selection; Arabic dataset

I. Introduction

The 21st century is truly the information age. With the introduction of ubiquitous computing via electronic/mobile devices in an unprecedented way in just about all aspects of live resulted in the production of data in Zetta scales. Today, one can move across the four corners of the earth without carrying a dime of cash. Also we can communicate live with video with just about anyone that has an average quality mobile phone. People are exchanging messages, documents, recipes, life experiences, medical advice, all through the now fully connected world.

The resulting mass of data needs to be analyzed and classified so that useful information is extracted. Data mining is an area of computer science that deals with developing ways to extract information out of existing masses of data. One of data mining applications is text classification (TC). Text classification is a data mining technique that is used to assign a given document in a set of documents to a given class or category.

There has been many works in the literature that addresses the topic of text classification in various languages such as English [1], Chinese [2], Russian [3] and many other languages.

Arabic language is one of the most widely spoken languages in the world. It is the fifth most spoken language in the word and the fifth most used language on the Internet. More than 6.0% of the world's population speaks Arabic language (more than 422 million speakers) [4]. Arabic language has rich morphology and an intricate orthography. The span of highlight feature vectors may increment and that make the undertaking of highlight choice progressively imperative to maintain a strategic distance from the insignificant data [5] that produces different words with different meanings. On contrary of Arabic text, there are many benchmarking corpus that can be used for Latin base language, Japanese and Chinese text classification [6] [7]. This language has specific letters known as Arabic vowels (waw, yaa, alf) that require a special system of morphology and grammars. The others are called consonant letters [8]. What also distinguishes Arabic is the huge amount of vocabularies and concepts [9].

There are 28 letters in Arabic language, in addition to the Arabic hamza (ء) which is considered as a letter by some Arabic linguistics and it is written from right to left. It has two genders: feminine and masculine. Numerical are singular, dual, and plural numbers. Grammatical are three cases: nominative, accusative, and genitive. A noun has three linguistic cases: nominative case when it is subject; accusative case when it is the object of a verb; and the genitive case when it is the object of a preposition [10] [11].

Arabic doesn't employ an upper/lower case. It also employs diacritics which represent a small vowel letters such as "fatha, kasra, damma, sukun, shadda, and tanween" [9].

The process of Text-Classification is concerned with assigning a given document or a group of documents to a given class or category. As an example, the task of classifying various news of a news website to several categories (Weather, Politics, Sport...). The explosion of Internet contents in various languages including Arabic presented a pressing need for effective classification. Classification algorithms are developed for this sole purpose which are called Classifier) [12] [13].

The remainder of the paper is organized as follows: Section II presents a short outline of Text Classification. Section III presents a brief description of the process of Arabic text classification stages. Section IV represents an overview of the previous work. Section V presents a tabulated summery of survey results. Section VI provides conclusion and recommendation.

II. Overview of Text Classification

Text Classification (TC) is the process of assigning given text in a document to preset categories [14]. Another definition describes the classification process as an assignment of category labels to natural language documents with the possibility that a one document be included in more than one category [15].

Manual TC is the process of classifying documents by some trained individuals. This process is time consuming and is prone to human errors. In addition, with the latest surge in data sizes and variety of classes of documents, the manual process is certainly not scalable and impractical [16] [17].

Automatic TC is defined as the assignment of documents in a collection to a predefined class or category. Therefore, text classification can be achieved through machine learning techniques by training the algorithms with a training dataset [18] [17]. Automatic TC approaches have been reported in the literature during the past three decades. Numerous algorithms have been introduced and evaluated in some surveys [19] that address the most common TC algorithms and a lot of works have been achieved to evaluate and compare between these algorithms.

Text classification can be performed with two main techniques/methodologies; the statistical techniques, and the machine learning techniques. The following two sections presents some details about those techniques.

A. Statistical Techniques

Statistical text classification techniques are based on mathematical foundations. These techniques have been developed relatively earlier than the machine learning techniques and are more suitable for relatively small datasets. Some of them are also more suitable for binary classifications rather than multiclass classification. Examples of those techniques are: the Frequentist procedures, Bayesian procedures, and the Binary and multiclass procedures.

B. Machine Learning Techniques

Due to the surge in the size of data for the past two decades, automation process is required to achieve the goals of information extraction and classification/clustering of data for a variety of purposes. Those include email filtering and routing; news observing; Spam filtering and search engines [20]; newsgroups classification, and survey data grouping [17]. Depending on the nature of the available data, machine learning can be classified to three main categories [10] [21].

1) Supervised learning where training of the model is required a priori using a previously labeled data. Data labeling can be a difficult task when data size is significantly large as it is done by humans. Some of the known supervised machine learning techniques include Naive Bayes, Logistic Regression, Support Vector Machine, Artificial Neural Networks and Decision Trees.

2) Unsupervised learning where initial training data is not required. The model rather groups the data in clusters (clustering) without labeling based on some feature similarity. Principal component analysis and Self-Organizing Maps (SOM) are some of the popular unsupervised learning algorithms.

3) Semi-Supervised machine learning: is a combination of the above mentioned two techniques. It employs a relatively small amount of labeled data with a significantly larger one of unlabeled data. Examples of the semi-supervised learning approaches are generative and the graph-based models.

There are two issues of Arabic language are the high dimensionality of the feature space and the rate of the precision is approximately low. The complexity of many learning algorithms increases in parallel with the increase in data dimension.

There are several reports in the literature that claim to present a review study on Arabic text classification. As an example, the work presented in [9] claims to be survey of the Arabic text classification. The author however provided a very broad background and definition of the term and gave a brief background of the popular four techniques, i.e. the Naïve Bayesian, the K-Nearest Neighbor, the Support Vector Machine and the Artificial Neural Network. The author did not provide a detailed or comprehensive treatment of the subject nor of the various techniques or algorithms used nor the results of those algorithms.

The nature of the data source may influence the execution of a classification algorithm; the insignificant and repetitive highlights of data may lessen the nature of the outcome [22]. The span of feature vectors may increment by wealth of the language that make task of feature selection vital to stay away from the immaterial component [5].

For Arabic language there is a lack of the studies on the classification of Arabic text documents with limitation of free benchmarking dataset [6] [23]. On the other hand, the richness in morphology of Arabic language significantly increases the length of the feature vector and that significantly influenced research and studies in the field of text classification [24].

The research deanship of Islamic University of Madinah.

There are three principle stages for the text classification as shown in Fig. 1 [24] [15]:

- Data pre-processing
- Text classification
- Evaluation

Classifying Arabic documents requires accomplishing some preprocessing steps for the documents through stemming the words; this process is quite a major issue in terms of reducing the number of related words in a document.

The text classification can be divided into other sub problems that have been investigated in the literature, for example, the document indexing, the weighting assignment, the document grouping, the dimensionality reduction, the threshold determination and the types of classifiers.

Document indexing is related with the method for extracting document's keywords. There are two fundamental ways to deal with the document indexing: the first methodology considers list terms as bag-of-words [25] and the second sees the file terms as phrases [7] [17] [26] [27].

A drawback of the first approach is that it complicates the extraction process of index term by increasing the number of words that must be dealt with in the document as well as dealing with irrelevant words (unrelated to any category). Classifying Arabic documents requires achieving some preprocessing steps for the documents through stemming the words, this process is quite significant in terms of eliminating the number related words in a document. The preprocessing of our dataset was an important step as it increased the accuracy of the classification and reduced the required memory size for the classification process.

Several techniques have been introduced to perform preprocessing tasks such as stemming, root extraction and thesaurus. Weight assignment procedures associate a real number that ranges from 0-1 for every term in the collection of documents [19], the loads will be required to classify new arrived documents. Different information retrieval models use different methodologies to compute these weights, for example the Boolean model assigns either 0 or 1 for each index term.

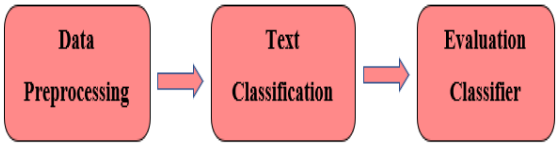

Fig. 1. Text Classification Stages.

Historically, the most widely famous Naïve Bayesian model was known as the binary independent classifier [26]. One of the first attempts to address clustering techniques for TC problem was stated through introducing a comparison between the k-means algorithms and hierarchical clustering algorithms [25]. The results showed better performance for the hierarchical algorithms although they were slower than the k-means algorithm. Distinctive TC strategies have been raised to categorize documents, for example, K-nearest neighbor KNN [28], the K-NN different models compute the separations between the document index terms and the known terms of every category by applying distance functions, for example, cosine, dice similarity or Euclidean functions, the returned classes are the kth classes with elevated scores.

III. Arabic Text Classification Stages

Text classification in general has five main stages: Data collection, Data preprocessing, Feature extraction/selection, Text classification, and Classifier evaluation [7]. Fig. 2 depicts those stages [29] [30].

Those main stages can be further broken to more fine grain tasks as given in the below list [23]:

- Data gathering;
- Data labeling;
- Data cleaning (removal of stop word, stemming);
- Feature extraction;
- Feature Selection;
- Classification Model Training;
- Classification algorithm testing.

1) Data Gathering: There are numerous standard data sets for English text classification that freely accessible. Unfortunately for Arabic language, we are not aware of a open access standard dataset [7]. The Open Source Arabic Corpus [31] is freely available but is not standardized. The majority of the researchers in the field of Arabic text classification built their test corpus from online Arabic news sites [32]. The scope of the chosen documents that introduced in the dataset varied from as low as 240 documents split into six categories [23] up to 17,658 documents partitioned into seven genres [33].

2) Data Pre-processing (labeling and cleaning): The purpose of this stage is to remove words that do not contribute to the semantics of the document such as stop words. Example of these words are the pronouns. This stage also removes suffixes and prefixes. It also combines words of the same root/origin. The main objective of this stage is to reduce the feature set of a given document and provide better classification accuracy.

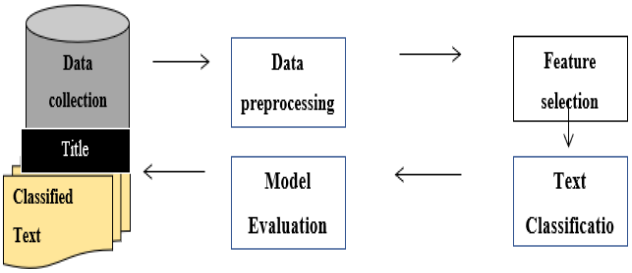

Fig. 2. Detailed Text Classification Stages.

To accommodate various writing styles, some normalization is also performed. This includes Hamza " ء " Taa Marboutah " ة", Alif " ا" and the Yaa "ى" . Combining words of similar root/origin is performed by stemming algorithms. There are three main types of stemming algorithms: the root-based stemmer; the light stemmer; and the statistical N-gram algorithm. These types of stemmers are summarized below.

1) Light Stemmer: In this method, the commonly used affixes (prefixes, infixes and suffixes). The commonly used affixes and stored in a separate file and are used in stem extraction [34]. This method is relatively easy and contributes positively to the efficiency of the classification algorithm.

2) Root Stemmer: In this method, the actual root of the word is found by applying some morphological rules. These rules are largely language dependent. A variant of this approach requires to build and search a lexicon, which is very time consuming and inefficient [35]. Another approach used by Al-Serhan [36] did not require a lexicon or to store roots in a separate file.

3) N-gram Stemmer: This stemming technique is classified as a statistical and is language independent. The algorithm forms a gram of two (bigram), or more, consecutive letters. The common di-gram between words is used to calculate a metric of similarity referred to as the Dice's coefficient [37].

One significant difference between the text classification of Arabic language and other commonly used languages such as English is the fact that stemming could significantly affect the result of the classification [38] [11] [39] [40]. Arabic is a highly inflectional and morphologically rich language and Arabic wards can come from a stem of three, four, five, and six letter words. Almost 80% of the Arabic words come from a three-letter root [38].

The work by [38] studied the effect of various stemming techniques on the accuracy (precision and recall) of the K-Nearest Neighbor (KNN) classifiers. The study concluded that light stemmer produces the best results as compared to root stemmer and statistical stemmers.

A. Feature Extraction and Selection

Two important states of text preprocessing prior to classification is feature extraction and feature selection. The process pf feature extraction is concerned with transforming the unstructured text into a structured representation and for removing redundancy which facilitates further processing and the application of machine learning techniques. Feature selection is an even further preprocessing step to exclude irrelevant features and reduce the high dimensionality of the result of the feature extraction prior step. There are three main categories of feature selection algorithms: the Wrapper, the Filter, and the Embedded [41]. Some of the most commonly used feature extraction is Chi-square, correlation coefficient scores, information gain, recursive feature elimination, and the Least Absolute Shrinkage and Selection Operator. Another approach for feature selection is using Rough Sets (RS) in various languages including Arabic [5] [42] [43] with satisfactory results.

B. Text Classification Methods and Algorithms

Text classification algorithm are the following: Decision Trees: which are used to classify documents through building a tree by computing the entropy function of the selected index terms such as ID3 and C4.5. Naive Bayesian models which have portrayed great outcomes in the text classification field. Historically, the most widely famous Naïve Bayesian model was known as the binary independent classifier. In the Clustering Techniques, the issue was stated through introducing a comparison between the k-means algorithms and hierarchical clustering algorithms. The conducted results showed better performance for the hierarchical algorithms although they were slower than the k-means algorithm. In the Distinctive text classification, strategies have been developed to sort documents, for example, K-Nearest Neighbor KNN, the K-NN different models compute the distances between the document index terms and the known terms of each category by applying distance functions, for example, cosine, dice similarity or Euclidian functions, the returned classes are the kth classes with most noteworthy scores. Support Vector Machine Support vector machines (SVMs) are considered one of the most well-known text classifiers. SVMs are one of the supervised machine learning techniques. In SVMs a training algorithm is used to build a model that will be used to assign a new unknown document to one category from a set of predefined categories. SVMs can be used to perform a linear and a non-linear classification. Fig. 3 explained Linear SVM vs Nonlinear SVM [44].

C. Model Evaluation

Text classification is assessed based on the efficiency and effectiveness of categorization. There are some techniques that have been utilized to quantify the advancement of the classifier. One of those techniques is F1, precision and recall that are utilized in the field of information retrieve and machine learning. There are different types of measurements to test the classifiers and this may not justify the result such as:

- F1-precision measure.
- Fallout and error rate as accuracy measure.
- K-fold cross-validation technique utilized to test the precision.

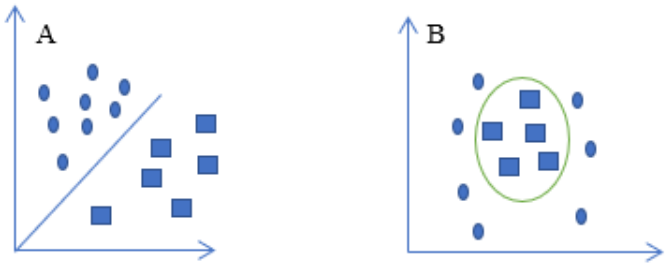

Fig. 3. Linear SVM (Left) vs. Nonlinear SVM (Right).

IV. Previous Work

There are some reports in the literature claiming to provide a review of the Arabic text classification work to date. To the best of our knowledge, none so far is reviewing the subject from the viewpoints considered in this research nor the works reported present a comprehensive treatment of the subject.

As an example, in [45], despite the fact that the author claims to present a review of the TC topic, they reviewed only four papers [46] [17] [32]. These works are relatively old and there has been numerous other works in the past two decades that the author did not include in their review.

The work presented by [29] was a relatively short overview of previous work and most of the cited work is a decade old. Despite the fact that authors cited various techniques used in classification process, they missed some significant recent works and did not consider and did not consider the effect of various stemming techniques on the accuracy of the result.

In [44] despite the fact that the title is claiming to review the topic, in fact the author presented a comparative study of Arabic text classification using various techniques such as NB, KNN, SVM and Artificial Neural Networks. The authors used their own dataset of size 4000 documents collected from Arabic websites.

There are many researches in classifying English documents. In addition, there are numerous researches in European languages, for example, German, Italian and Spanish, Asian languages as well, for example, Chinese and Japanese. In the past decade, the work on Arabic text classification started with four primary researches [47]:

1) El-Kourdi et. al. [17], used naïve bayes (NB) algorithm to automatic Arabic document classification. The average accuracy revealed was approximately 69%.

2) An Named Entity Recognition (NER) system is called Siraj from Sakhr [48]. The framework has no specialized documentation to clarify the strategy utilized in the framework and the exactness of the framework.

3) Sawaf et. al. [46] introduced a framework for statistical classification methods strategies, such as maximum entropy to classify and cluster news articles. The best classification accuracy obtained was approximately 63% with a precision of 50% which is a low reported precision in the research area.

4) El-Halees [49] depicted a technique that depends on affiliation principles to classify Arabic documents. The classification precision revealed was 74.41%.

The majority of authors used an in-house built Arabic dataset with various sizes and contents, with the Internet websites as their major source of data [50] [40]. The examinations in the field of Arabic text classification utilized Naïve Bayes [17] [51], Support vector machine [52], Decision Trees [24] as classifier algorithm. In [24], the author gathered in-house Arabic corpus that comprises of 1000 documents, were separated into 10 categories and utilized it to compare the performance of three Arabic text classifiers: Naïve Bayes, K-Nearest Neighbor and Distance Based. The recall, precision, error rate and fallout are utilizing to compare the accuracy of the classifiers. The data was preprocessed by expelled the stop words and extricated the root of the words. The results of the experimentation demonstrate that the Naïve Bayes classifier beats the other two (over 95%).

In [53] the authors' main objective was to classify the Internet content of the Arabic text. They used Internet data of size 40 Gigabytes and classified them into five categories, namely, religion, politics, culture, sports, and economics. They used the NB and K-NN machine learning algorithms.

In [16], the authors utilized in house corpus comprises of 242 documents which have a place with six categories to assess the execution of two classifiers K-NN and Naive Bayes to classify the Arabic content. The k-fold cross-validation strategy is utilized to test the accuracy. They extricated feature set of keywords to improve the execution. The outcome demonstrated that K-NN classifier would do well to execution.

In [26], the authors examined Naïve Bayesian and Support Vector Machine on various Arabic data indexes. The Exploratory outcomes against various Arabic text categorization data sets uncover that SVM algorithm outflanks the NB. While in [15] which thought about the Support vector machine and C5.0 classifier and C5.0 classifier gives better precision. By [23], Support vector machine had demonstrated the predominance in highlight determinations, weighting techniques, and classification algorithms, trailed by the decision tree algorithm (C4.5) and Naive Bayes. The best classification precision was 97% for the Islamic Themes dataset, and the least exact was 61% for the Arabic Poems dataset. Duwairi et al. [38], the examination makes a comparison between (stemming, light stemming, and word cluster). For training purposes, they pick K Nearest Neighbor KNN method, to demonstrate that light stemming accomplishes best performance and least time of model development. Another study [18] looked at 3 Feature Subset Selection (FSS) measurements. They did a relative report to look at the impact of the component choice measurements as far as accuracy. The results in general revealed that Odd Ratio (OR) worked superior to the others. A few examinations concentrated on different procedures like N-gram and distinctive distance measures and demonstrated their impacts on Arabic TC. El-Kourdi et al. [17] classified Arabic text documents automatically utilizing NB. The normal precision revealed was about 68.78%, and the best accuracy reported was about 92.8%. El-Kourdi utilized a corpus of 1500 text documents having a place with five categories; each category contains 300 text documents. All words in the documents are changed over to their foundations. The vocabulary size of

resultant corpus is 2,000 terms/roots. Cross-validation was utilized for assessment.

Sawaf et al. [46] utilized Maximum entropy (ME) to make a grouping to News articles. The investigation gives precision about 62.7%. Al-Zoghby [32] utilized Association Rules for Arabic text classification, furthermore, he utilized CHARM algorithm with soft-matching over hard big O exact matching. Data indexes comprise of 5524 records. Each record is a scrap of emails having the subject - nuclear. The vocabulary size is 103,253 words. Harrag et al. [54] utilized the feature selection dependent on hybrid approach for Arabic text classification. He utilized direct tree algorithm and the accuracy was of 93% for scientific data set, and 90% for literary data-set. Harrag gathered two data indexes; the first one is from the scientific encyclopedia. In [29], the authors addressed the issues of lacking free open Arabic corpora. In [30] centered around the distributed researches in the field of Arabic text classification and illustrates a logical view its procedure and camper the assessment of text classification strategies that were utilized. The authors in [12] gave a novel framework for text classification dependent on BPSO/REP-Tree hybrid. The first term refers to the "Binary Particle Swarm Optimization" that we use it for the feature selection process and the second term refers the classifier we used "Reduced Error Pruning Tree". We will show the results of the experiments on a data-set collected from the BBCArabic website using the Weka tool which specific for data classification. In [16] actualized the K-NN and Naïve Bayes algorithm so as to make a commonsense comparison among them and past studies. The algorithms are considered as probably the most renowned algorithms in the field of text classification. In [17] utilized a Naive Bayes (NB) algorithm which is a statistical machine learning algorithm. It is utilized to classify non-vocalized Arabic web documents (after their words have been changed to the corresponding canonical form, i.e., roots) to one of five predefined categories. In [4] utilized term Frequency-Inverse Document Frequency (TF-IDF), with the Convolutional Neural Network (CNN) on different sizes of the datasets (data_27k, data_55k, data_83k, data_111k). The accuracy was over 92%. In [55] Master-Slaves technique (MST) was updated and implemented on Arabic text classification. About 16757 Arabic documents are used 90% as training data and 10% as testing data. The accuracy was 86.495%. In [56] presented a parallel classification approach based on the Naïve Bayes algorithm for large volume of Arabic text using MapReduce with enhanced speedup, and preserved accuracy. The accuracy was close to 97%.

Master-Slaves technique (MST) consists of one classifier as a master and several other classifiers as slaves. The master classifier modifies its probability according to the results of slaves by multiplying each probability by a factor. This factor reflects the weight of those slaves. Multinomial Logistic Regression (MLR) is also known as Log-linear Models, and a conditional exponential classifier or logistic regression classifier.

Multinomial logistic regression modelling is a general and an intuitive way for estimating a probability from the data and it has been applied successfully in various natural language processing tasks [55]. Voting Technique building multiple models (typically of different types), and simple statistics (like calculating the mean) are used to combine predictions This work depends on NB, MW, MLR, and KNN in its results.

Maximum Weight (MW) is a new classifier suggested. It is a very simple method for text classification, which works by selecting the highest weight of the term among the categories and only these values are used to predict the best class for any input example [55]. Rocchio classifier is an information retrieval algorithm. Rocchio classifier is a linear classifier, and it is based on relevance feedback. Rocchio classifier developed based on Vector Space Model [44]. Artificial Neural Networks (ANNs) are one of the main tools used in machine learning. The concept of ANNs is inspired from biological human brains. In ANNs the system learns how to perform tasks after training stage. ANNs are available in different forms and shapes such as supervised and unsupervised learning. When ANNs are introduced the aim is to solve problems in the same way the human brain solves it. ANNs can be found in different shapes such as single layer perceptron, radial basis network (RBN), multi-layer perceptron (MLP) [44]. Arabic Text Classification using deep learning Technics: Term Frequency-Inverse Document Frequency (TF-IDF) with the Convolutional Neural Network (CNN) one of the most famous deep learning algorithms used especially in image processing and pattern recognition fields. CNNs is simple and efficient method to classify Arabic text from large dataset [4]. Fig. 4 shows a comparison between the traditional and deep learning techniques for various size dataset.

The accuracy is tested using the K-fold cross-validation strategy. In this strategy, the original sample is grouped into K sub samples. Of the K sub-samples, a solitary sub sample is held as the validation data for testing the model, and the rest of the K-1 sub samples are utilized as training data. The cross-validation process is then rehashed K times (the folds), with every one of the K sub samples utilized precisely once as the validation data. The K results from the folds at that point can be arrived at the midpoint of (or otherwise combined) to create a solitary estimation.

The advantage of this strategy over then rehashed arbitrary sub-sampling method is that all perceptions are utilized for both training and validation, and every perception is utilized for approval precisely once. The 10-fold cross-validation is generally utilized.

In stratified K-fold cross-validation, chosen with the goal that the mean reaction esteem is roughly equivalent in every one of the folds. Because of a dichotomous classification, this implies each fold contains generally similar extents of the two types of class labels.

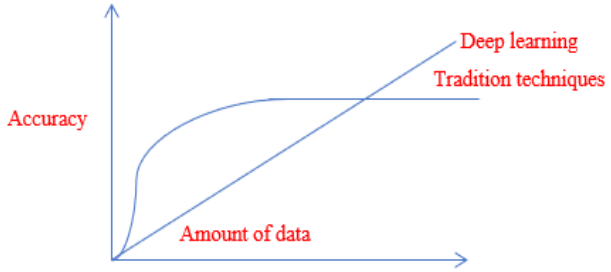

Fig. 4. Deep Learning vs Traditional Techniques.

V. TABULATED SUMMARY OF THE SURVEY WORK

The section shows a summary of the survey work in a tabular form. We present three tables, one showing the feature selection methods, the second shows the classification algorithms used, and the third shows the dataset chosen by various authors for testing.

Table I gives a comparison of various feature selection algorithms covered in this work and the best performing.

Table II gives a comparison of various text classification techniques used and the advantages and performance results of them.

Table III below shows the various datasets used by the referenced papers in this research. The table also shows the source of the dataset whether is in-house or open source. It also shows the size and categories of the datasets.

TABLE. I. FEATURE SELECTION ALGORITHM

Authors	**FS Technique**	**Conclusions**
[44]	Information gain, gain ration ,Chi square	Experiments show that chi-square is a little bit better than gain ratio and information gain
[27] [12]	Binary Particle Swarm Optimization (BPSO-KNN)	Results show that the BPSO-KNN produces classification results that compares well when other feature selection techniques are used
[57]	Term collocation	Results show that the classification outperforms the full and summery corpus techniques.
[55]	TF-IDF	Only one feature selection is presented by author due to its simplicity
[8]	None specified	–
[58]	Root and light stemmers	Light10 stemmer produced better results than the roo-based stemmer.
[33]	Authors used 10 feature selection methods that are included with the RapidMiner software. These include IG, TF, DF, CHI square, and Mutual Information (MI)	The CHI square showed best average accuracy when used with the SVM classifter.
[31]	They used the root-based and the light stemmer techniques.	The number of features for the light stemmer were larger than those of the root-based stemmer (15,000 as compared to 12,000).
[26]	Not specified	.
[16]	The TF-IDF term weighting	
[47]	The TF-IDF term weighting	.
[15]	The Chi-Square technique	The Chi-square is applied on document frequency and the top 30 terms in each class of documents.
[28]	TF, TF-IDF, WIDF	WIDF scheme provides the best performance on k-NN, while TF-IDF shows the best performance on Rocchio.
[10]	They have used six feature selection methods.	The ones used are Information Gain (IG), Chi-Square, Mutual Information (MI), NGL, GSS, and Odds Ratio (OR) feature selection. .
[52]	TF and IG	They used a combination of Term frequency to eliminate rare terms and Information Gain to leave the most valuable terms. They also used a combination of local an global features.
[54]	A combination of TF and DF is selected.	Authors tested various values of TF and analyzed the effect of the TF threshold on the efficiency of the classifier. An improvement of as high as 25% can be achieved by selecting the right value of the TF threshold.
[6]	They used the TF	Authors did not report any other feature selection methods.
[17]	They used TF-IDF	Authors tested their technique with and without root extraction selection. They concluded that the efficiency of the classifier is not sensitive to root extraction.

TABLE. II. TEXT CLASSIFICATION ALGORITHMS

Authors	TC Technique	Conclusion
[44]	Information square gain, gain ration,	Experiments show that chi-square is little bit better than gain ratio and information gain.
[27] [12]	Binary Particle Swarm Optimization (BPSO-KNN)	Results show that the BPSO-KNN produces classification results that compare well when other feature selection techniques are used.
[57]	SVM, NB, J45 and KNN	Results show that their AMLT technique based on term collocation (full-corpus bigram) outperforms other full-corpus and summery corpus.
[55]	Master-Slave technique where the NB is the Master classifier and the KNN, MLR and Maximum Weight (MW) are slave classifiers	Results show that using weighted voting of the slave classifiers will produce a better accuracy than the accuracy of each individual classifier.
[8]	NB, SVM, LSVC	It was found that the LSVC showed the best values of precision and recall
[58]	NB, SVM, KNN, Decision Trees (J48), and Decision Tables. Authors utilized machine learning software Weka and RapidMiner	SVM classifier showed the better results. Also the use of the light10 stemmer produced better results than the root based stemmer.
[33]	Conventional Neural Nets (CNN)	Experiments show with dataset of this size, the CNN outperforms traditional techniques such as SVM.
[31]	KNN, C4.5, NB, MLP, SVM	The results showed that the SVM classifier showed the best accuracy as compared to the other four classifiers with achieved average accuracy of 72% especially when using the Chi-Square term selection method. The NB classifier came very close to the SVM with an accuracy of 68%.
[26]	Artificial Neural Nets (ANN), SVM and the Hybrid Bee Swarm Optimization (BSO) BSO-Chi2-SVM	The authors compared the three techniques while using various stemming techniques such as light and root stemmers. Their results showed that the BSO-Chi2-SVM with light stemmer slightly out performs the other two techniques. The approximate accuracy for the three techniques was around 94% for the three techniques. They also showed that the model training time for the hybrid technique is more than the others [59].
[16]	SVM and NB	Authors used F1, Precision and Recall measures to evaluate the techniques. In most categories, the SVM technique outperformed the NB.
[47]	k-NN and NB	The authors compared their work to other work for the same algorithms that is reported in the literature and their results are in agreement. They also showed that the K-NN classifier outperform the NB. In addition, the performance results of the k-NN algorithm depend on the selection of k value. They showed that the effectiveness of the model declines after a value of k = 15.
[15]	They developed a tool called Arabic Text Classifier (ATC) based on NB and k-NN	The tool computes the accuracy of the two algorithms and selects the average accuracy.
[28]	SVM and Decision Trees C5.0	RapidMiner and Clementine software tools have been used to test the two algorithms. The results showed that the accuracy of the C5.0 algorithm shows a better accuracy by over 10% as compared to the SVM.
[10]	Authors used the k-NN, Rocchio and the NB	NB gave best measure results with Micro-F1 followed by k-NN and the Rocchio.
[52]	Authors used only the SVM algorithm	They used a publically available software called TinySVM. They tested the classifier for with and without feature selection. They concluded that the CHI, GSS and NGL showed best performance with the SVM classifier.
[54]	Tested SVM, NB, k-NN, and the Roc- chio classifiers	Their results showed that the SVM classifter outperform the other in high dimensional feature space.
[6]	Authors used the Decision Tree classifier (C4.5).	They compared their classifier with others such as the NB, and Vector Space Model (VSM). The F1 results of the comparison showed that the C4.5 classifier produces better values than other classifiers in the comparison.
[17]	The Maximum Entropy Classifier	Authors built their own software (ArabCat) using Java programming and based on an existing Arabic morphological analyzer. They compared their system to other existing systems such as those built by Sakhr, Al-Halees, Sawaf and El-Kourdi. The overall performance of their system was better than the others with some exception in the precision value.
[44]	NB algorithm	The average accuracy of all five categories were 68.8% which is comparable to other work by [60] that showed a result of 75%.

TABLE. III. DATASETS

Authors	Dataset	Conclusion
[44]	Authors built their own dataset.	Dataset consisted of 4000 documents collected from news websites including Al-Jazeerah, Al-Hayat and the Saudi press agency. The dataset is split into 8 categories including Economics, Politics, Arts, Culture, TEchnology, Science and Education.
[27] [12]	They have used three readily available datasets.	the Akhbar-Alkhaleej Arabic dataset that consists of 5690 documents, the Al-jazeerah dataset (Alj-News) that consists of 1500 documents, and the Alwatan Arabic dataset of 20,291 documents. These datasets are split into 4-5 categories including Sports, Arts, Economy and Religion.
[57]	800 Arabic text documents divided into four categories Economy, Politics, Religion, and Science, 200 documents for each class	Results show that the classification is better when using the proposed technique of term collocation.
[55]	16757 Arabic documents collected from Al-Sabah newspaper that are manually organized into five different categories	Authors used 90/10 percentage for the training and testing data and used a ten-fold cross-validation method.
[8]	Web content dataset but not mentioned specifically	Due to the limited size of the dataset the authors suspect the result could have a probability of inaccuracy due to insufficient training.
[58]	2700 online Arabic articles collected by Diab Abu Aiadh equally spread over nine categories.	Due to their hardware limitations, they managed to perform classification of 1000 documents spread over 5 categories.
[33]	Used three Arabic websites (Assabah, Hespress, Akhbarona) website to obtain around 111,000 documents	Authors claim to have collected over 300 million words covering five categories including sports, politics, culture, and economics. Experiments show that the training time for such a large dataset can be as high as 10 hours.
[31]	Authors built their own corpus that consisted of seven genres including Saudi press, Saudi newspapers, Writers, websites, Forums, and Islamic topics.	The total number of documents collected was over 17,000 with the number of wards of 11 million in total.
[26]	Authors used the Open Source Arabic Corpus (OSAC).	The corpus consists of 22,429 textual documents representing ten categories such as Economics, Religion, Health and Education.
[16]	The Saudi Newspapers (SNP)	Dataset consists of 5121 documents of various lengths that are categorized into seven categories (Culture, Economics, General, Sports, Information Technology, Politics and Social).
[47]	Used their own corpus	Corpus consisted of 242 documents grouped into 6 categories (which is a relatively small dataset).
[15]	Built their own corpus out of various newspaper websites including Al-Jazeera, Al-Nahar, Al-hayat, Al- Ahram, and Al-Dostor	The corpus consisted of 1562 documents in 6 categories.
[28]	They used the Saudi Press Agency (SPA) and The Saudi News Papers (SNP) and other genres including writers, Islamic topics and others.	The corpus consisted of 17,658 documents and over 11 million words. A 70/30 dataset split is used for training and testing results.
[10]	In house built from online available Arabic newspaper archives including Al-Jazeera, Al-Nahar, Al-hayat, Al- Ahram, and Al-Dostor	The corpus consisted of 1445 documents of various lengths that are spread over 9 categories including Medicine, Sports, Religion, and Politics.
[52]	in house built from online available Arabic newspaper archives including Al-Jazeera, Al-Nahar, Al-hayat, Al- Ahram, and Al-Dostor	The corpus consisted of 1445 documents of various length that are spread over 9 categories including Medicine, Sports, Religion, and Politics.
[54]	In house built collection of documents.	Authors collected 1,132 documents that contained about 95,000 words (22000 unique words). These documents were collected from the three news websites, the Ahram, Akhbar, and El-Gumhuria.
[6]	Authors build their own corpora but based on two existing ones	The two exiting datasets are the Arabic Scientiftc Encyclopedia ”Haal Taalam” or “Do you know” that contained 373 documents and 8 categories. The second is based on “Prophetic Hadeeth” and consists of 453 documents and 14 categories.
[17]	Built their own corpus from the Al-Jazeerah news website	The corpus contained six categories including arts, science and technology, politics, sports, culture, economics and health.
[44]	Authors built their own dataset collected from the Al-Jazeerah news website	They collected 1500 documents spread equally over five categories including sports, business, culture-art, science, and health. Each category had 300 web pages.

VI. Conclusions

In this work we reviewed over 50 papers reported on the literature regarding Arabic text classification. The majority of those papers were reporting technical content regarding various algorithms, datasets, and feature selection approaches while few other claimed to present a survey on the subject. We claim that our work presented herein provides a true comprehensive survey with a critique view. The following is a summary of our finding based on this survey.

- The majority of work on classifying the Arabic language is done in the past decade (2000 – 2010) with the exception of few incidences. They applied the mainstream classification techniques such as SVM, NB, k-NN, Decision Trees, and ANN. Few incidences proposed combined classifiers such as Maximum Entropy, master-slave and BSO-SVM.
- The datasets used in the work presented is mainly in-house built from news Arabic websites. Few works used datasets built by other researchers such as the Open Source Arabic Corpus. There are no popular Arabic corpses that is used by the majority of researchers.
- The size of the corpora used is relatively small as compared to the largest available English corpus (400 million words). The largest reported Arabic collected corpus in the literature contained about 310 million words collected from three websites only and covering five categories including a general one. This is only recently reported and is not standardized. Also, the number of categories are relatively small as compared to the nature and the richness of the Arabic language to other language.
- Some works used root-based stemming while others used light stemmers. It was shown that the use of light stemmer produces better performance measure than the root-based stemmers.
- The majority of work reported better performance measures of the SVM technique over other such as NB, k-NN, ANN, and Rocchio.
- Numerous feature selection techniques have been used including the TF, DF, TF-IDF, and Chi-Square. The majority of works used either TF or the TF-IDF due to their simplicity. The Chi-Square technique, however, proved to produce better results than other feature selection mechanisms.
- Some hybrid techniques have been used such as Master-Slave and BSO-SVM which showed better performance over the individual techniques.

VII. Future Work

We propose the importance of building a professional and diverse Arabic corpus to encourage further research on the Arabic language and to help generate benchmarks. We also recommend the adoption of one of the readily available light stemmer as a standard. Further research could be done on using the semi-supervised machine learning techniques to avoid the need for a large training dataset that is prepared with some human intervention that is usually prone to errors.

Acknowledgment

This research is funded by the Islamic University of Madinah Tamayoz research Fund number 1440/53. The authors would like to thank the Dean of Research, Dr. Saleh Al-Amri and the Vice-Dean, Dr. Tareq AlFraidy for their continuous help and support.

References

[1] R. Jindal, R. Malhotra, and A. Jain, "Techniques for text classification: Literature review and current trends." *Webology*, vol. 12, no. 2, 2015.

[2] X. Luo, W. Ohyama, T. Wakabayashi, and F. Kimura, "A study on automatic Chinese text classification," in 2011 International Conference on Document Analysis and Recognition. IEEE, 2011, pp. 920–924.

[3] A. Sboev, T. Litvinova, D. Gudovskikh, R. Rybka, and Moloshnikov, "Machine learning models of text categorization by author gender using topic-independent features," Procedia Computer Science, vol. 101, pp. 135–142, 2016.

[4] S. Boukil, M. Biniz, F. El Adnani, L. Cherrat, and A. E. El Moutaouakkil, "Arabic text classification using deep learning technics," International Journal of Grid and Distributed Computing, vol. 11, no. 9, pp. 103–114, 2018.

[5] Y. Caballero, R. Bello, D. Alvarez, and M. M. Garcia, "Two new feature selection algorithms with rough sets theory," in IFIP International Conference on Artificial Intelligence in Theory and Practice. Springer, 2006, pp. 209–216.

[6] A. M. El-Halees, "Arabic text classification using maximum entropy," IUG Journal of Natural Studies, vol. 15, no. 1, 2007.

[7] M. S. Khorsheed and A. O. Al-Thubaity, "Comparative evaluation of text classification techniques using a large diverse Arabic dataset," Language resources and evaluation, vol. 47, no. 2, pp. 513–538, 2013.

[8] M. A. Ahmed, R. A. Hasan, A. H. Ali, and M. A. Mohammed, "The classification of the modern Arabic poetry using machine learning." Telkomnika, vol. 17, no. 5, 2019.

[9] A. M. F. A. Sbou, "A survey of Arabic text classification models," International Journal of Electrical and Computer Engineering (IJECE), vol. 8, no. 6, pp. 4352–4355, 2018.

[10] A. M. A. Mesleh, "Support vector machines based arabic language text classification system: feature selection comparative study," in Advances in Computer and Information Sciences and Engineering. Springer, 2008, pp. 11–16.

[11] M. M. Syiam, Z. T. Fayed, and M. B. Habib, "An intelligent system for Arabic text categorization," International Journal of Intelligent Computing and Information Sciences, vol. 6, no. 1, pp. 1–19, 2006.

[12] H. Naji, W. Ashour, M. Al Hanjouri, and P. Gaza, "Text classification for arabic words using bpso/rep-tree," In: International Journal of Computational Linguistics Research, vol. 9, no. 1, pp. 1–9, 2018.

[13] F. Sebastiani, "Machine learning in automated text categorization," ACM computing surveys (CSUR), vol. 34, no. 1, pp. 1–47, 2002.

[14] Internet World Stats, "Internet world users by language-top 10 languages," 2015. [Online]. Available: http://www.internetworldstats.com/stats7.htm

[15] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M. Khorsheed, and A. Al-Rajeh, "Automatic arabic text classification," Proceedings of the 9th International Conference on the Statistical Analysis of Textual Data, Lyon-France, 2008.

[16] M. J. Bawaneh, M. S. Alkoffash, and A. Al Rabea, "Arabic text classification using k-nn and naive Bayes," Journal of Computer Science, vol. 4, no. 7, pp. 600–605, 2008.

[17] M. El Kourdi, A. Bensaid, and T.-e. Rachidi, "Automatic Arabic document categorization based on the naïve Bayes algorithm." The 20th International Conference on Computational Linguistics, Geneva, 2004.

[18] A. Moh'd Mesleh, "Feature sub-set selection metrics for Arabic text classification," Pattern Recognition Letters, vol. 32, no. 14, pp. 1922–1929, 2011.

[19] D. Said, N. M. Wanas, N. M. Darwish, and N. Hegazy, "A study of text preprocessing tools for Arabic text categorization," in The second international conference on Arabic language, 2009, pp. 230–236.

[20] S. C. Dharmadhikari, M. Ingle, and P. Kulkarni, "Empirical studies on machine learning based text classification algorithms," Advanced Computing, vol. 2, no. 6, p. 161, 2011.

[21] F. Sebastiani, "Text classification," 2005. [Online]. Available: http://nmis.isti.cnr.it/sebastiani/Publications/EDTA05.pdf.

[22] A. Khan, B. Baharudin, L. H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification," Journal of advances in information technology, vol. 1, no. 1, pp. 4–20, 2010.

[23] A. Karima, E. Zakaria, T. G. Yamina, A. Mohammed, R. Selvam, V. VENKATAKRISHNAN et al., "Arabic text categorization: a comparative study of different representation modes," Journal of Theoretical and Applied Information Technology, vol. 38, no. 1, pp. 1–5, 2012.

[24] R. Duwairi, "Arabic text categorization," Int. Arab J. Inf. Technol., vol. 4, pp. 125–132, 2007.

[25] S. Al-Saleem, "Associative classification to categorize Arabic data sets," International Journal of ACM Jordan, vol. 1, no. 3, pp. 118–127, 2010.

[26] S. Alsaleem et al., "Automated arabic text categorization using svm and nb." Int. Arab J. e-Technol., vol. 2, no. 2, pp. 124–128, 2011.

[27] H. K. Chantar and D. W. Corne, "Feature subset selection for Arabic document categorization using bpso-knn," in 2011 Third World Congress on Nature and Biologically Inspired Computing. IEEE, 2011, pp. 546–551.

[28] G. Kanaan, R. Al-Shalabi, S. Ghwanmeh, and H. Al-Ma'adeed, "A comparison of text-classification techniques applied to Arabic text," Journal of the American society for information science and technology, vol. 60, no. 9, pp. 1836–1844, 2009.

[29] M. Ahmed and R. Elhassan, "Arabic text classification review," International Journal of Computer Science and Software Engineering, vol. 4, no. 1, pp. 1–5, 2015.

[30] M. Ali and R. Elhassan, "Arabic text classification process," International Journal of Computer Science and Software Engineering, vol. 6, no. 11, pp. 258–265, 2017.

[31] M. K. Saad and W. M. Ashour, "Osac: Open source Arabic corpora," Osac: Open source arabic corpora, vol. 10, 2010.

[32] A. Al-Zoghby, A. S. Eldin, N. A. Ismail, and T. Hamza, "Mining Arabic text using soft-matching association rules," in 2007 International Conference on Computer Engineering & Systems. IEEE, 2007, pp. 421–426.

[33] B. Al-Shargabi, W. Al-Romimah, and F. Olayah, "A comparative study for arabic text classification algorithms based on stop words elimination," in Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications. ACM, 2011, p. 11.

[34] M. Aljlayl and O. Frieder, "On arabic search: improving the retrieval effectiveness via a light stemming approach," in Proceedings of the eleventh international conference on Informa- tion and knowledge management. ACM, 2002, pp. 340–347.

[35] M. Gheith and T. El-Sadany, "Arabic morphological analyzer on a personal computer," in Arabic Morphology Workshop, Linguistic Summer Institute, Stanford, CA, 1987.

[36] H. M. Al-Serhan, R. Al Shalabi, and G. Kannan, "New approach for extracting Arabic roots," 2003.

[37] S. Gupta, D. Kumar, and A. Sharma, "Data mining classification techniques applied for breast cancer diagnosis and prog- nosis," Indian Journal of Computer Science and Engineering (IJCSE), vol. 2, no. 2, pp. 188–195, 2011.

[38] R. Duwairi, M. N. Al-Refai, and N. Khasawneh, "Feature reduction techniques for Arabic text categorization," Journal of the American society for information science and technology, vol. 60, no. 11, pp. 2347–2352, 2009.

[39] R. M. Duwairi, "A distance-based classifier for Arabic text categorization." in DMIN, 2005, pp. 187–192.

[40] R. Mamoun and M. Ahmed, "Arabic text stemming: Comparative analysis," in 2016 Conference of Basic Sciences and Engineering Studies (SGCAC). IEEE, 2016, pp. 88–93.

[41] L. C. Molina, L. Belanche, and À. Nebot, "Feature selection algorithms: A survey and experimental evaluation," in 2002 IEEE International Conference on Data Mining, 2002. Proceedings. IEEE, 2002, pp. 306–313.

[42] Q. A. Al-Radaideh and G. Y. Al-Qudah, "Application of rough set-based feature selection for arabic sentiment analysis," Cognitive Computation, vol. 9, no. 4, pp. 436–445, 2017.

[43] Q. A. Al-Radaideh and L. M. Twaiq, "Rough set theory for Arabic sentiment classification," in 2014 International Conference on Future Internet of Things and Cloud. IEEE, 2014, pp. 559–564.

[44] A. H. Mohammad, "Arabic text classification: A review," Modern Applied Science, vol. 13, no. 5, 2019.

[45] A. M. Al Sbou, "A survey of arabic text classification models." International Journal of Electrical & Computer Engineering (2088-8708), vol. 8, 2018.

[46] H. Sawaf, J. Zaplo, and H. Ney, "Statistical classification methods for arabic news articles," Natural Language Processing in ACL2001, Toulouse, France, 2001.

[47] Z. S. Zubi, "Using some web content mining techniques for arabic text classification," in Proceedings of the 8th WSEAS international conference on Data networks, communications, computers, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society. Citeseer, 2009, pp. 73–84.

[48] A. Gelbukh, "Computational Linguistics and Intelligent Text Processing" 7th International Conference, CICLing 2006, Mexico City, Mexico, February 19-25, 2006, Proceedings. Springer, 2006, vol. 3878.

[49] A. M. El-Halees, "A comparative study on arabic text classification," A comparative study on Arabic text classification, vol. 30, no. 2, 2008.

[50] J. Ababneh, O. Almomani, W. Hadi, N. K. T. El-Omari, and A. Al-Ibrahim, "Vector space models to classify Arabic text," International Journal of Computer Trends and Technology (IJCTT), vol. 7, no. 4, pp. 219–223, 2014.

[51] M. K. Saad and W. M. Ashour, "Arabic text classification using decision trees," Arabic text classification using decision trees, vol. 2, 2010.

[52] T. F. Gharib, M. B. Habib, and Z. T. Fayed, "Arabic text classification using support vector machines." IJ Comput. Appl., vol. 16, no. 4, pp. 192–199, 2009.

[53] M. Abdeen, A. Elsehemy, T. Nazmy, and M. C. Yagoub, "Classifying the arabic web—a pilot study," in 2011 24th Canadian Conference on Electrical and Computer Engineering (CCECE). IEEE, 2011, pp. 865–868.

[54] F. Harrag, E. El-Qawasmeh, and P. Pichappan, "Improving arabic text categorization using decision trees," in 2009 First International Conference on Networked Digital Technologies. IEEE, 2009, pp. 110–115.

[55] Z. A. Abutiheen, A. H. Aliwy, and K. B. Aljanabi, "Arabic text classification using master-slaves technique," in Journal of Physics: Conference Series, ser. 1032, no. 1. IOP Publishing, 2018, pp. 012–052.

[56] M. M. Abushab and R. S. Baraka, "Large-scale Arabic text classification using mapreduce," Master, Islamic University of Gaza, 2017.

[57] F. OLAYAH and W. ALROMIMA, "Automatic machine learning techniques (amlt) for Arabic text classification based on term collocations." Journal of Theoretical & Applied Information Technology, vol. 96, no. 12, 2018.

[58] I. Hmeidi, M. Al-Ayyoub, N. A. Abdulla, A. A. Almodawar, R. Abooraig, and N. A. Mahyoub, "Automatic arabic text categorization: A comprehensive comparative study," Journal of Information Science, vol. 41, no. 1, pp. 114–124, 2015.

[59] R. Belkebir and A. Guessoum, "A hybrid bso-chi2-svm approach to arabic text categorization," in 2013 ACS Inter- national Conference on Computer Systems and Applications (AICCSA). IEEE, 2013, pp. 1–7.

[60] M. Yahyaoui, "Toward an arabic web page classifier," Master's Project, 2001.