Journal of Engineering and Applied Sciences 11 (3): 528-536, 2016
ISSN: 1816-949X

Arabic Text Classification: Review Study

[1]Musab Mustafa Hijazi, [2]Akram M. Zeki and [1]Amelia Ritahani Ismail
[1]Department of Computer Science,
[2]Department of Inforamtion Technology, Faculty of Communication and Tchnology,
International Islamic University, 53100 Kuala Lumpur, Selengor, Malaysia

Abstract: An enormous amount of valuable human knowledge is preserved in documents. The rapid growth in the number of machine-readable documents for public or private access requires the use of automatic text classification. Text classification can be defined as assigning or structuring documents into a defined set of classes known in advance. Arabic text classification methods have emerged as a natural result of the existence of a massive amount of varied textual information written in the Arabic language on the web. This study presents a review on the published researches of Arabic text classification using classical data representation, Bag of Words (BoW) and using conceptual data representation based on semantic resources such as Arabic WordNet and Wikipedia.

Key words: Advanced, Arabic, human, amount, rapid

INTRODUCTION

Documentation is the most effective way to clarify thoughts, ideas and expertise which means that documents are the fundamental depositories of knowledge (Khorsheed and Al-Thubaity, 2013). Owing the swift growth of the internet, the number of digital documents is increased which imposes a flexible and efficient methods to access, organize, extract a useful information (Al-Shalabi and Obeidat, 2008) and maximize the benefit of the knowledge that they have (Khorsheed and Al-Thubaity, 2013), such as text classification, text clustering (Al-Shalabi and Obeidat, 2008). Text Classification (TC) can be defined as assigning or structuring documents into a defined set of classes known in advance. Text categorization deal with sorting documents based on their content, while text classification is used to classify documents based on any kind of assignment to classes, by content, author, publisher, or by language (Elhassan and Ahmed, 2015). TC has been utilized for different applications such as documents organization, mail routing, automatic documents indexing, spam filtering, text filtering, news monitoring, word sense disambiguation and hierarchal catalog of web resources (Al-Shalabi and Obeidat, 2008).

The Arabic language is one of the Semitic languages, it is the mother language of almost 300 million people, it is widely spoken language in the world and it is used by about 1 billion Muslims in religious acts such as reciting Holy Quran, prayers. There is a growing interest of the Arabic language to present new techniques for processing this language and evaluates the effectiveness of the current techniques (applied to other languages) with Arabic language (Khorsheed and Al-Thubaity, 2013; Ghareb *et al*., 2014; Al-Tahrawi and Al-Khatib, 2015; Al-Saleem, 2010; Mamoun and Ahmed, 2014; Yousif *et al*., 2015; Hmeidi *et al*., 2015; Elberrichi and Abidi, 2012). Most of the researchers have used classical text representation which is called Bag of Word (BoW) in their studies which ignores the semantic relations between words while few researchers have utilized semantic resources such as Arabic WordNet and Wikipedia for Arabic text classification (Yousif *et al*., 2015a-c; Elberrichi and Abidi, 2012; Alahmadi *et al*., 2013).

ARABIC TEXT CLASSIFICATION USING CLASSICAL DATA REPRESENTATION (BoW)

There are a considerable amount of research studies that have been conducted for Arabic text classification.

Corresponding Author: Musab Mustafa Hijazi, Department of Computer Science, Faculty of Communication and Technology, International Islamic University, 53100 Kuala Lumpur, Selengor, Malaysia

Most of the researchers have been trying to find out the most effective and accurate system by comparing classification algorithms, studying the effect of document preprocessing, feature selection methods and term weighting methods (Al-Khorsheed and Thubaity, 2013).

Dataset: Unlike the case of English, there is no free benchmarking dataset for Arabic text classification so that, most of Arabic text classification researchers have collected their own datasets, mostly from online news sites which are in range from 175 texts divided into five classes (Fodil *et al*., 2014) to 33K divided into 33 classes (Sawaf *et al*., 2001). However, the performance of a classification algorithm may be affected by the quality of the data source (Elhassan and Ahmed, 2015; Abuaiadah *et al*., 2014; Said *et al*., 2009).

Preprocessing: Preprocessing is an attempt to improve text classification by removing of worthless information. Most researchers of Arabic text classification took into their account the importance of preprocessing either fully or partially but some research did not such as (Sawaf *et al*., 2001; Ta'amneh *et al*., 2014; Thabtah *et al*., 2008). El-Halees (2007) studied the effect of preprocessing and Part of speech tagging and found that they increased significantly the classification accuracy while Abuaiadah *et al*. (2014) found that preprocessing slightly improves the performance of classification. Elhassan and Ahmed (2015) investigated the effectiveness of the data preprocessing on a full word in the accuracy of training model and classifier. They used two approaches for data preprocessing: the observation of data set content and stop words estimation technique. Another study by (Al-Molegi *et al*., 2015) studied the effect of text preprocessing when N-gram was used and found that there was no significant improvement on the overall accuracy. Hmeidi *et al*. (2015) also investigated the effect of preprocessing on Arabic text classification and concluded that the accuracy varies from one algorithm to another depending on the nature and size of data.

Data division: There is no ideal ratio of training data to testing data so that different ratios have been used for Arabic text classification research ranging from 25% for training and 75% for testing up to 80% for training and 20% for testing (Al-Tahrawi, 2015; Sawaf *et al*., 2001; Kanaan *et al*., 2009). Kanaan *et al*. (2009) found that as the number of documents available in the training set increases and the number of categories decrease, the precision and recall approach a perfect value of 1. In (Al-Khorsheed and Thubaity, 2013; Al-Molegi *et al*., 2015; Harrag *et al*., 2009), they studied the effect of training and testing data set size on Arabic text classification and concluded that the best classification accuracy were achieved when the training data size was larger than testing data size while Abuaiadah *et al*. (2014) found in their study that there was only a marginal improvement on the performance of classification when the size of the training set exceeds 50 documents.

Feature extraction: In feature extraction for Arabic text classification, Most of the researchers concentrated on the simplest of lexical features, the word (Al-Khorsheed and Thubaity, 2013) which was addressed as a feature on three levels: using words in their orthographic form (Al-Khorsheed and Thubaity, 2013; Al-Saleem, 2010; Thabtah *et al*., 2008; Elhassan and Ahmed, 2015; Ababneh *et al*., 2014; Al-Diabat, 2012; Al-Hindi and Al-Thwaib, 2013; Al-Salem and Aziz, 2011; Al-Shargabi *et al*., 2011; Al-Thwaib, 2014; Al-Thwaib and Romimah, 2014; Halees, 2008), word stems in which the suffix and prefix were removed from the orthographic form of the word (Ghareb *et al*., 2014, 2015; Kanaan *et al*., 2009; Harrag *et al*., 2009; Aly *et al*., 2013; Bawaneh *et al*., 2008; Duwairi *et al*., 2009; Saad and Ashour, 2010) and the word root which is the primary lexical unit of a word (Al-Tahrawi, 2015; Odeh *et al*., 2015). Some of the researchers used character n-grams which usually convey no meaning. In this method, a certain number of consecutive characters are extracted and considered as features (Al-Shalabi and Obeidat, 2008; Sawaf *et al*., 2001; Al-Thubaity *et al*., 2015).

Several studies were investigated the effect of stemming for Arabic text classification (Yousif *et al*., 2015; Hmeidi *et al*., 2015; Abuaiadah *et al*., 2014; Said *et al*., 2009; Duwairi *et al*., 2009; Harrag *et al*., 2011; Hmeidi *et al*., 2014; Belkebir and Guessoum, 2013; Haralambous *et al*., 2014; Syiam *et al*., 2006; Al-Kabi *et al*., 2013 Azara *et al*., 2012; Chantar, 2013; Kanan and Fox, 2015; Alhutaish and Omar, 2015; Al-Salem and Aziz, 2011; Kechaou and Kanoun, 2014), all of them agreed on that in general, stemming reduces vector size which improves the time and accuracy of text classification except Al-Kabi *et al*. (2013) found in their study that stemming had a negative effect on the accuracy of Arabic text classification.

Saad (2010) found that light stemming and term pruning with a threshold of five words had the highest reduction of the number of features. Another study by Chantar (2013) found that light stemming had led to a significant reduction in the number of distinct features and slightly improved the classification accuracy.

The effect of using N-gram as a feature investigated by Al-Shalabi and Obeidat (2008), Al-Molegi *et al*. (2015), Al-Thubaity *et al*. (2015), Al-Salem and Aziz (2011) and

Sharef *et al*. (2014) with contradictory results, Al-Shalabi and Obeidat (2008) found that using N-gram enhanced the accuracy of text classification while Al-Thubaity *et al*. (2015), Al-Salem and Aziz (2011) and Sharef *et al*. (2014) they found that the use of single word as a feature was more effective than N-gram for Arabic text classification. In addition, Al-Thubaity, *et al*. (2015) found that the accuracy decreased when the number of N-grams increased. However, they used different dataset and different classification algorithms. Al-Molegi *et al*. (2015) found that for Arabic text categorization, it is best to use 3-letters N-gram then 5-letters N-gram and finally 4-letters N-gram but they did not compare with the single word as a feature. Another study by Alhutaish and Omar (2015) compared between single word and trigram as a feature. They found that single word outperformed trigram in all their experiments.

Feature selection: Feature selection aims to improve the classification accuracy and computational efficiency of classification techniques by removing irrelevant and redundant terms (features) from the corpus. It is also used to select features that contain sufficient information. FS has two wider approaches: wrapper and filter. In the wrapper approach, a subset of the features is selected based on the accuracy of the classifiers while, in the filter approach, a subset of features is selected or filtered using feature scoring metric (Chantar, 2013).

The widely used filter ranking methods for Arabic text classification are Chi-squared (CHI) (Al-Khorsheed and Thubaity, 2013; Al-Tahrawi, 2015; Sawaf *et al*., 2001; Kanaan *et al*., 2009; Al-Diabat, 2012; Syiam *et al*., 2006; Zahran and Kanaan, 2009), Term Frequency (TF) (Al-Khorsheed and Thubaity, 2013; Harrag *et al*., 2009; Al-Thwaib, 2014; Al-Thwaib and Romimah, 2014), Document Frequency (DF) (Al-Khorsheed and Thubaity, 2013; Al-Shalabi and Obeidat, 2008; Ghareb *et al*., 2014; Al-Thabtah *et al*., 2008; Zahran and Kanaan, 2009) and information gain (Al-Khorsheed and Thubaity, 2013; Sawaf *et al*., 2001; Halees, 2008; Al-Hmeidi *et al*., 2014; Syiam *et al*., 2006; Zahran and Kanaan, 2009). The word stems or roots were also used as feature selections where words with the same stem or root are considered as one feature and features with higher frequency are used (Al-Khorsheed and Thubaity, 2013; Kanaan *et al*., 2009; Aly *et al*., 2013; Bawaneh *et al*., 2008; Duwairi *et al*., 2009). Table 1 presents set of comparative studies that studied the effect of filter feature selection methods on Arabic text classification. In the effect of using Singular Value Decomposition SVD as FS Method with ANN was investigated (Harrag *et al*., 2010; Harrag and El-Qawasmah, 2009). They found that SVD enhanced the performance of ANN. Hawashin *et al*. (2013) proposed an efficient chi-squared based feature selection method. There are few researches have been studied the effect of using wrapper feature selection methods on Arabic text classification, Table 2 presents these researches.

Table 1: Comparative studies of the effect of feature selection methods on Arabic text classification

References	FS methods used	Classifier	Findings
Harrag *et al*. (2010)	Stemming, light stem, DF, TFIDF, LSI	Back-propagation neural network	DF, TFIDF and LSI techniques are favorable for BPNN
Saad *et al*. (2011)	CHI, IG, MI, TFIDF, DF	Dewey based classification	CHI and IG
Moh'd Mesleh (2011)	17 FS methods	SVM	Chi and Fallout FSS metrics work best for Arabic TC tasks
Khorsheed and Al-Thubaity (2013)	TF, DF, CHI, IG, RS, GSS, MI, NGL, DIA, RS, OddsR	SVM	GSS/TF as the term selection base
Al-Thubaity *et al*. (2013)	CHI, GSS, IG, NGL, RS and their combinations	NB	CHI followed by IG as single FS. Combining FS methods showed insignificant improvement
Haralambous *et al*. (2014)	TFIDF, dependency syntax	SVM, CAR	Dependency grammar feature selection
Adel *et al*. (2014)	CHI, correlation, GSS, IG, relief F and their combination	SVM, NB	IG as single FS combinations of FS methods performed better
Alhutaish and Omar (2015)	CHI, GSS, OR, MI	KNN (inew, cosine, jaccard, dice)	MI for trigram CHI for BoW
Ayadi *et al*. (2015)	LDA, LSI	SVM	LDA

Table 2: Researches that studied the effect of using wrapper feature selection methods on Arabic text classification

References	FS methods used	Classifier	Findings
Mesleh (2008)	Ant Colony Optimization based	SVM	ACO-based outperformed (CHI, GSS, IG, NGL, MI, OR)
Zahran and Kanaan (2009)	FS based on Particle Swarm Optimization (PSO)	Radial basis function networks classifier	PSO-based outperformed CHI, DF, TFIDF
Chantar (2013)	FS based on binary Particle Swarm Optimization (PSO) combined with KNN, and SVM	SVM, C5.0, NB	Hybrid approach was effective as selection method,BPSO/KNN may be favored if BPSO-SVM outperformed BPSO-KNN categories tend to be quite distinct
Belkebir and Guessoum (2013)	FS that combined with Bee Swarm Optimization algorithm (BSO) with CHI	SVM, ANN	The proposed hybrid approach has proven its ability to improve accuracy of SVM

Table 3: Comparative studies of the effect of term weighting on Arabic text classification

References	FS methods used	Classifier	Findings
Saad (2010)	Bool, TF, WC, TFIDF, normalization, TFIDF-norm-minfreq	C4.5, KNN, SVM,NB, NBM, CNB , DMNB	TFIDF-norm-Minfreq
Azara *et al*. (2012)	Bool, TF, TFIDF, IDF	Learning vector quantization	TFIDF
Khorsheed and Al-Thubaity (2013)	Bool, TF, TFIDF, RF, TFC, LTC, ENTROPY	SVM	LTC
Al-Thubaity *et al*. (2013)	Bool, TFIDF, LTC	NB	LTC
Zaghoul and Al-Dhaheri (2013)	TFIDF, TFIDF combined with Principal Component Analysis (PCA)	ANN	TFIDF combined with PCA (DF_CF threshold)
Al-Thubaity *et al*. (2015)	Bool, TFIDF, LTC	SVM, KNN, NB	Bool for NB LTC for SVM, KNN

Table 4: Comparative researches between classification algorithms for Arabic language

References	Classification algorithms	Findings (suitable one)
Saad (2010)	C4.5, KNN, SVM, NB, NBM, CNB, DMNB	SVM
Harrag *et al*. (2010)	Multilayer Perceptron (MLP) and the Radial Basis Function (RBF) classifiers	MLP
Al-Saleem (2010)	Associative Classification algorithm (AC), SVM, NB	AC
Alwedyan *et al*. (2011)	Multi-class Classification Based on Association Rule (MCAR), SVM, NB	MCAR
Moh'd Mesleh (2011)	SVM, KNN, NB, Rocchio	SVM
Al-Shargabi *et al*. (2011) and Al-Kabi *et al*. (2013)	SVM, NB, C4.5	SVM
Alsaleem (2011)	NB, SVM	SVM
Harrag *et al*. (2011)	SVM, ANN	ANN
Al-Salem and Aziz (2011)	NB, MBNB, MNB	MBNB
Al-Radaideh *et al*. (2011)	Classification Based on Association Rule ordered decision list, majority voting, and weighted rules	Majority voting (multiple rule perdition) was the best one and weighted rules were the worst
Al-Diabat (2012) and Thabtah *et al*. (2011, 2012)	Four rules based classification techniques (C4.5, PART, One Rule and RIPPER)	One Rule was least suitable one, whereas RIPPER, C4.5 and PART had the similar performance
Ghareb *et al*. (2012)	Classification Based on Association Rule: ordered decision list, majority voting	Majority voting (multiple rule perdition)
Wahbeh and Al-Kabi (2012)	SVM, NB, C4.5	NB
Khorsheed and Al-Thubaity (2012)	MLPs, SVM, KNN, NB, C4.5	SVM
Al-Thwaib and Al-Romimah (2014)	SVM, KNN	SVM
Hmeidi *et al*. (2014)	NB, SVM, KNN, Decision Tree and Decision Table	SVM
Haralambous *et al*. (2014)	MCAR, SVM	MCAR for small feature set and SVM for large feature sets
Ababneh *et al*. (2014)	KNN (Cosine, Dice, and Jaccard coefficient)	Cosine
Kechaou and Kanoun (2014)	Hidden Markov Model (HMM), NB, KNN	HMM
Al-Thubaity *et al*. (2015)	SVM, NB, KNN	SVM but NB gave better accuracy for N-grams
Elhassan and Ahmed (2015)	SVM, C5.0, NB, KNN	SVM
Hmeidi *et al*. (2015)	KNN, NB, NBM, Bayes net, Random forest, Kstar, DT	Accuracy vary from one algorithm to another depending on the nature and size of data
Kanan and Fox (2015)	SVM, NB, Random forest	SVM
Alhutaish and Omar (2015)	KNN (Inew, Cosine, Dice, and Jaccard coefficient)	Inew

Term weighting: Several methods have been used to assign the proper weight to the feature. The most-used weighting methods are, Term Frequency inverse Document Frequency (TFiDF) (Al-Khorsheed and Thubaity, 2013; Al-Shalabi and Obeidat, 2008; Sawaf *et al*., 2001; Al-Thabtah *et al*., 2008; Kanaan *et al*., 2009; Al-Shargabi *et al*., 2011; Bawaneh *et al*., 2008; Syiam *et al*., 2006; Zahran and Kanaan, 2009) and Term Frequency (TF) (Al-Khorsheed and Thubaity, 2013; Sawaf *et al*., 2001; Kanaan *et al*., 2009; Al-Hindi and Al-Thwaib, 2013; Al-Thwaib, 2014; Al-Thwaib and Romimah, 2014; Syiam *et al*., 2006). Table 3 presents comparative studies that investigated the effect of term weighting methods for Arabic text classification.

Classification algorithm: The state of the art text classification algorithms have been used in Arabic text classification are, Naive Bayes (NB) (Table 4)

(Al-Saleem, 2010), K-Nearest Neighbor (KNN) (Al-Shalabi and Obeidat, 2008; Ababneh *et al*., 2014; Bawaneh *et al*., 2008; Syiam *et al*., 2006) and Support Vector Machine (SVM) (Al-Khorsheed and Thubaity, 2013; Mamoun and Ahmed, 2014; Alsaleem, 2011; Halees, 2008; Hmeidi *et al*., 2014). There are a lot of comparative studies between classification algorithms to find out the most accurate one for the Arabic language Table 4 presents them.

Atlam *et al*. (2011) presented a new methodology for building a comprehensive Arabic dictionary using linguistic methods to extract relevant compound and single FA terms from domain-specific corpora using Arabic POS. Hattab and Hussein (2012) studied the effect of applying misspelling detection and correction algorithm with NB classifier. Classification of unstructured documents was addressed by Aly *et al*. (2013). Dawoud (2013) studied the effect of combining several classifiers using different combination methods. Text summarization was used to reduce the dimensionality of document representation by Al-Hindi and Al-Thwaib (2013) and Al-Thwaib (2014). A Compression-based TC (CTC) was investigated by Ta'amneh *et al*. (2014). They concluded the applicability of using CTC for Arabic TC. Ghareb *et al*. (2014) proposed an Arabic text classification model based on Associative Classification approach (AC) which integrated noun extraction, feature selection methods and Associative Rule Mining (ARM).

Abu-Errub (2014) proposed a dual-stages Arabic text classification algorithm using TFIDF measurement for categorization stage and chi-square measurement for classification stage. Fodil *et al*. (2014) proposed two statistical approaches of text classification, Semi-Automatic Categorization Method (SACM) and Automatic Categorization Method (ACM), ended with the superiority of SACM. Sharef *et al*. (2014) applied Frequency Ratio Accumulation Method FRAM for Arabic text classification. In (Ghareb *et al*., 2015), a hybrid classification approach was proposed for Arabic text mining which combined the advantages of a statistical classifier and rule-based classifier NB and AC, respectively. Nehar *et al*. (2015) presented an approach for Arabic text classification and stemming. It is based on using transducer for stemming, rational kernels for calculating the distance between documents and SVM for classification. Al-Tahrawi (2015) found that LR is a competitive Arabic TC algorithm. Furthermore, Al-Tahrawi and Al-Khatib (2015) used Polynomial neural Networks (PNs) as an Arabic TC algorithm.

Multi-label Arabic classification was investigated by Alwedyan *et al*. (2011), Ahmed *et al*. (2015) and Ezzat *et al*. (2012). Ezzat *et al*. (2012) proposed training-less ontology-based multi-labeling topic categorization system which classifies large volumes of data when no training data and no classification scheme are available. Ahmed *et al*. (2015) studied the multi-label text classification problem for Arabic text. They considered different problem transformation methods coupled with different base classifiers and studied the effect of scaling up the dataset. Their study is ended with the superiority of SVM with Label Combination method (LC).

ARABIC TEXT CLASSIFICATION BASED ON SEMANTIC RESOURCES

There is a trend for studying the effect of utilizing semantic information and relationships between the words with Arabic text classification based on Arabic Word Net (AWN) and Wikipedia (Yousif *et al*., 2015a-c; Elberrichi and Abidi, 2012; Alahmadi *et al*., 2014).

Taxonomy was utilized by Zaki *et al*. (2010, 2014) such that in (Zaki *et al*., 2010), they used it with fuzzy entropy with radial basis function. Zaki *et al*. (2014), they used a hybrid method of N-grams-TF-IDF with radial basis indexing for classification. Saad *et al*. (2011) semantic approach was presented using synonym merge to preserve features semantic and prevent important terms from being excluded. Zrigui *et al*. (2012) used topic modeling approach such as Latent Dirichlet Allocation (LDA) to represent documents as random mixtures over latent topics, where each topic is characterized by a distribution over words.

Yousif *et al*. (2015a-c) and Elberrichi and Abidi (2012) presents, a conceptual representation for Arabic text representation using Arabic WordNet was proposed. They found that semantic dimension is one of most promising ways for Arabic text classification. Alahmadi *et al*. (2014) used Wikipedia as a knowledge base to solve some of the limitations of the classic BoW representation in Arabic TC. Yousif *et al*. (2015) proposed two novel features sets that use lexical, semantic and lexico-semantic relations of Arabic WordNet (AWN) ontology with superiority of LoPW.

CONCLUSION

There are a considerable amount of research studies that have been conducted for Arabic text classification. Most of the researchers have been trying to find out the most effective and accurate system by comparing classification algorithms, studying the effect of document preprocessing, feature selection methods and term weighting methods. However, based on the literature that has been done, it can be concluded:

- The accuracy for the different classification algorithms ranged from 67-98%. The hypothesis is that the dominant factors in accuracy are the characteristics of the different datasets and not the algorithms and in particular, the source of the data and the methodology of selecting the documents. However, the use of a standard dataset would eliminate these factors and enable researchers to make meaningful comparisons between the performances of the different algorithms (Hmeidi *et al*., 2015; Abuaiadah *et al*., 2014; Said *et al*., 2009; Elhassan and Ahmed, 2015)
- Preprocessing has a positive effect on Arabic text classification (Al-Khorsheed and Thubaity, 2013; Elhassan and Ahmed, 2015; Yousif *et al*., 2015; Elhassan and Ahmed, 2015)
- Stemming might have different effects on different TC (Abuaiadah *et al*., 2014; Harrag *et al*., 2011) most of the researchers agreed that light stemming is the most suitable one for the Arabic language
- Although, many feature selection methods exist in text categorization, it is hard to state one is generally superior to others since, the success of the methods depends on various variables. It is more likely that combining different feature selection methods obtains more effective performance in text categorization (Adel *et al*., 2014)
- It is difficult to compare the effectiveness of Arabic text classification approaches for various reasons. The first reason is that each author used different corpora. The second reason is that even those who have used the same corpus did not use the same documents for learning and testing their classifiers. The last reason is that each author used different evaluation measures: precision, recall and F-measure
- Semantic dimension is one of most promising ways for Arabic text classification (Yousif *et al*., 2015a-c; Elberrichi and Abidi, 2012; Alahmadi *et al*., 2014)

REFERENCES

Ababneh, J., O. Almomani, W. Hadi, N.K.T. El-Omari and A. Al-Ibrahim, 2014. Vector space models to classify Arabic text. Int. J. Comput. Trends Technol., 7: 219-223.

Abu-Errub, A., 2014. Arabic text classification algorithm using TFIDF and Chi square measurements. Int. J. Comput. Applic., 93: 40-45.

Abuaiadah, D., J. El Sana and W. Abusalah, 2014. On the impact of dataset characteristics on Arabic document classification. Int. J. Comput. Applic., 101: 31-38.

Adel, A., N. Omar and A. Al-Shabi, 2014. A comparative study of combined feature selection methods for Arabic text classification. J. Comput. Sci., 10: 2232-2239.

Ahmed, N.A., M.A. Shehab, M. Al-Ayyoub and I. Hmeidi, 2015. Scalable multi-label Arabic text classification. Proceedings of the 6th International Conference on Information and Communication Systems, April 7-9, 2015, Amman, Jordan, pp: 212-217.

Al-Diabat, M., 2012. Arabic text categorization using classification rule mining. Applied Math. Sci., 6: 4033-4046.

Al-Hindi, K. and E. Al-Thwaib, 2013. A comparative study of machine learning techniques in classifying full-text Arabic documents versus summarized documents. World Comput. Sci. Inform. Technol. J., 2: 126-129.

Al-Kabi, M., E. Al-Shawakfa and I. Alsmadi, 2013. The Effect of Stemming on Arabic Text Classi?cation: An Empirical Study. In: Information Retrieval Methods for Multidisciplinary Applications, Lu, Z. (Ed.). Idea Group Inc., Pennsylvania, ISBN: 9781466638990, pp: 207-225.

Al-Molegi, A., I. Alsmadi, H. Najadat and H. Albashiri, 2015. Automatic learning of Arabic text categorization. Int. J. Digital Contents Applic., 2: 1-16.

Al-Radaideh, Q.A., E.M. Al-Shawakfa, A.S. Ghareb and H. Abu-Salem, 2011. An approach for Arabic text categorization using association rule mining. Int. J. Comput. Process. Languages, 23: 81-106.

Al-Saleem, S., 2010. Associative classification to categorize Arabic data sets. Int. J. ACM Jordan, 1: 118-127.

Al-Salem, B. and M.J.A. Aziz, 2011. Statistical bayesian learning for automatic Arabic text categorization. J. Comput. Sci., 7: 39-45.

Al-Shalabi, R. and R. Obeidat, 2008. Improving KNN Arabic text classification with n-grams based document indexing. Proceedings of the 6th International Conference on Informatics and Systems, March 27-29, 2008, Cairo, Egypt, pp: 108-112.

Al-Shargabi, B., W. Al-Romimah and F. Olayah, 2011. A comparative study for Arabic text classification algorithms based on stop words elimination. Proceedings of the International Conference on Intelligent Semantic Web-Services and Applications, April 18-20, 2011, Amman, Jordan 10.1145/1980822. 1980833-.

Al-Tahrawi, M.M. and S.N. Al-Khatib, 2015. Arabic text classification using polynomial networks. J. King Saud Univ.-Comput. Inform. Sci., 27: 437-449.

Al-Tahrawi, M.M., 2015. Arabic text categorization using logistic regression. Int. J. Intell. Syst. Applic., 6: 71-78.

Al-Thubaity, A., M. Alhoshan and I. Hazzaa, 2015. Using Word N-Grams as Features in Arabic Text Classification. In: Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Lee, R. (Ed.). Springer International Publishing, Switzerland, ISBN: 978-3-319-10388-4, pp: 35-43.

Al-Thubaity, A., N. Abanumay, S. Al-Jerayyed, A. Alrukban and Z. Mannaa, 2013. The effect of combining different feature selection methods on Arabic text classification. Proceedings of the 14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, July 1-3, 2013, Honolulu, HI., pp: 211-216.

Al-Thwaib, E. and W. Al-Romimah, 2014. Support vector machine versus k-nearest neighbor for Arabic text classification. Int. J. Sci., 3: 1-5.

Al-Thwaib, E., 2014. Text summarization as feature selection for Arabic text classification. World Comput. Sci. Inform. Technol. J., 4: 101-104.

Alahmadi, A., A. Joorabchi and A.E. Mahdi, 2013. Combining bag-of-words and bag-of-concepts representations for Arabic text classification. Proceedings of the 25th IET Irish Signals and Systems Conference on China-Ireland International Conference on Information and Communications Technologies, June 26-27, 2013, Limerick, Ireland, pp: 343-348.

Alhutaish, R. and N. Omar, 2015. Arabic text classification using K-nearest neighbour algorithm. Int. Arab J. Inform. Technol., 12: 190-195.

Alsaleem, S., 2011. Automated Arabic text categorization using SVM and NB. Int. Arab J. e-Technol., 2: 124-128.

Alwedyan, J., W.M. Hadi, M. Salam and H.Y. Mansour, 2011. Categorize Arabic data sets using multi-class classification based on association rule approach. Proceedings of the International Conference on Intelligent Semantic Web-Services and Applications, April 18-20, 2011, Amman, Jordan. -.

Aly, W.M., W.H. Sharaby, M.W. Youssef and H.A. Kelleny, 2013. Improved classification of Arabic unstructured documents based on automated domain dictionary construction. Proceedings of the 23rd International Conference on Computer Theory and Applications, October 29-31, 2013, Alexandria, Egypt, pp: 68-73.

Atlam, E.S., K. Morita, M. Fuketa and J.I. Aoe, 2011. A new approach for Arabic text classification using Arabic field-association terms. J. Am. Soc. Inform. Sci. Technol., 62: 2266-2276.

Ayadi, R., M. Maraoui and M. Zrigui, 2015. LDA and LSI as a Dimensionality Reduction Method in Arabic Document Classification. In: Information and Software Technologies, Dregvaite, G. and R. Damasevicius (Eds.). Springer International Publishing, Switzerland, ISBN: 978-3-319-24769-4, pp: 491-502.

Azara, M., T. Fatayer and A. El-Halees, 2012. Arabie text classification using learning vector quantization. Proceedings of the 8th International Conference on Informatics and Systems, May 14-16, 2012, Giza, Egypt, pp: NLP-39-NLP-43.

Bawaneh, M.J., M.S. Alkoffash and A.I. Al Rabea, 2008. Arabic text classification using K-NN and naive bayes. J. Comput. Sci., 4: 600-605.

Belkebir, R. and A. Guessoum, 2013. A hybrid BSO-Chi2-SVM approach to Arabic text categorization. Proceedings of the ACS International Conference on Computer Systems and Applications, May 27-30, 2013, Ifrane, Morocco, pp: 1-7.

Chantar, H.K.H., 2013. New techniques for Arabic document classification. Ph.D. Thesis, Heriot-Watt University, Edinburgh, Scotland.

Dawoud, H.M., 2013. Combining different approaches to improve Arabic text documents classification. M.Sc. Thesis, Islamic University, Gaza, Egypt.

Duwairi, R., M.N. Al-Refai and N. Khasawneh, 2009. Feature reduction techniques for Arabic text categorization. J. Am. Soc. Inform. Sci. Technol., 60: 2347-2352.

El-Halees, A., 2008. A comparative study on Arabic text classification. Egypt. Comput. Sci. J., Vol. 30.

El-Halees, A.M., 2007. Arabic text classification using maximum entropy. Islamic Univ. J. (Ser. Nat. Stud. Eng.), 15: 157-167.

Elberrichi, Z. and K. Abidi, 2012. Arabic text categorization: A comparative study of different representation modes. Int. Arab J. Inform. Technol., 9: 465-470.

Elhassan, R. and M. Ahmed, 2015. Arabic text classification on full word. Int. J. Comput. Sci. Software Eng., 4: 114-120.

Elhassan, R. and M. Ahmed, 2015. Arabic text classification review. Int. J. Comput. Sci. Software Eng., 4: 1-5.

Ezzat, H., S. Ezzat, S. El-Beltagy and M. Ghanem, 2012. Topicanalyzer: A system for unsupervised multi-label Arabic topic categorization. Proceedings of the International Conference on Innovations in Information Technology, March 18-20, 2012, Abu Dhabi, pp: 220-225.

Fodil, L., H. Sayoud and S. Ouamour, 2014. Theme classification of Arabic text: A statistical approach. Proceedings of the 11th International Conference on Terminology and Knowledge Engineering, June 19-21, 2014, Berlin, pp: 10-.

Ghareb, A.S., A.R. Hamdan and A.A. Bakar, 2012. Text associative classification approach for mining Arabic data set. Proceedings of the 4th Conference on Data Mining and Optimization, September 2-4, 2012, Langkawi, Malaysia, pp: 114-120.

Ghareb, A.S., A.R. Hamdan and A.A. Bakar, 2014. Integrating noun-based feature ranking and selection methods with Arabic text associative classification approach. Arabian J. Sci. Eng., 39: 7807-7822.

Ghareb, A.S., A.R. Hamdan, A.A. Bakar and M.R. Yaakub, 2015. Hybrid statistical rule-based classifier for Arabic text mining. J. Theor. Applied Inform. Technol., 71: 194-204.

Haralambous, Y., Y. Elidrissi and P. Lenca, 2014. Arabic language text classification using dependency syntax-based feature selection. arXiv preprint arXiv:1410.4863,2014.http://arxiv.org/abs/1410.4863.

Harrag, F. and E. El-Qawasmah, 2009. Neural network for Arabic text classification. Proceedings of the 2nd International Conference on the Applications of Digital Information and Web Technologies, August 4-6, 2009, London, pp: 778-783.

Harrag, F., A.M.S. Al-Salman and M. BenMohammed, 2010a. A comparative study of neural networks architectures on Arabic text categorization using feature extraction. Proceedings of the International Conference on Machine and Web Intelligence, October 3-5, 2010, Algiers, Algeria, pp: 102-107.

Harrag, F., E. El-Qawasmah and A.M.S. Al-Salman, 2010b. Comparing dimension reduction techniques for Arabic text classification using BPNN algorithm. Proceedings of the 1st International Conference on Integrated Intelligent Computing, August 5-7, 2010, Bangalore, India, pp: 6-11.

Harrag, F., E. El-Qawasmah and A.M.S. Al-Salman, 2011. Stemming as a feature reduction technique for Arabic text categorization. Proceedings of the 10th International Symposium on Programming and Systems, April 25-27, 2011, Algiers, Algeria, pp: 128-133.

Harrag, F., E. El-Qawasmeh and P. Pichappan, 2009. Improving Arabic text categorization using decision trees. Proceedings of the 1st International Conference on Networked Digital Technologies, July 28-31, 2009, Ostrava, Czech Republic, pp: 110-115.

Hattab, A. and A.K. Hussein, 2012. Arabic content classification system using statistical Bayes classifier with words detection and correction. World Comput. Sci. Inform. Technol. J., 2: 193-196.

Hawashin, B., A. Mansour and S. Aljawarneh, 2013. An efficient feature selection method for Arabic text classification. Int. J. Comput. Applic., 83: 1-6.

Hmeidi, I., M. Al-Ayyoub, N.A. Abdulla, A.A. Almodawar, R. Abooraig and N.A. Mahyoub, 2015. Automatic Arabic text categorization: A comprehensive comparative study. J. Inform. Sci., 41: 114-124.

Hmeidi, I., M. Al-Shalabi and M. Al-Ayyoub, 2015. A comparative study of automatic text categorization methods using Arabic text. Proceedings of the International Technology Management Conference, May 26-28, 2015, Mevlana University, Konya, Turkey, pp: 73-82.

Kanaan, G., R. Al-Shalabi, S. Ghwanmeh and H. Al-Ma'adeed, 2009. A comparison of text-classification techniques applied to Arabic text. J. Am. Soc. Inform. Sci. Technol., 60: 1836-1844.

Kanan, T. and E.A. Fox, 2016. Automated Arabic text classification with P-Stemmer, machine learning and a tailored news article taxonomy. J. Assoc. Inform. Sci. Technol. 10.1002/asi.23609

Kechaou, Z. and S. Kanoun, 2014. A new-Arabic-text classification system using a hidden Markov model. Int. J. Knowl. Intell. Eng. Syst., 18: 201-210.

Khorsheed, M.S. and A.O. Al-Thubaity, 2013. Comparative evaluation of text classification techniques using a large diverse Arabic dataset. Language Resour. Eval., 47: 513-538.

Mamoun, R. and M.A. Ahmed, 2014. A comparative study on different types of approaches to the Arabic text classification. Proceedings of the 1st International Conference of Recent Trends in Information and Communication Technologies, September 12-14, 2014, Johor, Malaysia -.

Mesleh, A.M.A., 2008. Support vector machine text classifier for Arabic articles: Ant colony optimization-based feature subset selection. Ph.D. Thesis, Arab Academy for Banking and Financial Scienses, Amman, Jordan.

Moh'd Mesleh, A., 2011. Feature sub-set selection metrics for Arabic text classification. Pattern Recognit. Lett., 32: 1922-1929.

Nehar, A., D. Ziadi and H. Cherroun, 2015. Rational kernels for Arabic stemming and text classification. arXiv preprint arXiv:1502.07504. https://arxiv.org/pdf/1502.07504.pdf.

Odeh, A., A. Abu-Errub, Q. Shambour and N. Turab, 2014. Arabic text categorization algorithm using vector evaluation method. Int. J. Comput. Sci. Inform. Technol., 6: 83-92.

Saad, E., M. Awadalla and A. Alajmi, 2011. Dewy index based Arabic document classification with synonyms merge feature reduction. Int. J. Comput. Sci Issues, 1: 46-54.

Saad, M.K. and W. Ashour, 2010. Arabic text classification using decision trees. Proceedings of the 12th International Workshop on Computer Science and Information Technologies, September 13-19, 2010, Russia, Moscow-St.Petersburg, pp: 75-79.

Saad, M.K., 2010. The impact of text preprocessing and term weighting on Arabic text classification. M.Sc. Thesis, The Islamic University, Gaza, Egypt.

Said, D., N.M. Wanas, N.M. Darwish and N. Hegazy, 2009. A study of text preprocessing tools for Arabic text categorization. Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, April 22-23, 2009, Cairo, Egypt, pp: 230-236.

Sawaf, H., J. Zaplo and H. Ney, 2001. Statistical classification methods for Arabic news articles. Proceedings of ACL/EACL 2001 Workshop on Arabic Language Processing: Status and Prospects, July 2001, Toulouse, Germany -.

Sharef, B.T., N. Omar and Z.T. Sharef, 2014. An automated Arabic text categorization based on the frequency ratio accumulation. Int. Arab J. Inform. Technol., 11: 213-221.

Syiam, M.M., Z.T. Fayed and M.B. Habib, 2006. An intelligent system for Arabic text categorization. Int. J. Intell. Comput. Inform. Sci., 6: 1-19.

Ta'amneh, H., E.A. Keshek, M.B. Issa, M. Al-Ayyoub and Y. Jararweh, 2014. Compression-based Arabic text classification. Proceedings of the IEEE/ACS 11th International Conference on Computer Systems and Applications, November 10-13, 2014, Doha, Qatar, pp: 594-600.

Thabtah, F., O. Gharaibeh and H. Abdeljaber, 2011. Comparison of rule based classification techniques for the Arabic textual data. Proceedings of the 4th International Symposium on Innovation in Information and Communication Technology, November 29-December 1, 2011, Amman, Jordan, pp: 105-111.

Thabtah, F., O. Gharaibeh and R. Al-Zubaidy, 2012. Arabic text mining using rule based classification. J. Inform. Knowl. Manage., Vol. 11.10.1142/S0219649 212500062.

Thabtah, F., W. Hadi and G. Al-Shammare, 2008. VSMs with K-nearest neighbour to categorise Arabic text data. Proceedings of the World Congress on Engineering and Computer Science, October 22-24, 2008, San Francisco, USA., pp: 778-781.

Wahbeh, A.H. and M. Al-Kabi, 2012. Comparative assessment of the performance of three WEKA text classifiers applied to arabic text. Abhath Al-Yarmouk: Basic Sci. Eng., 21: 15-28.

Yousif, S.A., V.W. Samawi, I. Elkaban and R. Zantout, 2015a. Enhancement of Arabic text classification using semantic relations of Arabic WordNet. J. Comput. Sci., 11: 498-509.

Yousif, S.A., V.W. Samawi, I. Elkabani and R. Zantout, 2015b. The effect of combining different semantic relations on Arabic text classification. World Comput. Sci. Inform. Technol. J., 5: 112-118.

Yousif, S.A., V.W. Samawi, I., Elkabani and R. Zantout, 2015c. Enhancement of Arabic Text Classification Using Semantic Relations with Part of Speech Tagger. In: Advances in Electrical and Computer Engineering, Mastorakis, N.E. and I.J. Rudas (Eds.). WSEAS Press, USA., pp: 195-201.

Zaghoul, F.A.L. and S. Al-Dhaheri, 2013. Arabic text classification based on features reduction using artificial neural networks. Proceedings of the UKSim 15th International Conference on Computer Modelling and Simulation, April 10-12, 2013, Cambridge, pp: 485-490.

Zahran, B.M. and G. Kanaan, 2009. Text feature selection using particle swarm optimization algorithm. World Applied Sci. J., 7: 69-74.

Zaki, T., D. Mammas, A. Ennaji and F. Nouboud, 2010. Classification of Arabic documents by a model of fuzzy proximity with a radial basis function. Int. J. Future Generat. Commun. Networking, 3: 31-42.

Zaki, T., Y. Es-Saady, D. Mammass, A. Ennaji and S. Nicolas, 2014. A hybrid method N-grams-TFIDF with radial basis for indexing and classification of Arabic documents. Int. J. Software Eng. Applic., 8: 127-144.

Zrigui, M., R. Ayadi, M. Mars and M. Maraoui, 2012. Arabic text classification framework based on latent dirichlet allocation. J. Comput. Inform. Technol., 20: 125-140.