Topic Identification of Arabic Texts based on Statistical Techniques.

L. Fodil, S. Ouamour
TDE, Laboratory of Speech Communication and Signal Processing.

***Abstract*—** One of the main factors that characterize a text is its content. Nowadays, the number of documents scattered online by private and public sectors are in the orders of millions. The rapid growth in the number of documents necessitates the use of automatic text classification. While a lot of effort has been put into manifold languages, minimal experimentation has been done with Arabic. Arabic language is highly inflectional and derivational language which makes text mining a challenging task.
In this paper, we propose two statistical approaches for topic identification. In the first approach we have developed two techniques ACM (Automatic Classification Method) and SACM (Semi-Automatic Classification Method) for the keywords extraction. In the second approach, we have used Centroid Classifier Models to classify the text documents by employing several distances (Euclidean, Manhattan, chebychev, etc.). The tests of evaluation are conducted on an Arabic textual corpus containing 5 different topics: Economics, Politics, Sport, Medicine and Religion. Results show the efficiency of the proposed approaches on topic identification.

Keywords— Arabic Language, Topic Identification, Text Categorization.

I. Introduction

Documents represent the principal repositories of knowledge and the most effective way to illustrate ideas, thoughts, and expertise (Khorsheed, 2013). Nowadays, the volume of document available on the World Wide Web and databases is increasing. The process of discovering and producing the hidden and useful information, embedded inside these documents, manually by domain experts is extremely hard and time consuming. This is since the numbers of online textual data are numerous and these data have large dimensionality. Therefore, using intelligent way such, as through text classification, to discover the benefit of the knowledge they contain

automatically from textual documents may give companies the right decisions that work for improving their competitive advantages (Diabat, 2012).
In this paper, we propose a method of topic identification on Arabic texts using statistical techniques. In the first stage, some techniques of keywords extraction is employed, by two different methods and weighting schemes. In the first method we use another dstinct dataset for the automatic extraction of the keywords, but in the seconds methods, we use an in-house dictionary of keywords from each domain. In the second stage, we have employed centroid based classification models with seven different statistical measures such as "Manhattan distance", "chebychev distance", "cosine distance", etc., in a purpose of Arabic texts classification.

II. Methodology

The methodology that has been adopted to develop the proposed system of Arabic Text Categorization (Duwairi, 2007) (Sebastiani, 2002) (ATC**)** is based on statistical techniques of automatic text categorization (TC). The main system consists in one block for feature selection and another one for the main step of statistical topic classification (Sawaf, 2001). We also recall that we proposed an automatic approach (ACM), a semi-automatic approach and a distance based technique.

III. Experiments and Results

The corpus used in this paper is collected from three sources: news (france24, BBC, Al-Ahram, CNN, Al-Jazeera..), newspaper (Al-Ahram, Al-watan, Elmoudjahid, Elkhabar Sport-Al-fagr..), and books (الطاغية, الوراثة , الإنسان, رياض الصالحين ..).
It consists of five different topics, which are sport, politics, economics, medicine and religion. Two datasets have been built, namely: ADTC1 and ADTC2.
ADTC1 is used in order to get automatically the keywords. There are 25 texts (5 texts for each topic) corresponding to 38605 words.
ADTC2 is used for the testing step, containing 150 texts (30 texts for each topic) corresponding to 46060 words.

In order to evaluate our methods, a recognition score R is calculated for each category and each method.

$$\text{Score R}=\frac{Number\ of\ \text{Correctly classified queries}}{\text{Total queries}} \quad (1)$$

Table 1 shows the different results obtained by the Automatic (ACM) and Semi-Automatic (SACM) methods using the TFIDF and TF_R (True Frequency).

Table 1: Scores of good classification by Topic in %.

Categories	SACM		ACM	
	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	100.00	100.00	96.67	76.66
Sport	93.33	80.00	90.00	86.67
Religion	90.00	83.33	96.67	86.66
Global Score	94.67	88.00	92.00	87.33

According to table 1, the Semi Automatic Method (SACM) using TF-IDF presents the best performances with a score of about 95%, followed by the Automatic Method (ACM) using the TF-IDF with a score of 92%. The other methods using the relative term frequency (TF_R) come in last positions with a score not exceeding 88% of good identification.
In the second experiment, we have computed the performance 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 topic). The numbers of these keywords are taken from 10 to 160 (figure 2).

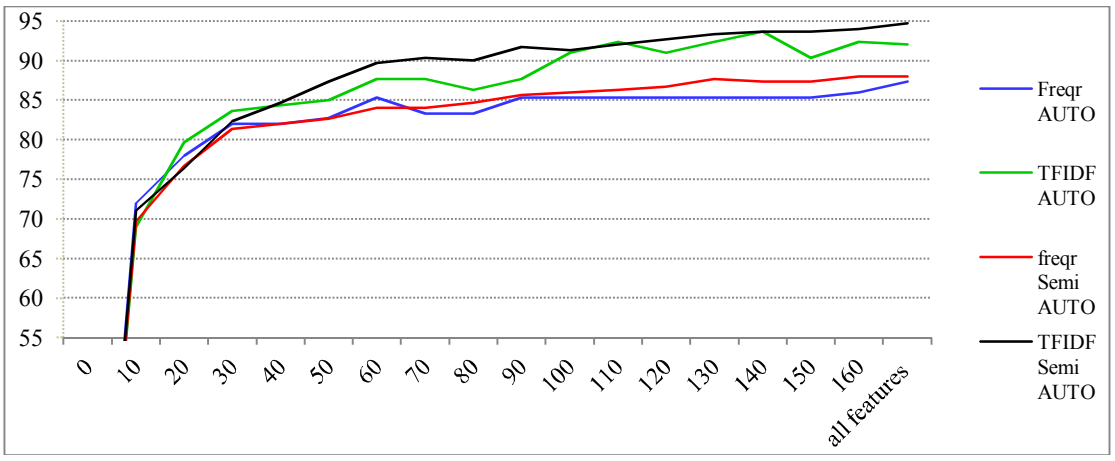

Figure 2: Score of good topic identification vs number of keywords.

Figure 2 shows that semi automatic method using TF-IDF presents the best performances. We also notice that, when the number of features (keywords) increases, the topic identification 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.

Table 2 shows the different results obtained by the Distance measures using the ACM Method.

Table2: Distance measures performances per topic

Topics	Euclidean	Manhanttan	Chebychev	Cosine	Correlation	Minkowski
Economics	100	75	88	92	100	81.55
Medicine	100	100	56	100	95	100
Politics	82.2	65	88.66	88	78.66	82.8
Sport	90.5	100	65	75	75.5	100
Religion	96	83.33	100	92.5	100	100
Global score	94.55	84.67	79.53	89.50	89.83	92.87

According to table 2, the Euclidean distance measure provides the best performances with a score of about 94.55%, followed by Minkowsky distance measure with a score of 92.87%. The other used distances come in last positions with a score not exceeding the 90% of good classification.

IV. 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 first proposed model in this paper adopts the keywords or “pertinent features” principle, which are selected according to two approaches. The proposed approaches extract those features statistically from the text and then the required theme is deduced from these selected features.
The comparison between the two proposed approaches shows that the Semi-Automatic Method using TF-IDF achieves the best classification (score of about 95%), followed by the Automatic Method 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 execution time. 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 to use according to his needs and constraints before choosing which method to employ.
In the second approach, we have used a centroid based classification technique and compared several similarity measures. The empirical results and analysis revealed that the proposed scheme for similarity measure is efficient and it can be used in real time applications. In

perspective, we plan to expand the number of themes and increase the size of our Arabic textual corpus.

V. References

Diabat,A. M. “Arabic Text Categorization Using Classification Rule Mining” , Applied Mathematical Sciences, Vol. 6, 2012, no. 81, 4033 – 4046.

Duwairi.R., “Arabic Text Categorization”, the international Arab Journal of Information Technology,Vol.4,No.2,pp 125-131, April,2007.

Khorsheed, M. S. and Al-Thubaity, A. O., “Comparative evaluation of text classification techniques using a large diverse Arabic dataset”, DOI 10.1007/s10579-013-9221-8, March, 2013.

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.

Sebastiani, F. , “Machine learning in automated text categorization” , ACM Publication: ACM Computing Surveys, Vol. 3(1) PP.1-47, 2002.