Theme Classification of Arabic Text: A Statistical Approach

Leila Fodil, Halim Sayoud, Siham Ouamour

Theme Classification of Arabic Text: A Statistical Approach

Leila Fodil, Halim Sayoud, Siham Ouamour

USTHB University

fodilleila@gmail.com, halim.sayoud@uni.de, siham.ouamour@uni.de,

Abstract. The huge amount of textual documents that is stored in a lot of domains continues to increase at high speed; there is a need to organize it in the right manner so that a user can access it very easily. Text-Mining tools help to process this growing big data and to reveal the important information embedded in those documents. However, the field of information retrieval in the Arabic language is relatively new and limited compared to the quantity of research works that have been done in other languages (eg. English, Greek, German, Chinese …).
In this paper, we propose two statistical approaches of text classification by theme, which are dedicated to the Arabic language.
The tests of evaluation are conducted on an Arabic textual corpus containing 5 different themes: Economics, Politics, Sport, Medicine and Religion.
This investigation has validated several text mining tools for the Arabic language and has shown that the two proposed approaches are interesting in Arabic theme classification (classification performance reaching the score of 95%).

Keywords: Arabic Language, Text classification, Theme classification, Term weighting, TF-IDF.

1. Introduction

It is known that the volume of information available in Arabic on the World Wide Web and databases is continuously increasing. However, it has become almost impossible to take the full advantage of information embedded inside these documents without the help of various data mining techniques and tools.

Data mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decision (Gharib 2009). Data mining can be applied on a variety of data types (*structured data, multimedia data, free text, and hypertext*).

Text mining usually involves the process of structuring the input text (*usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database*), deriving patterns within the structured data, and finally evaluation and interpretation of the output [1]. Text mining is well motivated, due to the fact that much of the world's data can be found in text form (*newspaper articles, emails, literature, web pages, etc.*). Text mining tasks include text classification, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling

TC (*Text Classification or Categorization*) is the task of automatically structuring a set of text documents into different categories [2], from a predefined set, according to a group structure that is known in advance. Text classification techniques are used in many applications, including e-mail filtering, mail routing, spam filtering, news monitoring, sorting through digitized paper archives, indexing of scientific articles, classifying news by subject or newsgroup, and searching for interesting information on the web.

There are two main approaches for text classification: the manual approach, and the automatic approach. In the manual approach, human experts classify the document manually or use classifiers [3]. Although it gives quite accurate results, it is still very difficult to continuously update the information. Furthermore, it is expensive to maintain the classifier.
In the automatic approach, some well-known classification methods exist such as decision trees Support Vector Machines (*SVMs*), *K* Nearest Neighbor (*KNN*), Neural Networks (*NN*), Naïve Bayes (*NB*), Decision Trees (*DT*) Maximum Entropy (*ME*) *N-Grams* and Association Rules [3].
Many of the TC researches have been implemented and tested with the English language. In addition to the English language, there are many studies in other European languages such as French, German, Spanish and Asian languages such as Chinese and Japanese. However, there is little current research in the automatic classification of text documents in Arabic, due to the specific morphological and structural changes in the language: polysemy, irregular and inflected derived forms, various spelling of certain words, various writing of certain combination character, short (diacritics) and long vowels, most of the Arabic words contain affixes [4].
In this paper, we propose a method of Arabic text classification in terms of theme using a statistical approach and want to compare the impact of text preprocessing, which has not been addressed before, on the classification performances.
Our approach is based on the extraction of different keywords by two different methods and weighting schemes. In the first methods we use a different dataset for extracting automatically the keywords (*pertinent terms*), but in the second method, we use an in-house dictionary of keywords from each domain.

This paper is organized as follows: Section 2 presents an overview on related works, section 3 presents the proposed methodology, section 4 shows the experiments and results and finally a conclusion is given in section 5.

2. Related works

Different studies address the problem of text classification using different techniques to classify text documents, and different metrics to evaluate the accuracies of these techniques. We presents in this section a number of studies and experiments of text classification conducted on Arabic language using different datasets, different features and different classification algorithms.

For instance, Sawaf et al. [5] used a statistical approach based on the maximum entropy technique to classify the Arabic NEWSWIRE corpus of the Linguistic Data Consortium (LDC) of the University of Pennsylvania which covers four classes: politics, economics, culture and sports. The main objective was to simplify Arabic classification difficulties by

avoiding morphological analysis and to use subword units (character n-grams). The best accuracy reported was 62.7%.

Also, in the same research field, El Kourdi et al. [6] used Naïve Bayes algorithm to classify 300 web documents into five classes (health, business, culture, science and sport). The average accuracy achieved was 68.78% and the best accuracy reported was about 92.8%.

On the other hand, Syiam et. al. [7], in their experimental results, show that the suggested hybrid method of statistics and light stemmers is the most suitable stemming algorithm for Arabic language and gives a general accuracy of about 98%.

Mesleh, A.A. [8] have applied a support vector machine (SVM) to classify Arabic articles with Chi-Square feature selection. The corpus contains 1445 documents that vary in length. They are collected from online Arabic newspaper archives and divided into nine categories: Computers, Economics, Education, Engineering, Law, Medicine, Politics, Religion, and Sport.
They were focused on Feature Subset Selection (FSS) methods for TC tasks and, in particular, to the usage of an ant colony optimization algorithm to optimize the process of FSS. The author have compared Ant colony Optimization Based-Feature Subset Selection Method (ACO Based-FSS algorithm) with six state-of-the-art feature subset selection methods: Chi-square, Informafion Gain, Mutual Information, NGL coefficient, CSS score and Odds Ration in the classification of Arabic articles. Compared to the six classical FFS methods, the ACO Based-FSS algorithm achieved better text classification and was adapted to handle the large number of features in theme classification tasks.

In another attempt, Harrag and El-Qawasmah [3] applied neural networks (NN) on Arabic text. Their experimental results show that using NN with Singular Value Decomposition (SVD) as a feature selection technique gives better result (88.3%) than the basic NN (without SVD) (85.7%). They also experienced scalability problem with high dimensional text dataset using NN. Harrag collected his corpus from Hadith encyclopedia (موسوعة الحديثُ الشريفُ) from the nine books. It contains 435 documents belonging to 14 categories. He applied light stemming and stopwords removal on his corpus.

In 2010, Hammouda et al. [9] have proposed a system of indexing and classification of Arabic texts enabling the text classification using a heuristic approach. The Arabic dataset is collected from Internet, newspapers and magazines, the collected documents were 25 documents, to form manually a corpus composed of around 12500 words, 5 documents for each of the following 5 categories: sports, medicine, politics, economics and agriculture. The system performance was evaluated on the different chosen domains, achieving 90% of precision and 85% of recall.
In this investigation, we propose two theme classification methods: a Semi-Automatic method (which we called: SACM) and an Automatic method (which we called: ACM) that are tested on an Arabic corpus containing five different themes.

3. Proposed Methods

This section presents our overall methodology for Arabic text classification. Hence, in our approach, we have proposed two methods: a Semi-Automatic Categorization Method (SACM) where the set of keywords (*pertinent terms*) is constructed manually for each

domain and an Automatic Categorization Method (ACM) where the set of keywords is extracted automatically.

In our TC methods, the text document pass through five main steps: data processing, stemming, feature selection, classification and evaluation. Figure 1 shows the different steps of this approach.

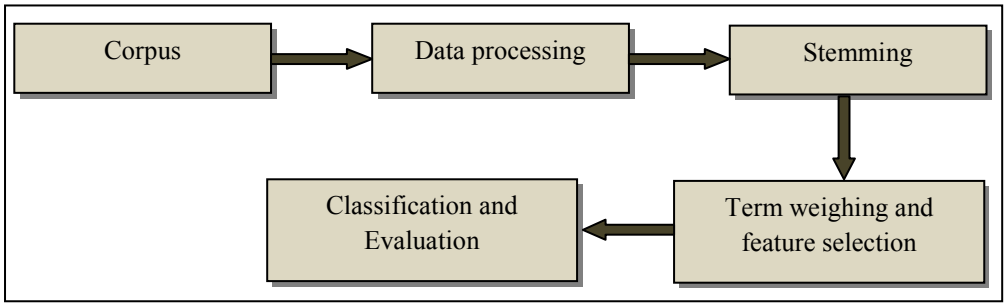

Figure 1: The different steps of our text classification approach.

3.1 Text Preprocessing

Preprocessing is actually a way to improve text classification by removing worthless information.
After converting our text corpus into UTF-8 encoding, in the first step, It is necessary to clean the texts by removing punctuation marks, diacritics, numbers, non Arabic letters [6], and removing kashida except in the term Allah.

In addition, Arabic texts need the following processes:

- Normalizing some writing forms that include “ ء” “ ى”, “ ة” to “ ا”, “ ي” and “ه. The reason for this normalization is that all forms of hamza are represented in dictionaries as one form and people often misspell different forms of aleph.

- A Tokenizer is employed. It is responsible for scanning each document, word by word, and extracting words (token) in this document. In this step, the Arabic Tokenizer uses white space tokenization because the space is the only way to separate words in Arabic language. The documents were tokenized to produce bag of words (*or vector of word*) for each text to classify.

- Removing Arabic function words (*stop words*) (*such as* ”على ”, “لكن” , “في ” *etc.*). Stop words are common terms that provide only a little of meaning and serve only as syntactic function without indicating any important subject or matter. The remove of stop words changes the document length and reduce the memory of the process.

3.2 Stemming

The main goal of the stemming is to improve the efficiency of the classification by reducing the number of terms being introduced to the classifier [10]. This step aims to make all the terms, which have the same root, in a unique form (*stem*) to facilitate the treatment.
We can take as an example the terms: 'اقتصادها'; 'اقتصادي'; 'الاقتصادية'; 'الاقتصادي'.
During the stemming operation, all those will be transformed into the term **اقتصاد** to calculate the frequency with an efficient way.

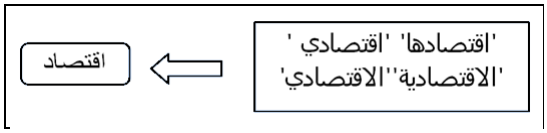

Figure 2: Example of stemming.

During this step, we compare every word (*or token*) of the text to classify with the list of predefined stems. If there is a correspondence, then the token will be replaced by the stem automatically.

3.3 Feature selection
3.3.1 Term weighting

After selecting the most significant terms in the super vector, each document is represented as a weighted vector of the terms found in this vector [11]. Term weighting is one of the important and basic steps in text classification based on the statistical analysis approach.
In our experiments, we attempt to use two weighting schemes to reflect the relative importance of each term in a document (*and a category*) and to reduce the dimensionality of the feature space. In the first time, we have calculated the relative frequency TF of a term (*denoted by* ***t***) in a text document (*denoted by* ***d***) by the following arithmetical formula:

TF (*t* , *d*)= [*number of occurrence of* ***"t"***] / [*number of word in* ***"d"***] (1)

In the second approach, we have used the TF-IDF: the Term Frequency (TF) of a word t is determined by the frequency of the word ***t*** in the document ***d***. The Document frequency (DF) of a word t is the number of documents in the dataset where the word ***t*** occurs in at least once. The inverse document frequency (IDF) of the word t is generally calculated as follows:

$$IDF = \log (N / n) \quad (2)$$

where ***N*** is the total number of documents in the dataset and ***n*** is the number of documents that contain the concerned word.

The weight of word ***t*** in document ***d*** using TF-IDF is:

$$TF\text{-}IDF = TF(t , d)* \log(N / n) \quad (3)$$

Thus, a term that has a high TF-IDF value must be simultaneously important in this document and must appear few times in the other documents. It is often the case where a term correlates to an important and unique characteristic of a document.

3.3.2 Feature Extraction

In this step, we propose two methods: in the first one called, which we called SACM (Semi-Automatic Classification Method), we use the local (*in-house*) dictionary; and in the second one, which we called ACM (Automatic Classification Method), we automatically extract this dictionary by a preliminary training from a specific dataset called ADTC1 dataset.

In the SACM method, we have constructed a dictionary for each theme and we have collected the specific keywords for each domain (*see table 1*).

Table 1: Number of SACM keywords per theme

Theme of the dictionary	Number of words
Economics	35
Politics	41
Sport	31
Medicine	31
Religion	56
Total	194

In the other method (ACM), this dictionary has been extracted automatically. We have used a corpus composed of 25 texts (*ADTC1*), which undergoes the same steps described previously (*processing, stemming and feature selection*).
After calculating the frequency (*or TF-IDF*) of all the terms and in every domain, we extract the most pertinent terms. We have kept only the terms for which the value of the frequency exceeds a certain threshold (*determined by experiment*). Finally, the pertinent terms for every domain will be grouped in different subsets.

3.4 Classification

The main goal is the assignment of an Arabic text to one or more predefined categories (*themes*) based on their content.
For that purpose, the Apparition Probability AP of a keyword ***k*** in each text ***d*** will be computed by:

AP (***k***)= [*number of occurrence of keyword* ***"k"***] / [*number of words in* ***"d"***] (3)

Once the probability of apparition is computed, we will be able to deduce the class of that text. Hence, the sum of apparition probabilities with regards to each domain *"**x**"* is determined.
This sum is called the Cumulative Thematic Probability or CTP.

CTP (***x***) = $\sum_k AP(k),\ \ where\ the\ keyword\ "k"\ \in domain\ \boldsymbol{x}$ (4)

According to the CTP probability, the system will classify the text according to the appropriate domain: in other words, the theme having the highest CTP should correspond to the real theme of the unknown text.

4. Experiments and results

4 .1 Dataset

The corpus used in this paper is collected from three sources: news (*Al-Jazeera, BBC, CNN, France24, Al-Ahram,..*), newspapers (*Al-Ahram, Al-watan, Elmoudjahid, Elkhabar Sport-Al-fagr..*), and books (الطاغية, الإنسان الوراثة, رياض الصالحين ..). It consists of five categories, which are sport, politics, economics, medicine and religion.

Two datasets have been built: ADTC1 (*Arabic Dataset for Theme Classification, subset 1*) and ADTC2 (*Arabic Dataset for Theme Classification, subset 2*).
ADTC1 is used in order to get automatically the keywords. There are 25 texts (*5 texts by theme*) from Internet and newspapers, corresponding to 38605 words.
ADTC2 is used for the testing. There are 150 texts (*30 texts by theme*) from three types of sources: internet, newspapers and books (*50 texts by source*), corresponding to 46060 words. Table 2 shows the number of documents present in the different categories.

Table2 : Specifications of the Arabic textual datasets ADTC1 and ADTC2.

Categories	**Number of texts**		**Number of words**	
	ADTC1	ADTC2	ADTC1	ADTC2
Economics	5	30	7786	9369
Sport	5	30	7705	9028
Medicine	5	30	7622	9303
Religion	5	30	7812	9017
Politics	5	30	7680	9343
Total	25	150	38605	46060

4.2 Evaluation Criterion

In order to evaluate our methods, a recognition score R is calculated for each category and each method. This recognition score measures the percentage of documents that are correctly assigned in each category.

4.3 Results

We recall that the primary steps in our Arabic text classification system is preprocessing the document, tokenization and stemming. Thereafter, the selection of the keywords (*features*) is performed.
Table 3 shows the number of selected features in the automatic method.

Table 3: Number of pertinent features that are selected

Categories	**Automatic method**	
	TF.IDF	frequency
Economics	34	31
Politics	45	30
Medicine	32	31
Sport	30	42
Religion	38	30
Total	179	164

Table 4 shows the different results obtained by the Automatic and Semi-Automatic methods and by using the two pertinence measures (*TF-IDF and relative frequency*). According to table 4, the Semi Automatic Method using TF-IDF presents the best performances with a score of about 95%, followed by the Automatic Method using the TF-IDF with a score of 92%. The other methods using the relative frequency come in last positions with a score not exceeding 88% of good classification.

Table 4: Scores of good classification by theme.

Categories	**Semi Automatic method**		**Automatic method**	
	TF.IDF	$Freq_R$	TF.IDF	$Freq_R$
Economics	93.33	83.33	93.33	93.33
Politics	96.66	93.33	83.33	93.33
Medicine	100	100	96.67	76.66
Sport	93.33	80	90	86.67
Religion	90	83.33	96.67	86.66
Global score	94.67	88	92	87.33

Figure 3 is a graphical representation of the precedent table (*table 4*).
We can see that, generally, the category medicine gets the highest score (*100%*); while the Religion category gets the lowest one (*83.33%*), in our case.

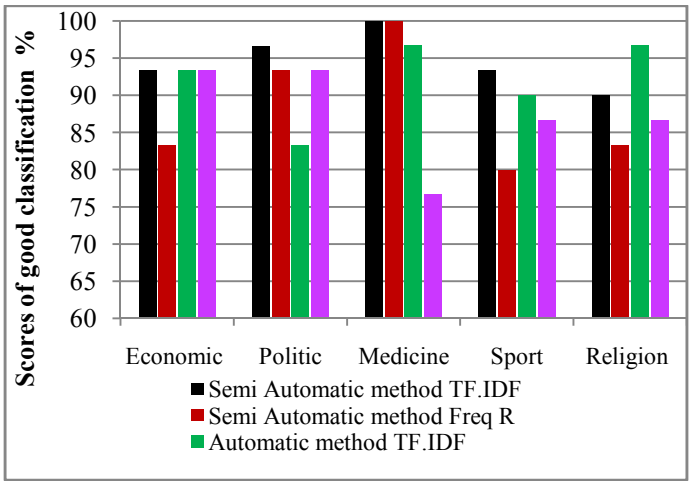

Figure 3: Scores of good classification by theme

In a second experiment, we have computed the performances of each method (SACM and ACM) by varying the number of keywords (*number of the most frequent consistent words that are kept for characterizing a theme*). The numbers of these keywords are taken as follows: 10, 20, 30, 40, 50…., until 160 (*see figure 4*).

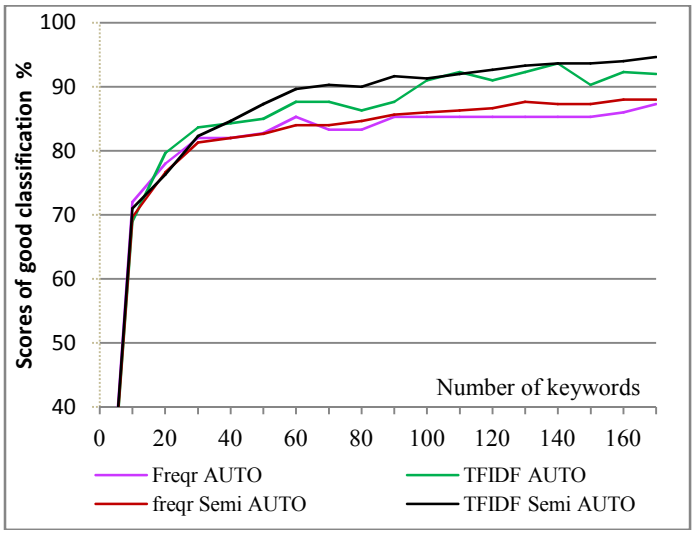

Figure 4: Score of good classification vs number of keywords selected.

Figure 4 shows that the semi automatic method using TF-IDF presents the best performances. We also notice that, when the number of features (*keywords*) increases, the classification accuracy increases too. An important point to note is that with 100 keywords (*and more*), the different methods begin to be powerful and accurate enough.

5. Conclusion

In this investigation, we have presented two approaches of Arabic text classification by theme. Five categories of themes were proposed and a special corpus has been built for that purpose.

During the text classification process, the document is coded into a vector of words (*bag of words*); this fact leads to a huge feature space and semantic loss. The proposed model in this paper adopts the keywords or "pertinent features" selected according to two approaches. These approaches extract these features statistically from the text and then the required theme is deduced from these selected features.

The comparison between the two proposed methods shows that the Semi-Automatic Method using TF-IDF achieves the best classification (*score of about 95%*), followed by the Automatic Methods using TF-IDF (*score of 92%*). On the other hand, the amount of time taken to build the dictionary of keywords is relatively greater for the semi automatic method, which makes the automatic method more interesting with regards to the time of execution.

These results show that there are some differences between these approaches according to two aspects, the classification score and execution time (*to build the models*).

Consequently, the user should decide which approach is suitable for him, according to his needs and constraints, before choosing the method to employ for the task of text categorization.
In perspective, we plan to expand the number of themes and increase the size of our Arabic textual dataset.

Reference

1. Morshed A., "Towards the automatic classification of documents in user-generated classifications*", PhD Thesis Proposal, University of Trento, Trento, Italy, Italy, January 2006*
2. Tripathi, N. Two-Level Text Classification Using Hybrid Machine Learning Techniques. PhD thesis, University of Sunderland, 2012.
3. Harrag F., El-Qawasmeh E., "Neural Network for Arabic text classification", In the *2nd Int. Conf. of Applications of Digital Information and Web Technologies, ICADIWT* '09, pp. 778 – 783, 2009.
4. Fridman, J.H.,"Data Mining and statistics: What's the connection?", *Stanford University.*
5. Sawaf, H., Zaplo, J., and Ney, H. "Statistical classification methods for Arabic news articles," *In Processing of the Arabic Natural Language Workshop (ACL2001), Toulouse, France.*
6. Elkourdi, M., Bensaid, A., and Rachid, T. ,"Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," *in Proc. of COLING 20th Workshop on Computational Approaches to Arabic Script-based Languages, Geneva, 2004.*
7. Syiam M., Fayed Z., Habib M., "An Intelligent System for Arabic Text Categorization", *In IJICIS, 6(1), pp. 1-19, 2006.*
8. Meslah, A.A.,"Support Vector Machine Text Classifier for Arabic Arctices: Ant Colony Optimization based feature subset selection", June ,2008.
9. Hammouda, F.K.,and Almarimi,A.A., "Heuristic Lemmatization for Arabic Texts Indexation and Classification", Journal *of Computer Science 6 (6):* 660-665, 2010, pp. 660-650.
10. Diabat, A. M. "Arabic Text Categorization Using Classification Rule Mining" , *Applied Mathematical Sciences, Vol. 6, 2012, no. 81,* pp. 4033 – 4046.
11. El-khair, I.A.,"Effects of stop Words Elimination for Arabic Information Retrieval: A comparative Study", *Internationnal Journal of Computing and Information Sciences, Vol.4,No.3,* on -Line ,December 2006, pp. 119-133.