Arabic Text Classification: An Improved Model using New Relations-Based Features

Ahmed T. Abdulameer[a], Israa S. Ahmed[b], Dalia A. Abdulameer[c]

[a] IT Dept., Technical College of Management, Middle Technical University, 10047 Bab Al-Muadham, Baghdad, Iraq
[b] Computer Dept., Informatics Institute for postgraduate studies, University of Information Technology and Communication, Baghdad, Iraq
[c] University of Information Technology and Communication, Baghdad, Iraq

Abstract

As a result of increasing Arabic text documents' warehouses on local PC storage as well as on the Web, various tools are emerged to process this type of documents. Text classification and categorization are the most important tools to classify documents in order to save, sort and retrieve these documents later. Accordingly in this paper, an improved model to classify Arabic text documents is proposed. In this model, relations between concepts in the Arabic WordNet dictionary are utilized to propose five new features. These new features are compared with the state of the art features using three quantitative metrics, three evaluation datasets, and three classification algorithms. In the results, the new Proposed Relation-based Features (PRF) show their superiority on the state of the art features in most cases.

Keywords: Arabic Text Classification, Arabic WordNet, Classification Algorithms, Relations-based Features.

1- Introduction

With rapid growth of text documents in electronic form, automatic text analysis is increased noticeably. One task of interest in this area is text classification or categorization (TC), which is a technique for organizing and understanding the text data. Text categorization is the task of classifying or labelling natural language documents into one or more pre-defined classes or categories (such as medicine, politics or sport) based on their content. There are many applications for TC, largely in the context of searching and/or browsing large collections of documents [2, 3].

Arabic is one of the most widely used languages in the world. It is spoken by more than 500 million people as a first or second language. Despite Arabic is widely used language, there are relatively few studies that concern with analysis of Arabic text documents in the literature. Arabic is a challenging language for some reasons [3-5] as follows:

A) Certain combinations of characters can be written in different ways; B) Arabic has a very complex morphology format as compare to English language; C) Irregular plurals are common; D) In Arabic, there are short vowels which give different pronunciation; E) Arabic synonyms are widespread.

Some attempts for automatic Arabic documents classification are accomplished. Most of these attempts are depended on statistical approaches that produce imprecise results. This is because; the lack of semantic information that needed to enhance TC. Therefore, there is an imperious need to use semantic methods/algorithms to classify Arabic documents [3, 6]. Arabic WordNet dictionary is one of the tools that are used to develop semantic-based Arabic TC systems for Arabic language.

Arabic WordNet is one of the semantic and lexical dictionaries for Modern Standard Arabic. It is utilized in Arabic natural language processing applications [7, 8]. Arabic WordNet contains: (1) words (nouns, pronouns, verbs, adjectives and adverbs); (2) their roots; (3) concepts (synsets) and (4) relations among these concepts. Relations among concepts in Arabic WordNet provide semantic information among concepts and the original words. These relations are utilized in this study to enhance Arabic text categorization/classification [1, 9].

Arabic WordNet dictionary is rarely used to enhance Arabic text classification, where these researches have dedicated on enrichment this dictionary itself to enhance classification. These enrichments were either by (a) extending the names' synsets [6, 10, 11] or (b) enriching the existing relations of this dictionary [7]. Few researches only that used the components of Arabic WordNet dictionary (such as relations, synonyms and concepts) for improving TC processes [1, 3, 8].

From other view, a lot of researches focus only on numerous classifiers (classification algorithms) to improve Arabic TC [4, 12, 13]. No one focus on semantics in Arabic TC. This fail in semantic text classification will motivate us to work on enhancing TC based on concepts and semantics. In this paper, new semantic and lexical relations (that extracted from Arabic WordNet) are proposed to improve TC.

This paper is organized in six sections. In Section 2, the literature that related to Arabic text classification is reviewed, especially about features. Section 3 describes parts of Arabic WordNet dictionary. Section 4 clarifies the five new proposed relation-based features. Section 5 shows evaluation of experiments of the proposed model as well as displays and discusses the results. Conclusions are presented in section 6.

2- Related Works

TC is allocating a document to a predefined class or category based on its content. This section will survey previous researches that utilized various feature extraction (FE) methods to enhance TC results. The widely used feature that have been used by most researches is Bag of Words (BoW) [14, 15]. BoW feature measures the frequency of each word in the document but this feature does not reflect semantic information of document which leads to inaccurate TC.

Other improvements have been achieved in this topic such as the researches that proposed by [16] and [17] that used character n-gram feature. In this feature, sequences of characters instead of separate words are utilized to represent text documents. Results of this feature do not enhance TC results over the BoW Noticeably [6]. Word stemming technique is also used to enhance TC (function of this technique is to extract the root of the word) as in [13, 18].

A few researchers utilize Arabic WordNet to enhance TC such as research in [19] that enhance TC by concept enrichment whereas research in [7] used Wikipedia for relations enrichment in Arabic WordNet. In these methods, worthy efforts are achieved to improve TC but with slightly improvement. Lastly, research in [6] used Arabic WordNet's concepts (synsets) instead of original words to enhance Arabic TC. In this research, the first concept from the list of concepts (synsets) is selected as a disambiguation method because they claim that the first concept is the most accurate one. Other researches such as [3, 8] propose Bag of Concepts (BoC) method as a disambiguation method. In this method, all synsets (concepts) are used.

Accordingly, concepts are used in previous works to improve text classification. TC accuracy is increased up to 7 percent. In this research، semantic relations between concepts in Arabic WordNet are used to enhance TC.

3- Arabic WordNet

Arabic WordNet dictionary contains four tags:

- Word: The terms (words) which considered as basic item in the dictionary.
- Item: The concepts of terms which represent the general concept of word.
- Form: The root of the words in the dictionary.
- Link: The relationships between concepts in the dictionary.

To get semantic relations between these components, the four connections must be understood as illustrated in Fig. 1 [1, 3].

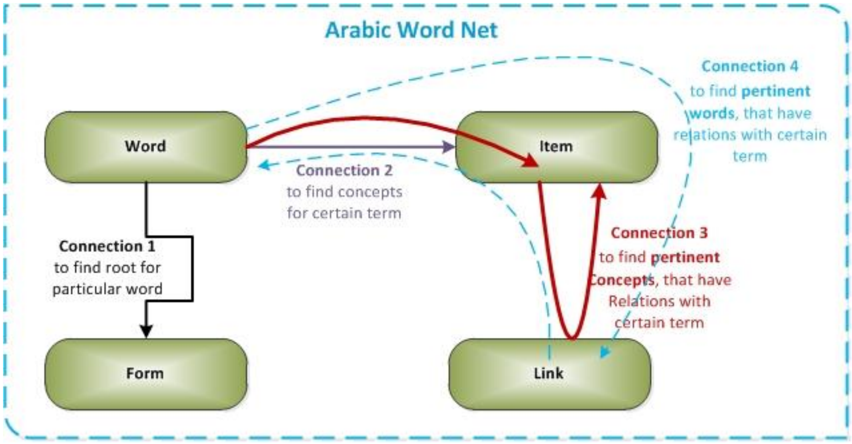


Figure 1: Connections of Arabic WordNet Components (adopted from [1])

Connection1, connection2, Connection 3, and connection 4 are used to find related root, concept or words of certain words in the text document. Connection 4 is used in this research, where it will use to find words that have relations with other words in the Arabic WordNet dictionary. In other words, the contribution of this research is how to find words that are semantically related with original document words (by using different schemes) to improve text classification.

4- The Proposed Relations-Based Features

In Arabic WordNet dictionary, relations between words are used to improve text classification accuracy as in [1, 9]. According to their results, some relations proved their effectiveness for improving text classification whereas the others did not (or just slightly) improve classification accuracy. Accordingly in this paper, weight to each relation can be assigned. High weight is assigned to some relations while low weight is assigned to other relations depending on their percentages in Arabic WordNet. But according to poorness of Arabic WordNet dictionary in covering all words in Arabic language, relations of Arabic WordNet dictionary cannot be fully dependent. Therefore, existence of related words and their frequency in dataset are considered the complementary of relations. An example is explained in Table 1.

Table 1: An example of related words with their frequencies to certain word 'فيلم' in ***BBC Arabic*** dataset

	Original word (term) =”فيلم” , belong to class “ **world news**”			
index	**Related Word**	**Relation Name**	**Class that the term belong to**	**Frequency in dataset**
1	عرض	has_hyponym	Middle east news	202
2	صور	related_to	world news	117
3	مشهد	has_holo_part	Middle east news	57
4	استعراض	has_hyponym	world news	8
5	فن	category_term	Middle east news	6
6	معرض	related_to	---	٠

In this research, the proposed features include replacing the original word in the document with list of related words but depending on their final weight. Thus, several settings can be used for selecting this list of related words; these settings can be explained as follows:

Setting 1: List of related words can be extracted by assign weight to each relation, then this weight multiply with frequency of the word in the dataset to get final weight. Then according specific threshold, list of certain related words will be selected, as explained in Table 2. This first Proposed Relation-based Feature can be denoted as (PRF1).

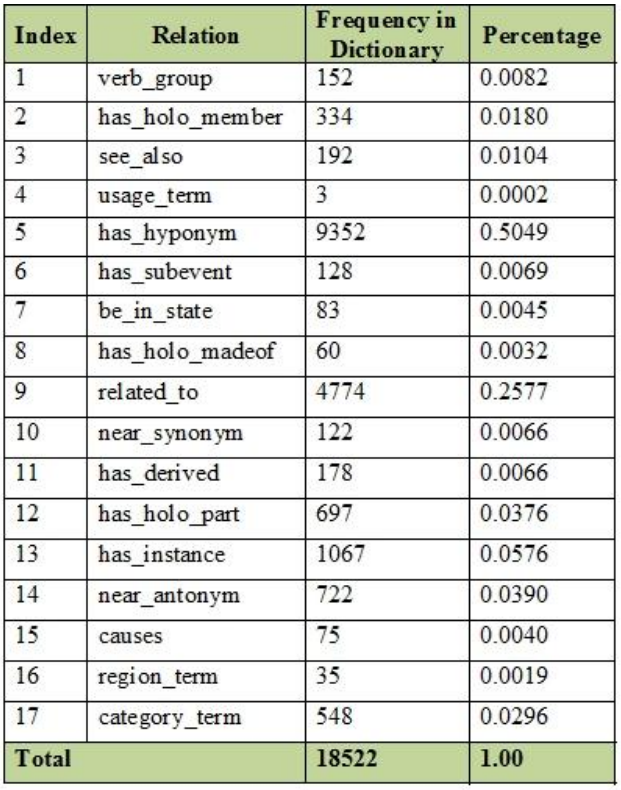
Table 2: Computing final weights of related words to word '**فيلم**'

	Original word =”**فيلم**”				
	Related Word	**Relation Name**	**Relations Weights**	**Frequency of word in dataset**	**Final weight**
1	عرض	has_hyponym	0.5049	202	0.5049*202= 101.99
2	صور	related_to	0.2577	117	0.2577*117= 30.15
3	مشهد	has_holo_part	0.0376	57	0.0376*57= 2.14
4	استعراض	has_hyponym	0.5049	8	0.5049*8= 4.03
5	فن	category_term	0.0296	6	0.0296*6= 0.177
6	معرض	related_to	0.2577	٠	0.2577*0= 0

Weight of relations is determined according to their frequencies in Arabic WordNet dictionary as shown in Table 3.

Setting 2: List of related words can be extracted by select the words that their classes are more frequently. As example in Table 1, the most frequent classes are “Middle east news” where it appeared 3 times. List of words are [عرض, فن, مشهد]. This second Proposed Relation-based Feature can be denoted as (PRF2).

Table 3: Weight (Percentage) of Relations in Arabic WordNet



Index	Relation	Frequency in Dictionary	Percentage
1	verb_group	152	0.0082
2	has_holo_member	334	0.0180
3	see_also	192	0.0104
4	usage_term	3	0.0002
5	has_hyponym	9352	0.5049
6	has_subevent	128	0.0069
7	be_in_state	83	0.0045
8	has_holo_madeof	60	0.0032
9	related_to	4774	0.2577
10	near_synonym	122	0.0066
11	has_derived	178	0.0066
12	has_holo_part	697	0.0376
13	has_instance	1067	0.0576
14	near_antonym	722	0.0390
15	causes	75	0.0040
16	region_term	35	0.0019
17	category_term	548	0.0296
Total		**18522**	**1.00**

Setting 3: List of related words can be extracted by choose all words except the words of frequency zero, because they do not exist in Arabic WordNet dictionary. This third Proposed Relation-based Feature can be denoted as (PRF3).

Setting 4: List of related words can be extracted by choose all words but without removing zero-frequency words, because these words may connect the original words semantically. This fourth Proposed Relation-based Feature can be denoted as (PRF4).

Setting 5: List of related words can be extracted by choosing only the words that have the same class of original word (“world news”), thus list of related words will be [استعراض,صور]. This fifth Proposed Relation-based Feature can be denoted as (PRF5).

5- Experiments Evaluation and Results

This section identifies several parameters that can be used to setup the experiments as follows:

5.1 The first parameter, ***An Evaluation Datasets:*** various datasets have been utilized to test Arabic text classification. (i) The BBC Arabic dataset is the widely used datasets [5, 20]. It contains 7 classes and 4,763 documents. (ii) Saudi Press Agency (SPA) dataset [21], it contains 6 classes and 1,562 documents. (iii) Newspapers Websites (NW) dataset [12], it contains 7 classes and 5,121 documents.

5.2 The second parameter, ***Testing Features:*** are used to test the proposed model by comparing the state of art and proposed features. These features are: (i) Bag-of-Words (BOW) feature [3]; (ii) Bag of Concepts (BOC) features [3]; (iii) Two conceptual features that introduced by [1] list of pertinent synsets (LOPS) and list of pertinent words (LOPW). (iv) Five proposed features (PF1, PF2, PF3, PF4, and PF5) as mentioned in Section 4.0.

5.3 The third parameter, An Evaluation Metrics: Three quantitative metrics are used to measure the performance of the competing features: Precision, Recall and F1-measure [4, 15]. These metrics is computed for each class in the dataset (as illustrated in Eq.1, Eq.2 and Eq.3). Thereafter, the average of these metrics is computed.

$$Precision\ (P) = \frac{number\ of\ correctly\ classified\ document (TP)}{total\ number\ of\ predicted\ documents\ (TP + FP)} \ldots (1)$$

$$Recall\ (R) = \frac{number\ of\ correctly\ classified\ document (TP)}{total\ number\ of\ documents\ in\ the\ class\ (TP + FN)} \ldots (2)$$

$$F1 - measure = \frac{2\ (Precision * Recall)}{Precision + Recall} \ldots (3)$$

5.4 The fourth parameter, Classification Algorithms: The Naïve Bayes (NB), SVM, and K-Nearest Neighbor (KNN) classifiers are employed to choose the best semantic features to improve the Arabic test classification. These classifiers are applied using Weka Software (Waikato Environment for Knowledge Analysis) [22]. It is a popular machine learning software written in Java, developed at the University of Waikato.

5.5 The Results: The classification results for the three classifiers (NB, SVM, and KNN) are presented in Tables 4 to Table 12 respectively. In each table, three metrics (Precision, Recall, and F1) are measured to the three evaluation datasets (BBC Arabic, Saudi Press Agency, and Newspapers Websites datasets).

From the results of classification on three evaluation datasets using NB classifier (in Table 4, 5, and 6), reader can notice that the proposed features outperform on the old features in most cases.

t

Table 4: Results of **NB Classifier** for Three Quantitative Metrics of **BBC dataset**

Features		Precision	Recall	F1-measure
The Old Features	**BOW**	0.68	0.66	0.66
	BOC	0.70	0.68	0.68
	LOPS	**0.72**	0.69	0.70
	LOPW	0.71	**0.76**	0.73
The Proposed Features	**PF1**	0.74	0.75	0.74
	PF2	0.75	**0.77**	0.76
	PF3	**0.76**	0.75	0.75
	PF4	0.73	0.71	0.72
	PF5	**0.76**	0.74	0.74

Table 6: Results of **NB Classifier** for Three Quantitative Metrics of **NW dataset**

Features		Precision	Recall	F1-measure
The Old Features	**BOW**	0.76	0.75	0.75
	BOC	**0.82**	0.76	0.788
	LOPS	0.81	0.77	0.789
	LOPW	0.80	**0.81**	0.80
The Proposed Features	**PF1**	0.85	0.81	0.829
	PF2	0.84	**0.84**	0.84
	PF3	**0.86**	0.83	**0.844**
	PF4	0.80	0.79	0.794
	PF5	0.83	0.82	0.824

From the results of SVM classifier on three evaluation dataset (as shown in Table 7, 8 and 9), reader can notice that also the proposed feature are outperformed the state of the art features in most cases. Additionally, the results show that SVM classifier accuracy in the three metrics (*P, R, and F1*) is higher than the accuracy of NB classifier.

Table 7: Results of **SVM Classifier** for Three Quantitative Metrics of **BBC dataset**

Features		Precision	Recall	F1-measure
The Old Features	**BOW**	0.71	0.72	0.71
	BOC	0.79	0.74	0.76
	LOPS	**0.80**	0.72	0.75
	LOPW	0.79	**0.76**	**0.77**
The Proposed Features	**PF1**	0.84	**0.79**	**0.81**
	PF2	**0.85**	0.77	0.808
	PF3	0.83	0.78	0.804
	PF4	0.83	0.76	0.793
	PF5	0.84	0.75	0.792

Table 8: Results of **SVM Classifier** for Three Quantitative Metrics of **SPA dataset**

Features		Precision	Recall	F1-measure
The Old Features	**BOW**	0.76	0.67	0.71
	BOC	**0.81**	0.65	0.72
	LOPS	0.80	0.68	0.73
	LOPW	**0.81**	**0.83**	0.81
The Proposed Features	**PF1**	0.82	0.81	0.815
	PF2	**0.84**	**0.79**	0.81
	PF3	0.79	0.77	0.77
	PF4	0.80	0.78	0.78
	PF5	0.82	0.86	0.83

It is evident that KNN classifier is weakest classifier among other (NB and

Table 9: Results of **SVM Classifier** for Three Quantitative Metrics of **NW dataset**

Features		Precision	Recall	F1-measure
The Old Features	**BOW**	0.79	0.76	0.77
	BOC	**0.84**	0.75	0.79
	LOPS	0.82	0.76	0.78
	LOPW	0.84	**0.82**	0.82
The Proposed Features	**PF1**	0.88	0.80	0.83
	PF2	0.87	**0.85**	0.86
	PF3	**0.89**	0.82	0.85
	PF4	0.82	0.77	0.79
	PF5	0.84	0.79	0.81

Table 10: Results of **KNN Classifier** for Three Quantitative Metrics of **BBC dataset**

Features		Precision	Recall	F1-measure
The Old Features	**BOW**	0.62	0.63	0.62
	BOC	0.66	0.64	0.64
	LOPS	**0.69**	0.65	**0.67**
	LOPW	0.63	**0.71**	0.66
The Proposed Features	**PF1**	0.70	0.72	0.71
	PF2	0.72	**0.74**	**0.73**
	PF3	0.71	0.73	0.72
	PF4	0.68	0.69	0.68
	PF5	**0.74**	0.70	0.718

SVM) classifiers from the results that shown in Table 10, 11, and 12. To show the results in more clear method, Figure 2 displays the classifiers results of the old and proposed feature on BBC Arabic dataset. The results show superiority the proposed features over the state of the art features. Additionally, the SVM classifier outperforms the other classifiers; NB follows the SVM whereas the KNN was the weakest one.

Table 12: Results of **KNN Classifier** for Three Quantitative Metrics of **NW dataset**

Features		Precision	Recall	F1-measure
The Old Features	**BOW**	0.71	0.66	0.68
	BOC	**0.75**	0.65	0.69
	LOPS	0.76	0.67	0.71
	LOPW	0.77	**0.71**	0.73
The Proposed Features	**PF1**	0.81	0.78	0.794
	PF2	0.82	**0.80**	0.80
	PF3	**0.83**	0.77	0.798
	PF4	0.76	0.73	0.74
	PF5	0.77	0.78	0.77

Table 11: Results of **KNN Classifier** for Three Quantitative Metrics of **SPA dataset**

رون

Features		Precision	Recall	F1-measure
The Old Features	**BOW**	0.63	0.60	0.61
	BOC	**0.70**	0.63	0.663
	LOPS	0.69	0.64	0.664
	LOPW	0.66	**0.70**	0.68
The Proposed Features	**PF1**	0.73	0.66	0.69
	PF2	0.72	**0.70**	0.71
	PF3	**0.74**	0.71	0.72
	PF4	0.70	0.65	0.67
	PF5	0.71	0.69	0.69

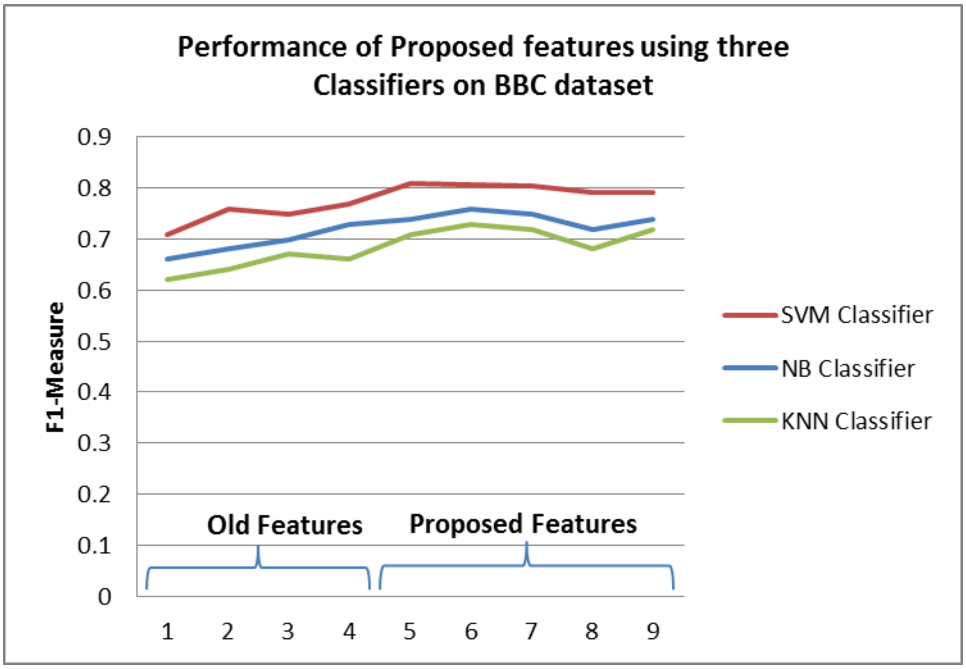


Figure 2: Graphical results of the Proposed Features on BBC dataset using three classifiers

6- Conclusion

According to improve the processes on the Arabic language, an improvement of Arabic text classification was an attempt of this paper. In this research, Arabic WordNet dictionary is used as a base to propose new features. In this context, relations-based features are proposed to improve Arabic text classification where new words that have relation with original document's words will be added to improve classification process. Three datasets and three classifiers are employed to test the proposed features. The results show that the proposed features outperform the state of the art features. SVM offers the best accuracy over NB and KNN classifiers where the F-measure of SVM starts from 0.7 to 0.8 whereas the other classifiers' F-measure ranged from 0.61 to 0.73.

References

[1] S. A. Yousif, V. W. Samawi, I. Elkaban, and R. Zantout, "Enhancement of Arabic text classification using semantic relations of Arabic WordNet," *Journal of Computer Science*, vol. 11, p. 498, 2015.

[2] M. El Kourdi, A. Bensaid, and T.-e. Rachidi, "Automatic Arabic document categorization based on the Naïve Bayes algorithm," in *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, 2004, pp. 51-58.

[3] A. Alahmadi, A. Joorabchi, and A. E. Mahdi, "Combining Bag-of-Words and Bag-of-Concepts representations for Arabic text classification," 2014.

[4] E. Alaa, "A comparative study on arabic text classification," *Egypt. Comput. Sci. J*, vol. 2, 2008.

[5] L. Fodil, H. Sayoud, and S. Ouamour, "Theme classification of Arabic text: A statistical approach," in *Terminology and Knowledge Engineering 2014*, 2014, p. 10 p.

[6] A. Karima, E. Zakaria, T. G. Yamina, A. Mohammed, R. Selvam, and V. VENKATAKRISHNAN, "Arabic text categorization: a comparative study of different representation modes," *Journal of Theoretical and Applied Information Technology*, vol. 38, pp. 1-5, 2012.

[7] M. M. Boudabous, N. C. Kammoun, N. Khedher, L. H. Belguith, and F. Sadat, "Arabic WordNet semantic relations enrichment through morpho-lexical patterns," in *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, 2013, pp. 1-6.

[8] Z. Elberrichi, A. Rahmoun, and M. A. Bentaallah, "Using WordNet for Text Categorization," *Int. Arab J. Inf. Technol.*, vol. 5, pp. 16-24, 2008.

[9] S. A. YOUSIF, V. W. SAMAWI, I. ELKABANI, and R. ZANTOUT, "Enhancement of Arabic Text Classification Using Semantic Relations with Part of Speech Tagger," *W transactions Advances In Electrical And Computer Engineering*, pp. 195-201, 2014.

[10] M. Alkhalifa and H. Rodríguez, "Automatically extending NE coverage of Arabic WordNet using Wikipedia," in *Proc. Of the 3rd International Conference on Arabic Language Processing CITALA2009, Rabat, Morocco*, 2009.

[11] L. Abouenour, K. Bouzoubaa, and P. Rosso, "Improving Q/A using Arabic wordnet," in *Proc. The 2008 International Arab Conference on Information Technology (ACIT'2008), Tunisia, December*, 2008.

[12] S. Al-Saleem, "Associative classification to categorize Arabic data sets," *Int. J. Acm Jordan*, vol. 1, pp. 118-127, 2010.

[13] 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, pp. 1836-1844, 2009.

[14] R. M. Duwairi, "Arabic text categorization," *Int. Arab J. Inf. Technol.*, vol. 4, pp. 125-132, 2007.

[15] 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, pp. 513-538, 2013.

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

[17] L. Khreisat, "Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study," *DMIN*, vol. 2006, pp. 78-82, 2006.

[18] 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, pp. 2347-2352, 2009.

[19] L. Abouenour, K. Bouzoubaa, and P. Rosso, "Using the Yago ontology as a resource for the enrichment of Named Entities in Arabic WordNet," in *Proceedings of The Seventh International Conference on Language Resources and Evaluation (LREC 2010) Workshop on Language Resources and Human Language Technology for Semitic Languages*, 2010, pp. 27-31.

[20] M. K. Saad and W. Ashour, "Osac: Open source arabic corpora," in *6th ArchEng Int. Symposiums, EEECS*, 2010.

[21] F. Thabtah, M. Eljinini, M. Zamzeer, and W. Hadi, "Naïve Bayesian based on Chi Square to categorize Arabic data," in *proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt*, 2009, pp. 4-6.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, pp. 10-18, 2009.