LDA and LSI as a Dimensionality Reduction Method in Arabic Document Classification

Rami Ayadi[1(✉)], Mohsen Maraoui[2], and Mounir Zrigui[3]

[1] LaTICE Laboratory, University of Sfax, Sfax, Tunisia
ayadi.rami@planet.tn
[2] Computational Mathematics Laboratory,
University of Monastir, Monastir, Tunisia
maraoui.mohsen@gmail.com
[3] LaTICE Laboratory, Faculty of Science of Monastir, Monastir, Tunisia
mounir.zrigui@fsm.rnu.tn

Abstract. In this work, we made an experimental study for compare two approaches of reduction dimensionality and verify their effectiveness in Arabic document classification. Firstly, we apply latent Dirichlet allocation (LDA) and latent semantic indexing (LSI) for modeling our document sets OATC (open Arabic Tunisian corpus) contained 20.000 documents collected from Tunisian newspapers. We generate two matrices LDA (documents/topics) and LSI (documents/topics). Then, we use the SVM algorithm for document classification, which is known as an efficient method for text mining. Classification results are evaluated by precision, recall and F-measure. The evaluation of classification results was performed on OATC corpus (70 % training set and 30 % testing set). Our experiment shows that the results of dimensionality reduction via LDA outperform LSI in Arabic topic classification.

Keywords: LDA · LSI · Topics model · Comparing · OATC · Arabic TC

1 Introduction

The text classification is a classic problem of text mining. In recent years, classification is proved to be effective in summarizing a search result or in distinguishing different topics latent in search results. In this work, we use and compare LDA and LSI for dimensionality reduction of feature vectors in Arabic document classification and verify if LDA can substitute LSI for the task. The original feature vectors have occurrences of terms as their entries and thus are of dimension equal to the number of vocabularies. The tow approaches LDA and LSI reduce the dimension of document vectors to the number of topics, which is far less than the number of vocabularies. We can regard each entry of the vectors of reduced dimension as a topic frequency, i.e., the number of words relating to each topic. We inspect the effectiveness of dimensionality reduction by conducting a classification on feature vectors of reduced dimension.

The rest of the paper is organized as follows. Section 2 gives the previous work concerning modeling document for the task of classification. Section 3 includes a short description of LDA. We omit the details about LSI from this paper and refer to the

G. Dregvaite and R. Damasevicius (Eds.): ICIST 2015, CCIS 538, pp. 491–502, 2015.
DOI: 10.1007/978-3-319-24770-0_42

original paper [1]. The results of the evaluation experiment are presented in Sect. 4. Section 5 draws conclusions and gives future work.

2 Related Work

In automatic text classification, it has been proved that the term is the best unit for text representation and classification [2]. Though a text document expresses a vast range of information, unfortunately, it lacks the imposed structure of traditional databases. Therefore, unstructured data, particularly free running text data has to be transformed into a structured data. To do this, many preprocessing techniques are proposed in literature [3, 4]. After converting an unstructured data into a structured data, we need to have an effective document representation model to build an efficient classification system. Bag of Word (BoW) is one of the basic methods of representing a document. The BoW is used to form a vector representing a document using the frequency count of each term in the document. This method of document representation is called as a Vector Space Model (VSM) [5]. Unfortunately, BoW/VSM representation scheme has its own limitations. Some of them are: high dimensionality of the representation, loss of correlation with adjacent words and loss of semantic relationship that exist among the terms in a document [6, 7]. To overcome these problems, term weighting methods are used to assign appropriate weights to the term to improve the performance of text classification [8, 9]. Li et al. in [10] used binary representation for a given document.

The major drawback of this model is that it results in a huge sparse matrix, which raises a problem of high dimensionality. Hotho et al., in [11] proposed an ontology representation of a document to keep the semantic relationship between the terms in a document. This ontology model preserves the domain knowledge of a term present in a document. However, automatic ontology construction is a difficult task due to the lack of structured knowledge base.

Cavana, in [12] used a sequence of symbols (byte, a character or a word) called N-Grams, that are extracted from a long string in a document. In an N-Gram scheme, it is very difficult to decide the number of grams to be considered for effective document representation. Another approach in [13] uses multi-word terms as vector components to represent a document. But this method requires a sophisticated automatic term extraction algorithm to extract the terms automatically from a document. Wei et al., in [14] proposed an approach called Latent Semantic Indexing (LSI) which preserves the representative features of a document. The LSI preserves the most representative features rather than discriminating features.

Statistical topic models have been successfully applied in many tasks, including classification, information Retrieval and data extraction, etc. [15, 16] These models may capture the correlation word in the corpus with a low-dimensional set of multinomial distribution, called “topics” and provide a short description for documents.

Latent Dirichlet Allocation (LDA) [16] is a widely used generative topic model. In LDA, a document is viewed as a distribution over topics, while a topic is a distribution over words. To generate a document, LDA firstly samples a document-specific multinomial distribution over topics from a Dirichlet distribution; then repeatedly samples the words in the document from the corresponding multinomial distribution.

The topics discovered by LDA can capture the correlations between words, but LDA cannot capture the correlations between topics for the independence assumption underlying Dirichlet distribution. However, topic correlations are common in real-world data, and ignoring these correlations limits the LDA's abilities to express the large-scale data and to predict the new data.

The most important aspect in the text representation is the reduction of the dimension of the space features. There are two goals for the reduction of dimension

- Reducing feature dimension to the computable degree makes the classification task feasible and executable
- Feature set selected should ensure the validity of classification.

There are two classes of dimension reduction techniques, feature selection (FS), and feature extraction (FE). Feature selection selects a representative subset of the input feature set, based on some criterion. An estimate function is used to rank original features according to the calculated score value for each feature. This value represents the quality or importance of a word in the collection. The features then ordered in descending or ascending order for the values, and then select a suitable number of words of the higher orders.

Feature selection algorithms are widely used in the area of text processing due to their efficiency.

Duwairi in [17] compared three dimensionality reduction techniques; stemming, light stemming, and word cluster. Duwairi used KNN to perform the comparison. Performance metrics are: time, accuracy, and the size of the vector. She showed that light stemming is the best in term classification accuracy.

Fouzi in [18] compares five reduction techniques; root based and, light stemming, document frequency DF, TF-IDF, and latent semantic indexing LSI. Then it shows that DF, TFIDF, and LSI methods were superior to the other techniques in term of classification problem.

Thabtahin in [19] investigates different variations of VSM and term weighting approaches using the KNN algorithm. Her experimental results showed that Dice distance function with tf-idf achieved the highest average score.

Said in [20] provided an evaluation study of several morphological tools for Arabic Text Categorization using SVMs. Their study includes using the raw text, the stemmed text, and the root text. The stemmed and root text is obtained using two different preprocessing tools. The results revealed that using light stemmer combined with a good performing feature selection method such as mutual information or information gain enhances the performance of Arabic Text Classification.

In [21] a semantic approach is presented using synonym merge to preserve features semantic and prevent important terms from being excluded. The resulting feature space was then processed with five feature selection methods, ID, TF-IDF, CHI, IG and MI. Experiment shows that classification performance is increased after merging terms and yielding better performance for CHI and IG selection method.

Forman and Yang compare different selection methods based on various aspects, including efficiency, discriminatory ability to obtain optimal performance, etc. In the view of results, statistical indicators such as CHI -2 and the information gain (IG) show their superiority. Different classifiers tend to accept different reduction strategy [22, 23]

Many applications of LDA to real-world problems are proposed, however, these researchers do not compare LDA with other probabilistic model for task of Text Classication. In [24], Mikio conduct intensive experiments comparing LDA with pLSI and Dirichlet mixture. While we can learn important things about the applicability of LDA and other document models, their work compares these document models not from a practical viewpoint, but from a theoretical one.

Tomonari in [25] compare latent Dirichlet allocation (LDA) with probabilistic latent semantic indexing (pLSI) as a dimensionality reduction method and investigate their effectiveness in document clustering by using Japanese and Korean Web articles. For clustering of documents, Tomonari use a method based on multinomial mixture. The experiment shows that the dimensionality reduction via LDA and pLSI results in document clusters of almost the same quality as those obtained by using original feature vectors. Therefore, the vector dimension is reduced without degrading cluster quality. This result suggests that LDA does not replace pLSI at least for dimensionality reduction in document clustering for Japanese and Korean language.

In [26], authors perform a series of experiments using LSA, PLSA and LDA for document comparisons in AEA (Automatic Essay Assessor) and compare the applicability of LSA, PLSA, and LDA to essay grading with empirical data. The results show that the use of learning materials as training data for the grading model outperforms the k-NN-based grading methods.

In this work, we check the effectiveness of LDA as a dimensionality reduction method in Arabic text classification and the effectiveness of quality of classified documents from testing set is checked to evaluate its effectiveness. Although Blei in [15] use LDA for dimensionality reduction, the authors compare LDA with no other methods. Further, their evaluation task is a binary classification of the Reuters-21578 corpus, a slightly artificial task. In this paper, we use LDA as a dimensionality reduction approach to make clear its effectiveness to the classification task for text written in Arabic by comparing it with LSI. Further, we compare LDA and LSI to know if LDA can provide better results than LSI.

3 Latent Dirichlet Allocation

Formally, we define the following terms [15]:

- A word is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1, \ldots, V\}$. We represent words using unit-basis vectors that have a single component equal to one and all other components equal to zero. Thus, using superscripts to denote components, the *vth* word in the vocabulary is represented by a V-vector w such that $wv = 1$ and $wu = 0$ for $u \neq v$.
- A document is a sequence of N words denoted by $\boldsymbol{w} = (w_1, w_2, \ldots, w_N)$, where w_n is the *nth* word in the sequence.
- A corpus is a collection of M documents denoted by $D = \{\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_M}\}$.

LDA is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

LDA assumes the following generative process for each document ***w*** in a corpus D:

i. Choose $N \sim$ Poisson(ξ).
ii. Choose θ $\sim$ Dir(α).
iii. For each of the N words w_n:
   a. Choose a topic $Z_n \sim$ Multinomial (θ).
   b. Choose a word w_n from ($w_n \mid Z_n$) a multinomial probability conditioned on the topic Z_n.

Several simplifying assumptions are made in this basic model, some of which we remove in subsequent sections. First, the dimensionality k of the Dirichlet distribution (and thus the dimensionality of the topic variable z) is assumed known and fixed. Second, the word probabilities are parameterized by a $k \times V$ matrix β where $\beta_{ij} = (w_j = 1 \mid z_{\text{i}} = 1)$, which for now we treat as a fixed quantity that is to be estimated. Finally, the Poisson assumption is not critical to anything that follows and more realistic document length distributions can be used as needed. Furthermore, note that N is independent of all the other data generating variables (θ and z). It is thus an ancillary variable and we will generally ignore its randomness in the subsequent development.

A k-dimensional Dirichlet random variable θ can take values in the $(k - 1)$-simplex (a k-vector θ lies in the $(k - 1)$-simplex if $\theta_i \geq 0,$), and has the following probability density on this simplex:

$$p(\theta/\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1} \tag{1}$$

Where, the parameter α is a k-vector with components $\alpha_{\text{i}} > 0$, and where $\Gamma(x)$ is the Gamma function. The Dirichlet is a convenient distribution on the simplex—it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution.

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta, z, w/\alpha, \beta) = p(\theta/\alpha) \prod_{n=1}^{N} p(z_n/\theta) p(w_n/z_n, \beta) \tag{2}$$

Where is simply i for the unique i such that. Integrating over and summing over z, we obtain the marginal distribution of a document:

$$p(W/\alpha, \beta) = \int p(\theta/\alpha) \left(\prod_{n=1}^{N} \sum_{2z} p(z_n/\theta) p(w_n/z_n, \beta) \right) d\theta \tag{3}$$

Finally, taking the product of the marginal probabilities of single documents, we obtain:

$$p(D/\alpha,\beta)=\prod_{d=1}^{M}\int p(\theta_d/\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}/\theta_d)p(w_{dn}/z_{dn},\beta)\right)d\theta_d \tag{4}$$

The parameters α and β are corpus level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ_d are document-level variables, sampled once per document. Finally, the variables and are word-level variables and are sampled once for each word in each document.

The LDA model starts with a set of topics. Each of these topics has probabilities of generating various words. Words without special relevance, like articles and prepositions, will have roughly even probability between classes (or can be placed in a separate category). A document is generated by picking a Dirichlet distribution over topics and given this distribution, picking the topic of each specific word. Then, words are generated given their topics. The parameters of Dirichlet distribution are estimated by the variation of the EM algorithm.

4 Evaluation of LDA and LSI as a Dimensionality Reduction Method

4.1 Pre-processing Step

For the validation of our study, firstly, the pre-processing step is performed for normalizing each document *d* in the corpus.

Second, each document is stemmed using a stemmer describe in [27, 28]. Third latent topics are learned with LDA for topic numbers K = 50. Then a supervised classification is performed on the reduced document distribution over topics. We apply the SVM algorithm for document classification.

Let D be a corpus, he has been split into a training set and testing set. Input is training set and testing set and output is the class of documents in testing set.

The process of validation follows these steps: first, we apply a preprocessing stage for all documents in the corpus by performing some linguistic choices in order to reduce the noise in the document as well as to improve the indexation efficiency. Some of the most popular choices are:

1. Each article in the Arabic dataset is processed to remove digits and punctuation marks{., :,/, !,§,&,’,[,(,_,-,|,-,^,),],},=,+,$,*,…
2. Remove all vowels except “ ” (الشدة).
3. Duplicate all the letters containing the symbols “ ” (الشدة).
4. Convert letters “ء” (hamza), “آ” (aleph mad), “أ” (aleph with hamza on top), “ؤ” (hamza on w), “إ” (alef with hamza on the bottom), and “ئ” (hamza on ya) to “ا” (alef).
5. Convert the letter “ى” to “ي” and the letter “ة” to “ه”. The reason behind this normalization is that there is not a single convention for spelling “ى” or “ي”and “ة” or “ه” when they appears at the end of a word.
6. All the non Arabic words were filtered.
7. Arabic function words were removed.

8. Applied stemming Algorithm for each article in Arabic data set to obtain a stemmed text.

In the next step, after preprocess documents in the training set, we learn the parameters of LDA and get θ (matrix of "document*topic") and φ (matrix of "topic*word"). Then, we model documents in the testing set according to the parameter got from the first step, that is, transform documents in testing set into the form of matrix "document*topic", after this, we perform classification on corpus using SVM classifier, that is, input the matrix "document*topic" of training set and testing set into SVM classifier and finally evaluate on classification results by using various metrics.

In the Table 1, we can show the 20 most likely words for 2 topics (9 and 5).

Table 1. Most likely words for 2 topics

Topic 9th:			Topic 5th:		
English	Arabic word		English	Arabic word	
Higher	العالي	0.015508	Political party	النهضه	0.014895
Education	التعليم	0.010981	Movement	حركه	0.012978
University	الجامعيه	0.008299	Government	حكومه	0.012955
Certificate	شهاده	0.007566	Tunisia	تونس	0.012724
Search	البحث	0.007035	party	حزب	0.011570
Scientific	العلمي	0.006857	Political	السياسيه	0.009654
Students	الطلبه	0.006857	Politician	السياسي	0.009584
Baccalaureate	الباكالوريا	0.005492	party	الحزب	0.008384
Institute	المعهد	0.005365	Front	جبهة	0.008107
Respect	النسبه	0.004986	Coming	الاحزاب	0.007437
Year	السنه	0.004960	Revolution	الثوره	0.007368
Of Education	التربيه	0.004859	president	رايس	0.006929
Tunis	تونس	0.004859	Political party	نداا	0.006837
Visited	زاره	0.004581	National	الوطني	0.006167
Institute	المعهد	0.004530	Motion	الحركه	0.005659
Science	راضيه	0.004277	People	الشعب	0.005405
Section	اجل	0.004252	Coming	القادمه	0.005290
Satisfied	العلوم	0.004176	Parties	الاطراف	0.004874
First	الاولي	0.004126	Dialogue	الحوار	0.004782
Number	عدد	0.004126	Government	جمعه	0.004782

In addition, the evaluation is completed by applying SVM [29, 30] classification with LSI reduction. Sparse matrix is generated for each word to represent the corpus, with each column being the vector representation of documents in the original space. The detail about the coding method can be referred in [1]. Then the SVD toolkit [31] is used for Singular Value Decomposition. Each word in the vocabulary is represented as a 100-dimension vector in S-space.

4.2 Document Sets

In the evaluation experiment, we use a document set of Tunisian Web news articles. We start by building our data set. The OATC (Open Arabic Tunisian Corpus) contains 20.000 documents that vary in length and writing styles. These documents fall into 10 categories that equal in the number of documents. In this Arabic dataset, each document was saved in a separate file within the directory for the corresponding category, i.e., the documents in this data set are single-labeled. Tables 2 and 3 show more specified details about the collection.

Table 2. Number of documents in each category

OATC	NB of text	Average number of words per text	Number of words per category	Category Size (Mo)
Sport	2 000	141.261	282 522	2.99
regional	2 000	125.723	251 447	2.71
Culture	2 000	168.485	336 971	3.62
World	2 000	105.701	211 402	2.26
National	2 000	136.739	273 479	2.97
Political	2 000	164.356	328 712	3.53
Economic	2 000	148.922	297 845	3.27
Student	2 000	203.485	406 971	4.50
Investigation	2 000	253.602	507 205	5.43
Judicial	2 000	126.93	253 860	2.70

Table 3. Specified details about OATC

NB of text in the corpus	20.000
NB of words in the corpus	2 .523 .022
Size of corpus (Mb)	34.0 Mb
NB of category	10

The corpus is collected from online Arabic Tunisian newspapers, including attounissia, alchourouk, assabahnews and jomhouria, the Table 4 summarizes the percentage split between different sources. As we can show, for example, the “sport” category is composed of 25 % from Attounissia[1], 25 % from Alchourouk[2], 25 % from Assabahnews[3], 25 % from Jomhouria[4].

We adopt the open source of LDA [32] to model our corpus and we set topic number as K = 50 in LDA model.

[1] http://www.attounissia.com.tn/.

[2] http://www.alchourouk.com/.

[3] http://www.assabahnews.tn/.

[4] http://jomhouria.com/.

Table 4. Percentage split between different sources

Sources	Attou-nissia	Alchou-rouk	Assabah-news	Jom-houria
Sport	25 %	25 %	25 %	25 %
Regional	–	50 %	50 %	–
Culture	25 %	25 %	25 %	25 %
Word	25 %	25 %	25 %	25 %
National	25 %	25 %	25 %	25 %
Political	–	100 %	–	–
Economic	50 %	–	–	50 %
Student	100 %	–	–	–
Investigation	100 %	–	–	–
Judicial incidents	25 %	25 %	25 %	25 %

TC effectiveness is measured in terms of Precision, Recall, and the F1 measure [27]. Denote the precision, recall and F1 measures for a class Ci by Pi, Ri and Fi, respectively. We have:

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{5}$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{6}$$

$$F_i = \frac{2P_iR_i}{R_i + P_i} = \frac{2TP}{FP_i + FN_i + 2TP_i} \tag{7}$$

Where: TPi: (true-positive): number of documents correctly assigned.
FPi: (false positives): number of documents falsely accepted.
FNi: (false-negative): number of documents falsely rejected.

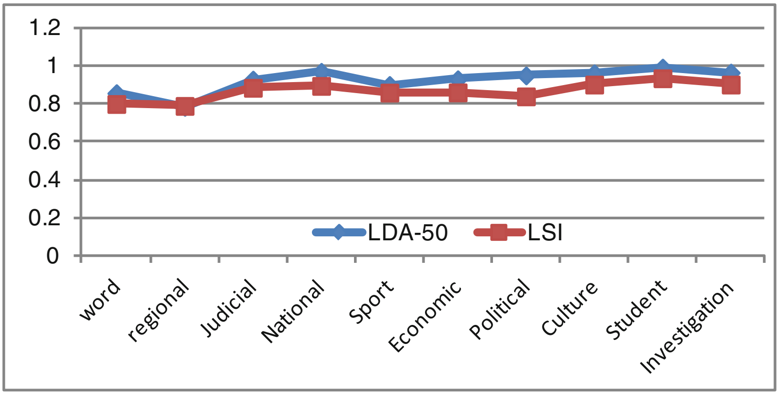

Fig. 1. Precision on each class

The results in Figs. 1, 2, 3 shows that the classification performances in terms of precision, recall and f-measure, in the reduced topics space (LDA-50) outperform those when using LSI reduction.

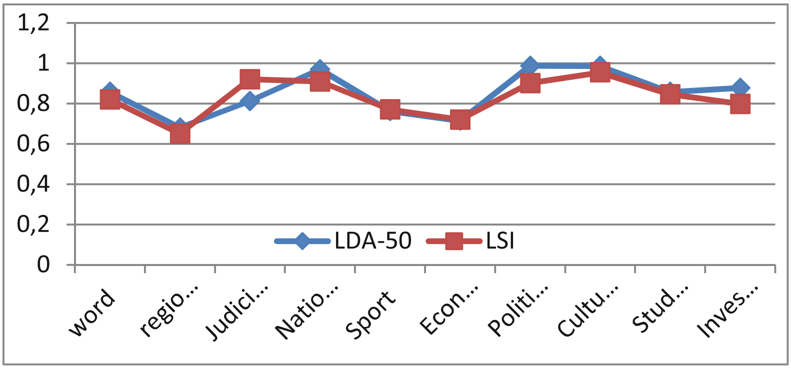

Fig. 2. Recall on each class

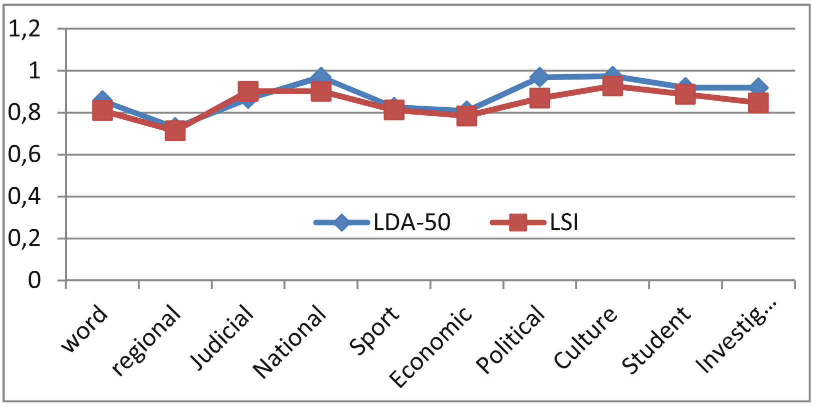

Fig. 3. F-measure on each class

5 Conclusion

In this paper, we have presented the results of an evaluation experiment for dimensionality reduction in Arabic text classification. We use LDA and LSI to reduce the dimension of document feature vectors which are originally of dimension equal to the number of vocabularies. We conduct an Arabic text classification experience based on SVM for the set of the vectors of reduced dimension. We also compare LDA and LSI and the results show that LDA can reduce the dimension of document feature vectors without degrading the quality of document clusters. Further, LDA is far superior LSI. However, our experiment tells no significant difference between LDA and LSI for the class with a small number of words.

References

1. Berry, M.W.: Large-scale sparse singular value computations. Int. J. Supercomputer Appl. **6** (1), 13–49 (1992)
2. Song, F., Liu, S., Yang, J.: A comparative study on text representation schemes in text categorization. Pattern Anal. Appl. **8**(1–2), 199–209 (2005)
3. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)
4. Hotho, A., Nürnberger, A., Paaß, G.: A brief survey of text mining. In: Ldv Forum, pp. 19–62 (2005)
5. Salton, G., Wong, A., Yang, C.-S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
6. Bernotas, M., Karklius, K., Laurutis, R., et al.: The peculiarities of the text document representation, using ontology and tagging-based clustering technique. Inf. Technol. Control **36**(2), 117–220 (2015)
7. Ayadi, R., Maraoui, M., Zrigui, M.: Intertextual distance for Arabic texts classification. In: International Conference for Internet Technology and Secured Transactions, ICITST 2009, pp. 1–6. IEEE (2009)
8. Lan, M., Tan, C.L., Su, J., et al.: Supervised and traditional term weighting methods for automatic text categorization. IEEE Trans. Pattern Anal. Mach. Intell. **31**(4), 721–735 (2009)
9. Altinçay, H., Erenel, Z.: Analytical evaluation of term weighting schemes for text categorization. Pattern Recogn. Lett. **31**(11), 1310–1323 (2010)
10. Li, Y.H., Jain, A.K.: Classification of text documents. Comput. J. **41**(8), 537–546 (1998)
11. Hotho, A., Maedche, A., Staab, S.: Ontology-based text document clustering. KI **16**(4), 48–54 (2002)
12. Cavnar, W.: Using an n-gram-based document representation with a vector processing retrieval model. NIST Special Publication SP, pp. 269–269 (1995)
13. Milios, E., Zhang, Y., He, B., et al. Automatic term extraction and document similarity in special text corpora. In: Proceedings of the Sixth Conference of the Pacific Association for Computational Linguistics, pp. 275–284 (2003)
14. Wei, C.-P., Yang, C.C., Lin, C.-M.: A latent semantic indexing-based approach to multilingual document clustering. Decis. Support Syst. **45**(3), 606–620 (2008)
15. Blei, D., Lafferty, J.: Correlated topic models. Adv. Neural Inf. Process. Syst. **18**, 147 (2006)
16. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
17. Duwairi, R., Al-Refai, M.N., Khasawneh, N.: Feature reduction techniques for Arabic text categorization. J. Am. Soc. Inform. Sci. Technol. **60**(11), 2347–2352 (2009)
18. Harrag, F., El-Qawasmah, E., Al-Salman, A.M.S.: Comparing dimension reduction techniques for Arabic text classification using BPNN algorithm. In: 2010 First International Conference on Integrated Intelligent Computing (ICIIC), pp. 6–11. IEEE (2010)
19. Thabtah, F., et al.: VSMs with K-Nearest Neighbour to categorise Arabic text data (2008)
20. Said, D., Wanas, N., Darwish, N., et al.: A study of Arabic text preprocessing methods for text categorization. In: The 2nd International Conference on Arabic Language Resources and Tools, Cairo, Egypt (2009)
21. Saad, E.M., Awadalla, M.H., Alajmi, A.F. Dewy index based Arabic document classification with synonyms merge feature reduction. In: IJCSI (2011)
22. Forman, G.: An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. **3**, 1289–1305 (2003)

23. Rogati, M., Yang, Y.: High-performing feature selection for text classification. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management. In: ACM, pp. 659–661 (2002)
24. Yamamoto, M., Sadamitsu, K.: Dirichlet mixtures in text modeling. University of Tsukuba, CS Technical report CS-TR-05-1 (2005)
25. Masada, T., Kiyasu, S., Miyahara, S.: Comparing LDA with pLSI as a dimensionality reduction method in document clustering. In: Tokunaga, T., Ortega, A. (eds.) LKR 2008. LNCS (LNAI), vol. 4938, pp. 13–26. Springer, Heidelberg (2008)
26. Kakkonen, T., Myller, N., Sutinen, E., et al.: Comparison of dimension reduction methods for automated essay grading. J. Educ. Technol. Soc. **11**(3), 275–288 (2008)
27. Zrigui, M., Ayadi, R., Mars, M., et al.: Arabic text classification framework based on latent dirichlet allocation. CIT. J. Comput. Inf. Technol. **20**(2), 125–140 (2012)
28. Ayadi, R., Maraoui, M., Zrigui, M.: SCAT: a system of classification for Arabic texts. Int. J. Internet Technol. Secured Trans. **3**(1), 63–80 (2011)
29. Joachims, T.: Making large scale SVM learning practical. Universität Dortmund (1999)
30. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Springer, Berlin, Heidelberg (1998)
31. Berry, M., Do, T., O'Brien, G., et al.: SVDPACKC (Version 1.0) User's Guide1 (1993)
32. Phan, X.-H., Nguyen, C.-T.: GibbsLDA++: AC/C++ implementation of latent Dirichlet allocation (LDA) (2007)