Combining Words and Concepts for Automatic Arabic Text Classification

Alaa Alahmadi, Arash Joorabchi, and Abdulhussain E. Mahdi

Electronic and Computer Engineering Department, University of Limerick, Limerick, Ireland
{alaa.alahmadi, arash.joorabchi, hussain.mahdi}@ul.ie

Abstract. The paper examines combining words and concepts for text representation for Arabic Automatic Text Classification (ATC) and its impact on the accuracy of the classification, when used with various stemming methods and classifiers. An experimental Arabic ATC system was developed and the effects of its main components on the classification accuracy are assessed. Firstly, variants of the standard Bag-of-Words model with different stemming methods are examined and compared. Arabic Wikipedia and WordNet were examined and compared for providing concepts for effective Bag-of-Concepts representation. Based on this, Wikipedia was then utilized to provide concepts, and different strategies for combining words and concepts, including two new in-house developed approaches, were examined for effective Arabic text representation in terms of their impact on the overall classification accuracy. Our experimental results show that text representation is a key element in the performance of Arabic ATC, and combining words and concepts to represent Arabic text enhances the classification accuracy as compared to using words or concepts alone.

Keywords: Arabic Text Classification, Text Representation Models, Bag Of Words, Bag Of Concepts; Wikipedia, WordNet.

1 Introduction

Automatic Text Classification (ATC) is an essential process for efficient organization of digital text. With the rapid growth of Arabic digital text, ATC has become one of the important tasks in Arabic text mining. The goal of ATC is to assign one or more predefined categories to a given textual document. The process involves three main components: text pre-processing, text representation and the classifier which is built using a generic Machine Learning (ML) algorithm. The classification begins by pre-processing the textual content of all the documents in the dataset in order to extract a set of well-defined features. These features are then passed to the text representation component where each document is represented as a set of features in a Vector Space Model (VSM) [1]. A document is often seen as a set of feature points in space of features, and different representation strategies are used to place these features in a VSM model. Finally, the VSM is fed to an ML based classification algorithm.

At its basic level, a text representation model expresses a piece of text or document in a compact representation of its textual content. Text representation models are commonly built using words as features, where text in a document is represented by the words it is composed of and the document is classified to a category based on the proportion of words that it has in common with other documents from the same category. In this case, the resulting representation is known as Bag-of-Words (BOW) model [2]. In addition to its simplicity, the BOW has proven its effectiveness particularly in English text classification [3], as well as many other languages. However, despite its efficiency, the BOW model has a number of limitations:

- The BOW model treats synonymous words as independent features. For example, "classification" and "categorization" are considered as two independent words with no semantic association. As a result, documents that discuss similar topics and contain synonymous words could be considered as unrelated.
- Words can have different meanings depending on their surrounding context, i.e., polysemy. The BOW model cannot recognize the meaning of a single word in different contexts even if the meaning is totally different. Take the word "bright" for example, depending on its context "bright" could mean shining or clever. Using the BOW as a representation model the word "bright" would be treated as a single feature irrespective of its intended meaning in different contexts.
- The BOW model representation breaks terms into their constituent words, e.g., it breaks "text classification" into the words "text" and "classification". As a result, the order of the words is lost and the unique meanings of the terms disappear. In addition, the BOW model tends to destruct the semantic relations between words and terms as it treats them as stand-alone units with. The semantic relation does not cover only synonymy and polysemy, it also reflects the relationship between words. For example, "text classification" is related to "text mining".

To address above limitations, a feature known as a concept has been introduced in text mining, giving raise to the Bag of Concepts (BOC) text representation model [4]. A concept is a unit of knowledge which provides a unique meaning. Synonymous words are mapped to the same concept which provides that unique meaning they share. Words with multiple meanings are mapped to different concepts based on the surrounding text. In order to use the BOC representation model in an ATC system, a knowledge base such as WordNet, Open Directory Project (ODP), or Wikipedia is needed to provide concepts. In recent years, concepts have been used to represent text for English ATC [5-11]. However, using concepts alone to represent text does not result in a significant classification improvement as confirmed by [4]. Therefore, a number of researchers have experimented with combining words and concepts to represent text based mainly on the following strategies:

- Adding Concepts (AC): this strategy involves forming a combined text representation model by adding concepts identified in the document to its BOW model as extra features using different knowledge bases. In particular, AC has been used for English ATC with Wikipedia [6, 9-11], WordNet [12], and ODP [5]. Furthermore, it has also been proposed for text clustering with WordNet [4] and Wikipedia [13].

- Replacing Terms with Concepts (RTC): this strategy is similar to AC, but words and terms in the document for which a concept has been identified are removed from the combined text representation. Only words which do not have corresponding concepts are added to the text representation model. The strategy was first proposed for English text clustering by Hotho et al. [14], who also showed that this strategy yielded less accurate clustering results compared to AC.
- Adding Concepts and Categories (ACC): in this strategy, which was proposed by Wang et al. [10, 11], the parent categories of the concepts are added to the combined representation model along with concepts and words.

Since its introduction, the AC strategy has been used by a number of researchers who demonstrated its improving impact on the performance of English ATC. For example, Gabrilovich et al. [5] used ODP as a knowledge base to provide concepts for text representation. They used a feature generating technique which searches for new features that describe the target document better than the ones contained in the training documents. The feature generator constructs new features from the ODP categories and adds them to the BOW model using AC strategy. Experimental results showed improved classification performance in comparison with the BOW model. However, the ODP has a number of drawbacks. Its categories are not equally covered, some categories are repeated in different branches of the categories hierarchy tree, and sometimes some are more influenced by the views of the editors in charge. Gabrilovich et al. [6] subsequently showed that using Wikipedia as a knowledge base instead of the ODP improved their classification results further.

Wang et al. [10] used Wikipedia synonyms, associated concepts and hyponyms (parent categories for the concept), by adding them to the BOW model. The study showed that adding synonyms to the BOW is not useful for ATC, whereas adding the top 5, 10, 15, 20, and 25 associated concepts improved the classification accuracy. In addition, the authors compared enriching the document representation with hyponyms of candidate concepts extracted from the first five levels of their relational hierarchy as provided by Wikipedia. Their results showed that adding hyponyms extracted from first three levels achieved better classification accuracy compared to adding hyponyms extracted from levels 1 to 5 of the hierarchy.

In this paper, we examine combining words and concepts for text representation for Arabic ATC and how this impacts the accuracy of the classification when used with various stemming methods and classifiers, compared to using words or concepts alone. To achieve this, an experimental Arabic ATC system has been developed and the effect of each main component on the classification accuracy is assessed. First, variations of the BOW model resulting from the application of different stemming methods at the pre-processing stage are examined and compared. Then two knowledge bases, namely Wikipedia and WordNet, are used to provide concepts to represent Arabic text using a BOC model. A comparison between these knowledge bases is conducted and the one yielding the best accuracy is used to provide concepts to build a combined text representation model using the AC, the RTC and the ACC strategies, as well as to develop two new combined model strategies. These combined models are then used in our Arabic ATC system and the classification accuracy

achieved by each is evaluated and compared to the use of the BOW or BOC alone. The paper is organised as follows: Section 2 reviews related work on Arabic ATC. Section 3 then describes the construction of our Arabic ATC system with a specific focus on Arabic text representation. The experimental setup and the datasets used in this work are described in Section 4. Section 5 presents and discusses our experimental results, and Section 7 concludes the work and highlights our main findings.

2 Arabic ATC - Related Work

Compared to English ATC, the field of Arabic ATC is underdeveloped. Text representation for Arabic ATC is therefore a relatively new field and, hence, limited work has been published in this field. To-date, most reported works focus on comparing different stemming methods, investigating the impact of pre-processing, applying different classification algorithms and evaluating their effects on the classification of Arabic text, as described below.

Researchers such as Harrag et al. [15] compared different stemming methods in ATC for Arabic text. The authors compared three stemming methods, namely Light Stemming (LS) [16], Root Extraction (RE) [16] and dictionary-lookup method [17]. The RE method works by removing the suffixes and prefixes attached to a given word and word pattern matching is used to extract the root of the word. The LS method only involves removing a small set of prefixes and suffixes. The LS does not deal with infixes or recognize patterns to find roots. The dictionary-lookup method uses dataset statistics to generate possible roots for a given word and estimates the probability of deriving the word from each of the possible roots. Harrag et al. [15] used the stemming methods to reduce the feature space of the VSM for two different classifiers, the Artificial Neural Networks (ANN) and the Support Vector Machine (SVM). To evaluate their ATC system's performance, an in-house Arabic dataset containing 453 documents distributed over 14 categories was used. Reported results showed that ANN yielded a better performance than the SVM. Furthermore, the dictionary-lookup stemming method performed better with ANN whereas the LS method performed better with an SVM classifier.

Al-Shammari et al. [18] proposed the local stemming method and compared it with the LS [19] and RE. This method selects the shortest form of a word among syntactically related words in a text. To evaluate the classification performance with different stemming methods, al-Shammari used a dataset that was constructed by merging the Saudi News Papers (SNP) dataset and the Saudi Press Agency (SPA) dataset as collected by Al-Harbi at al. [20]. Only 2,966 documents were selected which belonged to three categories: "cultural", "social", and "general". In the classification experiments, a 10-fold cross-validation was used with the Naive Bayes (NB), SVM and k-Nearest Neighbours (k-NN) ML algorithms. The experiment results showed that the local stemming method significantly improved text classification accuracy, in comparison to other stemming methods, and worked better with the SVM classifier.

Other researchers compared the accuracy of using different classification algorithms in Arabic ATC systems. For example, Mesleh et al. [21] used the original

words without using any stemming methods to build a BOW model. He used the Chi-squared (χ^2) as the Feature Selection (FS) technique to reduce the size of the feature space. The dataset used was collected from online Arabic newspaper archives and contained 1,445 documents that vary in length and fall into nine categories. The results showed that using an SVM classifier yielded a better classification performance compared to using the k-NN and NB classifiers. It yielded a macro-average F1 score of 88.11% when evaluated using the in-house complied Arabic dataset. Mesleh's dataset was also used by Kanaan et al. [22]. They applied the LS stemming method proposed by [23] to build a BOW model for Arabic ATC. A comparison was performed between the k-NN, Rocchio, and NB as classifiers with different weighting methods such as the Term Frequency (TF), the Term Frequency Inverse Document Frequency (TFIDF) and the Weighted Inverse Document Frequency (WIDF). Their results showed that the WIDF scheme yielded the best performance when used in conjunction with the k-NN, while TFIDF yielded the best performance when used in conjunction with Rocchio. Among the above three classifiers, the NB classifier was reported to be the best, yielding a macro-averaged F1 score of 84.53%.

Al-Harbi et al. [20] compared the SVM and C5.0 classification algorithms in Arabic ATC. The original text was used without any stemming to build the BOW model, and the Chi-squared was used as the FS technique to reduce the size of the feature space. The C5.0 provided a better performance than the SVM. On the other hand, Alsaleem et al. [24] compared the SVM and NB classification algorithms using the BOW model to classify the SNP dataset collected by [20], and his experiment results showed that the SVM algorithm outperformed the NB algorithm.

In terms of text representation models, most reported works in Arabic ATC have used the BOW model. For example, Khreisat et al. [25] used an N-gram frequency statistical technique to compare two similarity measures, the Manhattan distance and the Dice's coefficient. The authors used a dataset collected from four online Jordanian Arabic newspaper archives. Tri-grams were used to represent each document after pre-processing by removing punctuation marks, diacritics, non-letters and stop words. The Khoja RE stemming method [26] was applied to the remaining words. The chosen similarity measures were used with 40% of the dataset utilised for training and the rest for testing. Reported results showed that the best accuracy was obtained using the Dice coefficient in conjunction with the tri-gram frequency method.

Others have used statistical phrases to represent Arabic text for ATC. For example, Al-Shalabi et al. [27] compared the use of the BOW and phrases text representation models. The k-NN algorithm was used to classify documents from the dataset created by [21]. All Arabic documents were pre-processed by removing stop words, non-letters, and punctuation marks. Two independent text representation models were then built, namely a BOW model and a bag of phrases model with each phrase composed of two words. To train the classifier, 60% of the dataset was used for training and the rest for testing. The results showed that using phrases for text representation in conjunction with the k-NN classifier outperformed the BOW model and yielded an average accuracy score of 73.57% as compared to a score of 66.88% for the BOW model.

Elberrichi et al. [28] used the Arabic WordNet (AWN) to identify concepts appearing within the documents. A comparison between the use of different text representa-

tion models, utilizing words, N-grams and words and concepts combined using the RTC strategy, was conducted on the Arabic text dataset collected by [21]. The combined model, used in conjunction with the Chi-squared as the FS technique and a *k*-NN classifier, was reported to achieve higher performance results compared to other representations. Yousif et al. [29] developed two new representation models based on lexical and semantic relations extracted from the Arabic WordNet, namely the List of Pertinent Synsets (LoPS) and the List of Pertinent Words (LoPW). The LoPS is the list of concepts (synsets) that have relations with the documents' original terms, whereas LoPW is the list of words that have relations with the documents' original terms. The authors compared the developed representation models with the standard BOW and BOC representation models. In the classification experiments, a 10-fold cross-validation was used with an NB classifier to classify documents from the Arabic BBC dataset [30]. The experiment results showed that both developed representation models improved the accuracy of ATC as compared to the standard BOW and BOC models. In addition, it was found that the LoPW model outperforms the LoPS model.

3 Text Representation for Arabic ATC

In order to investigate and assess the effect of commonly used text representation models on the classification accuracy of Arabic ATC, an experimental ATC system has been built. As described in Section 1, the first component of this ATC system performs the pre-processing of the text, where the text is converted to a well-defined set of features. To achieve this, first the text is tokenised by breaking it up into individual and meaningful units known as tokens. Each token is separated from others by a particular character or symbol. In written Arabic, words are separated by a space and each word is considered as a single meaningful unit. Hence, the tokenisation involves keeping these words and removing all other remaining punctuation marks, digits and numbers as they are considered noise. Then, common words such as pronouns, prepositions and articles are removed from the text. These words are called "stop words" and occur frequently and, therefore, have no discriminatory significance. This is then followed by replacing all words with their roots or stemmed forms, where morphological information is used to merge various word forms, such as plurals and verb conjugations, into their distinct roots. In this work, we have focused on the Root Extraction (RE) and the Light Stemming (LS) as the two most commonly used stemming methods for Arabic text as indicated in our Related Work section.

Next, all rare words are identified and removed based on their frequency of appearance in the whole dataset, using a threshold of four that has been chosen by experimentation. This process also reduces the complexity of the text representation model and improves the training time of the classifier. Next, the remaining words are passed to the text representation component of the ATC system as valid features, which make the dimensions of the resulting VSM. Depending on the text representation strategy used, different types of features are employed to represent text. For example, in the case of the BOW model, the features are simply the remaining words as per above. Variations of the BOW model are built based on the stemming method of

the text pre-processing component of the system. Hence, three BOW models have been investigated; a BOW-RE, BOW-LS and a BOW-Original where no stemming method is applied.

In the case of the BOC model, different types of BOC are built depending on the knowledge base used to provide the concepts. In this work both Wikipedia and WordNet are employed as knowledge bases and the resulting BOC models are compared to find the best knowledge base. To identify concepts using Arabic WordNet, an open-source toolkit called Arabic WordNet (AWN) is used. For each word in a text, the AWN browser searches its database and returns an ordered list of synonyms and the first synonym in the list (i.e., the most commonly used sense) is used as the concept for the word. To build the Wikipedia-based BOC representation model, the Arabic Wikipedia XML dump files (http://dumps.wikimedia.org/arwiki/), consisting of 273,709 articles, is used in this work. An open source toolkit known as Wikipedia-Miner [31] has been used to process the dump files and create a database that contains a summarized version of the Arabic Wikipedia's content and structure.

When both words and concepts are used together as features, we get a combined text representation model which can be built using different strategies, such as the AC, the RTC or the ACC as described in Section 1. In our ATC system, we have applied and compared different strategies to build combined text representation models. These include the AC, RTC and AC, as well as two in-house developed strategies, which are our attempt at developing new combined text representation models with a relatively reduced vector size as described in the following section.

3.1 New Combined Text Representation Strategies

In this section, we describe two in-house developed strategies for building a combined text representation model, namely Adding Unmapped Concepts (AUC), and using Concepts for Terms which do not appear in the Document (CTD).

Adding Unmapped Concepts (AUC). This strategy first involves creating a "Concept-Words Map" for the whole dataset. To achieve that, we map each concept identified in the dataset to the corresponding word(s) that share the unique meaning provided by the concept. By doing this, the algorithm will resolve synonyms and capture different words which refer to the same concept in the documents of the training subset. These words are considered as alternative labels for the concept in the "Concept-Words Map". The concept's label provided by the Knowledge Base (KB) is considered as the preferred label. In order to build a combined model to represent a given document using the AUC strategy, the following tasks have to be performed. Firstly, a BOW vector has to be created for the document and all words that are considered as features in the BOW model are added to the AUC model's vector. In this way, words which have a significant frequency value in the BOW model and carry an important value for the classification task are kept in the AUC model. Then, for each concept in the BOC model of the document, the "Concept-Words Map" is checked for alternative labels of the concept. If one of the alternative labels appears in the text, the pre-

ferred label for the concept is added to the AUC model. Otherwise, nothing is added regarding that concept and we move to the next one. The difference between the AUC strategy and AC strategy is that only concepts mapped to different alternative labels are added to the combined representation model.

Using Concepts for Terms not appearing in the Document (CTD). This strategy first involves creating a "Concept-Words Map" for the whole dataset. To achieve that, we map each concept. All previously proposed combined model strategies use both concepts and words with their corresponding weights. As a result, the numbers of features in resulting model are larger than the number of features in a corresponding BOW or BOC model. The CTD strategy does not change the original dimensions of the BOW model. It is similar to the BOW model in two ways; the size of the VSM and the type of features that are used to represent a document, where each dimension in the VSM corresponds to a word from the BOW dictionary. The strategy works like the BOW model for words that appear in the document, where their corresponding dimensions in the VSM will have the value of the words' TFIDF weights. The only difference is related to those words that do not appear in the document but have a corresponding concept in the document's BOC model. For these words, the weight of their related concepts in the document will be used as corresponding dimensions in the VSM. Hence, the CTD works as follows. First, a "Word-Concepts Map" is created for the whole dataset. To achieve this, each word that has been identified as a concept(s) is mapped to its corresponding concept(s). By doing this, the algorithm resolves synonyms and capture different words which refer to the same concept in all the documents of the training subset. In addition, words that have multiple meaning are mapped to different concepts. To represent a document using the CTD strategy and build the combined model, first all words which are considered as features in the BOW model dictionary are checked for their appearance in the document. In the case that a given word appears in the document, the word's TFIDF weight in the document is used in the CTD model. If the word does not appear in the document, the "Word-Concepts Map" is checked to see if that word has a corresponding concept. If it does, the concept is retrieved from the "Word-Concepts Map" and checked for its existence in the document's BOC model. If the concept exists, the word's corresponding dimension in the CTD model is assigned to the weight of the concept. If the corresponding concept does not exist, the word will not be represented in the CTD combined model.

4 Datasets and Experimental Setup

In order to provide a baseline for an objective assessment and comparison of the performance of our Arabic ATC system and proposed text representations, we have used three of the most frequently employed datasets in all cited similar work. The documents in these datasets, which have been collected from on-line web sites, are all written in Modern Standard Arabic (MSA). The datasets used are:

- Arabic 1445 dataset: this dataset was created by Mesleh at al. [21] from online Arabic newspaper archives containing news articles from Al-Jazeera, Al-Nahar, Al-hayat, Al-Ahram, and Al-Dostor. The dataset contains 1,445 documents that vary in length and fall into nine categories: computer, economics, education, engineering, law, medicine, politics, religion, and sports.
- Saudi News Papers (SNP) dataset: this dataset consists of 5,121 Arabic documents of different length which belong to seven categories: culture, economics, general, information technology, politics, social, and sport. It has been collected by Al-Harbi et al. [20] and consists of articles and news stories from Saudi newspapers.
- Al-khaleej dataset: this dataset is a collection of 5,690 Arabic news documents gathered from the archive of the online newspaper Al-khaleej by Abbas et al. [32]. It consists of four categories: international news, local news, sport, and economy.

We conducted all the experiments using WEKA [33], which is a popular open source toolkit for ML. We first converted the textual documents into the format required by WEKA, i.e., ARFF format (Attribute-Relation File Format). We then used the data to train four separate classification algorithms, namely Support Vector Machines (SVM), Naive Bayes (NB), Decision Trees (DT), and Random Forest (RF). This was then followed by a 10-fold cross validation to evaluate the performance of the classifiers using the standard information retrieval measures of Precision, Recall, and F1.

5 Results and Discussion

In this section, results of our various experiments are presented and assessed, in terms of achieved F1 classification performance of the ATC system, with regards to:

- The effect of the pre-processing component on the performance the BOW model.
- The role of the characteristics of the knowledge base used to build the BOC.
- How the different BOW-BOC combined text representation models compare.

In the first stage of our experiments and evaluation, we focused on the effect of stemming on the overall classification accuracy. We conducted a comparison between the performance of our ATC system when the BOW is used with no stemming (i.e., Original-BOW) to that when stemming is applied in the pre-processing component using the LS and RE methods. Accordingly, three different BOW model representations have been built, BOW-LS, BOW-RE and the Original-BOW, for each dataset. Each of these BOW representation models has been used with four classification algorithms: SVM, NB, DT and RF. Our results here show that, in terms of classification algorithms, the SVM achieves the highest classification accuracy, as can be seen in Tables 1-3. In addition, the SVM shows higher classification accuracy when used with the Original-BOW model for the cases of the large datasets as illustrated in Tables 2 and 3. This can be attributed to the SVM's ability to deal successfully with high-dimensional data [34].The DT algorithm seems to perform relatively well with the Original-BOW only for two datasets which have a large number of categories as illustrated in Tables 1 and 2. The BOW-LS representation seems to achieve the best

average classification accuracy, particularly when applied to the two largest datasets, i.e., the SNP and Al-khaleej, as illustrated in Tables 2 and 3. The BOW-RE representation model on the other hand shows a good average performance with the Arabic 1455 dataset as can be seen in Table 1.

Table 1. F_1 scores for the three versions of BOW model for the Arabic1445 dataset with different classifiers.

Classifier	Original-BOW	BOW- based LS	BOW-based RE
SVM	0.90	0.92	0.91
NB	0.82	0.85	**0.88**
DT	**0.80**	0.79	0.77
RF	0.75	0.74	0.75
Average	0.81	0.82	**0.83**

Table 2. F_1 scores for the three versions of BOW model for the SNP dataset with different classifiers.

Classifier	Original-BOW	BOW- based LS	BOW-based RE
SVM	**0.81**	0.80	0.77
NB	0.63	**0.67**	0.66
DT	0.67	0.66	0.60
RF	**0.60**	0.59	0.57
Average	**0.67**	**0.68**	0.65

Table 3. F_1 scores for the three versions of BOW model for the Al-khaleej dataset with different classifiers.

Classifier	Original-BOW	BOW- based LS	BOW-based RE
SVM	**0.96**	0.94	0.92
NB	0.80	0.82	0.84
DT	0.86	0.87	0.83
RF	0.83	0.83	0.81
Average	0.86	**0.87**	0.85

The second stage of our experiments and evaluation focused on investigating the use of concepts as representation features for Arabic ATC. Two knowledge bases, namely WordNet and Wikipedia, were used to build the BOC representations model for our ATC system. Tables 4-6 show achieved F1 scores when the ATC system uses a WordNet-based BOC and a Wikipedia-based BOC with different classifiers. All classifiers achieved higher accuracy when Wikipedia was used as a BOC knowledge base for all the datasets, as compared to WordNet. We believe this can be attributed to the fact that WordNet provides a BOC model with fewer concepts for representing Arabic text. Arabic WordNet has only 9,228 concepts compared to Arabic Wikipedia which

has 273,709 concepts. Another reason for this is the ambiguity of the text as the documents in all the datasets are written in MSA and contain no diacritics. Arabic WordNet returns a ranked list of possible concepts for a word in the text, and the first ranked concept is the most commonly used, whereas Wikipedia selects concepts based on the surrounding text. In addition, WordNet mostly provides information about individual words rather than general conceptual knowledge [35].

Table 4. F_1 scores for the BOC model for the Arabic1445 dataset using Wikipedia and WordNet with different classifiers.

Classifier	WordNet-based BOC	Wikipedia-based BOC
SVM	0.87	**0.89**
NB	0.83	**0.89**
DT	0.70	**0.79**
RF	0.67	**0.84**
Average	0.76	**0.85**

Table 5. F_1 scores for the BOC model for the SNP dataset using Wikipedia and WordNet with different classifiers.

Classifier	WordNet-based BOC	Wikipedia-based BOC
SVM	0.72	**0.75**
NB	0.57	**0.71**
DT	0.56	**0.65**
RF	0.51	**0.66**
Average	0.59	**0.69**

Table 6. F_1 scores for the BOC model for the Al-khaleej dataset using Wikipedia and WordNet with different classifiers.

Classifier	WordNet-based BOC	Wikipedia-based BOC
SVM	0.89	**0.92**
NB	0.79	**0.83**
DT	0.79	**0.85**
RF	0.77	**0.87**
Average	0.81	**0.87**

In the final stage of evaluation, we experimented with different strategies to build combined text representation models and compared corresponding resulting classification accuracy. We have evaluated the use of five different strategies, namely the AC, RTC, ACC, AUC and CTD. For each dataset, five combined representation models were built using Wikipedia concepts and words stemmed using the LS method. Our experimental results are presented in Tables 7-9 in terms of obtained F1 scores. Regarding the classification algorithms, our results show that the SVM yield highest

performance with all combined models. The NB algorithm comes next in terms of its accuracy, followed by the DT and the RF. The results also show that the ACC combined model achieves the best classification accuracy for two datasets as illustrated in Tables 7 and 8. The AC strategy seems to be the second best in terms of the classification accuracy for two datasets, again as per Tables 7 and 8. The CTD strategy, on the other hand, seems to yield the best performance when used in conjunction with the NB algorithm for the two largest datasets as illustrated in Tables 8 and 9.

Table 7. F_1 scores for the different combined models for the Arabic1445 dataset with different classifiers.

Classifier	AC	RTC	ACC	AUC	CTD
SVM	**0.920**	0.912	**0.918**	0.917	0.919
NB	**0.872**	0.848	**0.873**	0.866	0.859
DT	**0.825**	0.807	0.815	**0.827**	0.792
RF	0.782	0.753	**0.805**	**0.783**	0.766
Average	**0.850**	0.830	**0.852**	**0.848**	0.834

Table 8. F_1 scores for the different combined models for the SNP dataset with different classifiers.

Classifier	AC	RTC	ACC	AUC	CTD
SVM	0.809	0.793	**0.824**	0.805	**0.824**
NB	0.651	0.626	**0.687**	0.650	**0.694**
DT	0.660	0.647	**0.676**	0.656	**0.668**
RF	0.598	0.564	0.614	0.581	0.593
Average	0.679	0.657	**0.700**	0.673	**0.694**

Table 9. F_1 scores for the different combined models for the Al-khaleej dataset with different classifiers.

Classifier	AC	RTC	ACC	AUC	CTD
SVM	0.809	0.793	**0.824**	0.805	**0.824**
NB	0.651	0.626	**0.687**	0.650	**0.694**
DT	0.660	0.647	**0.676**	0.656	**0.668**
RF	0.598	0.564	0.614	0.581	0.593
Average	0.679	0.657	**0.700**	0.673	**0.694**

From Table 10 it can be concluded that using the LS stemming method in the preprocessing stage provides better classification performance than employing the RE method, which is in line with the findings of other researchers [15, 36, 37]. We believe this is due to the fact that the RE method is harsher on words in comparison to the LS method. Using RE, two words with different meanings could be stemmed to the same root, which leads to misclassification. For example, the words "العالميه" (global) and "العلميه" (scientific) would share the same root, (علم) "science" if the RE

stemming method is used. Furthermore, Table 10 shows that using Wikipedia concepts to build a BOC model for Arabic ATC yields better classification accuracy than using a BOW representation. One of the reasons for this, we believe, is the broad categories of the documents of the Arabic datasets used in this work; all datasets contained documents from different newspapers and did not focus on specific topics. The other reason is the complex nature of the Arabic language and the poor morphological tools available, which make Wikipedia concepts better features for representing text compared to words.

Table 10. Average F_1 scores for all text representation models and for all datasets.

Original-BOW	BOW-based LS	BOW-based RE	WordNet-based BOC	Wikipedia-based BOC	AC	RTC	ACC	AUC
0.786	0.791	0.776	0.723	**0.802**	**0.804**	0.786	**0.812**	0.800

Our results have also shown that all the combined models we experimented with achieved higher classification accuracy than the BOW representation model, with the best performance achieved by the ACC strategy followed by the AC and the CTD. On the other hand, all combined models outperformed the BOC model. Finally, the results in Tables 1-9 suggest that the RF algorithm provides relatively higher classification accuracy, compared to other classifiers, when used with the BOC representation model for all datasets.

6 Conclusion

In this work we have built an ATC system for Arabic text and evaluated its performance using three different datasets, with the goal of identifying key elements of text representation that influence the classification accuracy. The evaluation involved using a number of different text representation models in association with different machine learning techniques. While work reported in the literature has mainly concentrated on the BOW based representation models, our study focuses on comparing the classification performance of the BOW and BOC models with those of various combinations of both. Furthermore, two knowledge bases (i.e., Wikipedia & WordNet) were examined for building a BOC model, making our study the first of its kind to utilize Wikipedia as a knowledge base for Arabic ATC.

In conclusion, each component in an ATC system plays an important role in the classification accuracy, with text pre-processing and representation being key elements as demonstrated by our experimental results. We believe the findings of this study pave the way for venues for further research. Among those is the use of Wikipedia concepts to represent Arabic text and its application for providing richer representation models for automatic classification of more specialized textual datasets.

References

1. Salton, G., Wong, A., Yang, C.-S.: A vector space model for automatic indexing. Communications of the ACM 18, 613-620 (1975)
2. McCallum, A., Nigam, K.: A comparison of event models for naive bayes text classification. AAAI-98 workshop on learning for text categorization, vol. 752, pp. 41-48 (1998)
3. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) 34, 1-47 (2002)
4. Hotho, A., Staab, S., Stumme, G.: Wordnet improves Text Document Clustering. (2003)
5. Gabrilovich, E., Markovitch, S.: Feature generation for text categorization using world knowledge. IJCAI, vol. 5, pp. 1048-1053 (2005)
6. Gabrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. AAAI, vol. 6, pp. 1301-1306 (2006)
7. Kehagias, A., Petridis, V., Kaburlasos, V.G., Fragkou, P.: A comparison of word-and sense-based text categorization using several classification algorithms. Journal of Intelligent Information Systems 21, 227-247 (2003)
8. Rodríguez, M.d.B., Hidalgo, J.M.G., Agudo, B.D.: Using WordNet to complement training information in text categorization. arXiv preprint cmp-lg/9709007 (1997)
9. Scott, S., Matwin, S.: Text classification using WordNet hypernyms. Use of WordNet in natural language processing systems: Proceedings of the conference, pp. 38-44 (1998)
10. Wang, P., Hu, J., Zeng, H.-J., Chen, L., Chen, Z.: Improving Text Classification by Using Encyclopedia Knowledge. 332-341 (2007)
11. Wang, P., Hu, J., Zeng, H.-J., Chen, Z.: Using Wikipedia knowledge to improve text classification. Knowledge and Information Systems 19, 265-281 (2008)
12. Benkhalifa, M., Mouradi, A., Bouyakhf, H.: Integrating external knowledge to supplement training data in semi-supervised learning for text categorization. Information Retrieval 4, 91-113 (2001)
13. Hu, J., Fang, L., Cao, Y., Zeng, H.-J., Li, H., Yang, Q., Chen, Z.: Enhancing text clustering by leveraging Wikipedia semantics. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 179-186. ACM (2008)
14. Andreas Hotho, S.S., Gerd Stumme: Wordnet improves Text Document Clustering. 541-544 (2003)
15. Harrag, F., El-Qawasmah, E., Al-Salman, A.M.S.: Stemming as a feature reduction technique for Arabic Text Categorization. Programming and Systems (ISPS), 2011 10th International Symposium on, pp. 128-133. IEEE (2011)
16. Syiam, M.M., Fayed, Z.T., Habib, M.B.: An intelligent system for Arabic text categorization. International Journal of Intelligent Computing and Information Sciences 6, 1-19 (2006)
17. Darwish, K., Oard, D.W.: Adapting Morphology for Arabic Information Retrieval*. Arabic Computational Morphology, pp. 245-262. Springer (2007)
18. Al-Shammari, E.T.: Improving Arabic document categorization: Introducing local stem. Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on, pp. 385-390. IEEE (2010)

19.Larkey, L.S., Ballesteros, L., Connell, M.E.: Light stemming for Arabic information retrieval. Arabic computational morphology, pp. 221-243. Springer (2007)
20.Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M., Al-Rajeh, A.: Automatic Arabic text classification. (2008)
21.Mesleh, A.M.d.A.: Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System. Journal of Computer Science 3, (2007)
22.Kanaan, G., Al-Shalabi, R., Ghwanmeh, S., Al-Ma'adeed, H.: A comparison of text-classification techniques applied to Arabic text. Journal of the American society for information science and technology 60, 1836-1844 (2009)
23.Larkey, L.S., Ballesteros, L., Connell, M.E.: Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275-282. ACM (2002)
24.Alsaleem, S.: Automated Arabic Text Categorization Using SVM and NB. Int. Arab J. e-Technol. 2, 124-128 (2011)
25.Khreisat, L.: A machine learning approach for Arabic text classification using N-gram frequency statistics. Journal of Informetrics 3, 72-77 (2009)
26.Khoja, S., Garside, R.: Stemming arabic text. Lancaster, UK, Computing Department, Lancaster University (1999)
27.Al-Shalabi, R., Obeidat, R.: Improving KNN Arabic text classification with n-grams based document indexing. Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt, pp. 108-112. Citeseer (2008)
28.Elberrichi, Z., Abidi, K.: Arabic Text Categorization: a Comparative Study of Different Representation Modes. International Arab Journal of Information Technology (IAJIT) 9, (2012)
29.Yousif, S.A., Samawi, V.W., Elkabani, I., Zantout, R.: The Effect of Combining Different Semantic Relations on Arabic Text Classification.
30.Saad, M.K., Ashour, W.: Osac: Open source arabic corpora. 6th ArchEng Int. Symposiums, EEECS, vol. 10, (2010)
31.Milne, D., Witten, I.H.: An open-source toolkit for mining Wikipedia. Artificial Intelligence 194, 222-239 (2013)
32.Abbas, M., Smaili, K.: Comparison of topic identification methods for arabic language. Proceedings of International Conference on Recent Advances in Natural Language Processing, RANLP, pp. 14-17 (2005)
33.Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 11, 10-18 (2009)
34.Ben-Hur, A., Weston, J.: A user's guide to support vector machines. Data mining techniques for the life sciences, pp. 223-239. Springer (2010)
35.Gabrilovich, E., Markovitch, S.: Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research 443-498 (2009)
36.Duwairi, R., Al-Refai, M.N., Khasawneh, N.: Feature reduction techniques for Arabic text categorization. Journal of the American society for information science and technology 60, 2347-2352 (2009)
37.Saad, M.K.: The impact of text preprocessing and term weighting on Arabic text classification. The Islamic University-Gaza (2010)