Utilizing Arabic WordNet Relations in Arabic Text Classification: New Feature Selection Methods

Suhad A. Yousif, Zainab N. Sultani, Venus W. Samawi, *Member, IAENG*

Abstract— **The availability of Arabic text documents on the Internet entails the use of convenient Arabic text classification (TC) techniques. Arabic TC requires extensive work in analyzing the content of valuable Arabic documents. Its rich vocabulary, semantic ambiguity, and words with semantic relations characterize the Arabic language. Therefore, using a bag-of-words (BoWs) text representation model may yield unsatisfactory results. This study is concerned with utilizing synsets and semantic relations from the original words to enhance Arabic TC accuracy. These relations are extracted using the Arabic WordNet (AWN) thesaurus as a lexical and semantic provenance. AWN provides various semantic relations for the original word. Some relations are more beneficial than others with respect to dataset content. Consequently, we suggest either assigning a weight for each relation, at which, the effect of weak relations can be minimized and the strong relations can be boosted, or selection of appropriate semantic relations. In this paper, two approaches are suggested, relation weighting scheme and relation grouping scheme. At the first approach, a developed weighting scheme for assigning weights to relations and their respective words, on the bases of Akhbar Al Khaleej dataset, is proposed. This method generates a large training file that contains the original words along with the corresponding relations extracted from AWN, as well as their weights. The second approach is based on relation grouping, at which two different types of relations are grouped based on one of three criterions (related semantic meaning, frequency occurrence (FO) of relations in AWN, and the ratio between the FO of relations in the dataset with respect to the FO of the corresponding relation in the AWN). Naive Bayes is used as a classifier, and F1 measure is used to assess the performance of the proposed methods. Tenfold cross-validation scheme is used to reduce the variability of the results. The efficiencies of the suggested methods are illustrated through the weighting scheme and semantic relation grouping. Results show that the proposed methods outperform the classic BoWs and statistical feature selection methods (Chi-Square and Information Gain). The grouping methods enhance classification accuracy and reduce feature dimensionality.**

***Index Terms*— Feature Selection, Machine Learning, Naive Bayes, Relations AWN, Semantic, Text Classification.**

Manuscript received March 26, 2019; revised July 17, 2019.

S. A. Yousif is with Department of Computer Science, College of Science, Al-Nahrain University, Baghdad, Iraq (corresponding author email: suhad.alezzi@gmail.com; say@sc.nahrainuniv.edu.iq).

Z. N. Sultani is with Department of Computer Science, College of Science, Al-Nahrain University, Baghdad, Iraq (email: zna@sc.nahrainuniv.edu.iq).

V. W. Samawi is with Department of Computer Science\Multimedia Systems, Faculty of Information Technology, Isra University, Amman, Jordan (email: venus.samawi@iu.edu.jo; dr_venus2004@yahoo.com).

I. Introduction

TEXT classification (TC) assigns a new text document to predefined categories according to the tenor of the document. The tremendous number of valuable documents available in the Internet, especially old documents or those lacking keywords, has intensified the complexity of information retrieval procedures[1]. Moreover, the contents of most documents, even those with keywords, cannot be readily determined, and document retrieval remains challenging. In this case, automatic text classification (ATC) system can accelerate and facilitate the assignment and retrieval of text documents. Accordingly, enhancing the accuracy of ATC presents an important opportunity for improving the document classification process. ATCs are applied in different applications, such as document clustering, automatic article indexing, email filtering, and natural language processing.

The nature of the Arabic language, with words affected by their diacritics[2], hinders the expansion of new Arabic text classification methods. Therefore, effective methods for manipulating the processing of Arabic text have received considerable attention. Statistical bag-of-words (BoWs) text representation model is commonly used in research about Arabic text classification. BoWs are sets of words extracted from text corpora (automatically)[3]. Each word signifies a feature in the document. BoWs features are proven to be effective in text classification. However, with BoWs, word sequence within a sentence is disregard, and thus information regarding the local context or syntactic structure is often lost. Moreover, the BoWs model lacks the necessary semantic information to facilitate an accurate text classification[4].

The Arabic language is characterized by its substantial vocabulary, semantic ambiguity, and words with semantic relations[5]. Some Arabic words have many synonyms. Likewise, the meaning of some words may differ according to their diacritics[6]. Most Arabic documents do not use diacritics, which makes semantic analysis significant in Arabic text understanding, summarization, and classification. Semantic analysis utilizes synsets, semantic relations (on word, phrase, or sentence level), and word disambiguation. In this work, we are interested in utilizing semantic relations (at word level), which is also called light semantic[7], to improve text document categorization (i.e., To improve Arabic ATC accuracy).

Semantic relations among words are rarely utilized in ATC algorithms, although it aids in clarifying document

meaning. Semantic relations are used to enhance a classification training file by adding synonyms and related words to the BoWs. However, challenges concerning the high dimensionality of features and selecting proper semantic relations that improve classification accuracy (i.e., some relations are more beneficial than the others) are identified[8, 9]. Semantic relations can be obtained by using lexical databases, such as WordNet. In Arabic language, Arabic WordNet (AWN) thesaurus, Wikipedia, among others, that depend on lexical, lexio-semantics, and semantic relations are used by few researchers [8-13]. Many studies have applied either the most used relation (e.g., synonym of words or concepts) or the AWN thesaurus to derive the semantic relations among words, such as usage terms, hyponyms, and verb groups[11]. Recent studies have examined all available semantic relations extracted from the AWN thesaurus to improve the accuracy of TC[8]. Based on their findings, the contribution of this paper examines semantic relations extracted from AWN and assigns different weights to these relations by proposing alternate feature selection weighting methods according to various criteria. These methods aim to generate an optimal training file that contains the original words along with their corresponding relations extracted from AWN as well as their weights calculated by different methods.

In this study, we address the problem regarding the selection of appropriate Arabic semantic relations to enhance text document features and reduce the effect of weak semantic relations. Two approaches are suggested to solve this problem: semantic relation weighting method and semantic relations grouping method. A developed semantic relation weighting method is proposed, where a weight is assigned for each semantic relation based on the frequency occurrence of the semantic relation emerging in the dataset. The second proposed approach (semantic relations grouping method), combines two robust semantic relations according to one of the three criterions. These criteria are as follows:

1) Select relevant semantic relations in terms of their semantic meaning
2) Select semantic relations with the highest frequency occurrence in AWN
3) Select semantic relations with the maximum ratio between the frequency occurrences of a relation in the dataset to the frequency occurrence of the corresponding relation in the AWN.

This approach exploits strong semantic relations to enhance the text features with reduced dimensionality in comparison with the weighting approach.

For a higher standard of evaluation, the performance of the proposed semantic-relations approaches is compared with that of the traditional statistical approaches. In this study, BoWs along with two feature selection methods (Information Gain (IG), and Chi-square) are utilized as traditional statistical approaches. Naive Bayes (NB) is used as a classifier, and F1-measure is used to measure the classification accuracy to assess the performance of the two suggested approaches. Tenfold cross-validation scheme is used to reduce the variability of the results.

Arabic has no benchmark dataset in contrast to English. Therefore, most researchers construct a dataset by collecting text from magazines freely available on the Internet. This study uses the Akbar Al Khaleej dataset, which is separable and an evenly distributed dataset (see [8] for more details). This source is freely available and extensively used in the literature. The component of this dataset can be grouped into defined classes.

For the remainder of the paper, Section 2 illustrates related works, and the dataset is exemplified in Section 3. Section 4 describes the AWN. The suggested Arabic ATC model utilizing semantic relation is illustrated in section 5, Section 6 presents the proposed feature extraction method based on semantic relations (weighting and grouping methods), Section 7 evaluates the experimental results, and Section 8 offers conclusions and directions for future work.

II. Related work

In automatic text classification, semantic relations can be used to enrich a training file, where synonyms and related words in addition to the BoWs are included in the training file. Most semantic relations between words or concepts can be represented by using ontologies. Semantic relations and bag-of-concepts (BoC) are utilized in many ATC methods for various languages. However, semantic relations, on word level, are seldom used in Arabic automatic text classification in spite of their ability of clarifying document sense due the ambiguous nature of Arabic language, in comparison with other natural languages. Moreover, studies that tests all semantic relations to evaluate the effect of these relations on text classification accuracy (especially in Arabic language) are rare. Scott, S. and Matwin[10] replace BoWs representation with lexical and semantic representation based on hypernym relation obtained from WordNet. Ripper system is used to generate text classification rules. Authors replace BoWs representation with lexical and semantic representation based on hypernym relation obtained from WordNet. They concluded that hypernym relation improves text classification for documents employing rare or diverse vocabulary but is inappropriate for concise well-written documents. In[14], the effect of combining semantic relations (synonyms, hypernyms, hyponyms meronyms, and holonyms) on text classification performance is studied. The five semantic relations were obtained from WordNet thesaurus. Authors found that synonym, hypernym, and holonym relations improve classification accuracy when utilizing coordinate matching classifier. Hyponym relation negatively affects the classification accuracy when Naive Bayes classifier is applied. Weighted hypernyms improve the accuracy when applying support vector machine (SVM) classifier. Hypernyms depth in the WordNet was used to calculate the weight of hypernyms. Trevor and Robert [15] also study the effect of integrating sematic relations, part of speech tags, and term weighting schemes (obtained from WordNet thesaurus) on text classification accuracy. Two text classifiers (Naive Bayes, SVM) are utilized. Authors deduced that integrating different WordNet semantic relations and part of speech tags does not remarkably improve classification accuracy of both classifiers (Naive Bayes and SVM). However, Boolean weighting in relation to term frequency weighting shows substantial improvement when applying both classifiers (SVM and Naive Bayes). In [12], using concept representation combined with hypernym words is proposed. The concepts (most frequently used

sense for the document terms) and hypernym relation between concepts are obtained from WordNet. The frequency of each concept is added to that of the corresponding hypernym relation used as the feature in text categorization. Multivariate Chi-square is used as feature selection technique to reduce the training file dimensionality because of the large dimensionality of training file, which resulted from adding hypernyms related to each concept in addition to the concept. Cosine distance is used as classifier. Authors claim that the proposed method is effective in improving the F1 measure. [16] suggested an online news classifier utilizing machine learning approach with SVM. BoC that deals with synonymy and polysemy is used for document representation. Experimental results indicate that the performance of BoC in text classification highly rely on the ability of the semantic annotator to extract concepts from documents. Authors also claim that BoC approach outperforms the BoWs approach. A semantic relation model is suggested in utilizing and extending abstract relations in DOLCE thesaurus. The resulted relations have the ability of expressing relations between words available in domain-specific corpora. Authors also propose combining relations to link between terms that have indirect relationship. Alaa (2014) [4] proposes an Arabic text classification system based on integrating terms and concepts for document representation. Various stemming and classification methods are usually used in studying the impact of combining BoWs and BoC on the classification accuracy. The performance of BoWs model is tested with various stemmers. BoC is extracted from Arabic Wikipedia and Arabic WordNet. Experimental results indicate that combining BoWs and BoC outperforms the classification accuracy of using one of them. In [8] authors studied the impact of semantic relations on Arabic text classification accuracy. Two sets of features are suggested to be used in Arabic text classification utilizing three types of relations (lexical, semantic, and lexico-semantic relations). These features are obtained from AWN thesaurus. The two feature sets are list of pertinent synsets (LoPS) and list of pertinent words (LoPW). LoPS is a list of synsets related to the original terms, while LoPW is a list of words related to the original terms. Fifteen different relations were extracted from AWN thesaurus and used to generate LoPS and LoPw. Naive Bayes is used as a classifier to study the impact of the proposed features on Arabic text classification accuracy. The experimental results indicate that utilizing LoPS as feature set outperforms both BoG and BoC (synset). LoPW, especially with related-to relation, outperforms LoPS, BoWs, and BoC. In [17] authors proposes the use of BoWs and semantic relations to improve Arabic text classification accuracy. Lexical and semantic relations are extracted from AWN thesaurus. Some relations, based on the experimental results, impact classification accuracy more than the others. Consequently, relation-weighting approach is proposed to assign weights to each relation utilizing the relation frequency in both the AWN and the corpus. The training file includes BoWs and their corresponding relation words (obtained from the AWN) along with the relation weights. Naive Bayes classifier is utilized to study the performance of the suggested approach. It was deduced that relation-weighting approach outperforms the BoWs approach.

III. DATASETS

A dataset in text categorization is a group of text documents classified into multiple classes. Since most researchers depend on independent datasets collected from Arabic magazines available on the Internet, it is indispensable to create a benchmark dataset for the Arabic language to help the researchers for using it as a standard.

In this work, the dataset used is AKHBAR-ALKHALIJ. The class that the AKHBAR-ALKHALIJ dataset belongs to is called separable dataset in which it is well subedited linguistically so that it can be recognized easily. Also, the AKHBAR-ALKHALIJ dataset is evenly distributed such that the number of documents is nearly the same [17]. Table (I) shows the separable group of the AKHBAR-ALKHALIJ dataset.

TABLE I.
AKHBAR-ALKHALIJ DATASET.

Class Name	Number of documents
International	346
Economy	273
Sport	429
Local	720
Total	1768

IV. ARABIC WORDNET (AWN)

AWN is a lexical dictionary or database used for finding synonyms and identifying different relations among words in the Arabic language. AWN includes several components, including adjectives, nouns, adverbs and verbs, and is diversified into sets of cognitive concepts (i.e., synsets) [13]. These synsets are connected by utilizing lexical and semantic relations that recognize the structure of AWN as a useful tool for linguistics, TC, and natural language processing. AWN groups the words depending on their definitions and connects them based on specific scenes [18, 19] with four components (tags):

- Item: the term concepts.
- Word: the terms (i.e., words).
- Form: the word roots.
- Link: the concept relationships.

The connections among this four-tuples are required to extract information from AWN. The connections among these four-tuples [20] are illustrated in Fig.1. The links between the AWN thesaurus components are defined as follows:

- *Connection 1*: from which we obtain the root of the solitary word (term).
- *Connection 2*: from which we connect a concept(s) to each word (term). Each word may be assigned with more than one related synset. For example, the word (Akal, عقل) (involves three concepts, namely "تذكر" Tathakar, (استنبط) Istanbata, and (استنتج) Istantaj
- *Connection 3*: from which we extract the related concepts pertinent to a specific word.

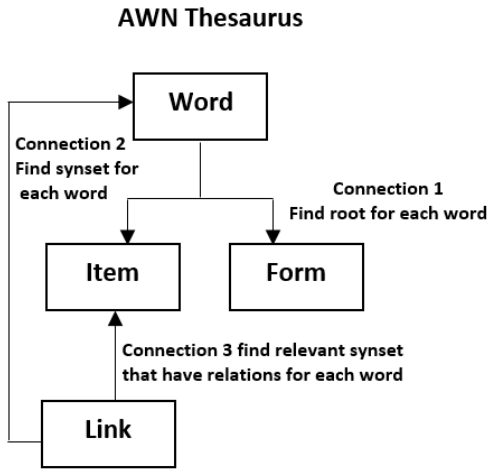
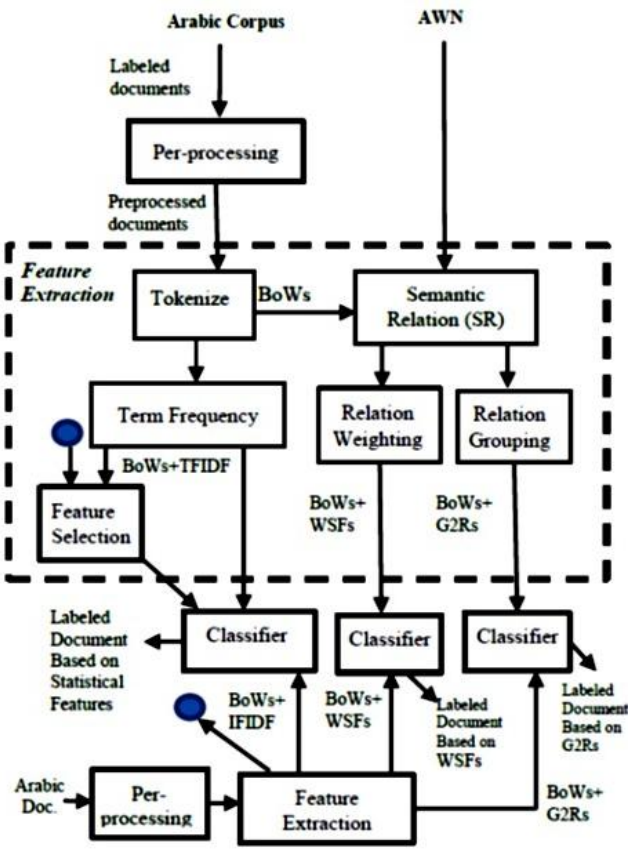


Fig 1: Links between AWN Thesaurus Components

V. ARABIC TEXT CLASSIFICATION UTILIZING SEMANTIC RELATION: PROPOSED MODEL

This work applies a Naive Bayes (NB) classifier[21], which is among the most popular and most straightforward supervised classifier algorithms utilized in the literature. Similar to other supervised classifier algorithms [22], the NB classifier divides the predefined dataset into the training and testing phases (as in Fig2):



*Weighted Semantic features (WSF); *Grouping two relations (G2R)

Fig 2: Classification model utilizing semantic relation: proposed model

- The training phase composes the labelled data, or the training file, and is fed into the NB algorithm. This algorithm is then trained on the labelled data to generate the desired training files for the predefined classes. After preprocessing, the generated training file contains the original words. The weighted frequency and relations among these words are extracted from the AWN[23].
- The testing phase feeds the NB algorithm with undefined data, which are grouped into classes depending on the training file generated during the training phase.

The training phase is divided into a preprocessing step, feature extraction step, and the construction of the training file after applying the previous two steps on the labelled data. Meanwhile, the first two stages in the testing phase are the same as those in the training phase, and the last step depends on the training file generated from the training phase[23]. Some of these stages are discussed in further detail in the following sections.

A. Preprocessing Steps and Cross-validation

Preprocessing is an essential phase in which all undesired data are treated that may affect the classification accuracy, the dimensionality of the training file, the memory and heap size, and the time required to obtain the results. These undesired data may include English or non-Arabic letters, words, numbers, special characters, and stop words (i.e., prepositions and pronouns) [17, 24]. We improve the preprocessing phase by increasing the number of stop words to be removed. We also apply normalization by replacing the characters "أ إ آ" with "ا", "ء ؤ" with "ء", and "ى" with "ا" as well as by increasing the occurrence frequency of these words to make the Arabic characters consistent[25]. We only save the substantial features in the newly generated training file.

To reduce variability and provide better generalization, a cross-validation method called k-fold[26] is used in which all entries in the original dataset for both training and testing data are included. Therefore, we reduce the variability and the prediction result that depends on the random selection of training and testing data. The value of k is chosen similar to the value used by most previous research and, at the same time, depending on the size of the used dataset. Here, we select k for AKHBAR-AL KHALIJ dataset to be ten.

B. Feature Extraction

The features we extract depend on the content of the documents (e.g., words) and their linguistic features, such as the lexical information and grammatical groups. We use two types of extracted features, including the BoWs document representation form[27] as well as those features that are extracted from AWN and are related to each word extracted in BoWs document representation (pertinent words). In each, these two types of extraction, the feature is represented as the words extracted from the documents[8] as illustrated in the following.

BoWs Document Representation

The BoWs document representation is one of the most accessible document representations for natural language processing. In this study, we use BoWs representation with Term Frequency Inverse Document Frequency (TFIDF) weighting in which the first step is to symbolize the text documents as vectors (i.e., the multi-set of words ignoring their grammar). In BoWs, the word is encoded as a feature vector with its frequencies as an element. Each feature has its frequency or weight calculated by TFIDF [4, 27].

Arabic Nets: AWN Semantic Relations

Semantic relation among word documents and concepts could be obtained by ontologies. AWN is considered as one of the most important lexical thesaurus available today [4, 10, 12, 19]. In this research, AWN thesaurus is used to extract concepts or semantic relations from the word documents. All relations established in the AWN thesaurus are illustrated in Fig.3, which form the basis of the proposed methods. From this concept and its semantic relations, we conclude that using semantic relations extracted from the AWN enriches the training file with many words (i.e., synsets or concepts and its semantic relation) and not only the original word.

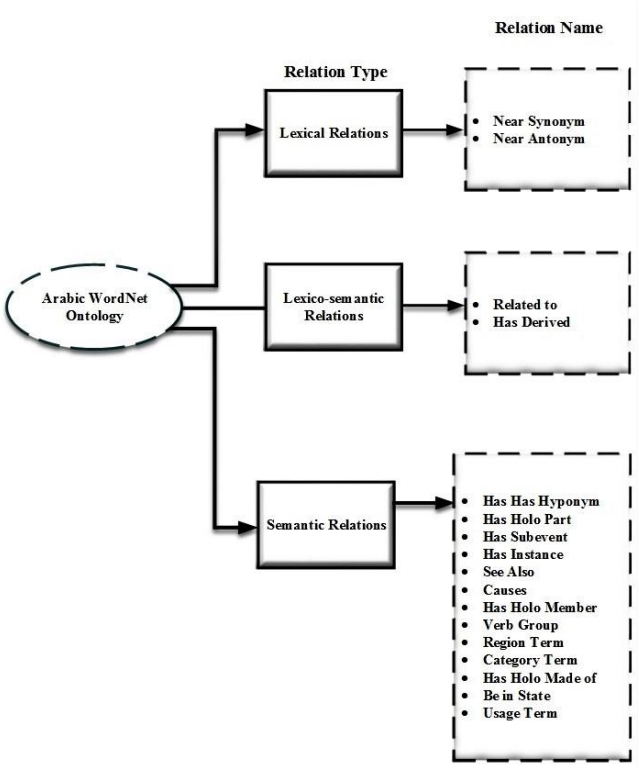


Fig 3: The relations and their types in the AWN thesaurus

C. Feature Selection

In text categorization, the major problem is the high dimensionality of features. Therefore, feature selection is needed to reduce the high dimensionality of features without affecting the categorization accuracy. Feature selection helps to choose the most valuable words that discriminate among different categories in the dataset[28]. As well known, high dimensionality of features may negatively affect the performance of text classification accuracy. To address the issue of high volume of data (i.e. dimensionality of data), many feature selection algorithms have been introduced to reduce the dimensionality of features[29]. In this study, two statistical algorithms are applied on BoWs to select valuable features (words), and two semantic feature selection algorithms based on relation weighting, and relation grouping, are proposed and utilized.

Feature Selection: Statistical Methods

The two statistical feature selection methods utilized in this study are:

1) Information Gain is considerably utilized as a term interest norm in the branch of machine learning algorithms[30]. The IG algorithm works on measuring the number of bits of information produced for category prediction according to the existence or absence of a term in a document. Assume $\{c_i\}_{i=1}^{m}$ signifies ***m*** categories in the dataset. The IG of term ***t*** is defined in (1), which is more general definition compared to the one used in binary categorization model[29].

$$G(t) = -\sum_{i=1}^{m} P_r(c_i)\log P_r(c_i) + P_r(t)\sum_{i=1}^{m} P_r(c_i \mid t)\log P_r(c_i \mid t) + P_r(t')\sum_{i=1}^{m} P_r(c_i \mid t')\log P_r(c_i \mid t')(x+a)^n \quad (1)$$

This general better suits text classification problems since such problems could has m-ary category space (i.e. m could be tens of thousands of categories). The goodness of a term with respect to all categories are need to be measured on average. The dataset is split into training and testing set. In the training set, for each term t, we computed the IG and extracted those features who's their IG less than TH (TH is a predefined threshold used to eliminate weak features).

2) In Chi-Square ($\chi 2$) statistic, mensuration the deficiency of independence between the term t and the category c. Using the two-way contingency table of both t and c, in which A represent the number of times that t and c co-occur, B represents the number of time the t occurs but not c, C is the number of times c occurs but not t, D is the number of times neither c or t occurs, and finally N is the total number of all documents in the corpus. The term-goodness measure is defined in (2).

$$\chi^2 = \frac{N \times (AD - CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (2)$$

The value of $\chi 2$ statistic equals to zero when *t* and *c* are independent. For each category, the $\chi 2$ statistic for each term t in a training set is computed. Finally, two scores ($\chi 2_{avg}$, and $\chi 2_{max}$) are calculated using (3) and (4) respectively[29].

$$\chi^2{}_{avg}(t) = \sum_{i=1}^{m} P_r(c_i)\chi^2(t,c_i) \quad (3)$$

$$\chi^2{}_{max}(t) = max_i^m\{\chi^2(t,c_i)\} \quad (4)$$

Feature selection: Proposed Semantic Relation Methods

In this study, two semantic feature selection methods are proposed:

1) Relation Weighting Method: Specifying semantic relation between words could be treated as a lightweight semantic interpretation paradigm [31]. Therefore, semantic relations could be utilized in Arabic text classification to enrich the training file, which might lead to enhance classification accuracy. However, some relations could negatively affect classification accuracy. To address this problem, we suggest assigning weights to semantic relationships based on certain criteria.
2) Relation Grouping Method: Another suggested approach is to combine two or more relations (ignoring

the relations that negatively affect classification accuracy). In this work, we suggest combining two relations based on three different criteria's.

D. Classifier

In this work, we apply the Naive Bayes classifier which is a simple "probabilistic classifiers based on satisfying Bayes theorem with robust independence postulate between the features[21, 32].

VI. THE PROPOSED FEATURE SELECTION METHODS

Relation weighting method is the process of assign different weights to each sematic relation based on the relations between the words in the AWN thesaurus and those in the dataset (AKBAR-ALKHALIJ). The proposed weighing method utilizes relation frequency (number of relation occurrences) in AWN (as shown in Table (II)). Given that most of the words available in AWN thesaurus include a has-hyponym relation, these words are deemed to have the highest frequency in the AWN. Relation weight is calculated using (5) [17].

$$Relation\_Weight = Relation\ Frequency\ /Total \quad (5)$$

TABLE II
RELATION FREQUENCY IN AWN THESAURUS [17].

#	Relation	Relation Frequency	Percentage
1	verb_group	152	0.00820646
2	has_holo_member	334	0.01803261
3	see_also	192	0.01036605
4	usage_term	3	0.00016197
5	has_hyponym	9352	0.50491308
6	has_subevent	128	0.0069107
7	be_in_state	83	0.00448116
8	has_holo_madeof	60	0.00323939
9	related_to	4774	0.25774754
10	near_synonym	122	0.00658676
11	has_derived	178	0.00961019
12	has_holo_part	697	0.03763093
13	has_instance	1067	0.05760717
14	near_antonym	722	0.03898067
15	causes	75	0.00404924
16	region_term	35	0.00188964
17	category_term	548	0.02958644
Total		**18522**	**1.00000000**

The final weight is computed depending on the results of running "all relations" method as discussed below.

All Relations Method utilizes the 17 relation (illustrated in Fig2) for each concept extracted from the AWN thesaurus. According to this, we propose to combine pertinent words (resulted from all relations 17 relation corresponding to original words) to enrich the training file with words and concepts along with its semantic relations. Table (III) shows the results of combining all the 17 relations available in AWN. However, combining all 17 relations is a time-consuming process, since training file is relatively large compared to the size of the training file that contains BoWs. Therefore, the suggested weighting methods depend on pruning the training file generated from the *All Relations (*17 relation) by choosing best features available in the training file. Pruning the training file will reduces file size, decreases runtime, and might improves classification accuracy since the weak features will be removed.

Based on the previously implemented results, we compute the final weight to be used in the proposed weighting methods. Taking a real example from the *All Relation* training file applied on the AKHBAR-AL KHALIJ dataset and apply (6) [17] to obtain the final weight.

$$W_{ij} = Freq(word_i, doc_j) \times AWN\_Weight(Relation(word_i)) \quad (6)$$

TABLE III.
ALL RELATION METHOD FOR AKHBAR-ALKAHIJ: 10-FOLDS CLASSIFICATION RESULTS

10-Folds	Averaged F-Measure
	AKHBAR-ALKHALIJ dataset
K1	0.9066
K2	0.8911
K3	0.8867
K4	0.8955
K5	0.8699
K6	0.8866
K7	0.899
K8	0.897
K9	0.8799
K10	0.869
Average	0.88813

Example: *related_to* relation weight of the $word_i$=0.2577 (Fakad, فقد) in a doc_j from the AKHBAR-AL KHALIJ dataset is calculated using (6) as relation weight (related-to) as extracted from Table II.

Freq.=90 (frequency of $word_i$ in doc_j) for related_to relation in AKHBAR-ALKHALIJ dataset

W ($word_i$, docj) = 0.2577 × 90 = 23.193, which equals the new final calculated weight as listed in Table IV.

TABLE IV
THE FINAL WEIGHT : A REAL EXAMPLE FROM AKHBAR-ALKHALIJ DATASET

Extracted term	**Relation name**	**AWN weight (Relation)**	**Freq.**	**Final weight**
تميز Tamyz	has_hyponym	0.5049	100	50.49
فقد Fakad	related_to	0.2577	90	23.193
معلم M'oalam	has_holo_part	0.0376	97	3.6472
قسمة Kasma	has_hyponym	0.5049	30	15.147
امتلك Amtalik	category_term	0.0296	70	2.072
حمل Hamal	related_to	0.2577	30	7.731

With the preparation of all required values from Table IV, the proposed methods are discussed below.

A) Dataset-based Weighting Feature Selection Method

Adding all words that are produced from "All Relation" method to the training file, is time-consuming. Therefore, it is beneficial to select set of features to build the training file. Here, mean function value in (7) is used as a threshold, where only the features (related words) with weights (w_i) geater-than the mean value are kept in the training file.

$$Mean = (w_1 + w_2 + \cdots w_n)/n \quad (7)$$

For example, in Table IV, the mean value of all weights calculated is 17.0467. Therefore, only the words [تميز Tamyz, فقد Fakad] are added to the training file (excluding the others) and the same concepts applied to the whole dataset. After applying this method (Dataset Weighting Method) to AKBHAR-ALKHALIJ dataset, the average F1-Measure is improved about 4%, as shown in Table (V).

TABLE V
DATASET WEIGHTING METHOD:
10-FOLDS CLASSIFICATION RESULTS

10-Folds	F1-Measure DS weighting 2
K1	0.957
K2	0.979
K3	0.927
K4	0.938
K5	0.9313
K6	0.9187
K7	0.8965
K8	0.9465
K9	0.892
K10	0.889
Average	0.9275

B) Semantically Grouped Relation Method

In this method, two relations are grouped either based on their semantic meanings, or based on the relation ratio (shown in table (IX), which will be explained in the third grouping method). The idea of using subsets of relations (grouping relations) instead of all relations is based on the fact that some relations are more beneficial than others regarding the dataset content[9], which may negatively affect the classification accuracy. The main problem in semantically grouping relations method is how to choose the group members? And how many relations are combined per group. In this work, integrating two semantic relations (i.e. two relations per group) is utilized. Integrating more relations is a future work. To choose group members, three methods of semantically group relations are suggested.

First Method: Semantically Grouped Relation

Semantically grouped relations focus on closely related relations only, as opposed to all relations, to improve word dispersion. Subsequently, the resultant words from these semantic relations are assumed to have similar meanings, as explained by the following [33]:

- Synset (near-synonym): a semantic relation of equivalence between literals belonging to the same part of speech. For example, if A is synonymous with B, then B is synonymous with A. The synonyms form a synonym set, which is also called a synset.
- Hyponym: in this semantic relation, a word has a more specific meaning than the general or superordinate term applied to it. For example, spoon is a hyponym of cutlery.
- Category term: in this relation, all related words under the same category are grouped under a "category_term" relation.
- Verb group: a relation similar to the category term relation, but the focuses on verbs, whereas "category_term" focuses on nouns.

The two groups of relations are synset and hyponym (representation of the semantic meaning) as well as category_term and related_to (representative of nouns). According to this and using the weights defined in Table 4, we construct the training files that contain only the two combinations of relations. Therefore, for the first combination only the related word [تميز Tamyz] are included in the training file (excluding the others), while for the second combination only the related word [امتلك Amtalik, فقد Fakad, حمل Hamal]. After applying this method to AKBHAR-ALKHALIJ dataset, the classification results are shown in Table VI, which represent the F1-measure in 10-Folds evaluations. By comparing the behavior of both relation combinations outperforms Dataset Weighting Method. This is because Dataset Weighting Method prunes words based on their weights regardless the importance of their relation. Many related words are pruned as illustrated in table (IV), at which, 4 words out of 6 are pruned. This pruning may negatively affect classification accuracy. On the other hand, in Semantically Grouped Relation method, most relations ignored because based on the chosen semantic relation meaning. On the other hand, the combination of Category_term and Related_to outperforms the combination of synset and has-hyponame. This is because the has-hyponym relation is not suitable for short well written documents[10], which is the case in AKBHAR-ALKHALIJ dataset.

TABLE VI.
SEMANTICALLY GROUPED RELATIONS(1ST METHOD):
10-FOLDS CLASSIFICATION RESULTS

10-Folds	Semantically Group First Method	
	Synset+ has hyponym	Category_term+ Related to
K1	0.9758	0.9865
K2	0.9487	0.9787
K3	0.9554	0.9554
K4	0.9655	0.9655
K5	0.9313	0.9413
K6	0.9187	0.9687
K7	0.8965	0.9765
K8	0.9465	0.9565
K9	0.8766	0.9466
K10	0.8654	0.9254
Average	0.92834	0.96011

Second Method: Semantically Grouped Relation

Here, we depend on the relation frequency of each relation in the AWN. Table (III) shows that hyponym and related-to have the highest frequencies, respectively. Thus, we combine these two relations after assigning to them the highest weights. The generated training file contains all the features extracted from combining these two relations, which enriches the training file. Table (VII) illustrates pertinent words between the AWN thesaurus and the datasets. Has-hyponym and related-to have the highest relation frequency with respect to the dataset. The classification results of tenfold are shown in Table (VIII). The first method outperforms the second because of the has-hyponym relation nature, which is incompatible with Akhbar Al Khaleej dataset.

Third Method: Semantically Grouped Relation

This method combines relations depending on the ratio

between relation frequencies in the dataset with respect to relation frequency in AWN (as illustrated in Table (VII)). The ratio is calculated using (8)):

Ratio = Dataset Relation-Freq./AWN Relation Freq. (8)

TABLE VII.
RATIO BETWEEN RELATION-FREQUENCIES IN AKBHAR-ALKHALIJ DATASET AND AWN

AKBHAR-ALKHALIJ Dataset	Relation Freq. In the Dataset	Relation Freq. in AWN	Ratio
Verb-group	254	152	1.671052632
Has_holo_part	1078	697	1.546628407
See_also	654	192	3.406250000
Usage_term	0	3	0.00000000
Has_hyponym	6576	9352	0.703165098
Be_in_state	164	83	1.975903614
Has_subevent	365	128	2.85156250
Related_to	4001	4774	0.838081274
Has_holo_made_of	73	60	1.216666667
Near_synonym	255	122	2.090163934
Has_Derived	621	178	3.488764045
Has_holo_member	321	334	0.961077844
Has_Instance	507	1067	0.475164011
Near_antonym	1654	722	2.290858726
Region_term	25	35	0.714285714
Category_Term	1754	548	3.200729927
Causes	154	75	2.053333333
Total Relations	**18456**	**18522**	

TABLE VIII.
SEMANTICALLY GROUPED RELATIONS (2ND METHOD): 10-FOLDS CLASSIFICATION RESULTS

10-Folds	F1-Measure Semantically Group "Has-Hyponym+Related-to"
K1	0.957
K2	0.9487
K3	0.927
K4	0.904
K5	0.9313
K6	0.9187
K7	0.8965
K8	0.9465
K9	0.8766
K10	0.889
Average	0.91953

Two relations with relatively comparable high ratios can be combined: For example, combining see_also with Has_subevent because their ratios are relatively high (3.63 and 4.12, respectively). The semantic relations that are integrated are:

- See_also : ("relation between concepts having a certain affinity (cold /frozen)"
- Category _term: all related words under the same category are grouped
- Near antonym: "Relation binding two opposite concepts (small /large)"
- Has derived : "A is a word or phrase that appears within B and contributes to B's meaning. pocketbook → book"[33]
- has-subevent: "A and B are events, and B happens as a subevent of A (eating → chewing)"[33]

The combination sets are See_also with Has_subevent, See_also with Category_term, See_also with Has_subevent, Near_antonym with Has_derived, Near_antonym with Category_term, and Has_derived with Has_subevent. Table (IX) shows that the best performance is reached when the relation Has_derived is part of the combined group (See_also with Has_derived, Has_derived with Has_subevent, and Near_antonym with Has_derived). This is due to the definition of Has_ derived relation and its high ratio (i.e., the dataset has many terms with Has_derived relation in comparison with the relation frequency in AWN). Combining of See_also with Has_derived relations outperforms all other methods utilized in this work (as shown in Fig4). From the definition of See_also with Has_derived relations, relations are not highly intersected.

TABLE IX.
CLASSIFICATION RESULTS OF SEMANTICALLY GROUPED RELATION (3RD METHOD)

	F1-Measure					
10-Fold	**Seealso+ has derived**	**Near antonym+ has derived**	**See also+ category term**	**Near antonym+ category term**	**Has derived+ has-subevent**	**See-also+ has-subevent**
K1	0.9865	0.9976	0.9665	0.9665	0.9876	0.9876
K2	0.9967	0.9899	0.9554	0.9675	0.9699	0.9965
K3	0.9844	0.9654	0.9611	0.9554	0.9954	0.9654
K4	0.9666	0.9854	0.9554	0.9777	0.9765	0.9432
K5	0.9532	0.9547	0.9547	0.9642	0.9643	0.9644
K6	0.9724	0.9654	0.9454	0.9588	0.9654	0.9587
K7	0.9654	0.9576	0.9576	0.9498	0.9423	0.9432
K8	0.9543	0.9443	0.9643	0.9547	0.9443	0.9765
K9	0.9343	0.9476	0.9576	0.9432	0.9498	0.9654
K10	0.9568	0.9568	0.9568	0.9589	0.9489	0.9653
Average	**0.96706**	0.96647	0.95748	0.95967	0.96444	0.96662

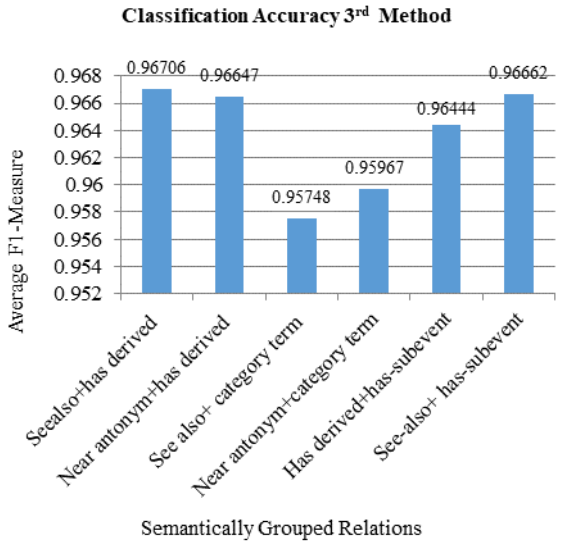


Fig 4 Classification results of semantically Grouped Relation Methods

Moreover, both have high ratio in the dataset with respect to AWN. This will support diversity and improve accuracy. The two groups (Near_antonym with Has_derived and Has_derived with Has_subevent, which enrich training file)

have the same performance. The strong performance of Near_antonym with Has_derived is due to the diversity of relation meanings and strong ratios. However, it is less than See_also with Has_derived as the ratio of See_also is higher. On the other hand, Has_subevent is more relative in meaning to Has_subevent despite having a very high ratio (i.e., Has_subevent has less diversity from Has_derived), which reduces the training file size.

VII. RESULTS AND ASSESSMENT

NB algorithm is among the simplest and most effective supervised machine learning algorithms used for ATC[22, 34, 35]. We evaluate the classification accuracy of the proposed methods with the three metrics of precision, recall, and F1-measure [36, 37] as shown in (9-11).

$$Precision = \frac{TP}{(TP + FP)} \quad (9)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (10)$$

$$F1\_Measure = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \quad (11)$$

The results obtained after running the NB classifier are used to construct a confusion matrix that represents the four parameters of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Based on these parameters, some documents are labelled as being classified correctly, while others are misclassified [38-40].

In this work, sematic relations are utilized in document representation as an attempt to improve Arabic text classification. To evaluate the performance of the suggest methods, they should be compared with traditional statistical approaches. The results of applying BoWs along with two selection methods, Chi-Square and IG, on the Akhbar-Alkhalij dataset are shown in table (X), which illustrates the results of tenfold cross-validation.

TABLE X.
THE AKHBAR-ALKHALIJ RESULTS USING BOWS.

10-Folds	F1-Measure		
	BoW	Chi-Square	IG
K1	0.9488	0.9350	0.9621
K2	0.9287	0.9310	0.9587
K3	0.8854	0.9265	0.9474
K4	0.9643	0.9140	0.924
K5	0.7943	0.9033	0.913
K6	0.7587	0.8830	0.9257
K7	0.8865	0.9124	0.9124
K8	0.8988	0.8921	0.9021
K9	0.6978	0.8861	0.8961
K10	0.8487	0.8798	0.8798
Average	0.8612	0.906323	0.92213

Fig.5 lists the average performance of the statistical selection methods, Chi-Square and IG, in addition to BoWs. One can observe that applying feature selection methods (IG and Chi-Square) outperforms BoWs. This is because, applying BoWs, without feature selection, may produce list of features (words) that are common among various categories (Weak features). In this case, the discrimination process among categories tends to be difficult due to large set of mutual features between categories, which negatively affects the classification accuracy. Therefore, feature selection methods need to be utilized to reduce weak features, which in turn helps in enhancing the classification accuracy.

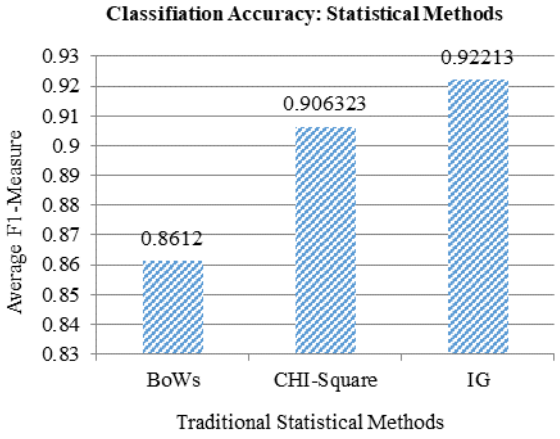


Fig 5 Classification results of applying Statistical Methods

Theoretically, measuring IG depends on the existence or absence of a term in a document, while Chi-Square measures "the lack of independence" between a term and a category[30]. In case of terms with few frequencies, Chi-Square is not considered reliable; this is not the case in IG. IG outperforms Chi-Square when applied on Akhbar-Alkhalij dataset (as shown in Fig5). This is due to low frequency of significant terms in some documents that belongs to certain category, which negatively affect Chi-Square performance with respect to IG. On the other hand, the main problem with IG occurs when applying IG to set of attributes that adopt considerable number of different values[28, 29] , which is not the case in this study.

Fig.6 lists that semantic relation methods outperform traditional statistical methods (BoWs, Chi-Square, and IG) representation as presented when classification results of the suggested semantic relation representation and traditional statistical methods representation when applied to the AKHBAR-ALKHALIJ Arabic dataset are compared. Using semantic representation highly improves the classification performance (0.8612 with BoWs to 0.967 in combined relations approach). In general, the best results are obtained when using the semantically grouped third method, which combines See_also with Has_derived relations (0.967). Semantically grouped first (related_to and category_term relations is (0.96011)) has comparable classification result with the third method. A highly improved performance is observed when semantic relation text representation is used because of the nature of Akhbar Al Khaleej database (documents of the dataset is short with limited vocabulary diversity) and the generated training file that contains valuable words enriched by semantic relations and its combinations. The difference between the performances of the suggested method can be attributed to the training file generated by using the new semantic features and the used weighting schemes. Some of the proposed feature selection methods (based on weighting or combining semantic

relations) might increase text classification accuracy owing to the frequencies of these relations in the dataset. Sometimes, it might be deceptive since enrichment of the training file with other new features will lead to an overlap in the semantic range with other concepts in addition to overlapping of the new features between the types of classes that lead to misclassification.

VIII. CONCLUSION AND FUTURE WORK

This study proposes several weighting schemes for enhancing Arabic TC by using the AWN thesaurus as a lexical and semantic source. Weighting schemes are proposed to assign weights to the relations in the AWN depending on their frequency (Table II) and the terms in the document corpus. The traditional statistical BoWs method is applied to the training file, as well as the weight-related terms extracted from AWN, where the terms are used as new features for the enrichment of the training file. The supervised machine learning of an NB classifier uses the training file to classify text documents. Semantic relations obtained from AWN enhances classification accuracy in comparison with the traditional statistical methods (BoWs, Chi-Square, and IG), as illustrated in Fig.6. This is because utilizing semantic relations will generate pertinent words that can enrich the training file and will improve the classification accuracy by increasing the diversity between classes. Sematic features with more benefits may cause classification degradation when all relations are used. This effect is evident when all relations are utilized to generate pertinent words. Such problem can be solved either by assigning weights to relations based on certain criteria or by using the group of relations (in this work, we tested combining two relations at a time). The findings proved that the approach of combining two relations greatly enhances the classification performance than statistical BoS, all relation, and the relation weighting methods. This is because BoWs ignores the sematic relevance between words, which limits the ATC ability. The proposed weighting scheme depends mainly on the relation frequency in the AWN. This might be deceptive since enrichment of the training file with new features will lead to an overlapping in the semantic range with other concepts, leading to misclassification. Finally, in semantic grouping approach, three different methods are suggested. The first two methods relay on the sense harmony between combined relations and relation frequency in the thesaurus (AWN). The second method showed poor performance because the criteria is based on the thesaurus and not the dataset, while it was found that the best criteria is the one based on the dataset and not the thesaurus (i.e., the third grouping method). The criteria of the first method relays on the harmony between sematic relations regardless of the thesaurus and the dataset. In this case, careful selection of relations must be observed to avoid conflict between classes due to dataset nature, as when we combine Hyponym with related_to. Hyponym does not suit short dataset with limited vocabulary (as the case in Akhbar Al Khaleej dataset). Therefore, the performance highly improved and become comparable to the best combinations of the third method in case of combining category_term with related_to. Fig.6 lists the performance of all suggested semantic relation methods in addition to Traditional statistical methods used in this study. Weighting scheme should be developed in the future to assign weight to relation in a thoughtful way and to study the effect of combining more than two relations on the classification performance.

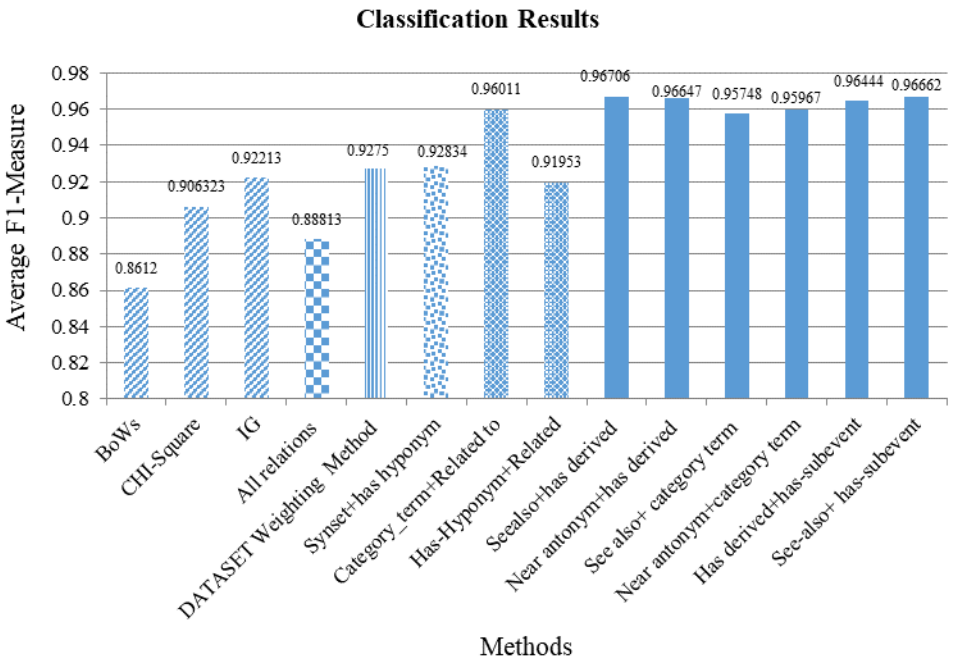


Fig .6 Classification results of traditional statistical methods (BoWs, Chi-Square, IG), and suggested semantic relations methods : applied on Akhbar-Alkhlij dataset

REFERENCES

[1] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining," in Ldv Forum, vol. 20, no. 1, pp 19-62, 2005.

[2] K. Saigh and N. Schmitt, "Difficulties with vocabulary word form: The case of Arabic ESL learners," System, vol. 40, no. 1, pp 24-36, 2012.

[3] F. S. Al-Anzi and D. AbuZeina, "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing," Journal of King Saud University-Computer and Information Sciences, vol. 29, no. 2, pp 189-195, 2017.

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

[5] T. I. Ababneh;, S. M. Ramadan;, and I. M. Abu-Shihab, "Perspectives on Arabic Semantics," International Journal of Humanities and Social Science, vol. 7, no. 7, pp 58-65, 2017.

[6] K. Saigh and N. J. S. Schmitt, "Difficulties with vocabulary word form: The case of Arabic ESL learners," vol. 40, no. 1, pp 24-36, 2012.

[7] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering using WordNet and lexical chains," Expert Systems with Applications, vol. 42, no. 4, pp 2264-2275, 2015.

[8] S. A. Yousif, V. W. Samawi, I. Elkabani, and R. Zantout, "Enhancement Of Arabic Text Classification Using Semantic Relations With Part Of Speech Tagger," Transactions Advances In Electrical Computer Engineering, pp 195-201, 2015.

[9] R. Florian and R. Wicentowski, "Unsupervised Italian word sense disambiguation using WordNets and unlabeled corpora," in Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions, vol. 8, pp 67-73, 2002.

[10] S. Scott and S. Matwin, "Text classification using WordNet hypernyms," Usage of WordNet in Natural Language Processing Systems, 1998.

[11] M. Sahlgren and R. Cöster, "Using bag-of-concepts to improve the performance of support vector machines in text categorization," in Proceedings of the 20th international conference on Computational Linguistics, p 487, 2004.

[12] Z. Elberrichi, A. Rahmoun, and M. A. Bentaalah, "Using WordNet for Text Categorization," International Arab Journal of Information Technology (IAJIT), vol. 5, no. 1, 2008.

[13] A. Abbache, F. Barigou, F. Z. Belkredim, and G. Belalem, "The use of arabic wordnet in arabic information retrieval," in Business Intelligence: Concepts, Methodologies, Tools, and Applications: IGI Global, pp 773-783, 2016.

[14] T. N. Mansuy and R. J. Hilderman, "Evaluating WordNet Features in Text Classification Models," in FLAIRS Conference, pp 568-573, 2006.

[15] T. Mansuy; and R. J. Hilderman, "A Characterization of WorldNet Features in Boolean Models for Text Classification," in Fifth Australasian Data Mining Conference, 2006.

[16] M. Mouriño-García, R. Pérez-Rodríguez, L. Anido-Rifón, and M. Gómez-Carballa, "Bag-of-concepts document representation for Bayesian text classification," in 2016 IEEE International Conference on Computer and Information Technology (CIT), pp 281-288, 2016.

[17] Suhad A. Yousif, Venus W. Samawi, and Islam Elkabani, "Arabic Text Classification: The Effect of the AWN Relations Weighting Scheme," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2017, 5-7 July, 2017, London, U.K., pp594-598.

[18] F. Harrag and E. El-Qawasmah, "Neural Network for Arabic text classification," in Applications of Digital Information and Web Technologies, ICADIWT'09, pp 778-783, 2009.

[19] 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), pp 1-6, 2013.

[20] T. Brasethvik and J. A. Gulla, "Natural language analysis for semantic document modeling," Data & Knowledge Engineering, vol. 38, no. 1, pp 45-62, 2001.

[21] B. Ning, W. Junwei, and H. Feng, "Spam Message Classification Based on the Naïve Bayes Classification Algorithm," IAENG International Journal of Computer Science, vol. 46, no. 1, pp 46-53, 2019.

[22] 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, pp. 51-58, 2004.

[23] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," Journal of machine learning research, vol. 9, pp 1871-1874, 2008.

[24] F. Dařena, "VecText: Converting Documents to Vectors," IAENG International Journal of Computer Science, vol. 46, no. 2, pp 170-177, 2019.

[25] S. A. Yousif, V. W. Samawi, I. Elkabani, and R. Zantout, "The effect of combining different semantic relations on Arabic text classification," World Comput. Sci. Inform. Technol. , vol. 5, no. 1, pp 12-118, 2015.

[26] J. Camacho, A. J. C. Ferrer, and I. L. Systems, "Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: practical aspects," vol. 131, pp 37-50, 2014.

[27] A. Hotho, S. Staab, and G. Stumme, "Ontologies improve text document clustering," ICDM 2003. pp 541-544, 2003.

[28] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," Data classification: algorithms and applications, p. 37, 2014.

[29] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in Icml, vol. 97, no. 412-420, p 35, 1997.

[30] L. Prasad and M. M. Naidu, "CC-SLIQ: Performance Enhancement with 2k Split Points in SLIQ Decision Tree Algorithm," IAENG International Journal of Computer Science, vol. 41, no. 3, pp 163-173, 2014.

[31] V. Santos, M. Hürliman, B. Davis, S. Handschuh, and A. Freitas, "Semantic relation classification: task formalisation and refinement," in Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V), pp 30-39, 2016.

[32] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," Appl. Environ. Microbiol., vol. 73, no. 16, pp 5261-5267, 2007.

[33] Jlowryduda,Relations in ConceptNet 5m. Available: https://github.com/commonsense/conceptnet5/wiki/Relations#relations-in-conceptnet-5, 2019.

[34] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," Expert Systems with Applications, vol. 36, no. 3, pp 5432-5435, 2009.

[35] S. Alsaleem, "Automated Arabic Text Categorization Using SVM and NB," Int. Arab J. e-Technol., vol. 2, no. 2, pp 124-128, 2011.

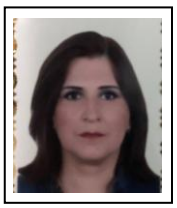
[36] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text classification using string kernels," Journal of Machine Learning Research, vol. 2, pp 419-444, 2002.

[37] G. Forman, "An extensive empirical study of feature selection metrics for text classification," Journal of machine learning research, vol. 3, pp 1289-1305, 2003.

[38] T. Mullen and N. Collier, "Sentiment Analysis using Support Vector Machines with Diverse Information Sources," in EMNLP, vol. 4, pp 412-418, 2004.

[39] M. Hadni, S. A. Ouatik, and A. Lachkar, "Effective arabic stemmer based hybrid approach for arabic text

categorization," International Journal of Data Mining & Knowledge Management Process, vol. 3, no. 4, p 1, 2013.
[40] F. Sebastiani, "Machine learning in automated text categorization," ACM computing surveys (CSUR), vol. 34, no. 1, pp 1-47, 2002.



SUHAD A. YOUSIF is an assistance professor at Al-Nahrain University/College of science/ department Computer science. She received her BSc from Al-Nahrain University in 1994, the MSc from Computer Science department/Baghdad university in 2005 and PhD degrees from Mathematics and Computer Science Department in Beirut Arab University/Lebanon in 2015 respectively. Dr. Suhad supervises MSc theses concerning cloud computing and Big Data analysis, and text classification (natural language processing). She also, leads and teaches modules at both BSc and MSc Levels in computer science. She is a reviewer in number of conferences and Journals. Her special area of research is in text classification, semantic analysis Lately, Dr. Suhad main research interest is in Big Data analysis.

Zainab N. Sultani is a lecturer at Al-Nahrain University, College of Science/ Computer Science department. She received her BSc in Computer Engineering from Al-Balqa'a University in 2006, the MSc from Middle East University in 2012 and PhD from Computer Science Department in University of Technology in 2016. Her special area of research is in machine learning, data mining and image processing.

Venus W. Samawi is an associative professor at Isra University, department of Computer science/Multimedia Systems. Venus Samawi became a Member of IAENG in 2016. She received her BSc from University of Technology in 1987, the MSc and PhD degrees from Computer Science Department in Al-Nahrain University (Saddam University previously) at 1992 and 1999 respectively. Dr. Samawi supervises PhD and MSc theses concerning system programming, pattern recognition, network security, and text classification (natural language processing). She also, leads and teaches modules at both BSc and MSc Levels in computer science. She is a reviewer in number of conferences and Journals. Her special area of research is in pattern recognition, neural networks, genetic algorithms, image processing. Lately, Dr. Samawi main research interest is natural language processing and sentiment analysis