Information Processing and Management 54 (2018) 105–115

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Beyond vector space model for hierarchical Arabic text classification: A Markov chain approach

Fawaz S. Al-Anzi*, Dia AbuZeina

Department of Computer Engineering, Kuwait University, Kuwait

ARTICLE INFO

Keywords:
Arabic text
Classification
Vector space model
Markov chain
Hierarchy

ABSTRACT

The vector space model (VSM) is a textual representation method that is widely used in documents classification. However, it remains to be a space-challenging problem. One attempt to alleviate the space problem is by using dimensionality reduction techniques, however, such techniques have deficiencies such as losing some important information. In this paper, we propose a novel text classification method that neither uses VSM nor dimensionality reduction techniques. The proposed method is a space efficient method that utilizes the first order Markov model for hierarchical Arabic text classification. For each category and sub-category, a Markov chain model is prepared based on the neighboring characters sequences. The prepared models are then used for scoring documents for classification purposes. For evaluation, we used a hierarchical Arabic text data collection that contains 11,191 documents that belong to eight topics distributed into 3-levels. The experimental results show that the Markov chains based method significantly outperforms the baseline system that employs the latent semantic indexing (LSI) method. That is, the proposed method enhances the F1-measure by 3.47%. The novelty of this work lies on the idea of decomposing words into sequences of characters, which found to be a promising approach in terms of space and accuracy. Based on our best knowledge, this is the first attempt to conduct research for hierarchical Arabic text classification with such relatively large data collection.

1. Introduction

The rapid growth of online textual data raises the need for efficient information retrieval (IR) methods in terms of both time and space complexities. Text classification is the process of finding the category of a document based on the contents. Hierarchical Arabic text classification has recently received noticeable attention that is also called multi-level text classification. However, few studies tackled this domain as most of the researchers focus on flat or one-level text classification. In general, text classification employs the vector space model (VSM) that was proposed by Salton, Wong, and Yang (1975) as a model for documents and queries representations. One of the limitations of VSM is the space problem, as each document has to be represented using the entire words in the dictionary (i.e. vocabulary). Despite the number of dimensionality reduction techniques to reduce the dimensions of the textual feature vectors, however, the research is still open to employ space efficient algorithms for text classification.

In this paper, we propose to overcome the space problem by performing text classification using a space-independent text classification algorithm. That is, a method that relaxes the condition of using all words in the dictionary when creating document feature vectors. The proposed method depends on the (first order) Markov chain theory in which the neighbour characters sequences

* Corresponding author.
E-mail addresses: fawaz.alanzi@ku.edu.kw (F.S. Al-Anzi), dia.abuzeina@ku.edu.kw (D. AbuZeina).

http://dx.doi.org/10.1016/j.ipm.2017.10.003
Received 15 April 2017; Received in revised form 17 August 2017; Accepted 11 October 2017

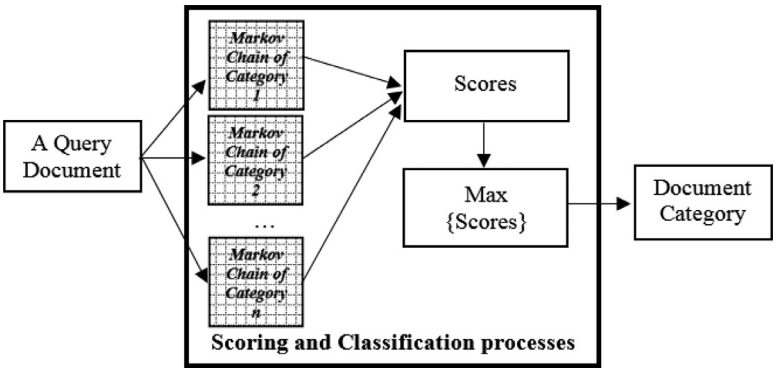

Fig. 1. The framework of the proposed method.

are used to create the probabilities transition matrices. Hence, each document is represented using a sequence of characters co-occurrences in the document. Hence, each category of the corpus is used to create a single probability transition matrix to be used in the classification process. The framework of the proposed work is described in Fig. 1. In this figure, each category represented by a Markov chain (a probability transition matrix) is used to score a query document for classification purposes. Once scored, a comparison process is performed to find the maximum score (i.e. the most likely category) among the collected scores. More details of the proposed method are found in Section 7, the proposed method.

In the next section, we present the literature review. In Section 3, the objectives is presented followed by the Markov chain background in Section 4. Section 5 presents the hierarchical text classification followed by the hierarchical data collection used in this work in Section 6. The proposed method is presented in Section 7 and the experimental results in Section 8. We present the discussion in Section 8 and, finally, we conclude in Section 9.

2. Literature review

This study includes two main topics. The first is the hierarchical Arabic text classification and the second is the Markov chain. For the first topic, the literature shows that the hierarchical Arabic text classification is still a limited research area. According to the authors' best knowledge, no research has been conducted to tackle this area for the Arabic language, however, many studies have been found to employ different classification methods for the flat Arabic text classification. For instance, Al-Anzi and AbuZeina (2017) have a thorough literature of the flat Arabic text classification. Alabbas, Al-Khateeb, and Mansour (2016) demonstrate a comprehensive study of the classifiers and the corpora of Arabic text classification. Uysal et al. (2014) presented a study for efficient use of the preprocessing tasks in text classification. In fact, most of the studies perform performance comparisons between the machine learning tools. Since the proposed method is based on the Markov chain, this literature focuses on the power of this modeling technique especially for linguistic applications. The literature shows that the Markov chain concept is used in many natural language processing (NLP) applications as shown in Table 1.

The Markov chain has also been used in other computing applications, such as software testing, DNA–sequence analysis, network path congestion, image compression, object segmentation, image retrieval, network traffic evaluation, frauds detection, evaluate groundwater quality, estimating phylogenetic trees using DNA, network service recognition, wind power, forecasting, etc.

Table 1
Some of Markov chain based linguistic applications.

Reference	Linguistic field
He, Li, and Chen (2012)	Morphological segmentation
Shen et al. (2014)	Digital document authentication
Haji et al. (2012)	Distributions of words in text lines
Goyal, Jadon, and Pujari (2013)	Documents clustering
Baomao et al. (2009)	Word segmentation
Sampathkumar, Chen, and Luo (2014)	Mining adverse drug reactions
Dowman et al. (2008)	Detecting topical structure
Meng et al. (2009)	Steganography detection
Li, Ding, and Huang (2008)	Recognizing names location
Osiek, Xexéo, and de Carvalho (2010)	Extracting acronyms and their meaning
Cai, Kulkarni, and Verdú (2006)	Text compression
Ahmed et al. (2015)	Authorship attribution
Rodrigues et al. (2013)	Inferring the location of Twitter users
Erkan et al. (2004)	Text summarization
Gao et al. (2011)	Detecting malicious attack

Table 2
Vector space model of N documents and 15 words.

No.	The translation using Google translator	Documents➔ Words	Doc1	Doc 2	…	Doc N
1	Wrote	كتب				
2	Book	كتاب				
3	Writes	يكتب				
4	the book	الكتاب				
5	Brochures	كتيبات				
6	Writer	كاتبة				
7	clerk	كاتب				
8	Writing	كتابة				
9	my books	كتبي				
10	booklet	كتيب				
11	Writing	كتابيب				
12	You write	تكتب				
13	Writing	تكاتب				
14	Takaful	تكتيب				
15	Type	اكتب				

3. Objectives

For a long time, the VSM has been applied in IR and NLP to represent documents as word vectors where each dimension corresponds to a separate word. Despite the popularity and the success of the VSM, however, it has some deficiencies. For instances, Wong et al. (1987) indicate that the main difficulty of VSM is that the explicit representation of term vectors is not known a priori, which leads to unrealistic assumptions. They also highlight the necessary to account for the correlations between terms. In fact, using the entire vocabulary to represent a document (even with a single word) is a challenge due to the demand space. In addition, VSM is a semantic loss method that does not consider the concepts due to the features independent assumption (i.e. the “bag of words”). As a result, it is unable to handle the synonyms, as one deficiency. In general, text classification tasks are characterized by huge feature dimensions, which adds more space and time complexity.

Therefore, we propose a space efficient method that attempts to alleviate the above-mentioned deficiencies. The proposed method employs smaller vectors compared the VSM method. In addition, the proposed method partially consider the synonyms cases (i.e. is not fully semantic rich). For illustration, if we consider the following dictionary that contains 15 words that appear in N documents: {كتب، كتاب، يكتب، الكتاب، كتيبات، كاتبة، كاتب، كتابة، كتبي، كتيب، كتابيب، تكتب، تكاتب، تكتيب، اكتب}. To represent the documents using these words, the VSM uses 15 dimensions for each document. Table 2 illustrate the VSM representation. The table shows that each document has to have a dimension for each single word. The problem is that if the data collection contains a huge number of words, then each document will be represented using that huge number. However, many of the dimensions contain zero value. Each cell in Table 2 is filled using either the term counts or the term frequency – inverse document frequency (TF-IDF).

The objective of this work is to reduce the size of the document vector as much as possible while maintaining the implicit correlations between the words. For the example shown in Table 2, each document is decomposed into a sequence of characters and that sequence is used in classification. Hence, no need to create a VSM vector for each document as we explain in Section 7. For instance, if a document has two words: {تكتب كتابة} then the document vector only contains this sequence: {تك كت تب كت تا اب تة}. In addition, this method will implicitly preserve a kind of semantic information. For clarification, the words appearing in Table 2 have some common words such as (كت) and (تب), which will have same representation in the training model and, therefore, will be helpful and important to classify the document in question even if it has different forms, which is not available in the VSM model. Arabic is morphologically rich language, hence, the proposed method is adequate to handle the similar meaning words as one word. Hence, our overall objective is to find a new method that perform text classification out of the constrains of the VSM such as space and time complexities.

4. Markov chain background

The Markov chain is increasingly being adopted in real-world computing applications since it provides a convenient way for modeling temporal, time-series data. At each clock tick, the system moves into a new state, which can be the same as the previous one. Markov chains are directed graphs (a graphical model) that are generally used for sequential data-mining tasks. Such tasks are characterized by relatively long sequences for many purposes such as prediction, classification, clustering, information sciences, internet applications, pattern discovery, etc. Rabiner (1989) indicates that there are two reasons for the Markov chains popularity, which it is very rich in mathematical structure and works well in practice for several important applications. In fact, there are many applications that employ the Markov chain, so it is worthy to take note of the study of Von Hilgers et al (2006) regarding the five greatest applications of the Markov chains. They indicated that the applications are: Scherr's application to computer performance evaluation, Brin and Page's application to PageRank and Web Search, Baum's application to Hidden Markov Models (HMMs), Shannon's application to information theory, and Markov's application to Eugeny Onegin. There are many reference for HMM, Leon-

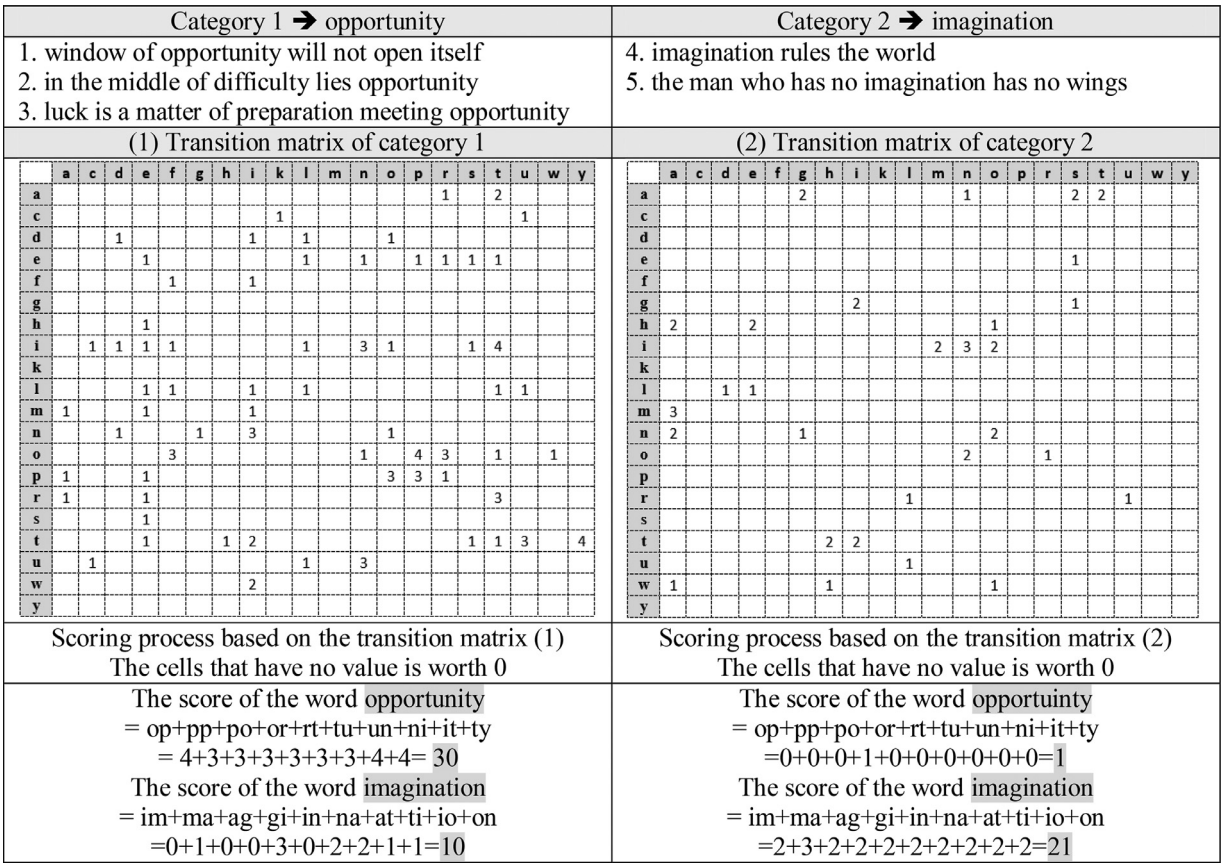

Fig. 2. Scoring words using transition matrices.

Garcia et al. (2008) demonstrate many concepts of HMM and the related applications.

To employ the Markov chain for text classification, a Markov chain model has to be created for each category. For further illustration, suppose that we have five sentences used to create the transition matrices for two categories: opportunity and imagination as shown in Fig. 2. The sentences are inspirational English quotes picked from California Indian Education (2016). In particular, the sentences on the left belong to the “opportunity” class and the sentences on the right belong to the “imagination” class. Fig. 2 also shows the corresponding Markov chain models for each category. We also score the words “opportunity” and “imagination” against the created models. The transition matrices are filled base on the count of transition from one character to another in the sentences.

Hence, the Markov chains can be used for text classification tasks by creating a probability transition matrix for each category. Then, a scoring process is performed for each document against all categories in the training set. Fig. 2 shows that if a document has a word “opportunity” then it will be scored better (i.e based on the maximum score) than the imagination word when using the transition matrix of the category 1. In summary, it is expected that the words belonging to a specific category will be modeled differently compared to other categories of the Markov chain models, and this is the essence of the proposed method.

5. Hierarchical text classification

Hierarchical text classification is characterized by the structure of data organization. Instead of flat or 1-level data, a hierarchical is designed as a multi-level data structure for many reasons, such as speeding up the retrieval time and narrowing the categories and subcategories. Examples of data collection are organized as a hierarchy in the evolutionary tree of species, the federal budget, and business organizational charts, Guerra-Gómez, Pack, Plaisant, and Shneiderman (2015). Hence, the hierarchical text classification aims at organizing the mass of information as a tree structure in which a document that belongs to a topic at a certain level also belongs to all of its parent topics, ancestors, etc.

Employing the hierarchical text classification technique is extremely important in today's huge online data. In the literature, there are many studies that discuss the importance of using hierarchical text classification in many aspects. Bi et al. (2012) indicated that many real-world classification problems already have hierarchical relationships between labels that naturally require hierarchical classification algorithms. The same was presented by Silla et al. (2011) as they indicated that the classes to be predicted in many real-world problems are organized into a class hierarchy such as tree or a Directed Acyclic Graph (DAG). Ying (2011) demonstrated that hierarchical structure can scale well and cope with changes to the category trees.

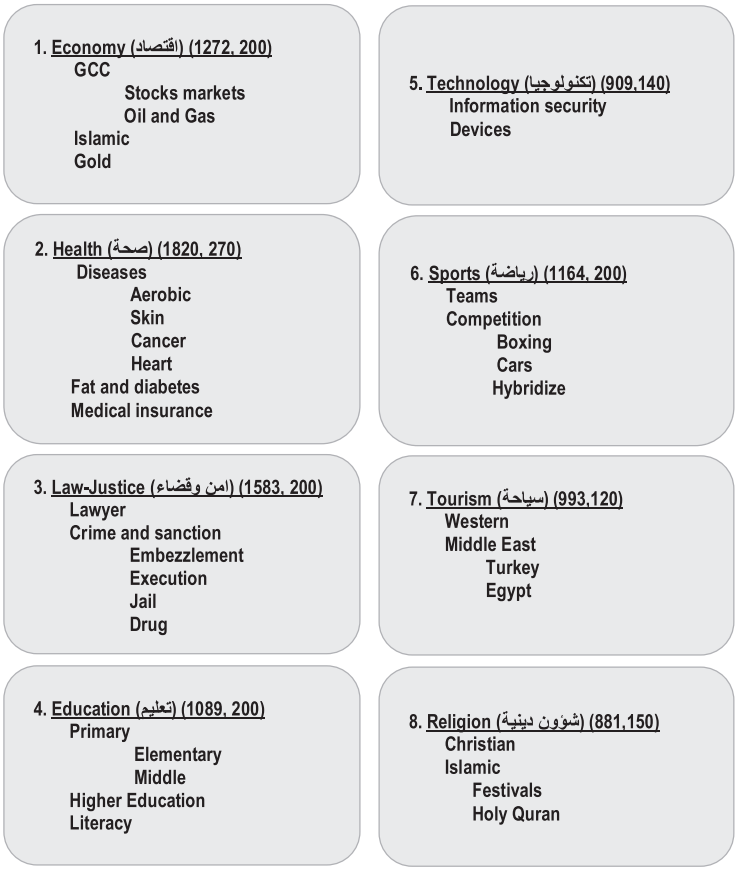

Fig. 3. The hierarchical data collection.

Regarding the performance, D'Alessio, Murray, Schiaffino, and Kershenbaum (1998)) demonstrated that precision and recall can be significantly improved when considering hierarchy in text categorization problem. Ruiz et al. (2002) indicated that the use of the hierarchical structure improves text categorization performance with respect to an equivalent flat model. Joshi et al. (2011) indicated that the hierarchical classification aims at increasing the classification accuracy as well as speed. Godbole (2002) indicated that hierarchies provide valuable navigational aid in browsing large text collections, such as Internet directories. Chakrabarti, Dom, Agrawal, and Raghavan (1997)) indicated that hierarchal databases are used for better searching and browsing digital libraries and patent databases. Dhillon, Mallela, and Kumar (2003) indicated that by using a hierarchy, the classifier utilizes a features set that is more relevant to the classifications sub-task at hand and requires only a small number of features.

Hierarchical text classification is performed using three main approaches, which are flat classification, local classifier (top-down), and global classifier (big-bang). Using the flat approach, each leaf in the hierarchy is represented as a class. However, if the data collection has a large number of leaves, this method will be computational costly. Top-down approach is based on using a classifier at each level of the tree. Hence, the documents of each level are used as training data for that level. While the local classifier approach uses a single classifier at each level, the global approach uses a single classifier to learn the whole information in the hierarchy, which adds more complexities since the entire class hierarchy is considered at once. In this study, we used top-down approach.

6. Hierarchical data collection

In our study, we used the directed acyclic category graph that allows documents to be assigned into both internal and leaf categories. To prepare the necessary corpus, we got a large data collection from Alqabas newspaper that has a sophisticated information center, Alqabas (2016). Initially, we filtered the collection and discarded all documents that had less than 800 characters length. Hence, the used corpus contained 11,191 documents that were split into two sets for training (9711 documents) and for testing (1480 documents). The overall documents of the training set were distributed into a 3-level hierarchy as following: 2978 documents in the first level, 4229 in the second level, and 2504 in the third level. The training set contains 201,982 unique words, and the testing set contains 82,651 unique words. Fig. 3 shows the used taxonomy that includes the categories and subcategories labels. We emphasize that the categories and the subcategories presented in Fig. 3 is just an example to be used for this research. However, in real-world applications, different and even complex categories might be raised according to the topics and domains. In the figure, GCC is shorthand for the Gulf Cooperation Council. The number beside each category label is regarding the total number of

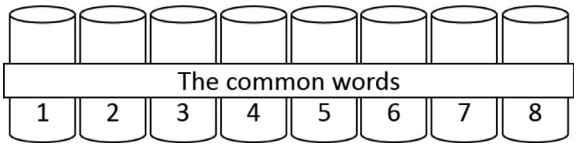

Fig. 4. Common words of the eight categories.

```
...
"المعلمون": "ن و م ل ع م *",
"المعلمين": "ن ي م ل ع م *",
"المعلوماتي": "ي ت ا م و ل ع م *",
"المعلوماتية": "ت ا م و ل ع م *",
"المعماري": "ي ر ا م ع م *",
"المعمارية": "ر ا م ع م *",
"المعوقات": "ت ا ق و ع م *",
...
```

Fig. 5. A part of the words in the dictionary.

training documents and the total number of testing documents.

7. The proposed method

The proposed method is a supervised learning method that requires a training set for learning. Hence, preparing a hierarchical Arabic corpus was the first task in this research. Mainly, three main steps were performed as following:

Step 1: preprocessing

1. The ignore characters set is defined. This set includes: à ® é \ ² … è ² ™ ⁻ › × — ô é ü ÷ ¹ × ‘ ° © ½ ¾ ë ٥ ' [] ” ـِ ـّ} {– : • = % ’ *, /: ! « ? » 9 8 7 6 5 4 3 2 1 0 ؟ - A B C D E F G H I J K L M N O P Q R S T U V W X Y Z a b c d e f g h i j k l m n o b q r s t u v w z y z > < | # @ $ + & ^ .) (¶.
2. The stoplist was defined to ignore the common words during the classification process. In this work, the stoplist is the common words among the eight categories (including all documents in the subcategories). Fig. 4 shows how to generate the stoplist. That is, we find the common words among the categories (including subcategories).
However, there are some words that are important to a particular category, and at the same time appear in all categories, such as the word:”اقتصاد” ➔”economy”. So, to avoid discarding the important words, any word that has more than 50% occurrences in a particular category is considered as an important word in that category, and therefore, removed from the stoplist. An example of words removed from the stoplist is “تكنولوجيا” ➔ “Technology”, in the Technology category. Removing words from the stoplist means considering the word as discriminative words and, hence, involved in the classification process.
3. After declaring the ignore characters and the stoplist, the words left normalized which include replacing some letters in other forms while maintaining the same meaning. In addition, all diacritics were eliminated. The replacement included: ا ➔آ,ا➔إ ,ا➔أ, ➔, ➔, ➔,ٔ ➔,ٓ ➔, ➔, ➔, ➔ . So, the diacritics replaced with nothing.
4. The Document Frequency (DF) textual feature is used that restrict the words to be added to the words dictionary. In this work, we chose DF as greater than 20. Only the words that appear in more than 20 different documents were added to the dictionary. Of course, a different DF threshold can be used.
5. The prepared dictionary was altered to remove two things: “ال” at the beginning of words and “ية” at the end of words. An example of removing “الابتكار” :”ال” ➔ “Innovation” is replaced to “ابتكار”, and an example of removing “الاجرامية” :”ية”➔” Criminal” is replaced to “اجرام”.

Step 2: Training

1. The previous step (preprocessing) ends up with the dictionary used in the training step. The Arabic alphabet characters used in this work include 34 characters that include:
{ا,آ,ب,ة,ت,ث,ج,ح,خ,د,ذ,ر,ز,س,ش,ص,ض,ط,ظ,ع,غ,ف,ق,ك,ل,م,ن,ه,و,ؤ,ى,ي,ء,ئ}

2. To create Markov chain matrices, a mapping process is used to denote a particular number for each Arabic character as follows:
{'ا':'0′,'آ':'1′,'ب':'2′,'ة':'3′,'ت':'4′,'ث':'5′,'ج':'6′,'ح':'7′,'خ':'8′,'د':'9′,'ذ':'10′,'ر':'11′,'ز':'12′,'س':'13′,'ش':'14′,'ص':'15′,'ض':'16′,'ط':'17′,'ظ':'18′,'ع':'19′,'غ':'20′,'ف':'21′,'ق':'22′,'ك':'23′,'ل':'24′,'م':'25′,'ن':'26′,'ه':'27′,'و':'28′,'ؤ':'29′,'ى':'30′,'ي':'31′,'ء':'32′,'ئ':'33'}

3. Each word in the dictionary is split for a sequence of characters. The sequence ends by “*” as indicator of the end of the word. The examples of words in the dictionary are shown in Fig. 5.

The sentence to be modelled:
"... مؤتمر المشروعات الصغيرة"
Using the dictionary, it translates to:
م ؤ ت م ر * م ش ر و ع ا ت * ص غ ي ر ة
Using the mapping table:
⇨ 25 29 4 25 11 * 25 14 11 28 19 0 4 * 15 20 31 11 27 *
The transitions:
25➔29, 29➔4, 4➔25, 25➔11, 25➔14, 14➔11, 11➔28, 28➔19, 19➔0, 0➔4, 15➔20, 20➔31, 31➔11, 11➔27
Each sequence is used as an index to add 1 in the corresponding cell in the transition matrix

Fig. 6. Finding transitions sequences of the training data.

	ا	آ	ب	ة	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص
ا	0.00	0.00	0.02	0.00	0.10	0.00	0.03	0.02	0.01	0.09	0.00	0.14	0.01	0.12	0.01	0.02
آ	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.01	0.00	0.00	0.00	0.00	0.63	0.00	0.00
ب	0.14	0.00	0.00	0.03	0.03	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.00	0.00	0.00
ة	0.05	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.05	0.00
ت	0.03	0.00	0.01	0.00	0.01	0.10	0.05	0.03	0.00	0.03	0.00	0.07	0.00	0.01	0.02	**0.23**
ث	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00
ج	0.18	0.00	0.06	0.02	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.02	0.09	0.01	0.00	0.00
ح	0.16	0.00	0.00	0.10	0.15	0.00	0.00	0.00	0.00	0.07	0.00	0.03	0.01	0.04	0.00	0.04
خ	0.25	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.21	0.01	0.03	0.00	0.03

Fig. 7. A part of the probability transition matrix of an economy documents.

4. For each category and subcategory, a probability transition matrix is created. The dimensions of the matrix is 34 × 34, the total number of unique characters appeared in the corpus. For each sequence occurrence, 1 is added in the corresponding cell in the transition matrix. An example how to fill the probability transition matrix is in Fig. 6, which has a short sentence on Economy.
5. An example of the probability transition matrix of the Economy category (first level) is shown in Fig. 7. The figure shows that the highlighted sequence ص➔ت that usually appears in the word "اقتصاد"➔"Economy" has a large weight compared to other sequences. It is also noted that the sequence س➔آ also has a large weight, this sequence mainly appears in the words "آسيا"➔ "Asia" which has large occurrences within the Economy category in the training set.

Step 3: Testing

1. The testing set (1480 documents) is used to evaluate the proposed method. As shown in the previous step (step 2), each document is converted to a list of sequences to be utilized in the transition matrices. We chose to find a document score by summation of the weights of sequence. As previously indicated, we use summation instead of multiplication to avoid handling small numbers.
2. Fig. 8 shows the scoring process. A testing document is initially compared with all probability transition matrices at the first level. Once finding the most related category, the process goes to the second level, and then to the third level.

The testing process is started by converting the query document into a sequence of characters. Of course, only the word in the dictionary contributes in the generated sequence. Hence, the query document sequence depends on the total words in the document that are also listed in the dictionary. Therefore, VSM is not employed in this method. An example of a generated sequence for a short sentence is shown in Fig. 9.

8. Experimental results

In this research, we present a novel approach for Arabic text classification. Therefore, it is necessary to compare the performance with other well-known text classification methods. Therefore, the well-known LSI method was used to validate the proposed method. The following subsections include a performance comparison with LSI, followed by the performance of hierarchical text classification using the proposed metho. We emphasize that we used Python for the proposed method and for generating the term-by-document LSI matrix. For singular value decomposition (SVD), we used Matlab.

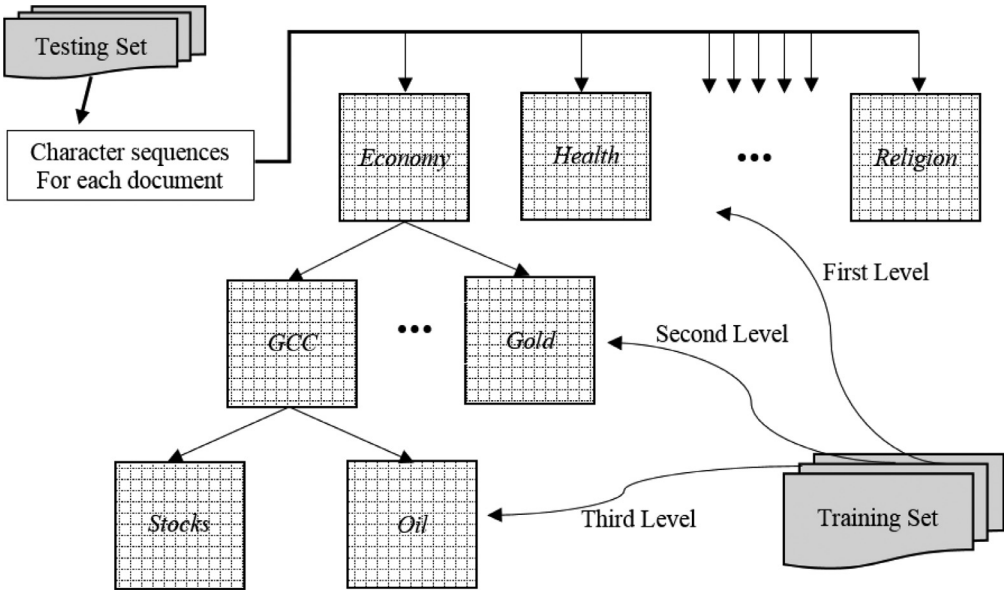

Fig. 8. Testing process.

A part of a document to be tested:
" ... ان سوق الكويت للاوراق المالية... "
The stop words discarded:
" ... ان سوق الكويت للاوراق المالية... "
The remaining words:
" ... سوق للاوراق المالية... "
س و ق * ل ل ا و ر ا ق * م ا ل *
Using the mapping table:
⇨ 13 28 22 * 24 24 0 28 11 0 22 * 25 0 24*
The score collected from these indices:
13➔28, 28➔22, 24➔24, 24➔0, 0➔28, 28➔11, 11➔0, 0➔22, 25➔0, 0➔24

Fig. 9. Finding transitions sequences of a testing document.

Table 3
Performance comparison of the Markov chain and the LSI method.

The proposed method (Markov Chain) (%)	The LSI method (%)
Precision = 87.98 Recall = 88.47 F1 measure = 88.23 Accuracy = 88.44	Precision = 85.31 Recall = 84.70 F1 measure = 85.00 Accuracy = 84.93

8.1. Validation of the proposed method

LSI is an information retrieval technique that is used to identify the relationships between words in the data collection. LSI, which is widely used in search engines, employs a linear algebraic method called SVD to extract semantic rich features for text classification purposes. For similarity measures, we used the cosine measure that also has been widely used with the LSI method. To validate the proposed method, we only compared the first level documents in the corpus. The first level contains 2978 documents distributed on the eight corpus topics as follows: (Economy: 418 documents, Health: 385 documents, Law-Justice: 341 documents, Education: 375 documents, Technology: 301 documents, Sport: 563 documents, Tourism: 226 documents, and Religion: 369 documents). The testing set contains 1480 document. The created dictionary contains 1739 unique words. The selected words are those words that appeared in more than 20 different documents (i.e DF > 20). The stoplist contained 10,744 words generated and described in Section 7(Step 1). The results shown in Table 3 indicates that the proposed method is a sign of future success for a space efficient text classification

Table 4
An economic article.

Apart of an article in Arabic	The translation using Google translator
"...وتقل بشكل كبير سرعة بيع العقار كلما ازداد سعره فهو يخرج من حيز السوق العام الى نطاق السوق الخاص وبالتالى يقل عدد المشترين المستهدفين حجم العقار وتعتبر العقارات الصغيرة الحجم المكونة من غرفة نوم وغرفتى نوم اكثر قبولا من المشترين المحليين الذين يشكلون اكبر سوق لاعادة البيع ويتفوق هذا الحجم من الشقق على الشقق والمنازل المكونة من ثلاث واربع غرف نوم فى سرعة البيع العوامل الخاصة بالمدينة قلة المشاريع العقارية فى المنطقة مقارنة بحجم الطلب جودة خدمات البنى التحتية فى المنطقة من كهرباء وماء وغاز ووسائل اتصال باحدث تقنيات الانترنت نمو قطاع الاعمال فى المدينة يؤثر ايجابيا فى الطلب على العقار وبالتالى يزيد من سرعة بيع العقار نمو قطاع التعليم الاكاديمى او الانشطة الاخرى كالاسواق او الاندية والحدائق..."	"... and less heavily selling the property the speed the greater the price is out of into the general market to the private market scope and thus fewer targeted buyers property size is small-sized real estate-bedroom and two-bedroom more acceptable to local buyers who make up the largest market for resale and outperform the size of the apartments on apartments and houses, consisting of three and four bedrooms in the sales for the city's lack of real estate projects in the region, factors speed compared to the demand for quality infrastructure services size in the area of electricity, water, gas and means of communication with the latest online techniques growth of the business sector in the city affect positive in demand for property and thus increases the speed of selling the property **academic education sector growth** or other activities such as markets or clubs ... "

$$\mathcal{E}_l = \frac{N}{N+z^2}\left(\hat{\mathcal{E}} + \frac{z^2}{2N} - z\sqrt{\frac{\hat{\mathcal{E}}(1-\hat{\mathcal{E}})}{N} + \frac{z^2}{4N^2}}\right) \qquad \mathcal{E}_u = \frac{N}{N+z^2}\left(\hat{\mathcal{E}} + \frac{z^2}{2N} + z\sqrt{\frac{\hat{\mathcal{E}}(1-\hat{\mathcal{E}})}{N} + \frac{z^2}{4N^2}}\right)$$

Fig. 10. Confidence interval calculation formula.

Table 5
The performance of the proposed method (three levels).

1st level performance (%)	2nd level performance (%)	3rd level performance (%)
Precision = 90.29, Recall = 90.69, F1 = 90.49, Accuracy = 90.29	Accuracy = 77.09	Accuracy = 63.33

method.

Both Markov chain methods and LSI has advantages and disadvantages. Markov chain does not consider words relationships in classification process. However, LSI does consider the strong relationships. In the case shown in Table 4, Markov chain successfully classified the document that belongs to the Economy category while the LSI wrongly classified it as educational document. From LSI point of view, the existence of educational words such as the bold underlined words "قطاع التعليم الاكاديمي" ➔" academic education sector" supports the decision that the query document belongs to the Education category. However, the existence of many economic words supported the weights and, therefore, the decision that this is an economic document.

To investigate whether the Markov chain method has significantly outperforms the LSI method; the performance detection method proposed by Plötz (2005) was used. The confidence interval $[\mathcal{E}_l, \mathcal{E}_u]$ has to be compute at the first place. Fig. 10 shows how to find the confidence interval. N is the number of documents in the Testing set (1480 documents). If the changed classification error rate is outside the confidence interval, these changes can be interpreted as statistically significant. Otherwise, they were most likely caused by chance. We used 95% as a level of confidence. We also used the error probabilities of the LSI method, as 15.00% (100% - 85.00%) as reported in Table 3. Since we used 95% as a level of confidence, z is equal 1.96 from the standard normal distribution. It might be interpreted as a 95% probability that a standard normal variable, z, will fall between -1.96 and 1.96.

The confidence interval is found to be [15.00% − 1.72%, 15.00% + 1.90%]➔ [13.28%, 16.90%]. The error probabilities of the proposed method 11.77% (100–88.23%). Hence, the Markov chain method is significantly outperforms the LSI method as the changed classification error rate is outside of the confidence interval (i.e 11.77% out of [13.28%, 16.90%]).

8.2. Performance for hierarchical text classification

The experimental results shows that the performance is relatively good in the first level while it is poor in the third level. This is natural, since the documents at the third level are more specific than the documents in the first level that are more general. For illustration, the accuracies of the Law-Justice > > Crime and sanction > > Embezzlement, Execution, Jail subcategories are extremely low since the documents contents are very similar. The accuracies are 0.35%, 55%, 37.5% for Embezzlement, Execution, Jail, respectively. In addition, the errors that occurred at the first level are spread out to the lower levels that might produce more errors in the deeper levels. Table 5 shows the performance achieved for the first level in terms of precision, recall, F1 value, and accuracy. Since the F1 measure and the accuracy are very close, we just measure the accuracies for the second and the third level.

Table 6
A Law-Justice article.

Apart of an article in Arabic	The translation using Google translator
" السجون العراقية تسمح للعائلات السعودية بزيارة ابنائها الرياض د ب ا قال محامى المعتقلين السعوديين فى العراق ان ادارة سجن الناصرية ترحب بزيارة العائلات السعودية لابنائها المعتقلين ونقلت صحيفة الوطن السعودية عن المحامى العراقى حامد احمد قوله ان ادارة السجن اكدت له عدم ممانعتها فى تمكين العائلات من زيارة ابنائها وتاتى تلك الانباء فى وقت تمر فيه العلاقات السعودية العراقية بمرحلة تحسن حملت انفراجا فى مسالة التواصل السياسى بين قيادتى البلدين وصولا الى بدء اولى الخطوات الفعلية لاعادة افتتاح سفارة الرياض لدى بغداد بعد اغلاق استمر ..."	"Iraqi prisons allow families of Saudi Arabia to visit her sons **Riyadh** (dpa) said Saudi detainees in Iraq, said the lawyer Nasiriyah prison administration welcomes the visit of Saudi families for their children detainees quoted Saudi Al-Watan newspaper, the Iraqi lawyer Hamid Ahmed as saying that the prison administration had assured him it has no objection to allow families to visit their children and the news comes at a time when the Saudi-Iraqi relations undergoing improvement brought a breakthrough in the issue of political communication between the leaderships of the two countries down to the actual start of the first steps to re-open an embassy **Riyadh** to Baghdad after the shutdown continued … "

Table 7
A misrecognized economic article.

Apart of an article in Arabic	The translation using Google translator
" اعتمد تعليمات معيار صافي التمويل المستقر للبنوك التقليدية والاسلامية المركزي يستكمل تطبيق حزمة اصلاحات بازل صرح محافظ بنك الكويت المركزي الدكتور محمد يوسف الهاشل بانه وفى اطار استكمال المعايير الصادرة عن لجنة بازل للرقابة المصرفية والمعروفة بحزمة اصلاحات بازل فقد اعتمد مجلس ادارة بنك الكويت المركزي بجلسته المنعقدة بتاريخ تعليمات معيار صافي التمويل المستقر لكل من البنوك التقليدية والبنوك الاسلامية بما في ذلك فروع البنوك الاجنبية العاملة في دولة الكويت ومشيرا الى ان مجلس الادارة كان قد اعتمد في شهر يونيو تعليمات معيار كفاية راس المال بازل بشكلها النهائي كما اعتمد في شهر اكتوبر تعليمات معيار الرفع المالي واعتمد ايضا في شهر ديسمبر تعليمات معايير تغطية السيولة ... "	"Adopted **instructions** standard net stable funding traditional Central and Islamic banks completed the application of reforms of the Basel package, said Governor of the Central Bank of Kuwait, Dr. Mohammad Yousuf Alhashl that in the framework of the completion of the Basel Committee on Banking Supervision, known bundle reforms Basel Standards Board of the Central Bank of Kuwait Directors held on was adopted Help **standard** net stable funding for both conventional banks and Islamic banks, including branches of foreign banks operating in the State of Kuwait, noting that the board of Directors had adopted in June **instructions** standard capital adequacy Basel in its final form, as adopted in October **instructions** leverage and adopted standard also in the month of December **instructions** cover the liquidity standards … "

9. Discussion

Despite the good performance achieved using the proposed method, however, it has some deficiencies. In fact, any text classification method should utilize the semantic information among the documents. Many studies illustrated the importance of using sematic rich methods for text classification. For example, Kantardzic (2011) indicated that LSI gives better results when used in text classification as it provides better representation of document's semantics. Eeven though the Markov chain method can handle some common sequences for different words, however, some cases shows misclassified results due to these common words. Table 6 shows a Law-Justice document that classified as Sport document. The reason is that the document contains a word "الرياض" ➔ "Riyadh" that has nothing related to the Sport category. But the sequences of the word are almost identical to the word "رياضة" ➔ " sport", which is a keyword in the Sport category. No doubt, Arabic langue is a rich language that has many (almost) similar words forms with different meaning. For illustration, the words "الرياض" is an example. It has the following forms: "الرياض" ➔ " Riyadh" the capital of Saudi Arabia", "الرياض" ➔ " paradises", "الرياض" ➔ " schools or classes that prepares children for first grade.", and "الرياضة" ➔ " sport" .

Another example of the deficiencies of the proposed method is shown in Table 7. An economy document was classified as an educational document. The article contains the word "تعليمات" ➔ " instructions" that has almost similar character sequence as the word "تعليم" ➔ " education" that leads to wrong classification result.

In fact, there are many cases like what we presented in Table 6 and Table 7. Hence, it is highly recommended to continue this work for supporting true semantic and words correlations.

Finally, this research opens a new direction to handle the textual features based on small words (i.e. the two consecutive characters). Using this method, no need to use more than 34 * 34 entries that is less than the typical VSM, which sometime contains thousands of words. Due to the use of small words, many different words but with the same concepts can be commonly represented, which increases the overall performance as we see in the obtained results.

10. Conclusion and future work

This work initiates a new research direction for space efficient text classification. We employed Markov chain for hierarchical Arabic text classification. The results are very encouraging as the proposed method is found to be better than the LSI method. The textual features generated in this work are semantic loss; it is worthy considering new semantic methods are based on the outcomes of this study. Finally, it is beneficial to conduct a thorough study to compare the time and the space complexities of the proposed method and other text classification methods. It is also worthy to consider adding log probabilities instead of adding the

probabilities. In this work, we compared the proposed method with the LSI, however, it might be better to consider other text classification methods. This work does not consider diacritized text; hence, we recommend to utilize the proposed method with diacritized text for different NLP tasks. More information about Arabic text challenges is found in Al-Anzi, Fawaz and AbuZeina (2015).

Acknowledgements

This work is supported by Kuwait Foundation of Advancement of Science (KFAS), Research Grant Number P11418EO01. The authors would like also to thank Kuwait University - Research Administration for its support of this research work.

References

Ahmed, Al-Falahi (2015). Authorship attribution in Arabic poetry. *2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA)*. IEEE.
Al-Anzi, FawazS., & AbuZeina, Dia (2015). Stemming impact on Arabic text categorization performance: A survey. *2015 5th International Conference on Information & Communication Technology and Accessibility (ICTA)*. IEEE.
Al-Anzi, FawazS., & AbuZeina, Dia (2017). Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. *Journal of King Saud University-Computer and Information Sciences, 29*(2), 189–195.
Alabbas, Waleed, Al-Khateeb, HaiderM., & Mansour, Ali (2016). Arabic text classification methods: Systematic literature review of primary studies. *Information Science and Technology (CiSt), 2016 4th IEEE International Colloquium on*. IEEE.
Alqabas. (2016). September). Retrieved from http://www.alqabas.com.kw/Default.aspx.
Baomao, Pang, & Haoshan, Shi (2009). Research on improved algorithm for Chinese word segmentation based on Markov chain. *Information Assurance and Security, 2009. IAS'09. Fifth International Conference on. 1*IEEE.
Bi, Wei, & Kwok, JamesT. (2012). Mandatory leaf node prediction in hierarchical multilabel classification. *Advances in Neural Information Processing Systems*.
Cai, Haixiao, Kulkarni, SanjeevR., & Verdú, Sergio (2006). An algorithm for universal lossless compression with side information. *IEEE Transactions on Information Theory, 52*, 4008–4016.
California Indian Education. (2016). Retrieved from http://www.californiaindianeducation.org/inspire/world/.
Chakrabarti, S., Dom, B., Agrawal, R., & Raghavan, P. (1997,). Using taxonomy, discriminants, and signatures for navigating in text databases. *VLDB. 97. VLDB* (pp. 446–455).
D'Alessio, S., Murray, K. A., Schiaffino, R., & Kershenbaum, A. (1998, June). Category Levels in Hierarchical Text Categorization. *EMNLP* (pp. 61–70). .
Dhillon, InderjitS., Mallela, Subramanyam, & Kumar, Rahul (2003). A divisive information theoretic feature clustering algorithm for text classification. *The Journal of Machine Learning Research, 3*, 1265–1287.
Dowman, Mike (2008). A probabilistic model of meetings that combines words and discourse features. *IEEE Transactions on Audio, Speech, and Language Processing, 16*, 1238–1248.
Erkan, Günes, & Radev, DragomirR. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research, 2*, 457–479.
Gao, Yu-Xiang, & Qi, De-Yu (2011). Analyze and detect malicious code for compound document binary storage format. *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on. 2*IEEE.
Godbole, Shantanu (2002). Exploiting confusion matrices for automatic generation of topic hierarchies and scaling up multi-way classifiers. *Annual Progress Report, Indian Institute of Technology–Bombay*.
Goyal, Anil, Jadon, MukeshK., & Pujari, ArunK. (2013). Spectral approach to find number of clusters of short-text documents. *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013 Fourth National Conference on*. IEEE.
Guerra-Gómez, J. A., Pack, M. L., Plaisant, C., & Shneiderman, B. (2015). Discovering temporal changes in hierarchical transportation data: Visual analytics & text reporting tools. *Transportation Research Part C: Emerging Technologies, 51*, 167–179.
Haji, Mehdi (2012). Statistical Hypothesis Testing for Handwritten Word Segmentation Algorithms. *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*. IEEE.
He, Miantao, Li, Miao, & Chen, Lei (2012). Mongolian Morphological Segmentation with Hidden Markov Model. *Asian Language Processing (IALP), 2012 International Conference on*. IEEE.
Joshi, Shweta, & Nigam, Bhawna (2011). Categorizing the document using multi class classification in data mining. *Computational Intelligence and Communication Networks (CICN), 2011 International Conference on*. IEEE.
Kantardzic, M. (2011). *Data mining: Concepts, models, methods, and algorithms.* New York: John Wiley & Sons.
Leon-Garcia, Alberto, & Leon-Garcia, Alberto (2008). *Probability, statistics, and random processes for electrical engineering.* Upper Saddle River, NJ: Pearson/Prentice Hall.
Li, Lishuang, Ding, Zhuoye, & Huang, Degen (2008). Recognizing location names from Chinese texts based on max-margin markov network. *Natural Language Processing and Knowledge Engineering, 2008. NLP-KE'08. International Conference on*. IEEE.
Meng, Peng (2009). Linguistic steganography detection algorithm using statistical language model. *Information Technology and Computer Science, 2009. ITCS 2009. International Conference on. 2*IEEE.
Osiek, BrunoAdam, Xexéo, Geraldo, & de Carvalho, LuisAlfredoVidal (2010). A language-independent acronym extraction from biomedical texts with hidden Markov models. *IEEE Transactions on Biomedical Engineering, 57*(1), 2677–2688.
Plötz, T. (2005). Advanced stochastic protein sequence analysis, PhD Thesis, Faculty of Technology, Bielefeld University.
Rabiner, LawrenceR. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*, 257–286.
Rodrigues, Erica (2013). Uncovering the location of Twitter users. *Intelligent Systems (BRACIS), 2013 Brazilian Conference on*. IEEE.
Ruiz, MiguelE., & Srinivasan, Padmini (2002). Hierarchical text categorization using neural networks. *Information Retrieval, 5*, 87–118.
Salton, Gerard, Wong, Anita, & Yang, Chung-Shu (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(1), 613–620.
Sampathkumar, Hariprasad, Chen, Xue-wen, & Luo, Bo (2014). Mining adverse drug reactions from online healthcare forums using hidden Markov model. *BMC medical informatics and decision making, 14*, 91.
Shen, JauJi, & Liu, KenTzu (2014). A Novel Approach by Applying Image Authentication Technique on a Digital Document. *Computer, Consumer and Control (IS3C), 2014 International Symposium on*. IEEE.
Silla, CarlosN., Jr, & Freitas, AlexA. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery, 22*(1-), 31–72.
Uysal, AlperKursat, & Gunal, Serkan (2014). The impact of preprocessing on text classification. *Information Processing & Management, 50*, 104–112.
Von Hilgers, Philipp, & Langville, AmyN. (2006). The five greatest applications of Markov Chains. *Proceedings of the Markov Anniversary Meeting*Boston Press.
Wong, S. K. Michael (1987). On modeling of information retrieval concepts in vector spaces. *ACM Transactions on Database Systems (TODS), 12*, 299–321.
Ying, Cao (2011). Novel top-down methods for Hierarchical Text Classification. *Procedia Engineering, 2*, 329–334.