Journal of Computer Science

Original Research Paper

Enhancement of Arabic Text Classification Using Semantic Relations of Arabic WordNet

[1]Suhad A. Yousif, [2]Venus W. Samawi, [1]Islam Elkaban and [3]Rached Zantout

[1]*Department of Mathematics and Computer Science, Faculty of Science, Beirut Arab University, Lebanon*
[2]*Department of Computer Information Systems, Amman Arab University, Jordan*
[3]*Department of Electrical and Computer Engineering, Rafik Hariri University, Lebanon*

Article history
Received: 13-01-2015
Revised: 05-04-2015
Accepted: 07-04-2015

Corresponding Author:
Suhad A. Yousif
Department of Mathematics and Computer Science, Faculty of Science, Beirut Arab University, Lebanon
Email: Suhadmail75@Yahoo.Com

Abstract: Arabic text classification methods have emerged as a natural result of the existence of a massive amount of varied textual information (written in Arabic language) on the web. In most text classification processes, feature selection is crucial task since it highly affects the classification accuracy. Generally, two types of features could be used: Statistical based features and semantic and concept features. The main interest of this paper is to specify the most effective semantic and concept features on Arabic text classification process. In this study, two novel features that use lexical, semantic and lexico-semantic relations of Arabic WordNet (AWN) ontology are suggested. The first feature set is List of Pertinent Synsets (LoPS), which is list of synsets that have a specific relation with the original terms. The second feature set is List of Pertinent Words (LoPW), which is list of words that have a specific relation with the original terms. Fifteen different relations (defined in AWN ontology) are used with both proposed features. Naïve Bayes classifier is used to perform the classification process. The experimental results, which are conducted on BBC Arabic dataset, show that using LoPS feature set improves the accuracy of Arabic text classification compared with the well-known Bag-of-Word feature and the recent Bag-of-Concept (synset) features. Also, it was found that LoPW (especially with related-to relation) improves the classification accuracy compared with LoPS, Bag-of-Word and Bag-of-Concept.

Keywords: Arabic Text Classification, Naïve Bayes, Arabic WordNet, Semantic Relations

Introduction

The evolution of the Internet has led to increased availability of digital textual information and documents written in different languages. Contemporary Internet users should be able to locate the desired information quickly and efficiently. Therefore, improving the information retrieval process has become essential. Although most new documents contain keywords that are used to locate and retrieve related documents quickly and accurately, still, there exist many old documents that do not have keywords. In order to make such old documents locatable, Automatic Text Classification Systems (ATCS) can be used to categorize them based on their content.

Text classification is the process of assigning a text document to a predefined category, or set of categories, depending on its content. ATCS can be used in several applications, such as web page and e-mail filtering, automatic article indexing and clustering and natural language processing (Abouenour *et al*., 2008; Alkhalifa and Rodríguez, 2009; Boudabous *et al*., 2013; Elberrichi and Abidi, 2012). Because English is one of the dominant languages on the World Wide Web, in addition to some other European and Asian languages, most text classification systems are designed for categorizing documents written in one of these languages (Alahmadi *et al*., 2014; El-Halees, 2008). Few attempts have been made to develop an ATCS for documents written in other languages, including Arabic. Most of these attempts are based on statistical approaches (applied on bag of words) that produce inaccurate results. This is due to the lack of semantic

information which is needed to improve text classification. As a result, there is an urgent need to develop ATCS which use semantic and conceptual approaches to classify Arabic documents (Alahmadi *et al*., 2014; Elberrichi and Abidi, 2012). Many tools are available to aide in creating semantic and concept-based ATCS for Arabic such as WordNet.

Arabic WordNet (AWN) is considered one of the best semantic and lexical thesauruses for Modern Standard Arabic. It is widely used in Arabic natural language processing applications (Boudabous *et al*., 2013; Elberrichi and Abidi, 2012). AWN is composed of words (nouns, verbs, adjectives and adverbs), which are listed with their roots, their concepts (synsets) and relations among these concepts. Because the relations between words outlined by AWN provide semantic information among the concepts and their original words, they are exploited in this research to improve Arabic text classification process.

Very few research utilized AWN to improve Arabic text classification. Some of the existing research have focused on the enrichment of the AWN itself to improve classification by either (i) extending the named entities (synsets) (Abouenour *et al*., 2008; Alkhalifa and Rodríguez, 2009; Elberrichi and Abidi, 2012) or (ii) enriching the relations already present in AWN (Boudabous *et al*., 2013). There are limited amount of research, however, that have tried to improve Text Classification (TC) processes using AWN components, such as using n-grams, synonym and concepts (Alahmadi *et al*., 2014; Elberrichi and Abidi, 2012). Many attempts have focused on using various classification algorithms to improve Arabic text classification (Al-Saleem, 2010; Bawaneh *et al*., 2008; El-Halees, 2008; Kanaan *et al*., 2009). All existing Arabic classification methods are not comparable to human classification since most of them do not consider text semantics. This work tackles Arabic text classification based on the enhancement of concepts and semantics. Two new semantic and lexical relations are suggested by means of AWN.

The remainder of this paper is organized in eight sections. In section two, the available literature regarding text classification methods are reviewed, particularly Arabic text classification. Section three explains components of Arabic WordNet thesaurus. Section four focuses on the main phases of the suggested text classification system. The main emphasis is on feature-extraction phase. Section five identifies evaluation metrics of the suggested text classification system. The dataset used in this study is described in section six. The experiment results are assessed in section seven. Finally, conclusions and implications of the experiment are illustrated in section eight.

Related Works

Text classification is the process of assigning text documents to a predefined category or class depending on its content. To improve Arabic TC, we suggest using lexical, semantic and lexico-semantic relations of AWN ontology for text classification improvement. This section will examine past experiments and research that used different feature extraction methods to improve the text classification result. The most basic tool used by most researchers is Bag of Words (BoW) (Duwairi, 2007; Khorsheed and Al-Thubaity, 2013), which uses the frequency of the documents' words as features. This feature lacks semantic information to classify text accurately. Other researchers suggest several improvements such as Sawaf *et al*. (2001; Khreisat, 2006) who used character n-grams as features. In character n-gram method, sequences of characters are used instead of words to represent text documents. This method does not significantly improve TC results over the BoW method (Elberrichi and Abidi, 2012). Extracting the root of the word using stemming methods has also been used to enhance TC results (Duwairi *et al*., 2009; Kanaan *et al*., 2009; Syiam *et al*., 2006). Still other researchers have attempted to use words in their orthographic form (without stemming) in TC (Mesleh, 2007; Thabtah *et al*., 2009). A few researches concerning Arabic TC used AWN for improving TC such as Abouenour *et al*. (2010), who uses Yago ontology for concept enrichment. Boudabous *et al*. (2013) used Wikipedia to enrich the relations of AWN. In these enrichment methods, great efforts are made to improve text classification but the improvement ratio raised by only 8%. Finally, Elberrichi and Abidi (2012) used AWN's components to improve TC. They used AWN's concepts (synsets) instead of original words to improve Arabic text classification. Elberrichi and Abidi (2012) selected the first synset from the list of synsets as a disambiguation method, arguing that the first one is the most accurate synset. Other work used all concepts, called the Bag of Concepts method (BoC), as a disambiguation method (Alahmadi *et al*., 2014; Elberrichi *et al*., 2008). Additionally, the concept in conjunction with the original term was considered to improve TC (Elberrichi *et al*., 2008). As a result, using concepts to improve TC can be achieved using three distinct methods: (i) adding the concept to the original term; (ii) replacing the original word with the concept; (iii) using the list of concepts (BoC) only (Alahmadi *et al*., 2014; Elberrichi *et al*., 2008; Mansuy and Hilderman, 2006). The classification results of using concepts are improved by a ratio up to 7%. In this study, we use semantic relations between concepts to improve text classification accuracy.

Arabic WordNet (AWN)

Before explaining the contribution of this work, AWN components need to be illustrated first. Arabic WordNet has the four components (tags):

- Item: The concepts of terms
- Word: The terms (words)
- Form: The root of the terms and
- Link: The relationships between concepts

Figure 1 clarifies the connections of AWN components and how to find particular information from AWN. Three of these connections (connections 2, 3 and 4) are used in this study. Connection 2 has been used in previous researches (Alahmadi *et al*., 2014; Elberrichi and Abidi, 2012) and is used in this study for comparison purposes.

The four connections illustrated in Fig. 1 are:

- Connection 1 (from word to form): This connection is not used to find the root of a certain term. Instead, the documents' terms are used in their orthographic form.
- Connection 2 (from word (term) to Item): This connection is used to find the concept(s) (synsets) of a specific term. Usually, many synsets are connected with certain term. Some examples of list of concepts are shown in Table 1. In this study, the list of synsets is used in three different forms to compare and evaluate their effectiveness in text classification. The first method is using the simple disambiguation method (Elberrichi and Abidi, 2012), where the first synset from the list is chosen as a replacement for the terms in the document. The second synset disambiguation method chooses the root and the first synset as are placement for the terms in the document. In the third method, all the synsets are selected as a replacement for the terms of the document (Alahmadi *et al*., 2014)
- Connection 3 (Word-Item-Link-Item): is used to find synsets that have relations with words. In other words, connection 3 is used to find the list of pertinent synsets that are closely related to synsets of documents' terms. The third column of Table 2 shows some examples of the lists of pertinent synsets of all relations in AWN. The "usage_term" relation "near_antonym" are not implemented in this study because there are very few instances of it in AWN (occur only three times). Because of the limited usage of the "usage_term" relation, using it will not affect the results. The "near_antonym" relation is also not used because it has the opposite meaning of the original term and thus degrades the results of text classification
- Connection 4 (Word-Item-Link-Item-Word): Used to find relations between terms (words). In other words, it finds the List of Pertinent Words (as shown in column 4 Table 2) in AWN that is closely related to documents' terms. This can be achieved indirectly via synsets, as shown in the blue dashed line in Fig. 1

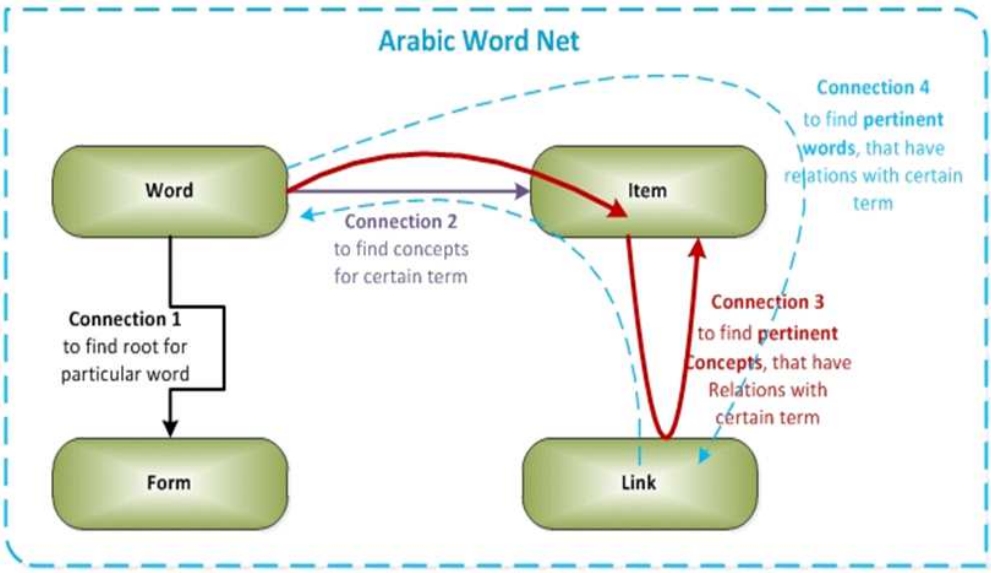

Fig. 1. Connections between the components of Arabic WordNet lexicon

Table 1. Examples of list of concepts (synsets) of some words

Term	List of Synsets
Hokom حكم	مرسوم Marsoom, رأي Ra'i, حكم Hukom, حكم قضائي HukomQatha'i, سيادة Seyada, قرار Kara, نظم Nuthum, عزم Azem, ساد Sad
Ssowar صور	مثل Mathal, رسم Rasem, صور Suwar عرف Urf, صورفوتوغرافيا SuwarFotoghrafia
Aleph ألف	ألف Allafa, كون Kawana, شكل Shakkala, كتب Kataba

Table 2. Examples of lists of pertinent concepts and words of all relations in AWN

Doc. terms	Relation	Pertinent synsets	Pertinent words
Aq'il عقل	related_to	تذكر, استنبط, استنتج Tathakar, Istanbata, Istantaj	استنبط, انتبه الى,تذكر,استنتج, فكر, استدل, توصل استخرج, وصل, توصل إلى Estanbata, EntabehEla, Tathakar, Estantaja, Fakara, Istadala, Tawasala, Istakhraja, Wasala, TawasalaAla
Kas قص	has_hyponym	فرق, شق, أسقط, نشر, نقش, ثلم, ثقب, خدد, تشويه, عملية جراحية Faraqa, Shaqa, Askata, Nashara, Naqasha, Thalama, Thaqaba, Khadada, Tashweeh, Amaliya Jerahiya	فصل, فرق, شق, بضع, حز, أسقط, نشر, نقش, حفر, نمنم, شظى, ثلم, ثقب , نقب, خدد, تشويه, مسخ, جراحة, عملية, عمليةجراحية, قص Fasala, Faraqa, Shaqa, Batha'a, Askata, Nashara, Naqasha, Hafara, Namnama, Shatha, Thalama, Thaqaba, Naqaba, Khadada, Tashweeh, Maskh, Jeraha, Amaliya, AmaliyaJerahiya, Kas
Alseen الصين	has_holo_part	شانغهاي, بكين Bakeen, Shangahi	شانغهاي, بكين, عاصمة الصين الشعبية, الصين Shangahi, Bakeen, AsematAlseenAlsh'biya, Alseen
Amala عمل	verb_group	فعل ,مثل,اجتهد,شغل ,أنجز ,عمل ,شكل Amala, Anjaza, Shaghala, Ejtahada, Mathala, Fa'al, Shakala	عمل, فعل, مثل, لعب دور, اجتهد, شغل, اتخذ, صنع, أنجز, نفذ, اشتغل, شكل, كون Amal, Fa'al, Mathala, La'baDawr, Ejtahada, Shaghala, Etakhatha, Sana'a, Anjaza, Nafatha, Shakala, Kawana
Kataba كتب	has_subevent	راسل ,أكمل ,تهجى Tahaja, Akmala, Easala	كاتب, تراسل, راسل, عبأ, ملأ, أكمل, تهجى Katib, Trasala, Rasala, Aba'a, Mala'a, Akmala, Tahaja
Kahd'am'a قدم	see_also	أخذ ,أعطى ,أعاد A'da, A'ta, Akhatha	جلب, أخذ, أحضر, سلم, أعطى, سلم باليد, أعاد دفع, رد, أعاد, أعاد المال Jalaba, Akhatha, Ahthara, Sallama, A'ta, Salama- Bel-yad, A'daDafe'a, Rada, A'da, A'daAlmal
Eqtesad اقتصاد	category_term	إنتاج ,احتكارالسوق ,استهلاك Estahlaq, EhtekarAlsouq, Entaj	نظرية اقتصادية, نظرية الألعاب, احتكار السوق, استهلاك Nathariya Eqtesadya, Nathariyat Alala'ab, Ehtekar Alsouq, Estehlaq
Adiryatic أدرياتيك	has_instance	Bahar بحر	Bahar, Balteeq, Eeja بحر ,بلطيق, ايجة
Mutawasit متوسط	near_synonym	Tabee'e طبيعي	طبيعي, قياسي, نظامي,عادي Tabee'e, Qyasee, Nethami, Adi
Welada ولادة	has_derived	مخاضي, ولادي Makhathi, Weladi	مخاضي, ماخض, ولادي, متمخض, ولادة Makhathi, Makhath, Weladi, Mutamakheth, Welada
sabab سبب	Causes	حدث Hadath	حصل, جرى, جد, وقع, حدث, دار, سبب Hasala, Jara, Jada,Waqa'a, Hadatha, Dara, Sabab
Qadir قادر	be_in_state	قدرة, إمكان Qudra, Emkan	قدرة, إمكانية, إمكان, مقدرة Qudra, Emkanyia, Emkan, Maqdera
Afreqyia إفريقيا	region_term	ناميبيا Namebia	جمهورية ناميبيا, جنوب غرب أفريقيا, ناميبيا JemhouriyatNamebia, Janoub Gharb Afreqyia, Namebiya
Ma'a ماء	has_holo_made of	الأكسجين, هيدروجين, ماء, ثلج Aloksejeen, Hedroujeen, Ma'a, Thalej	الأكسجين, العنصرالثامن هيدروجين, إيدروجين, العنصرالاول, جليد, ثلج Aloksejeen, AlonsurAlthamenHedroujeen, Eydroujeen, AlonsurAlawal, Jaleed, Thalej
Ispania اسبانيا	has_holo_membe	الاتحاد الأوروبي, منظمة حلف شمال الأطلسي AlitehadAlorupi, Munathamat Helf Shamal Alatlasi	الاتحاد الأوروبي, المجموعة الاقتصادية الأوروبية, السوق الأوروبية, أوروبا, منظمة حلف شمال الأطلسي, حلف الناتو AlItehad Alorupi, Almajmoua'a Aliqtesadiya Alorupiya, Alsouq Alorupia, Aorupa, Munathamat Helf Shamal Alatlasi, Helf Alnato

The Proposed Arabic Text Classification (ATC) Model

The text classification model, mainly, involves three phases. At the first phase, text preprocessing (stop word removing, stemming, normalization, etc.) is needed to prepare the document for features extraction phase. Phase two is concerned with extracting features to be used in the classification phase. Finally, the classification phase, which categorizes the document based on their features. The flow graph of the proposed text classification system using thesaurus (Arabic WordNet is used in this study) is illustrated in Fig 2.

In this study, supervised classification is used. Therefore, the proposed system needs to be trained to produce the knowledge source. The knowledge source contains all the important concepts (with their semantic relations and concept-frequencies) that are part of each class. To specify the class of a new document, the documents need to be pre-processed. Then, features should be extracted and fed to the classifiers along with the knowledge source, which is resulted from the training part.

Text Pre-Processing Phase

In text Pre-Processing, all text redundancies and insignificant information, that affects the text classification accuracy, are removed. The main pre-processing steps (Elberrichi and Abidi, 2012; Torunoglu *et al*., 2011) are:

Text encoding: Avoid any distortion of characters during the text reading process. In this study, all documents are encoded using Unicode (UTF-8).

Removing stop words (determinants, auxiliaries, etc.): Removing all insignificant words from the text (such as "fee في", "illa الى", "lakin لكن", etc…) to

avoid accuracy degradation during the TC process. These words are considered as general words (they do not belong to any text category). Therefore, removing them will not affect classification accuracy whereas considering them lead to downgrade classification accuracy. Stop words also include prepositions, single letters, auxiliary words and formatting tags (Al-Kabi and Al-Sinjilawi, 2007; Khoja and Garside, 1999). In this study, stop words suggested by (Al-Kabi and Al-Sinjilawi, 2007; Khoja and Garside, 1999) are removed.

Text Normalization process: In Arabic language, it is important to transform some characters to single canonical form. This is because some Arabic characters could be written in different forms depending on context. This process includes morphological standardization of some characters and lemmatization of Arabic words. It is the process of grouping the different modified forms of a word together so that they can be analyzed as a single word. The process depends on linguistic concepts as demonstrated in the following examples:

- Delete El-Tanwin "ً"
- Replace all forms of Alef "إ","آ","أ" with "ا"
- Replace all Alif-maksura"ى" withYa"ي"
- Replace all Ha' "ه"with Ta' marbota "ة"

Features Extraction Phase

The performance of any TC model depends on the text representation and classification algorithm (learning algorithm) (Amine *et al*., 2010; Elberrichi and Abidi, 2012). Extracting suitable features from the text can significantly affect the TC performance. Currently, the most popular text representations and features to be extracted are:

Term Frequency (TF)

It reflects the relative importance of certain words (term t) in the document (d). It is used in most TC research (Duwairi, 2007; Elberrichi and Abidi, 2012; Elberrichi *et al*., 2008; Fodil *et al*., 2014; Khorsheed and Al-Thubaity, 2013). It can be computed using Equation (1):

$$TF(t,d) = \#\ occurence\ of\ term\ t\ in\ document\ d \ldots\ldots \quad (1)$$

TF can be extracted from one of the simplest representations of text, Bag of Words (BoW). The basic idea of this representation is to convert the text into a vector of words with their frequencies.

Concept-Based Feature

Concept-based retrieval is a method of retrieving information that is conceptually (or semantically) similar to the information provided in a search query. Extracting concepts from the text cannot be done directly. Instead, the extraction process is achieved by using a lexicon or thesaurus, which serves to connect the semantic concepts to the words. In these lexicons, the word or group of words may relate to their concept (synsets) by different relations like (Has hyponym, near synonym, Related to, Has Derived, etc.). Both WordNet (GWA, 2014) and Arabic WordNet (AWN) lexicons are used for this purpose (Black *et al*., 2006; Elkateb *et al*., 2006). In this study AWN is used. In AWN, words may be associated with their semantic concepts by different relations (Boudabous *et al*., 2013) as illustrated in Fig. 3.

TC can be greatly improved by using these synsets rather than original words. There are three primary concept-based features that are used: Concept Frequency (CFc), term with concept (CF_{t+c}) and Bag of Concepts CFBoC. Assume that *l is* a lexicon, *t* is a term in the document, *s* is a Synset, *SL* is Synset List of the term t and *d* is a document. Equation 2, 3 and 4 illustrate the equations for *CFc,* CF_{t+c} and CF_{BoC} respectively:

$$CFc(t,d) = \frac{\#\ occurance\ of\ concept(t,l)\ in\ d}{\#\ of\ concepts\ in\ d} \ldots\ldots \quad (2)$$

$$CF_{t+c}(t,d) = \frac{\#\ occurrence\ of\ t + \#\ occurrence\ of\ Concept(t,l)}{\#\ words\ in\ d + \#\ Added\ concepts} \ldots \quad (3)$$

In Eq. (3), "# Added Concepts" term refers to the frequency of concepts (synsets related to the original term), added to the frequency of the original term (i.e., computes the frequency of the term plus the frequency of its related concepts (CF_{t+c})):

$$CF_{BoC}(t,d) = \#Synsets\ of\ t\ in\ document\ d \ldots \quad (4)$$

Using synsets will reduce the dimensions of features because many terms may have the same concept and therefore, will be used as one concept. Table 3 above shows that there is a single general concept for several terms.

Finding the concept enhances the classification accuracy because all these terms will be mapped to the related concept when the CF is computed. This will increase the frequency of certain concepts when any of their related words are found in the document.

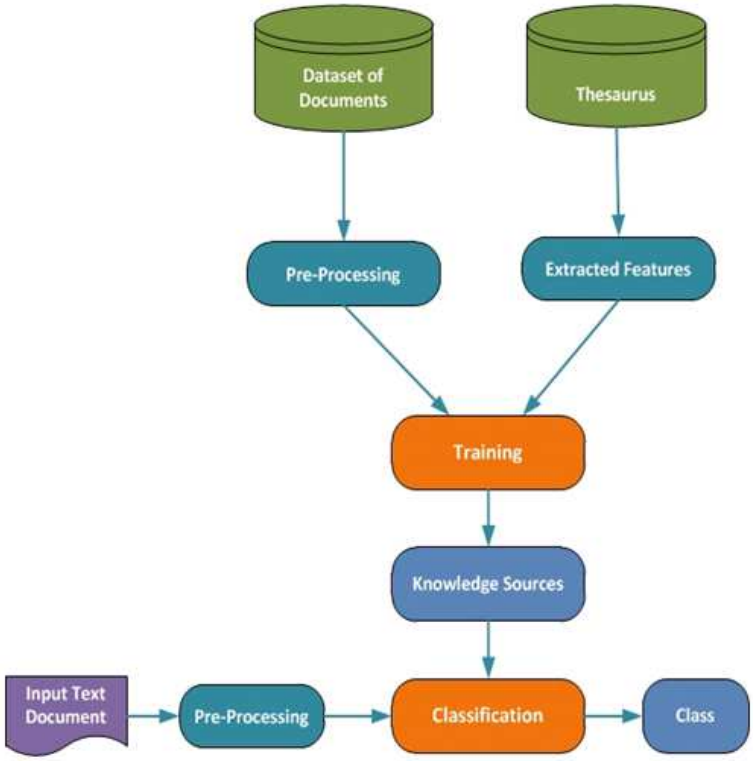

Fig. 2. Flow graph of the proposed Arabic Text classifier

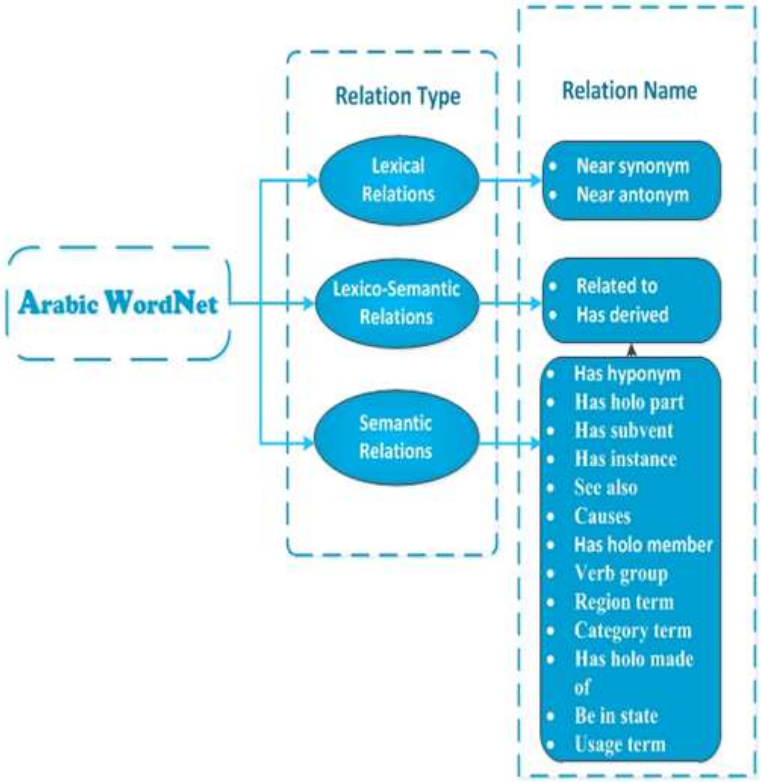

Fig. 3. Relations and their types in AWN (Boudabous *et al*., 2013)

Table 3. Examples of some terms and their concepts

Concept	Lists of terms
سينما Cinema	صور Suar, ابداع فني Ebedaa' Fan'ny, إنتاج فني Enetaa'geFan'ny, خلق فني KhelkFan'ny, أعمال Aa'mal, وسيلة عرض WaseelatArth, فن Fann, مشهدMash'had قدرة Qodra
معرفة Ma'arifa	أدرك Adraka, إدراك Idrak, ملكة Malaka, ملكةعقلية MalakaAqliya
جيش Jaysh	عرف Arafa, سمع Sama'a, بلغ Balagha, تلقى Talqa, وجد Wajada, علم Alema, اكتشف Iktashafa, سمع Sama'a, بلغ Balagha, تلقى Talqa
	عسكر Askar, قوات مسلحة QuwatMusalaha, خدمة عسكرية KhedmaAsqariya, الة الحرب AlatAl-harb

Classification Phase

Text classification can be achieved by one of two approaches, manual or automatic. The manual approach is accomplished by human experts, while the automatic approach is accomplished by well-known classifiers such as Naïve Bayes (NB), Support Vector Machine (SVM), Decision Trees, K-Nearest Neighbor (KNN) and Neural Network (Fodil *et al*., 2014; Harrag and El-Qawasmah, 2009). Recently, due to the massive amount of documents that need to be classified, the automatic approaches have been more widely used. Naïve Bayes and SVM achieved the best results, especially in the text classification (Al-Saleem, 2010; Bawaneh *et al*., 2008; El-Halees, 2008; El Kourdi *et al*., 2004; Kanaan *et al*., 2009; Khorsheed and Al-Thubaity, 2013; Thabtah *et al*., 2009).

Naïve Bayes (NB) classifier has been proven effective in Arabic text classification (Al-Kabi and Al-Sinjilawi, 2007; Bawaneh *et al*., 2008; Kanaan *et al*., 2009; Khorsheed and Al-Thubaity, 2013). We have, therefore, selected it to classify Arabic texts. NB is a supervised machine learning algorithm which involves a learning (training) stage and a testing stage. The learning stage aims to train the NB using samples of already classified data to enable it to predict the classes of unclassified documents. NB depends in its prediction on Bayes' probabilistic rule (Duda and Hart, 1973) illustrated in Equation 5, where c_j represents the class or category of the document d_i that NB needs to predict. The document is assigned to the class that has the highest probability (Duda and Hart, 1973; Duwairi, 2007):

$$P(c_j|d_i) = \frac{P(c_j)*P(d_i \mid c_j)}{P(d_i)} \ldots \quad (5)$$

System Evaluation and Effectiveness Measure

In this study, the main contribution is to determine the proper conceptual features that improve the ATC process, especially with non-linearly separable datasets. The Naïve Bayes classifier is used with competing features to choose the best conceptual features to improve the ATC accuracy. Four features are competed, two old features (Bag-of-Words features and synsets (Bag of Concepts) features) and two newly suggested conceptual features (list of pertinent synsets that have relations with original terms LoPS and list of pertinent words that have relations with original terms LoPW).

To construct a classifier, the system must be trained using the training set. To validate the trained system performance, it must be tested using testing set. Therefore, the dataset must be partitioned into training set and testing set. To reduce variability in results, cross-validation is used. In cross-validation, multiple rounds of dataset partitioning are performed using different random partitioning. The average of the validation results of all rounds is used to evaluate the classification performance of the trained classifier. K-fold cross-validation is used in this research, where K is set to 10 in keeping to the precedent established in prior research (Dai *et al*., 2007; Genkin *et al*., 2007; Mullen and Collier, 2004). The advantage of K-fold cross validation is that all dataset samples are used and nominated for both training and testing. This ensures that the system produces reliable results (Zhang and Yao, 2003). Usually, text documents are represented as a vector of words (terms). Classification of documents therefore depends on these terms and their frequencies in the documents.

To evaluate the performance of the proposed TC system, three quantitative metrics are used: Precision, recall and F1-measure (Forman, 2003; Lodhi *et al*., 2002). Since the output of NB classifier is a confusion matrix that shows the number of documents assigned to each class. Some documents are assigned correctly while others are misclassified, as the confusion matrix demonstrates in Table 4:

$$Precision(P) = \frac{number\ of\ Correctly\ Classified\ documents(TP)}{The\ total\ number\ of\ Predicted\ documents(TP+FP)} \ldots \quad (6)$$

$$Recall(R) = \frac{The\ number\ of\ Correctly\ Classified\ documents(TP)}{\begin{array}{c}The\ total\ number\ of\ documents\ that \\ actually\ belong\ to\ the\ class(TP+FN)\end{array}} \ldots \quad (7)$$

Although these metrics measure classification performance accurately, they are inadequate when used alone. Using the trade-off metrics between them, called F1-measure, is therefore essential:

$$F1-measure = \frac{2(Precision*Recall)}{Precision+Recall} \ldots \quad (8)$$

First, the F1-measure is computed for each class (category) in the dataset. Then, the average of the F1-measures of the 10 rounds is used (known as the F1-measurevalue).

Table 4. Confusion matrix

Predicted class	Class (C)	Class (not C)
Class (C)	TP	FP
Class (not C)	FN	TN

Table 5. Number of documents in each category of Arabic BBC dataset

Category	Number of documents
Middle East news	2356
World news	1489
Business and economics	296
Sports	219
Magazine	49
Science and technology	232
Collection (art and culture)	122
Total	4763

Evaluation Dataset

Several datasets have been used for Arabic text classification. The BBC Arabic dataset is one of the most widely used datasets (Fodil *et al*., 2014; Saad and Ashour, 2010). It is free and public and contains a suitable number of documents for the classification process (Dawoud, 2013). Therefore, it is widely used in previous and current research. The BBC Arabic dataset is downloaded from (Saad and Ashour, 2010). It includes 7 classes and 4,763 text documents. The corpus contains 1,860,786 words (approximately 1.8 million words).

The type of dataset strongly affects the TC results. The datasets can be divided into two types: Linearly separable datasets and non-linearly separable datasets. The non-linearly separable type of datasets has a high percentage of intersection between its categories. In other words, there is a group of words that belongs to more than one class at the same time. Accordingly, this degrades the accuracy of the classification results of such datasets compared with the accuracy of the results of the linearly separable type datasets. In this study, the BBC Arabic dataset is used to test the classification ability of the suggested system. It is a large non-linearly separable dataset. Table 5 shows the number of documents in each category of BBC Arabic dataset.

Assessment of Experimental Results

In this study, four different features are used. These features covers both the traditional Bag-of-Words with term frequency features and the synsets (Bag of Concepts BoC) recently used by few Arabic language researchers (Alahmadi *et al*., 2014). In addition to the conceptual features proposed in this study. The proposed features are based on lexical, semantic and lexico-semantic relations of AWN ontology. The proposed features are: List of Pertinent Synsets (LoPS) (list of synsets that have specific relation with original terms *t*) and List of Pertinent Words (LoPW) (list of words that have specific relation with original terms). LoPS and LoPW are illustrated in Equations 9 and 10 respectivly:

$$LoPS(t) = \#\,Synsets\ that\ related\ to\ term\ t\ldots \quad (9)$$

$$LoPW(t) = \#\,Words\ that\ relate\ to\ Synsets\ that\ relate\ to\ t\ldots \quad (10)$$

In both proposed features, the 15 different relations from AWN (listed in Table 7) are used. The classification results (illustrated in Table 6) shows that LoPS outperforms BoW in all AWN's relations.

Table 7 illustrates the improvement ratio of LoPS and LoPW over BoW and BoC for all relations. The improvement ratio of LoPS and LoPW over BoW reached is about 12 and 13.1% respectively in the "related_to" relation. In most cases, the proposed LoPS and LoPW outperformed the BoC (the most recently developed method of TC) with an improvement ratio up to 6.2 and 7.4% respectively. They outperform BoC since the substitutions of certain terms with concepts that are closely related to them increase the probability of finding similar terms in the same category.

The (LoPW) proposed feature improved the classification results compared to the results produced by the (LoPS) proposed feature in all relations, as listed in Table 8. This improvement of results is achieved using words and synsets (that related to the original term) in most cases instead of concepts (synsets) only. The ratio is not improved (or slightly improved) in 3 relations and degraded in 2 relations. This is because the relations return words different than the original words (according to the relation type), as illustrated in Table 9. Accordingly, the LoPW with relation "related-to" is outperforms the other features (LoPS, BoW and BoC). The improvement in the proposed methods is explained in the following example:

- Term= ' عقل'
- LoPS= "تذكر, استنبط, استنتج"
- LoPW= " انتبه الى, تذكر, استنتج, فكر, استنبط, استدل, توصل"

Assume that the TF of term (akel عقل) in certain document is 3. When LoPS is used, in this case, the frequency of the 3 concepts (istiantagea, tatha'akarah, istanbata) (استنتج, تذكر,استنبط) is added to the term 'akel عقل'. In this case, the TF in the document becomes 7. LoPW contains 7 pertinent concepts (using the relation "related-to" with the term 'akel عقل'). The 7 pertinent concepts and the term (akel عقل) appeared 12 times in the document (i.e., the TF of the term 'akel عقل' becomes 12). From this example, it is clearly seen why LoPW outperforms BoC and LoPS.

Table 6. Classification results of competing features on Arabic BBC datasets

Index	Feature	Results (Average F1-measure)	
Old features			
0	Bag of Words (orthographic form)	0.68008	
1	1st Synset (synonym)	0.71489	
2	Word + 1stSynset	0.72165	
3	Bag of concepts (list of synsets)	0.72538	
Proposed features			
	Relation name	List of pertinent Synsets	List of pertinent words
4	Related-to	0.77323	0.783010
5	Has- hyponym	0.75437	0.758450
6	Has-holo-part	0.75185	0.758177
7	Verb-group	0.73319	0.739601
8	Has-subevent	0.72574	0.730921
9	See-also	0.72574	0.752368
10	Category-term	0.76961	0.769981
11	Has-instance	0.73742	0.754569
12	Near-synonym	0.73072	0.731720
13	Has-derived	0.72656	0.731167
14	Causes	0.71916	0.732273
15	Be-in-state	0.72521	0.728807
16	Region-term	0.72866	0.728670
17	Has-holo-madeof	0.72117	0.725337
18	has_holo_member	0.73118	0.746392

Table 7. Improvement ratio of the proposed features

		Improvement ratio of LoPS		Improvement ratio of LoPW	
Index	Relation name	Over BoW (%)	Over BoC (%)	Over BoW (%)	Over BoC (%)
1	Related-to	12.0	6.2	13.1	7.40
2	Has- hyponym	9.8	3.8	10.2	4.30
3	Has-holo-part	9.4	3.4	10.2	4.30
4	Verb-group	7.2	1.1	7.9	1.90
5	Has-subevent	6.2	0.0	6.8	0.60
6	See-also	6.2	0.0	9.5	3.70
7	Category-term	11.5	5.7	11.5	5.70
8	Has-instance	7.7	1.6	9.8	3.80
9	Near-synonym	6.8	0.6	6.9	0.80
10	Has-derived	6.3	0.1	6.9	0.80
11	Causes	5.4	-0.8	7.1	0.95
12	Be-in-state	6.2	0.0	6.5	0.40
13	Region-term	6.5	0.4	0.1	0.40
14	Has-holo-madeof	5.6	-0.5	0.5	0.04
15	has_holo_member	6.9	0.8	2.0	2.80

Table 8. Improvement ratio of LoPW over LoPS

Index	Relation name	Improvement ratio of LoPW over LoPS (%)
1	Related-to	1.2
2	Has- hyponym	0.5
3	Has-holo-part	0.9
4	Verb-group	0.8
5	Has-subevent	0.6
6	See-also	3.5
7	Category-term	0.1
8	Has-instance	2.2
9	Near-synonym	0.1
10	Has-derived	0.6
11	Causes	0.7
12	Be-in-state	0.4
13	Region-term	0.1
14	Has-holo-madeof	0.5
15	has_holo_member	2.0

Table 9. Results of using list of synsets versus list of pertinent synsets in represent the original term

Original term	Relation	Synsets (BoC)	Pertinent synsets LoPS
قص Kasa	has_hyponym	قطع, قص, بتر Kata'a, Kasa, Batara	فرق, شق, أسقط, نشر, نقش, ثلم, ثقب , خدد, تشويه, عملية جراحية Faraka, Shaqa, Askata, Nashara, Naqasha, Thalama, Thaqaba, Khadada, Tashweeh, AmaliyaJerahiya
ولادة Welada	has_derived	ولادة, انجاب Welada, Enjab	مخاضي , ولادي Makhathi, Weladi
سبب Sabab	Causes	أنتج, أدى إلى, سبب, أقنع, أداة, مسبب, وسيلة Antaja, Ada, Ela, Sabab, Akna'a, Adat, Mosabeb, Waseela	حدث Hadatha

Table 10. Examples of (LoPW) outperforming (LoPS) in representation the original terms

Original term	Relation	Pertinent synsets LoPS	Pertinent words LoPW
الصين Alseen	has_holo_part	شانغهاي, بكين Shangahi, Bakeen	شانغهاي ,بكين ,عاصمة الصين الشعبية , الصين Shangahi, Bakeen, AsematAlseenAlsha'abiya, Alseen
سبب Sababa	Causes	حدثHadatha	حصل, جرى, جد, وقع, حدث, دار , سبب Hasala, Jara, Jada, Waqa'a, Hadatha, Dara, Sababa
قدم Kidam	see_also	أخذ ,أعطى ,أعاد A'da, A'ata, Akhatha	جلب, أخذ, أحضر, سلم, أعطى, سلم باليد, أعاد دفع , رد, أعاد, أعاد المال Jalaba, Akhatha, Ahthara, Salama, A'ata, Salamabelyad, A'daDafa'a, Rada, A'da, A'daAlmal

LoPS and LoPW are considered as extended (more inclusive) version of BoC (BoC uses only synsets, while LoPS and LoPW uses synsets with the 15 relations). Therefore, LoPW and LoPS outperform BoC (Table 9 and 10).

Conclusion and Future Work

In this study, two novel features based on lexical, semantic and lexico-semantic relations of Arabic WordNet (AWN) ontology are used with Naïve Bayes classifier to classify Arabic documents. The first feature is List of Pertinent Synsets (LoPS), which is the list of concepts (synsets) that have relations with documents' original terms. The second proposed feature is List of Pertinent Words (LoPW), which is the list of words that have relations with original documents' terms. In this study, 15 different relations of AWN are used with each of the two proposed features. The experimental results indicate that the introduction of adapted semantic features enhances the ATC. It was found that using LoPS improves the accuracy ATC over statistical methods. The improvement is about 12% over BoW and 6.2% over BoC. The results of using LoPW feature increase the classification accuracy up to 13.1% over BoW and up to 7.4% over BoC. According to the obtained results, we recommend using the Naïve Bayes classifier with LoPW (especially *Related-to* relation) to improve Arabic text classification accuracy.

This research lends itself to further work to improve ATC. One opportunity for further research is to use a stemming algorithm to find roots of original documents' terms instead of using terms in their orthographic form. This would improve classification results. This research could also be expanded by analyzing the effect when merging two or more AWN relations. Additionally, using term frequency-inverse document frequency (tf-idf) for text representation is one of the important work need to be done. Using tf-idf could improve TC accuracy since this metric will ignore terms that appear frequently in several categories (i.e., ignores general terms that are not specific for certain class). Actually, we are studying the use of concept-inverse document frequency (cf-idf) instead of tf-idf since our interest is in concepts not terms.

Funding Information

The authors have no support or funding to report.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abouenour, L., K. Bouzoubaa and P. Rosso, 2008. Improving Q/A using Arabic Wordnet. Proceedings of the Arab Conference on Information Technology, (CIT'08), IBTIKARAT Research Group, Tunisia.

Abouenour, L., K. Bouzoubaa and P. Rosso, 2010. Using the Yago ontology as a resource for the enrichment of Named Entities in Arabic WordNet. Proceedings of the 7th International Conference on Language Resources and Evaluation Workshop on LR & HLT for Semitic Languages, May 17-23, Text-Mess Repository, Malta.

Alahmadi, A., A. Joorabchi and A.E. Mahdi, 2014. Combining Bag-of-Words and Bag-of-Concepts representations for Arabic text classification. Proceedings of the 25th IET Irish Signals and Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communications Technologies, Jun. 26-27, IEEE Xplore Press, Limerick, pp: 343-348. DOI: 10.1049/cp.2014.0711

Al-Kabi, M.N. and S.I. Al-Sinjilawi, 2007. A comparative study of the efficiency of different measures to classify Arabic text. Univ. Sharjah J. Pure Applied Sci., 4: 13-26.

Alkhalifa, M. and H. Rodríguez, 2009. Automatically extending NE coverage of Arabic WordNet using Wikipedia. Proceedings of the 3rd International Conference on Arabic Language Processing, May 4-5, Rabat, Morocco, pp: 23-30.

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

Amine, A., Z. Elberrichi and M. Simonet, 2010. Evaluation of text clustering methods using WordNet. Int. Arab J. Inform. Technolo., 7: 349-357.

Bawaneh, M.J., M.S. Alkoffash and A.I. Al Rabea, 2008. Arabic text classification using K-NN and Naive Bayes. J. Comp. Sci., 4: 600-605. DOI: 10.3844/jcssp.2008.600.605

Black, W.J., S. Elkateb, H. Rodriguez, P. Vossen and A. Pease *et al*., 2006. Introducing the Arabic WordNet project. Proceedings of the 3rd Global WordNet Conference, (GWC'06), Jeju Island, Korea.

Boudabous, M.M., N.C. Kammoun, N. Khedher, L.H. Belguith and F. Sadat, 2013. Arabic WordNet semantic relations enrichment through morpho-lexical patterns. Proceedings of the 1st International Conference on Communications, Signal Processing and their Applications, Feb. 12-14, IEEE Xplore Press, Sharjah, pp: 1-6. DOI: 10.1109/ICCSPA.2013.6487245

Dai, W., G.R. Xue, Q. Yang and Y. Yu, 2007. Transferring naive bayes classifiers for text classification. Proceedings of the 22nd National Conference on Artificial Intelligence, (CAI '07), AAAI Press.

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

Duda, R.O. and P.E. Hart, 1973. Pattern Classification and Scene Analysis. Ist Edn., John Wiley and Sons, New York, ISBN-10: 0471223611, pp. 512.

Duwairi, R., 2007. Arabic Text Categorization. Int. Arab J. Inform. Technol., 4: 125-131.

Duwairi, R., M.N. Al-Refai and N. Khasawneh, 2009. Feature reduction techniques for Arabic text categorization. J. Am. Society Inform. Sci. Technol., 60: 2347-2352. DOI: 10.1002/asi.21173

El Kourdi, M., A. Bensaid and T.E. Rachidi, 2004. Automatic Arabic document categorization based on the Naïve Bayes algorithm. Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, (ASL'04), Stroudsburg, PA, pp: 51-58.

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

Elberrichi, Z., A. Rahmoun and M.A. Bentaalah, 2008. Using WordNet for Text Categorization. Int. Arab J. Inform. Technol., 5: 16-24.

El-Halees, A., 2008. A comparative study on Arabic text classification. Egypt. Comput. Sci. J., 20: 57-64.

Elkateb, S., W. Black, H. Rodríguez, M. Alkhalifa and P. Vossen *et al*., 2006. Building a WordNet for Arabic. Proceedings of the 5th International Conference on Language Resources and Evaluation, (LRE' 06).

Fodil, L., H. Sayoud and S. Ouamour, 2014. Theme classification of Arabic text: A statistical approach. Proceedings of the Terminology and Knowledge Engineering, (TKE' 14).

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. J. Machine Learning Res., 3: 1289-1305.

Genkin, A., D.D. Lewis and D. Madigan, 2007. Large-scale Bayesian logistic regression for text categorization. Technometrics, 49: 291-304. DOI: 10.1198/004017007000000245

Harrag, F. and E. El-Qawasmah, 2009. Neural network for Arabic text classification. Proceedings of the 2nd International Conference on the Applications of Digital Information and Web Technologies, Aug. 4-6, IEEE Xplore Press, London, pp: 778-783. DOI: 10.1109/ICADIWT.2009.5273841

Kanaan, G., R. Al-Shalabi, S. Ghwanmeh and H. Al-Ma'adeed, 2009. A comparison of text-classification techniques applied to Arabic text. J. Am. Society Inform. Sci. Technol., 60: 1836-1844. DOI: 10.1002/asi.20832

Khoja, S. and R. Garside, 1999. Stemming Arabic text. Lancaster University.

Khorsheed, M.S. and A.O. Al-Thubaity, 2013. Comparative evaluation of text classification techniques using a large diverse Arabic dataset. Language Resources Evaluation, 47: 513-538. DOI: 10.1007/s10579-013-9221-8

Khreisat, L., 2006. Arabic text classification using N-gram frequency statistics a comparative study. Proceedings of the Conference on Data Mining, (CDM' 06).

Lodhi, H., C. Saunders, J. Shawe-Taylor, N. Cristianini and C. Watkins, 2002. Text classification using string kernels. J. Machine Learning Res., 2: 419-444.

Mansuy, T. and R.J. Hilderman, 2006. Evaluating WordNet features in text classification models. University of Regina.

Mesleh, A.M.A., 2007. Chi square feature extraction based SVMs Arabic language text categorization system. J. Comp. Sci., 3: 430-443. DOI: 10.3844/jcssp.2007.430.435

Mullen, T. and N. Collier, 2004. Sentiment analysis using support vector machines with diverse information sources. Proceedings of the Conference on Empirical Methods in Natural Language, (MNL' 04), CLAIR Group at the University of Michigan.

Saad, M.K. and W. Ashour, 2010. OSAC: Open Source Arabic Corpora. Proceedings of the 6th International Conference on Electrical and Computer Systems, Nov. 25-26, Lefke, North Cyprus.

Sawaf, H., J. Zaplo and H. Ney, 2001. Statistical classification methods for Arabic news articles. Proceedings of the Natural Language Processing in ACL, (ACL' 01), Toulouse, France, pp: 547-553.

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

Thabtah, F., M. Eljinini, M. Zamzeer and W. Hadi, 2009. Naïve Bayesian based on Chi Square to categorize Arabic data. Proceedings of the 11th International Business Information Management Association Conference, (MAC' 09), Cairo, Egypt.

GWA, 2014. The global wordnet association.

Torunoglu, D., E. Cakirman, M.C. Ganiz, S. Akyokus and M.Z. Gurbuz, 2011. Analysis of preprocessing methods on classification of Turkish texts. Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, Jun. 15-18, IEEE Xplore Press, Istanbul, pp: 112-117. DOI: 10.1109/INISTA.2011.5946084

Zhang, L. and T.S. Yao, 2003. Filtering junk mail with a maximum entropy model. Proceedings of the 20th International Conference on Computer Processing of Oriental Languages (POL'03), pp: 446-453, Shen Yang, China.