International Journal of Mathematics and
Computer Science, **17**(2022), no. 2, 937–946

Arabic Text Classification based on Semantic Relations

Musab Hijazi[1,3], **Akram Zeki**[2], **Amelia Ismail**[3]

[1]Department of Computer Science and Software Engineering
College of Engineering
Al Ain University
Al Ain, UAE

[2]Department of Information Technology
Kulliyyah of Information and Communication Technology
International Islamic University Malaysia
Kuala Lumpur, Malaysia

[3]Department of Computer Science
Kulliyyah of Information and Communication Technology
International Islamic University Malaysia
Kuala Lumpur, Malaysia

email: musab.hijazi@aau.ac.ae, akramzeki@iium.edu.my,
amelia@iium.edu.my

(Received October 13, 2021, Accepted January 29, 2022)

Abstract

Documentation is the best method to illustrate knowledge implying that the substantial repositories of information are documents. Due to the rapid expansion of the internet, there is a massive growth in the number of electronic documents which require flexible and effective ways to access, arrange, and extract useful information, such as text classification and text clustering. Text classification is the process of grouping or categorizing documents into pre-defined groups or classes based on pre-defined criteria. Bag of Words (BoWs) representation is

Key words and phrases: Text Classification, Arabic Language, Feature Selection, Arabic WordNet, Semantic Relations, Conceptual Representation.
AMS (MOS) Subject Classifications: 68T20.
ISSN 1814-0432, 2022, http://ijmcs.future-in-tech.net

considered the simplest representation of text that is used to represent data in text classification. Many Arabic researchers have been trying to find an accurate Arabic text classification based on the traditional Bag of Words (BoWs) for data representation which does not consider the semantic relationships between the words, such as synonymy and hypernymy. This research aims to utilize Arabic WordNet (AWN) as a lexical and semantic resource to study the impact of the semantic relations between the words on Arabic text classification. Experiments showed that utilizing concepts and semantic relations between them enriches the text representation by adding semantic meaning which could improve the text classification performance.

1 Introduction

Documents contain a tremendous quantity of valuable human information. The use of automatic text classification is needed because of the significant increase in the number of machine-readable materials available for public or private use. Although most attention has been invested into Western languages, mostly English, the Arabic language has received less attention [1]-[5]. Concern in Arabic TC has been recently increased with the necessity for automatic Arabic TC systems throughout recent years due to many factors: First, Arabic language content on the internet exceeded more than 3of all online content (ranking eighth). This massive amount of data must be searched, exchanged, and retrieved as swift and accurately as possible. Secondly, the Arabic language is regarded as one of the United Nations' seven official languages. Finally, the majority of native Arabic speakers are unable to read English. [6]. Bag of Words (BoWs) representation is considered the simplest representation of texts. In this representation, the data is usually represented as a matrix with n rows and m columns with rows representing the training data texts and the columns representing the selected features. The weight of each feature in the text is represented by the value of each cell in the matrix. The training matrix, which comprises chosen features, is used to train the classification algorithm [7]-[9].

The conceptual representation of the text considers the semantics and relations between concepts related to words that appear in the document which may increase the semantic representation of the text which in turn could lead to a better interpretation and thus better accuracy of text classification. In this paper, the Arabic WordNet, as semantic information and knowledge base, is utilized to investigate the impact of using concepts and semantic relations between them in text representation on the Arabic Text Classification. The paper is organized as follows: In the first section, we review related works and discuss the background of Arabic WordNet. In the second section, we explain the proposed model to study the impact of conceptual text representation on Arabic Text Classification. After that, the results of the experiments were discussed. Finally, we conclude our work.

2 Literature Review and Background of Arabic WordNet

Many researchers have attempted to represent text using concepts and semantic relations between them instead of words only by utilizing semantic information and knowledge bases such as Arabic WordNet and Wikipedia. This approach considers the semantics and relations between concepts related to words that appear in the document which may increase the semantic representation of the text which could lead to a better interpretation and thus better accuracy of text classification [7]-[9].

2.1 Related works

Karima et al. [1] suggested a conceptual model representing Arabic text using Arabic WordNet. They compared their proposed conceptual model with BoWs and N-gram representations. They concluded that the proposed conceptual model outperformed other representations and that the semantic representation of text is one of the most promising approaches for categorizing Arabic texts. Alahmadi et al. [8] proposed different approaches based on combining the Bag-of-Words (BOW) and the Bag-of-Concepts (BOC) text representation schemes and utilizing Wikipedia as a knowledge base. Their experiments showed that SVM was the best classifier by all BoCs models and the BoCs model performed better than the traditional BOWs. Yousif et al [9] studied the impact of stemming and conceptual representation based on Arabic Wordnet on Arabic text classification. They concluded that root extractor and utilizing Has hyponym semantic relation gave better results especially when it combined with the position tagger. In [12], they suggested two new feature sets based on the Arabic WordNet (AWN). The first one is the List of Pertinent Synsets (LoPS), which is a list of synsets that have a specific relationship to the terms in the text. The second is the List of Pertinent Words (LoPW), which is a list of words that have a specific relationship to the terms in the text. They compared these feature sets for various AWN semantic relations. They concluded that LoPW outperformed LoPS, Bag-of-Word, and Bag-of-Concept in terms of classification accuracy. Alahmadi et al. [13] also studied the impact of a conceptual text representation on Arabic text classification. Arabic WordNet and Wikipedia were used as semantic knowledge. They found that utilizing Wikipedia had better performance compared with Arabic WordNet as a semantic knowledge source. They concluded that the text representation has an impact on text classification and combining words with concepts enhanced the text classification compared to using words and concepts individually. Yousif et al. [14] proposed a weighting scheme to assign a different weight for every semantic relation based on the frequency of the relation in the Arabic WordNet (AWN) and the dataset which will reduce the effect of weak relations and boost the effect of stronger relations. They concluded that the weighting scheme improved the classification accuracy and outperformed the traditional bag of words (BoWs) representations. Yousif et al. [15] proposed another approach to utilize semantic relations based on relation grouping. They grouped the semantic relations based on one of three criteria: the semantic meaning, the frequent occurrence of the semantic relations in the AWN, and finally based on the ratio between the frequency of the relations in the dataset and the frequent occurrence of the alike semantic relations in the AWN. They compared their proposed text representations with traditional BoWs representation and filter selection methods; namely: Chi-square and information gain. They concluded that their grouping approaches outperformed the other methods and reduced the feature dimensionality.

2.2 Arabic WordNet

Arabic WordNet (AWN) is a lexical database for the modern standard Arabic language which depends on Princeton WordNet and it is established based on the methods developed for Euro WordNet. AWN consists of verbs, nouns, adverbs, and adjectives that are organized into a set of synonyms (synsets). Synsets are linked based on lexical and semantic relations which makes the AWN a beneficial tool for text classification, linguistics, and natural language processing. AWN has four tuples or tags [14], [15]:

- Item: The concept of the term
- Word: The term (word)
- Form: The root of the word
- Link: The relationships between concepts

The semantic information and relations are extracted from the AWN using the connections between AWN's tuples.

Figure 1 illustrates the connections between AWN tuples [12].

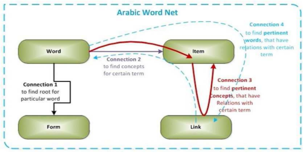
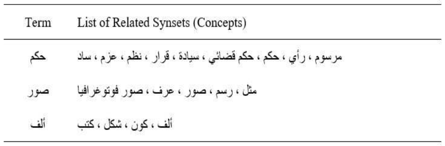

Arabic Word Net
Word
Item
Form
Link
Connection 1 to find root for particular word
Connection 2 to find concepts for certain term
Connection 3 to find pertinent Concepts, that have Relations with certain term
Connection 4 to find pertinent words, that have relations with certain term

Figure 1. Connections between the components of AWN.

Figure 2 presents some examples of a list of concepts connected to specific words [12].



Term	List of Related Synsets (Concepts)
حكم	مرسوم ، رأي ، حكم ، حكم قضائي ، سيادة ، قرار ، نظم ، عزم ، ساد
صور	مثل ، رسم ، صور ، عرف ، صور فوتوغرافيا
ألف	ألف ، كون ، شكل ، كتب

Figure 2. Examples of a list of synsets (concepts) connected to specific words.

Figure 3 shows the list of all AWNs relations [12].

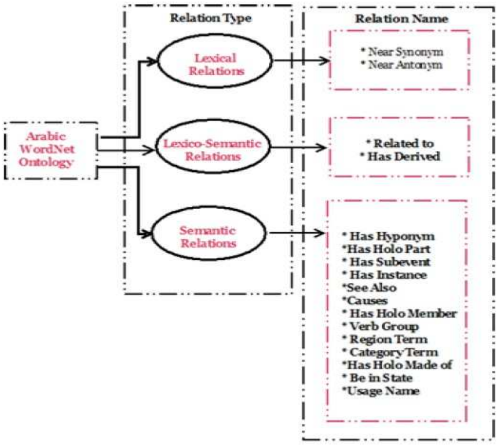


Figure 3. Relations in AWN and their types.

3 Impact of Conceptual Text Representation

Several studies showed out the positive impact of a conceptual representation of text on the accuracy of Arabic text classification. The impact of utilizing conceptual vector representation using concepts and semantic relations was studied in this research.

To study the impact of a conceptual representation of the text, several scenarios are applied: adding all related concepts to terms in the document, replacing terms with all related concepts, using a simple method of word sense disambiguation by adding the first related concept only to the term in the document, replacing term by first related concept only. In addition to that, the impact of utilizing semantic relations based on relations grouping, and utilizing a higher level of hyponym (2nd level of has-hyponym relation) on Arabic Text Classification were studied.

Table 1. Frequency occurrence of Semantic relations in BBC Dataset.

Semantic Relation	Relation Frequency
Has-hyponym	10175
Related-to	6171
has-holo-member	1717
Category-term	1313
Magazine	41
See-also	1129
Verb-group	932
Has-instance	862
Has-subevent	840
Has-derived	708
Causes	596
Be-in-state	560
Has-holo-part	551
Near-synonym	450
Has-holo-madeof	114
Region-term	51
Usage-term	1
Total	28248

Several datasets were used in Arabic text classification. One of the most widely used datasets is the BBC Arabic dataset which is free, public, and contains an appropriate number of documents and comprises of seven classes with 4763 documents: Middle East News 2356 documents, World News 1489 documents, Business and Economics 296 documents, Sport 219 documents, Magazine 49 documents, Science and Technology 232 documents, and Collection (Art and Culture) 122 documents. The study had applied to the original dataset with the complete feature set. The grouping was applied based on relations occurrence in the dataset. In Table 1, we present the relations occurrence in the BBC dataset. Only the list of related words (i.e., concepts and their semantic relations) that have the same class of the original word (document term) were added to the training file. Preprocessing is very beneficial because it decreases the index size, improves accuracy, and unifies the classification activities. Preprocessing includes the following which was used by most researchers [12], [14]:

- Removing digits, punctuations, non-Arabic letters, letters that do not belong to a word, special symbols, and diacritics.
- Removing stop words which are the terms that appear frequently in the text and are insignificant. Pronouns, prepositions are examples.
- Removing rare words: words that appears less than four times will be removed [13], [17].
- Replacing a congested collection of spaces and new line characters with a single space so separating each pair of words by one space.

Naïve Bayes (NB) classifier is a probability-based classifier based on Bayes theory. NB is considered one of the simplest and practical classifiers. Naïve Bayes (NB) classifier has been proven effective in Arabic text classification so it is used in this research. In our study stratified 10-fold cross-validation was used. Stratified K-fold cross-validation is a variation of K-fold cross-validation in which the dataset is divided into K-parts as K-fold cross-validation in a such way that ensures that the same class distribution in each K parts will be the same as in the complete dataset. Stratified K fold cross-validation is usually used with an imbalanced dataset to overcome the overfitting problem where k-classification models are built using a collection of partitions for training and testing [3], [9], [18].

Various performance measures could be used to evaluate Text classification effectiveness. F1-measure is the harmonic mean of precision (p) and recall (r) which provides a more realistic measure of a test's performance. Weighted-F1 which weights the F1-score of each class C_i by the number of documents N_i from that class. Weighted F1- Measure was used as a performance measure in this study:

$$F1 - Measure = 2 \ . \ \frac{p \ . \ r}{p+r} \qquad (1)$$

$$Weighted\ F1 - Measure = \frac{\sum_{i=0}^{n} F1 - measure(C_i) * N_i}{D} \qquad (2)$$

such that D is the total number of documents.

4 Experimental Results

In Table 2, we show the performance of Arabic text classification using BoWs, and various conceptual text representations. Notably, all conceptual representations outperformed BoWs text representation which demonstrated that text classification could be enhanced by incorporating related concepts into texts. Furthermore, it is shown that enriching document representation with either adding/ (replacing term with) the first synset had slightly better performance compared with adding/ (replacing the term with) all related synsets which will not increase the dimensionality of text vector space. Based on that in this study, only the first synset was considered as a simple word sense disambiguation strategy.

Table 2. Weighted F1-measure of BoWs and conceptual representations for Complete

	Method	Weighted F1-measure
Features set	BoW	72.6
	Adding all related concepts	75.7
	Replacing term with its related concepts	74.3
	Adding First related concept only	75.9
	Replacing term with its first related concept only	74.5

4.1 Semantically Grouped Relation based on relation occurrence method

Based on Table 1, the semantic relations that have the highest occurrence in the dataset; namely Has-hyponym Related-to, has-holo-member, Category-term were studied here. Semantic Relations were grouped as follows:
- Has-hyponym with Related-to.
- Has-hyponym with Has-holo-member.
- Has-hyponym with Category-term.
- Related-to with Category-term.
- Related-to with Has-holo-member.
- Category-term with Has-holo-member.

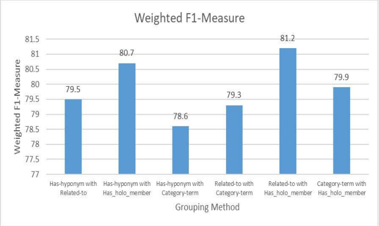


Figure 4. 10-fold classification results of grouped relations for NB classifier.

In Figure 4, we present the 10-fold classification results of the grouped relations for the NB classifier. The experiments results showed a positive impact of enriching document representation with semantic relations. The best classifier performance was when the has-holo-member relation is part of combined relations. The superior result was 81.2 for the combination of related-to with has-holo-part relations while the lowest result was 78.6 for the combination of has-hyponym with category-term relations.

4.2 Utilizing a higher level of hyponym relation

The impact of having more generalization on Arabic Text Classification was investigated by adding the 2nd level related concepts to the term which could be by applying the 2nd level of hyponym relation. Figure 5 presents the 10-fold classification results for has-hyponym and 2nd level of the relation. It is shown that 2nd level of hyponym relation outperformed the first level of hyponym which means that having more generalized concepts have a positive impact on Arabic text classification.

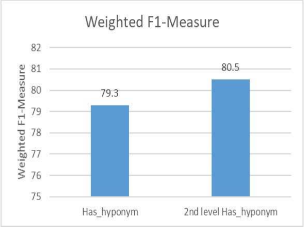


Figure 5. 10-fold classification results for has-hyponym and 2nd level of the relation

5 Conclusion

We investigated the impact of utilizing the conceptual representation and semantic relations and showed that adding concepts to the training files leads to a better classification performance. Moreover, we studied the impact of grouping semantic relations based on their occurrence in the dataset. The experiments showed a positive impact of enriching document representation with semantic relations on Arabic Text Classification. The best classification performance occurred when the has-holo-member relation is part of combined relations. The superior result was 81.2 for the combination of related-to with has-holo-part relations while the lowest result was 78.6 for the combination of has-hyponym with category-term relations. Finally, having more generalized concepts (2nd has-hyponym) could improve the classification results.

References

[1] A. Karima, E. Zakaria, T. G. Yamina, A. A. S. Mohammed, R. P. Selvam, V. Venkatakrishnan, Arabic text categorization: a comparative study of different representation modes, J. Theor. Appl. Inf. Technol., **38,** no. 1, (2012), 1–5.

[2] I. Hmeidi, M. Al-shalabi, M. Al-Ayyoub, A comparative study of automatic text categorization methods using Arabic text, International Technology Management Conference, **73,** (2015).

[3] M. S. Khorsheed, A. O. Al-Thubaity, Comparative evaluation of text classification techniques using a large diverse Arabic dataset, Lang. Resour. Eval., **47,** no. 2, (2013), 513–538.

[4] R. Mamoun, M. A. Ahmed, A Comparative Study on Different Types of Approaches to the Arabic text classification, Proc. 1st Int. Conference of Recent Trends in Information, **2,**, no. 3, (2014).

[5] S. A. Yousif, V. W. Saaawi, I. Elkabani, R. Zantout, Enhancement of Arabic Text Classification Using Semantic Relations with Part of Speech Tagger, W Trans. Adv. Electr. Comput. Engg., (2015), 195–201.

[6] M. M. Al-Tahrawi, Arabic text categorization using logistic regression, Int. J. Intell. Syst. Appl., **7,** no. 6, (2015), 71.

[7] A. Alahmadi, A. Joorabchi, A. E. Mahdi, A new text representation scheme combining bag-of-words and bag-of-concepts approaches for automatic text classification, 7th IEEE GCC Conference and Exhibition, (2013), 108–113.

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

[9] S. A. Yousif, V. W. Samawi, I. Elkabani, R. Zantout, The effect of combining different semantic relations on Arabic text classification, World Comput. Sci. Inform. Technol. J., **5,** no. 1, (2015), 12–118.

[10] S. A. Yousif, V. W. Samawi, I. Elkaban, R. Zantout, Enhancement of Arabic text classification using semantic relations of Arabic WordNet, J. Comput. Sci., **11,** no. 3, (2015), 498.

[11] A. Alahmadi, A. Joorabchi, A. E. Mahdi, Combining Words and Concepts for Automatic Arabic Text Classification, International Conference on Arabic Language Processing, (2017), 105–119.

[12] S. A. Yousif, V. W. Samawi, I. Elkabani, Arabic text classification: The effect of the awn relations weighting scheme, Proc. World Congress Engg., **2,** (2017).

[13] S. A. Yousif, Z. N. Sultani, V. W. Samawi, Utilizing Arabic wordnet relations in Arabic text classification: New feature selection methods, IAENG Int. J. Comput. Sci., **46,** no. 4, (2019), 1–12.

[14] H. K. H. Chantar, New techniques for Arabic document classification, Ph. D. Thesis, Heriot-Watt Univ., (2013), Accessed: Mar. 12, 2019. [Online]. Available: https://www.ros-test.hw.ac.uk/handle/10399/2669.

[15] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Stat. Surveys, **4,** (2010), 40–79.