World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 5, No. 6, 112-118, 2015

The Effect of Combining Different Semantic Relations on Arabic Text Classification

Suhad A. Yousif
Mathematics and Computer Science Department
Beirut Arab University
Lebanon

Islam Elkabani*
Mathematics and Computer Science Department
Beirut Arab University
Lebanon

Venus W. Samawi
Computer Information System Department
Amman Arab University
Jordan

Rached Zantout
Electrical and Computer Engineering Department
Rafik Hariri University
Lebanon

Abstract —A massive amount of documents are being posted online every minute. The task of document classification requires extensive background work on the content of documents, where keyword-based matching alone may not be sufficient. Much research has been carried out in several languages that has revealed significant results. However, Arabic documents still pose a great challenge due to the nature of Arabic language. Extracting roots or stems from the breakdown of multiple Arabic words and phrases are an important task that must be completed before applying text classification. The research at hand proposes an algorithm for classifying Arabic-Text documents using semantic relations between words based on an Arabic thesaurus, mainly synonyms, hyperonyms and hyponyms. The experiments conducted in this study evaluated the results using F1-Measure and compared them to results obtained via other existing methods, such as utilizing stemmers and part-of-speech taggers, where it indicated an increment of more than 12.6% for the novel method using semantic relation over other methods. Arabic-WordNet was utilized as a thesaurus for indicating possible relations to be examined. The obtained results indicate that the domain of the semantic web reveals a variety of options for enhancing text classifications, which are highly competitive with current methods. Future work will include identifying best relations to be utilized among the available 20 relations.

Keywords- Arabic Text classification; Stemmer; Part of Speech; Conceptual features; Semantic relations.

I. INTRODUCTION

The process entitled Text Categorization has a significant aim to classify a recent document into one or multiple categories. It is performed through the utilization of prearranged and already classified documents as a training set, thus making it a supervised classification technique. Such a technique in Text Classification has become an important tool to process huge amounts of data on the web [1, 2].

Text preprocessing for Arabic documents is considered a challenging task especially in information retrieval, text mining, and natural language processing where the processing task includes different stages including stop word removal and stemming. The reason behind these additional steps is that Arabic, a Semitic language, is considered a more complicated language compared to English, which is a highly inflected language. Due to this complexity, Arabic needs a set of preprocessing procedures to be ready for manipulation [3]. In fact, text processing techniques might have a positive or negative impact on the accuracy of any text categorization, thus the enhancement of preprocessing stage will necessarily lead to the improvement of any text categorization.

Scientists and researchers have developed many stemming algorithms for different languages including English, Malay, Latin, Indonesian, Swedish, Dutch, German, Italian French, Slovene, Turkish, Bangla, and Chinese [4]. Yet for Arabic, three main different stemming approaches are used: the root-based approach (Khoja [5]), the light stem-based approach (Larkey [6]), and the statistical stemmer approaches root extractor [7]. There is still no complete stemmer for Arabic, however.

The aim of this paper is twofold:1)to compare the accuracy of the three existing techniques used for stemming Arabic text and identify the technique that generates the best results in terms of accuracy, 2) to exploit the Arabic WordNet (AWN) and use it as a lexical and semantic resource in the conceptual representation

*On leave from Department of Mathematics and Computer Science, Faculty of Science, Alexandria University, Alexandria, Egypt.

approach. Moreover, we incorporate AWN in a comparative study with other representation modes in order to analyze its effect. A new relation "Has- Hyponym" is suggested to be used in addition to other previously used relations like “Synset,” “term+ Synset,” and “all Synsets.”The main contribution of this research is to answer whether it is enough to use preprocessing operations (find roots or stems) with bag of words to get good classification results, what the effect of using Part of Speech (PoS) tagger on the classification accuracy for Arabic language is, whether conceptual representation enhances the Arabic classification performance, and which semantic relation positively affects the classification accuracy.

The rest of this paper is organized in five sections. Section two discusses the related works. The suggested approach is explained in section three. System evaluation and effectiveness measure are illustrated in section four. Finally, we conclude in section five.

II. RELATED WORKS

There is a considerable amount of work that has recently been conducted to study [3] techniques specialized for Arabic text stemming that also enhance its accuracy. This new strategy merges three known stemmers known as Khoja, Light Stemmer, and n-Gram. In addition to that, they use the Naive Bayesian (NB) algorithm to stratify all the texts. Eventually, they ended up with a Macro F1average or classification of 0.83.

The authors [4] developed an effective hybrid approach where numerous stemming algorithms used in the pre-processing of the texts. It works by breaking down words into their roots and stems. The previously suggested hybrid approach turned out to be of a superior efficiency compared to the other stemming approaches.

As for the author [8], he proposed a feature reduction method that aims to enhance the effectiveness of the Arabic text classification through artificial neural network and support vector machines [17]. Its main goal is to reduce the number of features for the process. Multiple experiments were conducted yielding significant results that emphasize the superiority of artificial neural networks over support vector machines with an executive function computed by macro averaging F1 measure. By benchmarking these three stemming strategies according to their classification accuracy, the dictionary-lookup stemming surpassed the root-based stemming and light-stemming methods for ANN classifier. As for the SVM classifier, the light stemming turned out to be of a higher efficacy compared to the root based stemming and dictionary-lookup stemming methods.

In [9], the author examines an approach for document classification based on the WordNet notion if the text representation is limited to a set of words, it can omit possible terminologies. This strategy works by selecting the generic and nonexclusive terminologies from the text to integrate them at the next stage with the terms in different ways to form a new representation. This method was tested in multiple experiments using the multivariate chi-square test to reduce the dimensionality. It was concluded that this approach has a significant impact particularly on raising the macro-averaged F1 value.

In [10], the author presented multiple new methodologies for automated categorization of Arabic text documents. These methodologies combine the well-known Bag-of-Words (BOW) as well as the Bag-of-Concepts (BOC) text representation patterns alongside Wikipedia as a source of knowledge. Three distinctive instrument-learning based classifiers were used. The efficacy of the models was assessed by a standard BOW scheme and a concept-based scheme. In [11], the authors generated a conceptual framework of texts through the WordNet. Their model was constructed by clear and genuine terminologies derived from the documents. The manipulated the terminologies concepts of WordNet and their combination. To apply text categorization, they utilized three algorithms: SVM, Decision trees, and KNN. They tested their model on two distinctive corpora. The first one consisted of 11 categories of Reuters for a total of 21,578 articles. As for the second one, it consisted of 7 groups with 20 other documents. They concluded that a combination between terminologies and concepts yielded significant results concerning the three training algorithms. This conclusion is especially significant for the decision trees algorithm.

III. METHOD

The proposed approach is composed of several stages that follow the standard classification model. This model divides the dataset into mutually exclusive training files and testing files, where both parts execute preprocessing and feature extraction phases before the experiment is conducted on the testing files. The last phase is to evaluate the results of the classification after using different feature extraction methods. The schema of the research is depicted in figure 1. Following is a description of those phases and its components.

A. Dataset

A dataset or a corpus is a group of text documents that is categorized under various classes. Lately, it has become highly significant to create an Arabic Corpus as it provides help for all current and future researchers in linguistics topics [13].

Several datasets exist for testing Arabic text classifier systems [14]. However, these datasets can be categorized in two categories: documents that are well formulated linguistically and can be recognized under their respective classifications easily through domain-specific jargon used to write those documents (Separable datasets), and documents that are hard to distinguish due to highly mutual words used in different categories (non-separable)[15][16]. Research conducted on separable datasets tends to render higher accuracy ratios, which in turn make it difficult to distinguish performance measures between various approaches. Therefore, it is a challenging task to produce classifiers for non-separable datasets such as the one proposed in this paper. Among publically available separable datasets there are Al-Jazeera [15] and Al-Watan [16] datasets and among non-separable datasets there is the BBC dataset. The BBC dataset [15] was selected for experimentation purposes in this research due to the huge amount of documents included and the difficulty of classifying its documents using common existing methods.

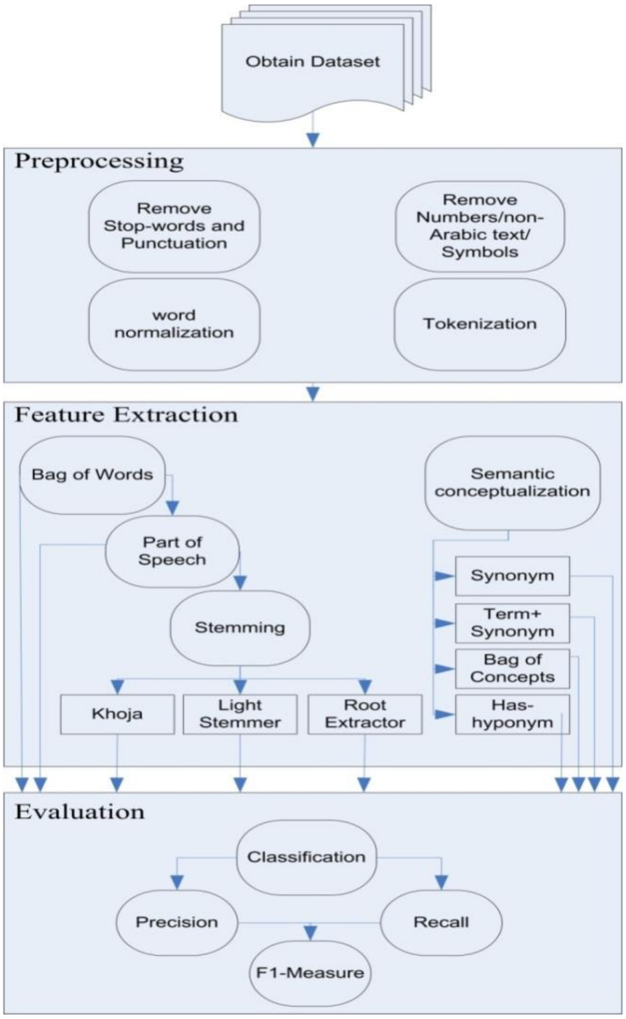

Figure 1. Research schema.

B. Preprocessing

In the preprocessing phase, the size of the document that must be classified is significantly reduced. The main preprocessing task is removing punctuation marks, numbers, and words written in different languages, in addition to stop words (prepositions and pronouns), in order to enhance the text classification technique.

This phase also includes normalizing the documents by replacing letters (" آ إ أ") with ("ا"), the letter ("ؤ ء ") with ("ا"), and the letter ("ى") with ("ا"). The rest of the words are kept and called “keywords” or “features.” However, in large files, the number of these keywords is usually large and must be filtered. Therefore, their number is reduced by removing redundancy wherever exists.

C. Feature extraction

There are two types of text features, external and internal features. External features are not related to the content of the text and include author name, publication date, author gender, and so on. On the other hand, internal keywords reflect the text content and are mostly linguistic features, such as lexical items and grammatical categories [1][18].In this work, words (internal features) are manipulated as a feature on four levels: a) using a Bag of Words form, b) using position taggers to minimize processed data, c) word stem (where suffix and prefix were removed) and word root (where suffix, prefix, and infixes are removed), and d) word concept using a lexical thesaurus.

a) The Representation "Bag of Words"

Bag of Words (BoW) representation originated from the vector model framework and it is considered the simplest representation of texts. Within this representation, the text is transformed into vectors of words that exclude any distance between the words [11]. The representation has two important deficiencies, however, namely, polysemy and synonymy. These occur due to the ambiguity of words and the insufficient information about word's relations. In order to remedy these weaknesses we used conceptual representation in this work.

b) Part-of-Speech Tagging (Position Tagger)

In this work, we consider terms having noun and adjective PoS tags only. Obviously, not all word forms affect the document's meaning in the same way. For instance, nouns contribute effectively to the meaning while adverbs do not. Hence, we make extensive analysis to every word in text categorization.

c) Arabic Stemming Algorithms

Stemming algorithms are extremely helpful to breakdown words to one form; this form can be termed as root or stem. Stemming can be explained as the process of removing prefixes, infixes, or/and suffixes from words to reduce these words to their stem or roots. In this case, resultant roots or stems are called terms. Three different techniques of stemming are applied and tested (Khoja stemmer, light stemmer, and root extractor).

To show different stemming examples, random samples of words from the BBC dataset are illustrated in Table I.

TABLE I. STEMMING EXAMPLES, FROM BBC DATASET

Original Terms	Khoja Stemmer	Light Stemmer	Root Extractor
(derubuk)دربك وستضر (wsetathur) (alqa'ah)القاعة	(derub)درب (therar)ضرر (qoa'a)قوع	(derubuk)دربك (setathur)ستضر (qa'a)قاع	(derub)درب (sther)سضر (laqaa)لقع

The results in Table I shows that there are some words stemmed to incorrect root, and this is due to the deficiencies in each type of stemmer as illustrated in Table II.

TABLE II. DEFICIENCIES OF STEMMERS

Stemmer Algorithm	Weakness
Khoja	• The root Dictionary requires an update to ensure that new detected terms are correctly stemmed • If the root contains a weak letter (i.e. alif, waw, or yah), the form of this letter may change during derivation. Stemmers deal with this by checking if the weak letter is in the correct form. If not, it produces the correct form of this weak letter, which then gives the correct form of the root. • The Khoja stemmer replaced a weak letter with (و ي ا) which produces an incorrect root. For example, the word (“munathammat” منظمات) is stemmed to (“thama” ظمأ)instead of (”nath'ama” نظم)
Light	• Light stemming removes all affixes, predefined in the list, without checking if the remainder is a stem. And in some cases, truncates it from the word and produces an erroneous stem (e.g.“bustan”بستانproduces “busta”بستا). • There is no standard algorithm for Arabic light stemming; all trails in this field were a set of rules to strip off a small set of suffix and prefixes. Also, there is no definite list of these strippable affixes [3]
Root Extractor	• Sometimes in the Root Extractor stemmer, the letter with three smallest product values represents the wrong root. For example (“amalieat”لعمليات) will produce the root (”la’al”لعل, while the correct root is “a'mal” عمل).

d) Representation Based on Concepts

We relied in this work on vector formalism in which vector elements are related to text concepts rather than to text terms. In order to use such representation, we needed to project the terms on a lexicon such as WordNet [11]. By definition, AWN is considered as a lexical reference system whose design was formed by modern psycholinguistic theories examining human semantic systems [11][19][20]. In AWN nouns, verbs, adverbs, and adjectives are arranged into sets of synonyms (synsets) in which every set represents a lexical concept. A synonym is a word that could replace another word without significant changes in meaning. Through conceptual associations, every single synset is connected to a different one.

From another point of view, the most common associations in WordNet can be known as hyperonymy and hyponymy. The hyperonymy class builds notions that enable the generalization of the associations. As for hyponymy, it is the exact opposite of hyperonymy [11] where if word1 is a hyperonym (parent class) of word2 then word2 is a hyponym (Child class) of word1.

d) Classifier

There is no single, perfect text classification algorithm; every algorithm has its own strengths and weaknesses. The most popular classifiers, however, are C4.5 decision tree, SVM, K-NN, and Naive Bayes algorithms, which are applied for text classification [1,14].

The Naïve Bayes classifier is chosen for text classification,. The Naïve Bayes algorithm is based on Bayes rule and conditional probability. Previous research has proven that the Naive Bayesian classifiers one of the most efficient and effective classifiers in terms of computation. It can be easily used in data mining applications [4].

IV. EXPERIMENTATION: EVALUATION AND RESULTS

This section is concerned with the experimental results and the assessment of these results. Two main document representations are used, BoW and concept-based. The BoW representation is sub-divided into three categories (full, with position taggers, and with stemmers)and used with documents classified using NB classifier. The classification accuracy of each representation form and combination of them is illustrated and discussed following.

A. Evaluation

To evaluate the proposed approach, first, a suitable dataset was needed. Second, effective measure had to be specified.

a) Dataset selection

A dataset, or a corpus, is a group of text documents that is categorized under various classes. Lately, it has become highly significant to create an Arabic Corpus as it provides help for all current and future researchers in linguistics topics. We therefore used the BBC dataset as a benchmark dataset for Arabic [15]. The BBC Arabic Corpus was collected from the BBC Arabic website (bbcarabic.com.).The corpus includes 5,258 text files. Each text file was assigned to 1 of 6 categories (Middle East News, World News, Business & Economy, Sports, Religions, Science and Law).The dataset is linearly non-separable.

b) Performance Evaluation Measures

With the previously mentioned features, it is important to extract and generate the frequency list of the dataset features (tokenized single words) and save it in a training file. As for feature extraction, the output result is a long list of features in which not all of them are necessary for the classification operation or might result in contradictory results, such as antonyms. Different techniques were suggested to solve such problems and to help select the most representative features for each class. The most popular methods for Arabic text classification are Term Frequency (TF), Chi Square ($\chi 2$), Document Frequency (DF), and Information Gain (IG). In this work, we used Term Frequency (TF) in feature selection by assigning the weight to be equal to the number of occurrences of term tin document d. This weighting scheme is referred to as term frequency and is denoted TFt,d with the subscripts denoting the term and the document in order[12].

The adopted efficient evaluation was Macro-averagedF1 test. It is a setup from the F1 measure, which combines recall and precision in an equally weighted manner. The measures are explained in the following formulas:

$$Precision = \frac{tp}{tp+fp} (1)$$

$$Recall = \frac{tp}{tp+fn} (2)$$

$$F1 = 2 * \frac{Precision * recall}{Precision + recall} (3)$$

K-fold cross-validation was used in this research to ensure that the system produces reliable results. K was set to 10 in keeping to the precedent established in prior research. The MacroF1 is the harmonic average of the F1 for all distinctive categories, where all the categories are tested in an equal manner. As a result, it is easily affected by the rare categories [5].

B. Classification Results

Following are the experimental results for the main representations based on the four levels discussed previously.

a) Bag of Words

In this work, BoW is adopted with multiple stemmers with and without position taggers to test the possibility of enhancing the accuracy in different cases.

Table III shows the results for applying BoW with the existing three types of stemmers (Khoja, Light, and Root Extractor). By comparing the performance of these three stemmers, we observed that Root Extractor is the best stemmer since it improves the accuracy by 6.2%.

TABLEIII. RESULTS OF BOW WITH AND WITHOUT STEMMERS

BoW with different stemmers	Macro F1 Average
BoW	0.68008
BoW+khoja stemmer	0.71357
BoW+Light stemmer	0.72123
BoW +Root Extractor stemmer	**0.74250**

Table IV illustrates the results for applying BoW with the three types of stemmers (Khoja, Light, and Root Extractor) and position tagger. Based on the achieved outcomes, we can clearly note that using the position tagger with the root extractor improves the classification results by 5.6% compared to using BoW with PoS tagger.

TABLE IV. RESULTS FOR BOW WITH POSITION TAGGER AND STEMMERS

BoW with stemmers and Tagger	Macro F1
BoW +PoS. tagger	0.69096
BoW +PoS.tagger+khoja stemmer	0.71134
BoW +PoS. tagger +Light stemmer	0.7140
BoW +PoS. tagger+ Root Extractor stemmer	**0.74698**

b)Concept Base Representation

The relation between concepts is considered very important in capturing the ideas in texts. Recent research shows that replacing terms with concepts without taking into consideration the relation does not improve the accuracy significantly [9]. Based on that, we used the "Has-hyperonym" relation, and then we added the frequency of "Hyponym" relation to the concept frequency to enhance text classification accuracy.

Table V summarizes the results of this approach with varying features, such as Synonym (concept), terms + synonym, set of all synonym (bag of concept), and the proposed feature Has-Hyponym. The results showed that the best performing feature is the new "Has-Hyponym" relation without the PoS tagger as it improves the accuracy by 7.4% compared to the BoW representation.

TABLE V..RESULTS OF CONCEPT RELATIONS

Semantic Features	Macro F1 Average
First Synset (Synonym)	0.71489
Term + First Synset	0.72165
Bag of Concepts (List of synsets)	0.74799
Has- Hyponym	**0.75437**

Table VI illustrates the results of applying different features such as synonym (concept), terms + synonym, set of all synonym (bag of concept) and Has-Hyponym with position tagger. By comparing the results, we observe that with position tagger, Has-Hyponym provided the best results and improved the accuracy by 7.8%.

TABLE VI. RESULTS OF CONCEPT RELATIONS WITH POSITION TAGGER

Semantic Features and Tagger	Macro F1 Average
First Synset(Synonym) +PoS tagger	0.72538
Term + First Synset+PoS tagger	0.718096
Bag of Concepts (List of synsets)+PoS tagger	0.72067
Has-Hyponym+PoStagger+light stemmer	**0.7589**

c) Extending the Method

The utilization of taggers and stemmers increased the accuracy of results with respect to the full BoW, and although the proposed method of utilizing semantics has shown a similar effect [21], the research was extended to include other semantic relations and measure the change of classification accuracy. The semantic features included synonyms initially which was then used with previous methods such as PoS and Stemmers.

The experiments then tested the Has_Hyponym relation individually and the results were significant. However, when adding the PoS taggers to the relation testing results indicated slightly more enhancement. For this reason we decided to extend the research to include additional experiments on other relations and hybridizations. Among these experiments, Has_Hyponym rendered the highest results. Figure 2 indicates the semantic relations tested initially and figure 3 shows the extended relations that were tested subsequently.

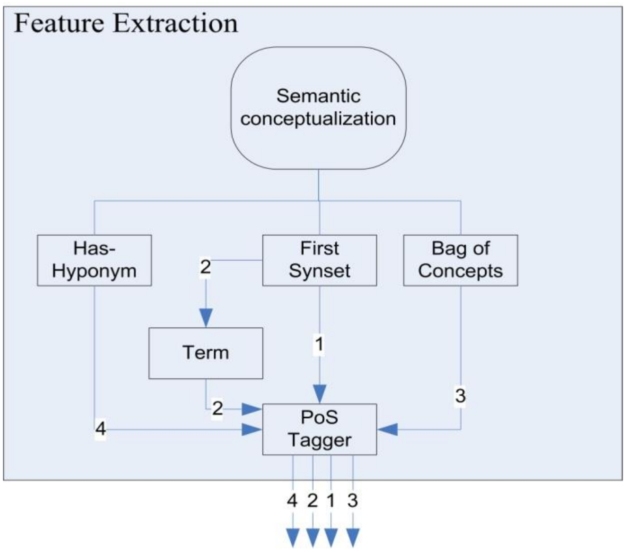

Figure 2. Initial semantic relations included in the classification testing

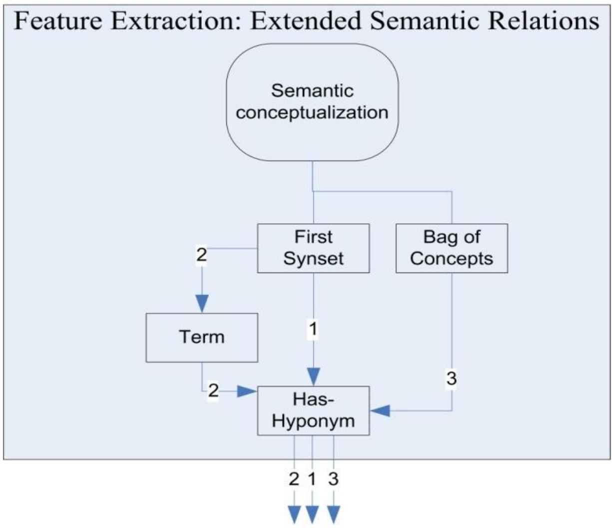

Figure 3. Extended semantic relations included in the classification testing

Table VII below indicates the results for combining different relations as shown in figure 3.

TABLE VII. RESULTS FOR HAS_HYPONYM COMBINATIONS

Combinng of Semantic Features BBC	Macro F1 Average
First Synset +Has_Hyponym	0.7184
Term + First Synset + Has_Hyponym	0.7456
Bag of Concepts (List of synsets) + Has_Hyponym	**0.7660**

As shown in Table VII, combining Has_Hyponym with the list of synsets resulted in the highest measure so far, even more than combining it with the original term. This result was not expected in comparison to synonyms, but indicated that non-separable datasets may include variations of writing styles that tend to detail the topics rather than keeping the same level of discussion. The increment of the last method in table 7 compared to the result of BoW in table 3 reveals an increase of 12.63%, which is a significant increase for the same dataset. This finding leads to expand the research at following stages to test all variation of combining same relation with various techniques or even combining several semantic relations at the same time to arrive at a high accuracy level for classifying similar data.

V. CONCLUSION AND FUTURE WORK

In this work, a comparison between different types of classification techniques was presented. Two main models were tested ,namely, the Bag of Words and lexical concepts models. In the first method, stemmers and part of speech methods were presented and tested. The root extractor with the position tagger showed the best performance among all other stemming approaches. In addition to that, the conceptual representation using WordNet concepts were included and tested. In this approach, the “Has-hyponym” relation outperformed other semantic relations, especially when position tagger was combined with it.

For future work, the research suggests that a new rooting approach based on Arabic WordNet would be greatly beneficial. Furthermore, it also suggests that generating and trying more combinations between conceptual representation relations may produce more accurate results.

REFERENCES

[1]. F. Sebastiani, "Machine Learning in Automated text categorization". ACM Computing Surveys (CSUR), 34(1), 2002, 1-47.

[2]. Genkin, D. D. Lewis, and D. Madigan, "Large-scale Bayesian logistic regression for text categorization," Techno metrics, vol. 49, 2007, pp. 291-304.

[3]. M.Hadni, A.Lachkar, and S. AlaouiOuatik, " A New and Efficient Stemming Technique for Arabic Text Categorization", IEEE, 2012 , pp. 791-796.

[4]. M. Hadni, S. AlaouiOuatik and A. Lachkar, "Effective Arabic Stemming Based Hybrid Approach For Arabic Text Categorization", International Journal of Data Mining & Knowledge Management, Vol.3, No.4, 2013, p1.

[5]. S. Khoja , "Stemming Arabic Text". Lancaster, U.K.,Computing department, Lancaster University, 1999,http://zeus.cs.pacificu.edu/shereen/research.htm.

[6]. L. Larkey., L. Ballestones, and M.E. Connell. "Improving Stemming for Arabic Information Retrieval", PGIR'02, 2002. PP 275-282.

[7]. R. Al-Shalabi, G. Kanaan, and H. Al-Serhan, “New approach for extracting Arabic roots,” Proc. of the International Arab Conference on Information Technology (ACIT’2003), Alexandria, Egypt, December 2003, pp 42-59.

[8]. F. Harrag, E. El-Qawasmah and A.MalikS.Al-Salman, "Stemming as a Feature Reduction Technique for Arabic Text Categorization", IEEE, 2011, pp. 128-133.

[9]. Alahmadi1, A. Joorabchi2, and A. E. Mahdi 3, "Combining Bag-of-Words and Bag-of-Concepts Representations for Arabic Text Classification", ISSC 2014/CIICT 2014, Limerick, June, 2014, 26-27.

[10]. H. Nezreg, H. Lehbab, and H. Belbachir, "Conceptual Representation Using WordNet for Text Categorization", Inernational Journal of Computer and Communication Engineering, Vol. 3, No. 1, January 2014.

[11]. Z. El berrichi and K. Albidi," Arabic Text Categorization: A comparativeStudy of different representation Modes", The International Arab Journal of Information Technology, Vol. 9, No.5, September 2012, pp.465-470.

[12]. 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, , 2013, pp. 513-538.

[13]. M. K. Saad, "OSAC: Open Source Arabic Corpora", Published at the 6th International Conference on Electrical and Computer Systems (EECS’10), Nov 2010, pp. 25-26.

[14]. El-Halees, "A comparative study on Arabic text classification," Egyptian Computer Science Journal, vol. 20, 2009, pp.110-115.

[15]. Mesleh, Support Vector Machines based Arabic Language Text Classification System: Feature Selection Comparative Study, 12th WSEAS Int. Conf. on Applied Mathematics, Cairo, Egypt, December 29-31, 2007, 228-233.

[16]. B.Hawashin, . M Mansour, ShadiAljawarneh",An Efficient Feature Selection Method for Arabic Text Classification", International Journal of Computer Applications (0975 – 8887), Volume 83, No.17, December 2013.

[17]. M. Ikonomakis, S.Kotsiantis, V. Tampakas “Text Classification Using Machine Learning Techniques” WSEAS Transactions On Computers, Issue 8, Volume 4, 2005, pp. 966-974.

[18]. S. D. and S. Horiguchi, "An Efficient Feature Selection using Multi-Criteria in Text Categorization for Naïve Bayes Classifier", WSEAS Transactions on Information Science and Applications, Vol.2, Issue 2, 2005, pp.98-103.

[19]. C..kavitha,.G.SuhdaSadasivam, S.kiruthika, "Semantic similarity based web document classification using Artificial Bee Colony

(ABC) algorithm", WSEAS Transactions on Computers, v 13, 2014, pp 476-484.

[20]. G.Lu, P. Huang, L. He, C. Cu, X. Li, "A new semantic similarity measuring method based on web search engines", WSEAS Transactions on Computers, v 9, n 1, 2010, pp 1-10.

[21]. 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.