Enhancement of Arabic Text Classification Using Semantic Relations of Arabic WordNet

Suhad A. Yousif, Venus W. Samawi, Islam Elkabani and Rached Zantout

Department of Mathematics and Computer Science, Faculty of Science, Beirut Arab University, Lebanon
Department of Computer Information Systems, Amman Arab University, Jordan
Department of Electrical and Computer Engineering, Rafik Hariri University, Lebanon

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 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; Alkhalfia and Rodriguez, 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 a replacement 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](image-url)
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; Tornogulu et al., 2011) are:

Text encoding: Avoid any distortion of characters during the text reading process. During this text reading process, in 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 الي”, “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 “إ”
- Replace all forms of Alef “أ،ة،ة” with “ا”
- Replace all Alif-maksura “ى” with Ya“ي”
- Replace all Ha’“ه” with Ta’marbota “ت”

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):

$$\text{TF}(t,d) = \# \text{ occurrence of term } t \text{ in document } d \ldots$$  \hspace{1cm} (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 (CFt+c) and Bag of Concepts CFB:\text{BoC}. 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, CFt+c, and CFB:\text{BoC} respectively:

$$\text{CFc}(t,d) = \frac{\# \text{ occurrence of concept}(t,l) \text{ in } d}{\# \text{ of concepts in } d} \ldots$$ \hspace{1cm} (2)

$$\text{CFt+c}(t,d) = \frac{\# \text{ occurrence of } t + \# \text{ occurrence of Concept}(t,l)}{\# \text{ words in } d + \# \text{ Added concepts}} \ldots$$ \hspace{1cm} (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 (CFt+c)).

$$\text{CF}_{\text{BoC}}(t,d) = \# \text{Synsets of } t \text{ in document } d \ldots$$ \hspace{1cm} (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)
**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 data 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; Khorshced and Al-Thubaity, 2013; Thabitah 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; Khorshced 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):

$$P(c_j|d_i) = \frac{P(c_j)P(d_i|c_j)}{P(d_i)}$$  \hspace{1cm} (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:

\begin{align*}
\text{Precision}(P) &= \frac{\text{number of Correctly Classified documents}(TP)}{\text{The total number of Predicted documents}(TP + FP)} \quad (6) \\
\text{Recall}(R) &= \frac{\text{The number of Correctly Classified documents}(TP)}{\text{The total number of documents that actually belong to the class}(TP + FN)} \quad (7)
\end{align*}

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:

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

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).
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 respectively:

\[
\text{LoPS}(t) = \# \text{Synsets that related to term } t …
\]

\[
\text{LoPW}(t) = \# \text{Words that relate to Synsets that relate to } t …
\]

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: \\
- LoPS = “نذكر. استنتاج.”
- LoPW = “ب强有力 إلى نذكر. استنتاج. فكر. استنتاج. استنتاج. استنتاج. استنتاج. استنتاج. استنتاج.”

Assume that the TF of term (اكمل) in certain document is 3. When LoPS is used, in this case, the frequency of the 3 concepts (istiantagea, tatha’akarah, istanbata) (استنتج. تذكر. استنتج.) is added to the term ‘اكمل’. In this case, the TF in the document becomes 7. LoPW contains 7 pertinent concepts (using the relation “related-to” with the term ‘اكمل’). The 7 pertinent concepts and the term (اكمل) appeared 12 times in the document (i.e., the TF of the term ‘اكمل’ 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

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

Table 7. Improvement ratio of the proposed features

<table>
<thead>
<tr>
<th>Index</th>
<th>Relation name</th>
<th>Improvement ratio of LoPS</th>
<th>Improvement ratio of LoPW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Over BoW (%)</td>
<td>Over BoC (%)</td>
</tr>
<tr>
<td>1</td>
<td>Related-to</td>
<td>12.0</td>
<td>6.2</td>
</tr>
<tr>
<td>2</td>
<td>Has-hyponym</td>
<td>9.8</td>
<td>3.8</td>
</tr>
<tr>
<td>3</td>
<td>Has-holo-part</td>
<td>9.4</td>
<td>3.4</td>
</tr>
<tr>
<td>4</td>
<td>Verb-group</td>
<td>7.2</td>
<td>1.1</td>
</tr>
<tr>
<td>5</td>
<td>Has-subevent</td>
<td>6.2</td>
<td>0.0</td>
</tr>
<tr>
<td>6</td>
<td>See-also</td>
<td>6.2</td>
<td>0.0</td>
</tr>
<tr>
<td>7</td>
<td>Category-term</td>
<td>11.5</td>
<td>5.7</td>
</tr>
<tr>
<td>8</td>
<td>Has-instance</td>
<td>7.7</td>
<td>1.6</td>
</tr>
<tr>
<td>9</td>
<td>Near-synonym</td>
<td>6.8</td>
<td>0.6</td>
</tr>
<tr>
<td>10</td>
<td>Has-derived</td>
<td>6.3</td>
<td>0.1</td>
</tr>
<tr>
<td>11</td>
<td>Causes</td>
<td>5.4</td>
<td>-0.8</td>
</tr>
<tr>
<td>12</td>
<td>Be-in-state</td>
<td>6.2</td>
<td>0.0</td>
</tr>
<tr>
<td>13</td>
<td>Region-term</td>
<td>6.5</td>
<td>0.4</td>
</tr>
<tr>
<td>14</td>
<td>Has-holo-madeof</td>
<td>5.6</td>
<td>-0.5</td>
</tr>
<tr>
<td>15</td>
<td>has_holo_member</td>
<td>6.9</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Table 8. Improvement ratio of LoPW over LoPS

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


DOI: 10.3844/jcssp.2007.430.435


