Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 6, No. 1, February 2015), Pages: 125-136
www.jacr.iausari.ac.ir

Building Semantic Kernel for Persian Text Classification with a Small Amount of Training Data

Amir H. Jadidinejad[1]✉, Venus Marza[2]

1) Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

2)Department of Computer Engineering, West Tehran Branch, Islamic Azad University, Tehran, Iran

amir.jadidi@qiau.ac.ir; v.marza@srbiau.ac.ir

Received: 2014/05/06; Accepted: 2014/06/23

Abstract

The original idea of semantic kernels is to use semantic features instead of terms appeared in the text document. In this article, the documents are transformed into a new k-dimensional feature space by applying Singular Value Decomposition on the Term-Document matrix and extracting k eigenvectors with higher energy. The suggested semantic kernel causes severe reduction of dimensions which leads to two main conclusions. First, the computational complexity of the classifier is severely reduced. Second, the trained classifier has less sensitivity on the input terms; therefore, it can classify documents effectively. Experiments on Persian documents indicate the absolute superiority of the suggested semantic kernel in comparison to well-known vector space (Bag-of-Words) kernel, especially under the circumstances in which external semantic resources are not available and the amount of available training data is not sufficient

Keywords: *Semantic Kernel, Vector Space Kernel, Support Vector Machine, Dimensionality Reduction, Text Classification*

1. Introduction

The main idea in kernel methods is transferring data to a new computational semantic space, in which the computations are done with “more accuracy” and “less computational complexity”[1]. Among the many issues considered in kernel methods, text classification has a special significance. Since the words in a text document are usually used as the features of the document (Bag-Of-Words), the number of features is large, therefore performing computations such as classification faces serious problems.

Considering the words as features results in three major problems in text processing: First, combined words like “Information Retrieval” are indicated as two separate features. Second, synonymous words such as “Vehicle” and “Automobile” are presented as two separate features and finally, only one dimension is dedicated to words with more than one meaning. The above problems lead BOW-based methods to be unable to present the meaning of text documents and as a result, semantic kernels are qualified as a serious approach in this field.

In semantic kernels we are looking for conceptual features to replace the words in text documents in a way that two main goals are achieved: First, the number of these

features must be considerably less than the words appeared in the document. Second, the classifier's efficiency can be improved due to the use of new semantic features.

Previous studies on kernel methods in English language have shown that even though most of the semantic kernels reduce the computational complexity, they do not improve the classification accuracy (especially for Support Vector Machine classifiers)[2-4]. In this paper, we primarily present a semantic kernel based on Singular Value Decomposition on the term-document matrix, then we show that the proposed semantic kernel is not only able to significantly reduce the computational complexity, but also it can effectively improve the classification accuracy. This is a result of the Persian's structure which has a complicated morphology and a lot of compound nouns with multiple meanings that are sharply captured in the proposed semantic kernel.

The rest of the paper is organized as follows: In Section 2, related works in the field of semantic kernels will be studied. Our proposed method has been presented in Section 3 and it is followed by experiments over standard benchmarks in the Section 4. Conclusion and discussion will be presented in Section 5.

2. Literature Review

Salton et al. [5]introduced the Bag-Of-Word model which was used as the representation model of text documents for many years. In this model, every document is described by a vector in a vector space of terms. Every term makes a unit of vector space. The corresponding weight of each term is determined with different patterns such as TFIDF[6]. In this way the similarities of documents d_1 and d_2 are defined by inner product, like the following:

$$k(d_1, d_2) = < \square(d_1), \square(d_2) > \quad (1)$$

One of the problems with vector space kernels is that it does not consider the meaning of words. Synonym words fill two distinguish dimension of the vector space; on the other hand, words with multiple meanings fill a shared dimension in the vector space. As a result, a semantic kernel is considered as an emerging approach among researchers[3, 4, 7, 8].

Previous studies in the field of semantic kernels can be divided into two major groups. The first group is the kernels which are formed based on an external knowledge source[9]. WordNet [3] and Wikipedia [4, 10] are two well-known external sources in the field of knowledge-based semantic kernels. The second group consists of kernels usually based on computational latent concepts [11] and made without any interference of external knowledge sources[8].

In languages like Persian [12] which suffers from the lack of external resources, semantic kernels based on computational latent concepts are of more importance[13, 14]. Therefore, among the proposed semantic kernels, Generalized Vector Space kernel [15] and Latent Semantic Kernel [16] are of more importance.

In Generalized Vector Space model [15], in order to apply the relatedness of words in vector space, the matrix $S$is made from the co-occurrence of terms in the documents of the test corpus:

$$\square(d) = \square(d)\acute{D} \quad (2)$$

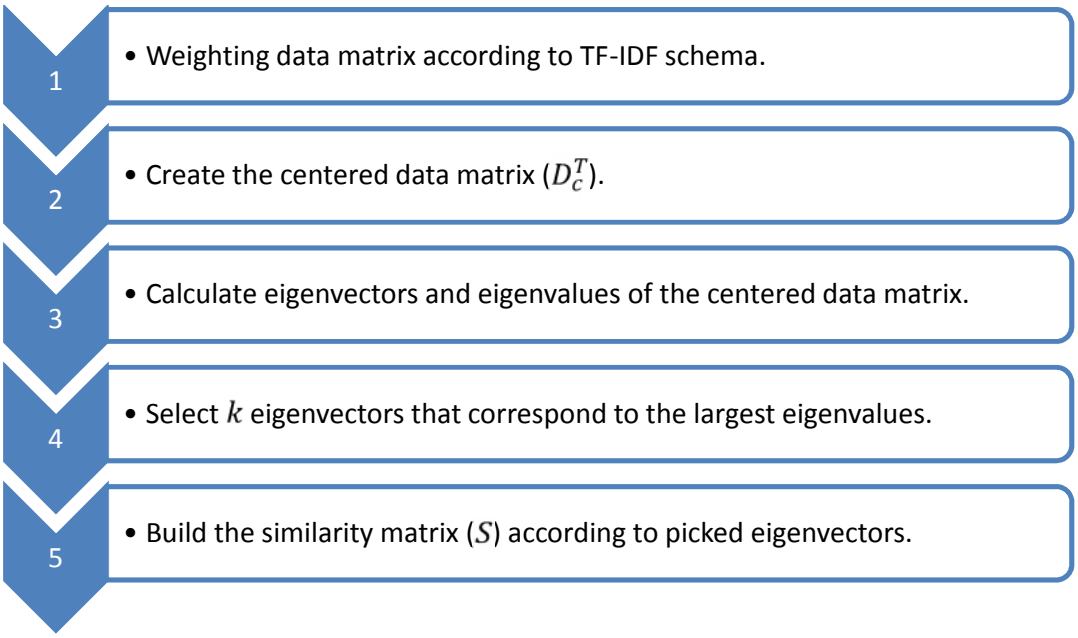

Figure 1: The overall picture of the proposed semantic kernel.

$\acute{D}$is the transpose of term-document matrix. In other words, in this model, the occurrence of the terms in the test corpus is used as a measure to determine the relatedness among terms. The similarity matrix (S) can be provided from external sources. For instance, in [3] and [10] WordNet and Wikipedia was used respectively.
Explicit Semantic Analysis [17, 18] is an approach to make features for text documents that works like Generalized Vector Space kernel [15, 19], except the semantic relatedness of terms ($\acute{D}D$) is provided with an external knowledge source [20].
On the other hand, Latent Semantic Kernel [8] is made based on the term co-occurrence matrix in the test corpus, such as Generalized Vector Space kernel [15, 19]; the only difference is that it can effectively reduce the dimension of the computational space by applying the singular value decomposition ofthe term-document matrix.

3. Proposed Method

In vector space model [21, 22], each document has been shown as:

$$\square: d \longmapsto \square(d) = \left(tf(t_1, d), tf(t_2, d), \dots, tf(t_N, d)\right) \in \mathbb{R}^N \quad (3)$$

Where $tf(t_i, d)$ represents the frequency of a term t_i in document$d$. In addition, a semantic kernel on vector space is defined as:

$$\tilde{\emptyset}(d) = \square(d)S \quad (4)$$

Where S is the similarity matrix and it is described in the form of $S = RP$; Where R is a diagonal matrix and P is the similarity matrix between corpus terms. Various semantic kernels differ from each other in a way which they determine the similarity matrix$S$. In this section we described a combined method to build the similarity matrix S.

Figure 1 shows needed steps to build the semantic kernel. In order to specify the similarity matrix (S), consider the term-document matrix which is defined as:

$$D = \begin{pmatrix} tf(t_1, d_1) & \dots & tf(t_N, d_1) \\ \vdots & \ddots & \square \\ tf(t_1, d_l) & \dots & tf(t_N, d_l) \end{pmatrix} \quad (5)$$

In the first step, Inverse-Document-Frequency (IDF) is calculated for each term in the dictionary [6]. This value indicates the importance of each term. Then, each row of the matrix D is multiplied to the corresponding IDF value. In the second step, we center the documents by subtracting the mean to each column of D. The centered doc-term matrix is shown as D_c^T in the following.

By applying Singular Value Decomposition on the centered doc-term matrix (D_c^T), we have $D_c^T = U \sum V$; Where $\sum$ is a diagonal matrix with the same size as D; U and V are identity matrix that their columns are equal to eigen-vectors corresponding to the centered doc-term matrix (D_c^T). The eigen-vectors can store data with higher energy but in lower dimensions. Now, the k eigen-vectors in matrix U can be utilized as a new similarity data matrix, and mapped data to a new k-dimensional space:

$$d \longmapsto \square(d) U_k \quad (6)$$

Where U_k consists of the first's k columns of matrix U. In other word, a new semantic kernel with the aid of U_k can be defined as:

$$\tilde{k}(d_1, d_2) = \square(d_1) U_k \acute{U}_k \emptyset(\acute{d}_2) \quad (7)$$

In fact, $P = U_k \acute{U}_k$ would be equal to the matrix in which semantic relatedness between terms has been presented by latent concepts. If we deal with a large number of documents, the above matrix can be applied to a variety of applications in information retrieval and natural language processing [23].

Using k eigen-vector with higher energy leads to not only reduce dimensions of computational space but also improve the classification quality, because the related terms are categorized in a same latent feature.

In addition, k parameter controls the amount of combination of different terms in making latent features, i.e. the more amount of k the harder to combine terms; therefore, various terms are categorized in more latent features and the number of space dimensions has been increased.

One of the key problems in semantic kernels is determining k parameter. For this matter, the amount of k is specified in a way that it can compensate more than half of variations in the input data. Based on this hypothesis and distribution of terms in the test corpus, the amount of k is calculated according to the input data ($k = 500$) and is used in experiments of section 4.

4. Experiments

In this section, experimental procedure, benchmark corpus, and evaluation criterion have been explained. Experimental results have shown that semantic kernel reduces the dimensional space and improves the classification accuracy, especially in situations where a few training data are available.

Weka [24] has been utilized as a powerful tool for our experiments. We aim to prove the semantic kernel advantages in comparison to vector space kernel. For this purpose, all documents were decomposed as terms and then existence terms in each document were given to classifier algorithm as features; in this step, the output has been evaluated. Besides, in a separate run, the same procedure was done by semantic kernel with the exception that k eigen-vectors instead of terms has been used as features. The results were compared with the vector space kernel [22].

4.1 Benchmark Data Set

Our experiments were performed on the second edition of Hamshahri corpus [25]. We used a Persian corpus because of several reasons:

- Persian is a complex language in terms of morphology [12], since it has a lot of synonyms and many words with multiple meaning.
- Despite English language that has proper external corpus for performing semantic processing such as WordNet, such semantic corpus is not available in Persians so using these semantic kernels for semantic processing i worthful.
- To our knowledge, there has not been any experiment on semantic kernels in Persian so far.

After pre-processing of Hamshahri corpus and removing documents that belongs to more than one class, hierarchal structure has been extracted as shown in Figure 2.Since 29 different categories in Figure 2 had not appropriate distribution and the number of documents in the some classes were so enormous, we randomly selected hundred documents of each class as benchmark dataset for future tests. Since categories had hierarchical structure, automatic classification procedure was harder than situations in which the classes were flat and independent [26, 27].

4.2 Evaluation Criterion

In order to evaluate text classification systems, various criterions can be used. Assume that the labelled dataset $D = \{d_1, d_2, \dots, d_n\}$ is composed of the categories$C^+ = \{C_1^+, \dots, C_k^+\}$, and we apply a classification algorithm to find classes $C = \{C_1, \dots, C_k\}$ in this dataset. Different validation measures are based on counting the pairs of points on which two sets are agree/disagree. The best-known measures are precision (π), recall (ρ) and F-measure (F_1).The precision is calculated as the portion of class$C_p$that includes the documents of categoryC_p^+. On the other hand, the recall is calculated as the proportion of documents from categoryC_p^+ that are included in class C_p, therefore measure the completeness of the classification algorithm. The most popular metric is $F1$ that is defined based on precision and recall [6, 28]:

$$F_1 = \frac{2PR}{P+R} \tag{8}$$

Moreover, in order to accurate evaluation, all samples are randomly divided into ten independent subsets; one of them has been used as test set and the other for training set. This procedure is repeated ten times by various partsand the average value is calculated as well. Therefore, represented results in this section are the outcome of 10 separate and independent runs.

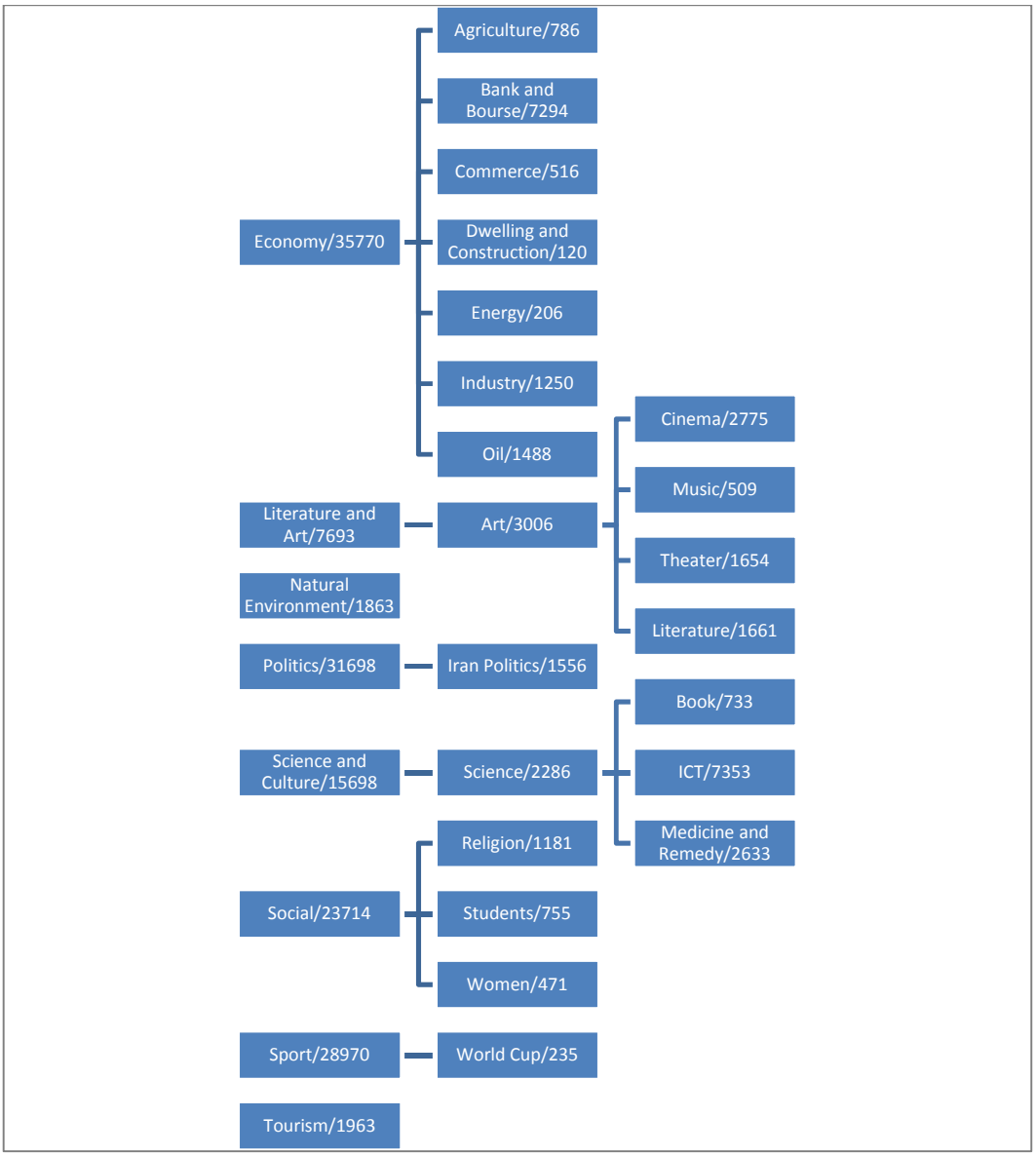

Figure 2: Hierarchical class structure of Hamshahri corpus [25] with the number of documents in each class

4.3 Classification Algorithm

So far, many algorithms have been presented in the field of machine learning [28]; among of them support vector machine (SVM) and Naïve Bayes algorithms are more successful and have been referred in many articles for text classification. In this paper, both algorithms have been utilized in our experiments separately.

Naïve Bayes is a well-known statistical classifier. The probability of a represented document $\overrightarrow{d_j}$ being of c_i class is calculated using Bayes formula as following:

$$P(c_i|\overrightarrow{d_j}) = \frac{p(c_i)P(\overrightarrow{d_j}|c_i)}{P(\overrightarrow{d_j})} \quad (9)$$

Table 1. The comparison of the proposed semantic kernel and the classic vector space kernel using two different classification algorithms

ProposedSemanticKernel					VectorSpaceKernel					ClassificationAlgorithm
Time	NO.	F_1	ρ	π	Time	NO.	F_1	ρ	π	
5.43/53.01	**500**	**0.73**	**0.72**	**0.75**	7.2/103	4603	0.66	0.64	0.71	***SVM***
0.8/0.71	**500**	**0.57**	**0.65**	**0.53**	8.1/6.7	4603	0.47	0.52	0.43	***NaïveBayes***

Assuming that the terms of document d_j are independent, the above equation would be simpler as following [28]:

$$P(\overrightarrow{d_j}|c_i) = \prod_{k=1}^{N} P(tf(t_k, d_j)|c_i) \quad (10)$$

In contrast, Support Vector Machine [29] is a well-known geometrical classifier [28]. It has been referred as the superior classifier for text documents [3, 4, 10, 28, 30]. In geometrical point of view, SVM classifier finds the support vector that can best separate the positive and negative samples in the training set among all the support vectors in the term space [29].

Being resistance against noises in high-dimensional data space made SVM to be the superior algorithm for text classification [28]. So it is expected that the proposed semantic kernel using Naïve Bayes has a better improvement than SVM.

It is important to mention that semantic kernel is independent from the classification algorithm and using two classifiers to approve our experiments is just for completeness. Of course both of the algorithms have had positive feedbacks against semantic kernel.

4.4 Experimental Results

In Table 1 the results of vector space kernel and the proposed semantic kernel are shown. In vector space kernel [22, 31], after tokenizing each document into terms and weighting those according to TFIDF [6] schema 4603 feature (Term) have been extracted from 2900 documents. Finally, every document has been presented as a vector in the vector space. This kernel has been used as the baseline algorithm in our experiments. Also, the experiments have been repeated for both classification algorithms (SVM and Naïve Bayes). In the proposed semantic kernel by applying singular value decomposition to the term-document matrix and exporting k eigen-vector with higher energy, documents have been presented in a new k-dimensional space.

In both conditions (vector space kernel as a well-known baseline and semantic kernel as the proposed method) the amount of accuracy (F_1), the number of features and the time needed for test/train have been evaluated in Table 1.

Our experiments showed that by leveraging k eigen-vector with higher energy and mapping the data items into a new feature space, the number of features has reduced effectively (500 features versus 4603) and as a result, the processing time needed for training and testing has reduced noticeably. Also, the accuracy in both classification algorithms has improved effectively.

Case by case study of the test samples in confusion matrix showed that in vector space kernel most of the system's errors were about the classes that were close in

meaning. For instance, in a sample running, 25 documents that belonged to "science and culture" class have been classified by mistake in "society".

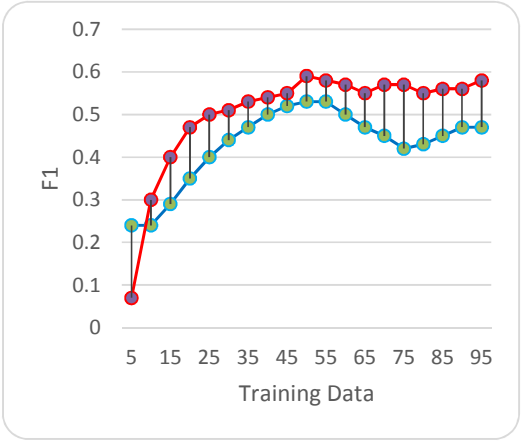

Figure 3: Comparing the practicality of Naïve Bayes classifier when using features generated by the proposed semantic kernel (the red series) and vector space kernel (the blue series) for a certain percentage of training data.

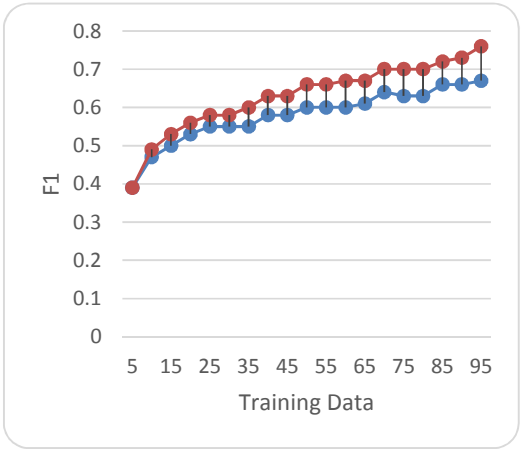

Figure 4: Comparing the practicality of SVM classifier when using features generated by the proposed semantic kernel (the red series) and vector space kernel (the blue series) for a certain percentage of training data.

Since in the vector space kernel [22, 31], synonymous terms or terms with multiple meaning were not properly covered; such a problem was possible to be predicted. By transferring data to the new feature space by the proposed semantic kernel, classification error rate between the two groups “knowledge and culture” and “community” was reduced to 19 cases; because many synonymous terms were

organized in the form of a new feature. Therefore, the classification algorithm wasmore efficient in learning the model.

As shown in Table 1, using the proposed semantic kernel improves both precision and recall. This is a valuable consequence in traditional Information Retrieval as precision and recall are the two balanced key factors and often by increasing precision, recall will decrease[32]. The reason refers to traditional bag-of-word model (vector space kernel), which is based on terms. Synonymy and polysemy are two fundamental problems when using terms as key features. Whereas ambiguous terms are often used in documents with different meanings, it may lead to the retrieval of irrelevant documents and eventually decreases the precision. Of course the occurrence of a term in the different documents would not be identified by bag-of-word model when synonym words have been questioned and this procedure ultimately leads to reduction of recall. Using latent concepts instead of terms appeared in a document causes problems of synonymy and polysemy to be solved through Singular Value Decomposition.

In order to confirm the association between the number of training documents and the learning process of the classifier, in both cases of using vector space kernel and semantic kernel, previous experiments for certain amount of training documents have been repeated. Figure 3 and Figure 4 show the performance of Bayes and Support Vector Machine classifiers for a given amount of training data. For each number of training data, the semantic kernel acts firmly better than the vector space kernel. Also, in the state of having a small number of training documents using semantic kernel can better contribute to improvement in the performance of the classifier. The proposed semantic kernel leverages Singular Value Decomposition to reduce the term-doc matrix. When very few number of documents are available (less than 5%), the co-occurrences of terms in documents will be invalid and applying Singular Value Decomposition is not effective.

5. Conclusion and Future Works

In this paper, a semantic kernel based on the extraction of k eigen-vectors from term-document matrix was proposed. Since the proposed kernel was formed without external knowledge sources interference, it was capable of being used when the external sources of knowledge were not available or would cost a lot to be made.

In terms of language structure, Persian is one of the complex languages which a good source of external knowledge is not available for [12, 25]. The experiments presented in section 4 illustrate this fact that using the proposed semantic kernel can improve the performance of Naïve Bayes and Support Vector Machine classifiers. Using semantic kernel in Persian documents will have three favorable outcomes: First, vector space dimensions will extremely get reduced (500 features versus 4603 features in the vector space kernel). It causes the computation volume to reduce dramatically to train and test the classifier.

Second, using k vector with higher energy will cause the classifier training process to get done more appropriately and finally will make the classifier more efficient in the testing phase.The obtained results indicated that the proposed method is able to both reduce the running time and increase the performance of average recall (ρ), precision (π) and F_1 criteria.

Eventually, when we have a small number of training documents, semantic kernel can best improve the classification performance. Semantic kernel is able to incorporate the semantic relationship between terms and organize the synonymous terms in one dimension of the vector space; therefore, the classifier is less sensitive to the input terms and is able to classify the Persian texts in a better way.

Text classification is a basic machine learning problem which has a wide range of applications in the field of Information Retrieval and Natural Language Processing [28, 33] such as: sentiment analysis [34] and plagiarism detection [35]. On the other hand, using latent concepts instead of words can be used to represent cross-language documents [36].

Acknowledgement

We thank the anonymous reviewers for their very useful comments and suggestions. Also, the authors would like to thank Qazvin Branch, Islamic Azad University for the financial support of this research, which is based on a research project contract.

6. References

[1] Cohen, W.W., "Learning To Classify Text", 2010, Carnegie Mellon University.

[2] Jurgens, D. and K. Stevens. "The S-Space package: an open source package for word space models". in *Proceedings of the ACL 2010 System Demonstrations*. 2010. Stroudsburg, PA, USA: Association for Computational Linguistics.

[3] Siolas, G. and F. d'Alche Buc. "Support Vector Machines based on a semantic kernel for text categorization". in *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*. 2000.

[4] Wang, P. and C. Domeniconi. "Building semantic kernels for text classification using wikipedia". in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008. New York, NY, USA: ACM.

[5] Raghavan, V.V. and S.K.M. Wong, "A critical analysis of vector space model for information retrieval".*Journal of the American Society for Information Science*, 1986. **37**(5): p. 279-287.

[6] Manning, C.D., P. Raghavan, and S. Hinrich, Introduction to Information Retrieval.*Cambridge University Press*, 2008.

[7] AlSumait, L. and C. Domeniconi. "Local semantic kernels for text document clustering". in *SIAM international conference on data mining workshop on text mining*. 2007.

[8] Cristianini, N., J. Shawe-Taylor, and H. Lodhi, "Latent Semantic Kernels".*Journal of Intelligent Information Systems*, 2002. **18**(2-3): p. 127-152.

[9] Li, J., Y. Zhao, and B. Liu, "Exploiting semantic resources for large scale text categorization".*Journal of Intelligent Information Systems*, 2012. **39**(3): p. 763-788.

[10] Wang, P., et al., "Using Wikipedia knowledge to improve text classification".*Knowledge and Information Systems*, 2009. **19**(3): p. 265-281.

[11] Landauer, T.K., P.W. Foltz, and D. Laham, "An introduction to latent semantic analysis".*Discourse Processes*, 1998. **25**(2-3): p. 259-284.

[12] Bijankhan, M., et al., "Lessons from building a Persian written corpus: Peykare".*Language Resources and Evaluation*, 2011. **45**(2): p. 143-164.

[13] Maghsoodi, N. and M.M. Homayounpour, "Improving Farsi multiclass text classification using a thesaurus and two-stage feature selection".*Journal of the American Society for Information Science and Technology*, 2011. **62**(10): p. 2055-2066.

[14] Zahedi, M. and A.G. Sorkhi, "Improving Text Classification Performance Using PCA and Recall-Precision Criteria".*Arabian Journal for Science and Engineering*, 2013. **38**(8): p. 2095-2102.

[15] Wong, S.K.M., W. Ziarko, and P.C.N. Wong. "Generalized vector spaces model in information retrieval". in *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*. 1985. New York, NY, USA: ACM.

[16] Deerwester, S., et al., "Indexing by latent semantic analysis".*Journal of the American Society for Information Science*, 1990. **41**(6): p. 391-407.

[17] Agichtein, E., E. Gabrilovich, and H. Zha, "The Social Future of Web Search: Modeling, Exploiting, and Searching Collaboratively Generated Content".*IEEE Data Engineering Bulletin*, 2009. **32**(2): p. 52-61.

[18] Gabrilovich, E. and S. Markovitch, "Wikipedia-based Semantic Interpretation for Natural Language Processing".*Journal of Artificial Intelligence Research*, 2009. **34**: p. 443-498.

[19] Tsatsaronis, G. and V. Panagiotopoulou. "A generalized vector space model for text retrieval based on semantic relatedness". in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*. 2009.

[20] Anderka, M. and B. Stein. "The ESA retrieval model revisited". in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009. New York, NY, USA: ACM.

[21] Erk, K., "Vector Space Models of Word Meaning and Phrase Meaning: A Survey".*Language and Linguistics Compass*, 2012. **6**(10): p. 635-653.

[22] Turney, P.D. and P. Pantel, "From frequency to meaning: vector space models of semantics".*Journal of Artificial Intelligence Research*, 2010. **37**(1): p. 141-188.

[23] ZHANG, Z., A.L. GENTILE, and F. CIRAVEGNA, "Recent advances in methods of lexical semantic relatedness – a survey".*Natural Language Engineering*, 2013. **19**: p. 411-479.

[24] Bouckaert, R.R., et al., "WEKA--Experiences with a Java Open-Source Project".*Journal of Machine Learning Research*, 2010. **11**: p. 2533-2541.

[25] AleAhmad, A., et al., "Hamshahri: A standard Persian text collection".*Knowledge-Based Systems*, 2009. **22**(5): p. 382 - 387.

[26] Murtagh, F. and P. Contreras, "Algorithms for hierarchical clustering: an overview".*Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2012. **2**(1): p. 86-97.

[27] Silla, C., Jr. and A. Freitas, "A survey of hierarchical classification across different application domains".*Data Mining and Knowledge Discovery*, 2011. **22**(1-2): p. 31-72.

[28] Sebastiani, F., "Machine learning in automated text categorization".*ACM Comput. Surv.*, 2002. **34**(1): p. 1-47.

[29] BARR, J.R., A. JAHEDI, and M. EHSANI, "SUPPORT VECTORS MACHINE: A TUTORIAL WITH R".*International Journal of Semantic Computing*, 2013. **07**(02): p. 185-203.

[30] Ahlgren, O., et al. "A Dimensionality Reduction Approach for Semantic Document Classification". in *SPIM*. 2011.

[31] Shawe-Taylor, J. and N. Cristianini, "Kernel Methods for Pattern Analysis". 2004, New York, NY, USA: *Cambridge University Press*.

[32] Buckland, M. and F. Gey, "The relationship between Recall and Precision".*Journal of the American Society for Information Science*, 1994. **45**(1): p. 12-19.

[33] Cardoso-Cachopo, A. and A. Oliveira, "An Empirical Comparison of Text Categorization Methods", in *String Processing and Information Retrieval*, M. Nascimento, E. Moura, and A. Oliveira, Editors. 2003, Springer Berlin Heidelberg. p. 183-196.

[34] Pang, B. and L. Lee, "Opinion Mining and Sentiment Analysis".*Found. Trends Inf. Retr.*, 2008. **2**(1-2): p. 1-135.

[35] Barrón-Cedeño, A., P. Gupta, and P. Rosso, "Methods for cross-language plagiarism detection".*Knowledge-Based Systems*, 2013. **50**(0): p. 211 - 217.

[36] Vulić, I., W. Smet, and M.-F. Moens, "Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora".*Information Retrieval*, 2013. **16**(3): p. 331-368.