Journal of King Saud University – Computer and Information Sciences 29 (2017) 189–195

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

# Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing

Fawaz S. Al-Anzi *, Dia AbuZeina

*Department of Computer Engineering, Kuwait University, Kuwait*

ARTICLE INFO

*Article history:*
Received 4 November 2015
Revised 28 March 2016
Accepted 2 April 2016
Available online 8 April 2016

*Keywords:*
Arabic text
Classification
Supervised learning
Cosine similarity
Latent Semantic Indexing

ABSTRACT

Cosine similarity is one of the most popular distance measures in text classification problems. In this paper, we used this important measure to investigate the performance of Arabic language text classification. For textual features, vector space model (VSM) is generally used as a model to represent textual information as numerical vectors. However, Latent Semantic Indexing (LSI) is a better textual representation technique as it maintains semantic information between the words. Hence, we used the singular value decomposition (SVD) method to extract textual features based on LSI. In our experiments, we conducted comparison between some of the well-known classification methods such as Naïve Bayes, *k*-Nearest Neighbors, Neural Network, Random Forest, Support Vector Machine, and classification tree. We used a corpus that contains 4,000 documents of ten topics (400 document for each topic). The corpus contains 2,127,197 words with about 139,168 unique words. The testing set contains 400 documents, 40 documents for each topics. As a weighing scheme, we used Term Frequency.Inverse Document Frequency (TF.IDF). This study reveals that the classification methods that use LSI features significantly outperform the TF.IDF-based methods. It also reveals that *k*-Nearest Neighbors (based on cosine measure) and support vector machine are the best performing classifiers.


## 1. Introduction

Recently, text classification (TC) for Arabic language has been widely investigated. Manning and Schütze (1999) defined text classification as the task of classifying texts into one of a pre-specified set of classes based on their contents. According to Sebastiani (2002), text classification is the activity of labeling natural language texts with thematic categories from a predefined set. With big data environment, researchers have been hard at work to address the text classification problem in this huge information era. With massive growth of text search transactions, effective algorithms are needed to satisfy efficient retrieval time and relevance constraints. In today's market, achieving user satisfaction within this astronomical growth of online data is becoming very appealing to business investment. Search engines, e.g., Google and other high traffic query processing portals, are expected to meet and satisfy today's user demands.

Supervised machine learning (ML) approaches are widely used for text classification. The most popular machine learning algorithms include Naïve Bayes (NB), *k*-Nearest Neighbor (*k*-NN), Support Vector Machines (SVM), Neural Networks (NN), Classification Trees (CT), Logistic Regression (LR), Random Forest (RF), and Maximum Entropy (ME). In addition, similarity or distance measures are used for text classification as well as the bases for some classifiers. For example, *k*-NN algorithm uses a similarity function such as Euclidean distance or cosine similarity to find neighbors, Torunoğlu et al. (2011).

In text classification problems, large feature sets are a challenge that should be handled for better performance. Therefore, utilizing feature reduction techniques are important for efficient representation of textual features. Harrag and Al-Qawasmah (2010) presented a number of dimensionality reduction techniques such as root-based stemming, light stemming, and singular valued decomposition (SVD). In this work, we use the SVD as a feature reduction technique as well as for producing semantic rich features. SVD is a linear algebra method that is used to truncate the term-document matrix that produced by Latent Semantic Indexing (LSI), a well-

* Corresponding author.
*E-mail addresses:* fawaz.alanzi@ku.edu.kw (F.S. Al-Anzi), abuzeina@ku.edu.kw (D. AbuZeina).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2016.04.001
1319-1578/

known indexing and retrieval method. Even though the vector space model (VSM) is widely used for textual features representation, however, it is a semantic loss while LSI-SVD is characterized by maintaining the semantic information. Rosario (2000) showed that SVD could be used to estimate the structure in word usage across the documents based on LSI that has the underlying structure in a word choice. Kantardzic (2011) indicated that LSI gives better results when using in text classification as it enables better representation of document's semantics.

This paper contains two parts. First, the LSI-SVD techniques were used to generate the textual features of a corpus that contains 4000 documents. The generated features were then used along with the cosine similarity measure to classify the testing set documents. Second, a number of classification methods were employed for a comparison purpose. The classifiers include NN, NB, *k*-NN, SVM, RF, CT, LR, and CN2 (induction rule). In the implementation, Gensim ("Gensim", 2016) and Orange tool ("Orange", 2016) were used. Gensim is a Python library for natural language processing (NLP) while Orange is an open source machine-learning tool for data visualization and analysis.

In next section, a literate review is presented. In Section 3, we present the singular value decomposition followed by the theoretical background of the cosine similarity in Section 4. The experimental setup is presented in Section 5, and the results are discussed in Section 6. Finally, the conclusion and future work are presented in Section 7.

## 2. Literature review

Cosine similarity measure has been widely used in pattern recognition and text classification. For example, Nguyen and Bai (2011) used cosine similarity measure for face verification. In this work, the focus will be on the cosine measure for linguistic applications. Among such applications, Silber and McCoy (2002) used cosine measure for text summarization. El Gohary et al. (2013) used cosine measure to detect the emotions in the Arabic language text. Takçı and Güngör (2012) indicated that cosine is the commonly used similarity measure in the language identification problem. Sobh et al. (2006) used cosine measure for Arabic language text summarization. Roberts et al. (2005) used cosine similarity for Arabic language concordance. Al-Kharashi and Evens (1994) used the cosine measure for indexing and retrieval processes for Arabic bibliographic data. Lin and Chen (1996) used cosine measure to extract concept descriptors (terms or keywords) from a Chinese-English bibliographic database. Elberrichi and Abidi (2012) indicated cosine similarity dominant measures in information retrieval (IR) and text classification.

For Arabic text classification domain, various feature extraction and classification methods were proposed in the literature as shown in Table 1. In this table, TF.IDF is the shorthand for Term Frequency Inverse Document Frequency, the well-known weighting scheme of text features. TF.IDF is a combination of two parts, (TF: the frequency of the word in the document, and IDF: the inverse of the frequency of the word throughout all documents). ANSI is the shorthand for American National Standards Institute.

Even LSI is a powerful feature representation for words' semantic, the literature provided in Table 1 shows that LSI has very little contribution for Arabic text classification. Therefore, an effort was made to address this deficiency by utilizing semantic information for Arabic text classification. The cosine similarity measure was chosen for classification process. The highlighted cells in Table 1 indicate that only two research works have the same scope as this research (LSI and cosine). However, the first work, i.e. Froud et al. (2013) was conducted for document clustering while our proposed research is for text classification. In addition, we used a larger data set containing 4000 documents while they used 278 documents. We also compared the results using eight well-known classifiers as well as exploring the performance of LSI using a wide range of rank approximation. Regarding the other work, i.e. Harrag and Al-Qawasmah (2010), they used NN for classification while we used cosine similarity measure.

As this research demonstrates a comparative study of the different text classification algorithms, we present a comparison between the supervised machine learning algorithms found in the literature. Table 2 shows that SVM outperforms most of the classification algorithms for Arabic language text classification. The information presented in Table 2 are arranged as the researchers, the classifiers used, the best performance classifier, and the corpus size. However, the information provided in Table 2 is not judgemental as we agree with Sebastiani (2002) that illustrated comparisons are only reliable when they are based on experiments performed by the same author under carefully controlled condi-

**Table 1**
Summary of Arabic text features and classifiers.

| References | Features | Classifier |
|---|---|---|
| Syiam et al. (2006), Thabtah et al. (2008), Gharib et al. (2009), Hmeidi et al. (2008), Ababneh et al. (2014), Kanaan et al. (2009), Duwairi (2007), Zrigui et al. (2012), Moh'd Mesleh (2011), | TF.IDF | *k*-NN |
| Elberrichi and Abidi (2012) | ANSI | |
| Al-Shalabi and Obeidat (2008) | N-gram | |
| Syiam et al. (2006), Gharib et al. (2009), Omar et al. (2013), Kanaan et al. (2009), Moh'd Mesleh (2011) | TF.IDF | Rocchio |
| Jbara (2010), Larkey et al. (2004), Al-Eid et al. (2010), Alghamdi and Selamat (2012), Ezzat et al. (2012), Al-Kabi and Al-Sinjilawi (2007), Erkan and Radev (2004) | TF.IDF | *Cosine* |
| *Froud et al. (2013)* | *LSI* | |
| Gharib et al. (2009), Omar et al. (2013), Hmeidi et al. (2008), Al-Shargabi et al. (2011), Zrigui et al. (2012), Alsaleem (2011), Khorsheed and Al-Thubaity (2013), Hadni et al. (2013), Moh'd Mesleh (2011), Al-Shammari (2010), Harrag et al. (2009), Raheel et al. (2009), | TF.IDF | SVM |
| Al-Harbi et al. (2008) | Chi-Squared | |
| Al-Kabi and Al-Sinjilawi (2007), Duwairi (2007) | TF.IDF | Dice distance |
| Gharib et al. (2009), Omar et al. (2013), Al-Shargabi et al. (2011), Al-Kabi and Al-Sinjilawi (2007), Kanaan et al. (2009), Duwairi (2007), Zrigui et al. (2012), Alsaleem (2011), Khorsheed and Al-Thubaity (2013), Hadni et al. (2013), Moh'd Mesleh (2011), Al-Shammari (2010), Harrag et al. (2009), Raheel et al. (2009), | TF.IDF | NB |
| Al-Shargabi et al. (2011), Khorsheed and Al-Thubaity (2013), Harrag et al. (2009), Raheel et al. (2009) | TF.IDF | CT |
| Al-Harbi et al. (2008) | Chi-Squared | |
| Harrag et al. (2009) | TF.IDF | ME |
| *Harrag and Al-Qawasmah (2010)* | *LSI* | *NN* |