ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

Arabic Semantic Text Classification Based on Wavelet Spectral Analysis

Ibtissam El Hassani[1], Tawfik Masrour[2]

Doctoral Studies Center, Moulay Ismail University ENSAM, Meknes, Morocco [1]

Research Laboratory (M2.I), Mathematical Modeling for Analysis and Decision Making Research team (M2APD),
Moulay Ismail University ENSAM, Meknes, Morocco [2]

Abstract: We propose in this paper a new document representation in Text Mining based on signal representation and spectral processing by Wavelets Transform. Our method gives a solution of syntactic and semantic descriptor dependency problem, without deleting information. This can be done by grouping dependent descriptors in clusters with a single representative. Thereafter each class is represented by a discrete signal giving different degrees of dependence between descriptors, we then take advantage of the Multi Resolution Analysis properties of the Wavelet Transform. We show that we are able to achieve higher precision when compared to Vector Space Model representation and Latent Semantic Analysis in the context of Arabic Text Classification.

Keywords: Arabic TextMining; Signal Analysis; Wavelet Transform; Descriptors dependency

I. INTRODUCTION

The representation of texts in the current methods of information extraction, and Text Mining in general, does not always reflect the dependencies between descriptors. In Vector Space Model (VSM), related descriptors (with some degree) are often considered to be either totally independent or totally similar (for example by putting them in a canonic form). So this type of approach (VSM) can be considered as a “coarse” resolution of the document. Many current systems convert documents to vectors and consider only the occurrences of the words in the document. Accordingly, the similarities between documents are functions of occurrences of each descriptor, and it ignores wholly other information such as dependence between words.

Our model offers a new approach using a spectral representation of documents and wavelet transform for semantic classification. As in Latent Semantic Analysis we are interested in semantic dependence between descriptors of texts. We will present in the following paragraph related work namely spectral representation based on the positions of words in the text and Latent Semantic Analysis. The third paragraph presents the methodology and the mathematical modeling. In the fourth one we justified the choice of the wavelet and the measure of similarity between the documents. Finally, we give an example of implementation in the context of Arabic language and experiences conducted on a corpus extracted from Wikipedia.

II. RELATED WORKS

A new spectral method of Text Representation providing different levels of document resolution based on descriptor positions was proposed by Park & al [1]. Rather than observing a single resolution, it has multiple resolutions depending on the descriptor positions. That is possible due to the Wavelet Transform, which is able to decompose a given signal to wavelets of position. Thaicharoen & al have applied this representation to the classification of text where each descriptor is represented by signals based on the structure of documents [3]. Wavelet transform have been also used in visualization systems Text [2].

Our model introduces a new signal representation of “semantic dependent descriptors”. In general, dependent descriptors are considered completely different or completely similar by putting them for example in a canonical form. But the dependence is not always binary: there is some fuzzy relation between two words [4].

Latent Semantic Analysis (LSA) is one of the best known methods that offer a solution to this problem. The LSA method is based on the fact that words that appear in the same context are semantically close. The principle of the approach is to reduce the projecting dimension of the original matrix components in a vector space reduced. It concise to reduce the dimensionality by Singular Value Decomposition (SVD) : the Salton matrix $A = [aij]$ where aij is the frequency of occurrence of the word i in the context j, is decomposed into a product of three matrices

$U\Sigma V'$ where A and V are orthogonal matrices and Σ is a diagonal matrix.

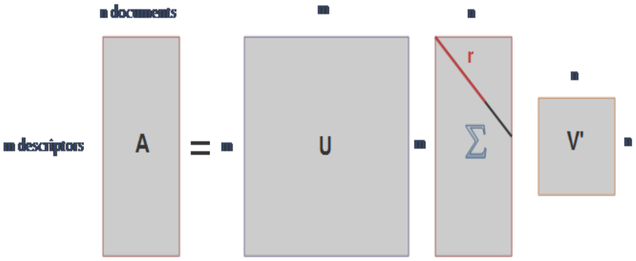


Fig. 1 Singular Value Decomposition (SVD). r is the rank of the matrix A

Our approach consists of representing documents in signals (of descriptors) that contain the dependency information. The signals are then analyzed by the wavelet transform. Relation of dependency $\Re$ is external to the corpus in our model. We propose first measuring the dependency in order to regrouping dependent descriptors in clusters, then each cluster is presented by one representative. The signal $s_{doc,desc}$ of descriptor $\mathrm{desc} \in \mathrm{Desc}$ in document $\mathrm{doc} \in \mathrm{Doc}$ is represented as follows:

$$s_{doc,desc} = [f_{doc,desc,0}, f_{doc,desc,1}, \dots, f_{doc,desc,B-1}] \quad (1)$$

Where $f_{doc,desc,b}$ is the value of the b^{th} component of the signal and B is the number of intervals of dependence. We calculate the value of the component $\mathrm{b} \in [0, \mathrm{B}-1]$ of the signal by counting the occurrences of desc such as $\Re(\mathrm{desc}, \mathrm{representative}) \in \left]\frac{\mathrm{bD}}{\mathrm{B}}, \frac{(\mathrm{b}+1)\mathrm{D}}{\mathrm{B}-1}\right]$, where $\mathrm{D} = |\mathrm{doc}|$ is total number of words in the document. Processing these signals by wavelets provides multi-resolution analysis according to the different "forms of appearances" of dependant words.

III. Methodology and Mathematical Modelling

A. The signal of descriptor based on dependency between descriptors

Models that have used the representation of signal descriptors were interested in the position of the descriptors in the text. The signal of descriptor is then a sequence of values which indicate the occurrence of a specific descriptor in a particular section of a document. Let Doc be the set of documents and let Desc be the descriptors of the corpus. Regarding the model that we propose, it is not the different positions in the text that interest us, but instead the different syntactic or semantic forms the descriptor may appear with. Let $\Re$ denote the function that associates to each pair of descriptors one value in $[0,1]$. This function measures the degree of dependence or similarity and can be estimated in several ways. Once this relation is defined and measured it is then possible to be considered as a dimension as well as the position of words in the text. We groups descriptors that have a dependency $\Re \in]0,1]$ in classes with a single representative descriptor (rep).

Example: Consider three documents where the words appear with different forms connected via a syntactic or semantic dependency. With the standard vector representation, no similarity will be detected between two documents because they have no common term.

TABLE I
An Example of Dependency Measure of Descriptors in three Documents

Desc	**Rep**	**$\Re$**	**Doc1**	**Doc2**	**Doc3**
word 1	rep1	**$\Re$(word1, rep1) = 0.5**	1	0	0
word 2		**$\Re$(word2, rep1) = 0.4**	0	1	0
word 3		**$\Re$(word3, rep1) = 0.9**	0	1	0
word 4		**$\Re$(word4, rep1) = 0.7**	0	0	1
word 5	rep2	**$\Re$(word5, rep2) = 1**	1	0	0
word 6		**$\Re$(word6, rep2) = 0.7**	0	1	0
word 7		**$\Re$(word7, rep2) = 0.6**	0	0	1
word 8		**$\Re$(word8, rep2) = 1**	0	0	1

$(\mathrm{word\ i})_{1\le i\le 4}$ are dependent and rep1 indicates the representative of their class. $(\mathrm{word\ i})_{5\le i\le 8}$ are dependent and rep2 indicates the representative of their class. We divide each measurement of dependence in B intervals (for example, if $\mathrm{B} = 4$ there will be 4 measurement intervals dependence $\Re \in]0,0.25]$, $\Re \in]0.25,0.5]$, $\Re \in]0.5,0.75]$, and $\Re \in]0.75,1]$). Then we calculate the value of the component b by counting occurrences of desc with dependency $\Re$ between $\frac{\mathrm{b}-1}{\mathrm{B}}$ and $\frac{\mathrm{b}}{\mathrm{B}}$.

TABLE II
Documents as a Function of Representatives

Rep	**Doc1**	**Doc2**	**Doc3**
rep1	$\Re$(word1, rep1)	$\Re$(word2, rep1) $\Re$(word3, rep1)	$\Re$(word4, rep1)
rep2	$\Re$(word5, rep2)	$\Re$(word6, rep2)	$\Re$(word7, rep2) $\Re$(word8, rep2)

Desc	**Doc1**	**Doc2**	**Doc3**
rep1	0.5	0.9 − 0.4	0.7
rep2	1	0.7	0.6 − 1

We represent then the descriptor as a signal which the values are the different "form of appearance". The weight of a descriptor in the document is no longer a scalar as was the case in the vector representation, but instead a discrete signal measured in degree of dependence. We then reduced the descriptor space since each class is represented by a single descriptor.

TABLE III
EXAMPLE OF SIGNAL REPRESENTATION OF DOCUMENTS

Signal	Doc 1	Doc 2	Doc 3
$s_{doc,\text{rep 1}}$	[0,1,0,0]	[0,1,0,1]	[0,0,1,0]
$s_{doc,\text{rep 2}}$	[0,0,0,1]	[0,0,1,0]	[0,0,1,1]

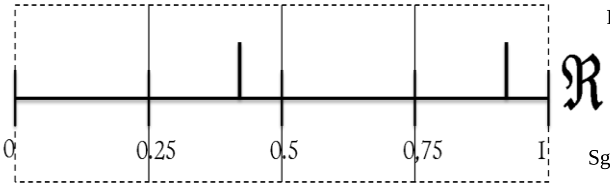


Fig. 2 Sgnal $s_{doc\,2,\text{rep 1}}$ in example of Table III

Note: If B = 1 (one signal component is required for each descriptor), we find the Vector Space Model.

The most convenient way to compare the signals is to by examining their wavelet spectrum, given by:

$\tilde{\zeta}_{\text{doc},\text{des}} = \left[\zeta_{\text{doc},\text{desc},0}, \zeta_{\text{doc},\text{desc},1}, \dots, \zeta_{\text{doc},\text{desc},B}\right]$ (2)

In the representation of documents as a position signal, several spectrum analyses were performed (Fourier transform, cosine transform …) [4] [5] [6] [7] [1]. These transforms decompose the signal into a series of sinusoids, they are able to extract information about the frequency of the signal, and however they focus only on the signal as a whole. The wavelet transform is able to focus on portions of the signal with different resolutions providing information on the frequency and the semantic form that the descriptor appears with.

B. Signal Wavelet Transformation

A Wavelet ψ is a function $\in L^2(\mathbb{R})$, s.t its average is zero and its energy is 1. The wavelet transform of a function $f(t) \in L^2(\mathbb{R})$ at the moment u and scale s is :

$$W(u,s) = \langle f, \psi_{u,s} \rangle = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s}\right) dt \quad (3)$$

With ψ^* is the complex conjugate of ψ. A wavelet can be scaled and translated by adjusting s and u , respectively [8].

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) \quad (4)$$

The advantage of the wavelet transform is its location on power parts of the signal. When comparing two or more signals having a similarity, we may consider the signals from top to bottom. When the signals differ to a certain level, we would know that we do not need to go further. This is an interesting property that can be used in information retrieval and TextMining in general.

C. Discrete Wavelet Transform

The continuous wavelet transform is an implementation of the wavelet transform using arbitrary scales and arbitrary wavelet. The wavelets used are not orthogonal and the data obtained by this transform are highly correlated. We may also use Wavelet Transform for discrete time series, with the limitation that the smaller translations of wavelets are equal to the sampling data. This is possible by the Multi Resolution Analysis (MRA) which is a recursive method generating a family of orthogonal wavelets.

A Multi Resolution Analysis (MRA) is a sequence $(\mathrm{V_j})_{\mathrm{j}\in\mathbb{Z}}$ of vectoriel spaces of $\mathrm{L}^2(\mathbb{R})$ satisfying:

- There are $\varphi \in \mathrm{V_0}$ as $\{\varphi(\mathrm{t}-\mathrm{n})\}_{\mathrm{n}\in\mathbb{Z}}$ representing a base $\mathrm{V_0}$.
- For $\in \mathbb{Z} : \mathrm{f(t)} \in \mathrm{V_j} \Leftrightarrow \mathrm{f}\left(\frac{\mathrm{t}}{2}\right) \in \mathrm{V_{j+1}}$.
- For $(\mathrm{j,k}) \in \mathbb{Z} : \mathrm{f(t)} \in \mathrm{V_j} \Leftrightarrow \mathrm{f(t-2^jk)} \in \mathrm{V_j}$.
- For $\mathrm{j} \in \mathbb{Z}$: $\mathrm{V_{j+1}} \subset \mathrm{V_j}$: $\{0\} \subset \cdots \subset \mathrm{V_2} \subset \mathrm{V_1} \subset \mathrm{V_0} \subset \mathrm{V_{-1}} \subset \mathrm{V_{-2}} \subset \cdots \subset \mathrm{L^2}(\mathbb{R})$ where $\lim_{\mathrm{j}\to-\infty} \mathrm{V_j} =$ L2ℝ and limj→+∞Vj= 0.
- The resolution of $\mathrm{V_j}$ is $2^{-\mathrm{j}}$:

$$\mathrm{f(t)} \in \mathrm{V_0} \Leftrightarrow \mathrm{f}\left(\frac{\mathrm{t}}{2^{\mathrm{j}}}\right) \in \mathrm{V_j} \quad (5)$$

$\mathrm{f(t)} \in \mathrm{V_j} \Leftrightarrow \mathrm{f}\left(\frac{\mathrm{t}}{2}\right) \in \mathrm{V_{j+1}}$ is a dilatation of 2 of $\mathrm{f(t)} \in \mathrm{V_j}$ gives approximation in $\mathrm{V_{j+1}}$ i.e. at the scale $2^{\mathrm{j+1}}$. Since $\mathrm{V_{j+1}} \subset \mathrm{V_j}$, an approximation to the scale $2^{\mathrm{j+1}}$ can also be obtained as an approximation at the scale 2^{j}.

IV. SIGNAL REPRESENTATION OF DOCUMENTS IN TEXT MINING

A. Choosing a Wavlet

Before applying a wavelet transform, we must first choose a wavelet from the many varieties that exist. To choose one, it is necessary to understand the properties that define each wavelet. The two main factors which will influence our choice of wavelet are the number of vanishing moments and the size of support. The k^th^ moment of a function $f(t)$ is defined as:

$$\nu_{\mathrm{k}} = \int_{-\infty}^{+\infty} \mathrm{t^k f(t)\, dt} \quad (6)$$

The support of a function f is the domain in which the function is nonzero. The wavelets with the smallest support size are the Haar wavelet [9] and the Daubechies-4 wavelet [10]. So we choose the Haar wavelet. The Haar wavelet's mother wavelet function $\psi(t)$ can be described as:

$$\psi(t) = \begin{cases} 1 & 0 \le t < 1/2 \\ -1 & 1/2 \le t < 1 \\ 0 & otherwise \end{cases} \qquad (7)$$

Its scaling function can be described as:

$$\varphi(t) = 1 \;\; if \;\; 0 \le t < 1, and\ 0\ otherwise \qquad (8)$$

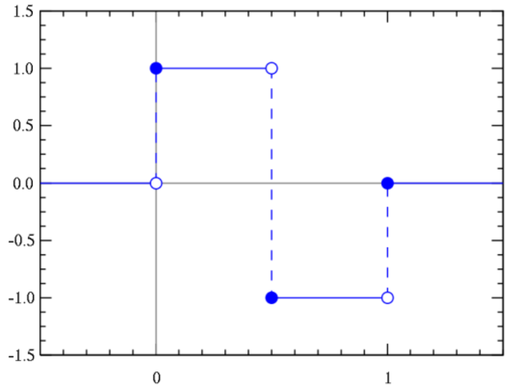


Fig. 3 The Haar scaling wavelet

We now give an example of how we can use the Wavelet Transform to provide different levels of resolution. Let us consider a signal $\mathrm{s}_{\mathrm{doc\,,desc}} = [2,0,0,1,1,1,0,0]$. Let W be the 8x8 Haar Wavelet Matrix:

$$\mathrm{W} = \begin{bmatrix} \sqrt{1/8} & \sqrt{1/8} & \sqrt{1/8} & \sqrt{1/8} & \sqrt{1/8} & \sqrt{1/8} & \sqrt{1/8} & \sqrt{1/8} \\ \sqrt{1/8} & \sqrt{1/8} & \sqrt{1/8} & \sqrt{1/8} & -\sqrt{1/8} & -\sqrt{1/8} & -\sqrt{1/8} & -\sqrt{1/8} \\ \sqrt{1/4} & \sqrt{1/4} & -\sqrt{1/4} & -\sqrt{1/4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{1/4} & \sqrt{1/4} & -\sqrt{1/4} & -\sqrt{1/4} \\ \sqrt{1/2} & -\sqrt{1/2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{1/2} & -\sqrt{1/2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{1/2} & -\sqrt{1/2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{1/2} & -\sqrt{1/2} \end{bmatrix} \quad (9)$$

Wavelet Transform of $\mathrm{s}_{\mathrm{doc\,,desc}}$ is :

$\mathrm{W}\overline{\mathrm{f}_{\mathrm{d,t}}}^{\mathrm{T}} = [5\sqrt{8}, 1\sqrt{8}\ , 1\sqrt{4}, 2\sqrt{4}, 2\sqrt{2}\,, -1\sqrt{2}\,, 0\,, 0\,]^{\mathrm{T}}$

With x^T is the transpose of the vector x. This transformed signal clearly shows the positions of the descriptors in many resolutions. The first component $(5/\sqrt{8})$ shows that there are five occurrences of the term. The second component $(1/\sqrt{8})$ shows that there is one more occurrence of the term in the first half of the signal than in the second half. The third component shows that there is one more occurrence of the term in the first quarter compared to the second quarter. The fourth component compares the third and fourth quarters. The next four components compare the eighths of the signal. Therefore, we can observe the signal at different levels of resolution by noting certain components of the transformed signal

B. *Spectral Similarity between Documents*

After transforming the documents, we can calculate the similarity between two documents by summing similarities descriptors for each resolution level, we use the cosine similarity:

$$\mathrm{SIM}\big(\tilde{\zeta}_{\mathrm{doc\,1}}\,, \tilde{\zeta}_{\mathrm{doc\,2}}\big) = \textstyle\sum_{\mathrm{b}=0}^{\mathrm{B}-1} \alpha_{\mathrm{b}} \frac{\overrightarrow{\tilde{\zeta}_{\mathrm{doc\ 1,b}}} \cdot \overrightarrow{\tilde{\zeta}_{\mathrm{doc\ 1,b}}}}{\left|\overrightarrow{\tilde{\zeta}_{\mathrm{doc\ 1,b}}}\right| \times \left|\overrightarrow{\tilde{\zeta}_{\mathrm{doc\ 1,b}}}\right|} \qquad (10)$$

Where $\overrightarrow{\tilde{\zeta}_{\mathrm{doc\,,b}}}$ is the vector containing the values of the component b of all descriptors, $\overrightarrow{\tilde{\zeta}_{\mathrm{doc\,,b}}}$ the norm and α_{b} coefficients giving levels of importance to different resolutions of the signal.

V. Experiments and Results

A. *Implementation in the Arab context*

The Arabic is an inflectional language. The derivation in Arabic is based on morphological patterns and the verb plays a greater inflectional role than in other languages. Furthermore, Arabic words are built-up from “roots” representing lexical and semantic connecting elements.

Let a, b be two words. The morphosyntactic analysis of a, b gives respectively the couples $\{x_1, y_1\}, \{x_2, y_2\} \in R_{corpus} \times S$ where R_{corpus} is the set of roots (جذور) in the corpus and S is the set of patterns (اوزان). We define a function that measures the dependence between these two words $f(\{x_1|y_1\}, \{x_2|y_2\}) = \mathbb{I}_{x_1=x_2} \times \psi(y_1|y_2)$ with $\mathbb{I}_{\mathrm{x}_1=\mathrm{x}_2}$ is the function that gives 1 if $x_1 = x_2$ and 0 otherwise. And $\psi(y_1, y_2)$ measures the semantic relation between the patterns $\mathrm{y}_1, \mathrm{y}_2$. Every two patterns $\mathrm{y}_\mathrm{i}, \mathrm{y}_\mathrm{j}$ have indeed a certain dependency $\psi_{ij} = \psi(\mathrm{y}_\mathrm{i}, \mathrm{y}_\mathrm{j}) \in [0,1]$. We can propose an automatic estimation using the difference between n_i, n_j the number of letters of prefixes and infix (without suffixes) between the two patterns [11] :

$$\mathfrak{R}(\mathrm{a,b}) = \psi_{\mathrm{ij}} = \frac{1}{1+|\mathrm{n_i}-\mathrm{n_j}|} \qquad (11)$$

This function satisfied the properties we need, namely $\psi_{ij} \in [0,1]$. If $\{x_1|y_1\} = \{x_2|y_2\} \Rightarrow f(a,b) = \psi_{ij} = 1$. Several other functions can also be proposed.

Example : Let us consider three documents which appear Arabic words with a syntactic and semantic dependency:

- الاستماع الى الدرس يساعد على فهمه.
- سمعت تقريرا عن دراسة تقنيات سماع نبضات الجنين.
- استمعت ودرست بتركيز ملخص دراسته.

Note that, in the vector representation, the documents were considered different because they did not share any common descriptor. And if we keep only the roots, they will then be considered completely equal and dependent and this is not the case. So we do a morphosyntactic analysis that will allow us to define classes of dependence with a single representative which is the root. It will also allow defining a measure of dependence between words.

TABLE IV
AN EXAMPLE OF DEPENDENCY MEASURE OF DESCRIPTORS IN THREE ARABIC DOCUMENTS

Desc	Root	Pattern S	Doc1	Doc2	Doc3
الاستماع	سمع	افْتِعَال	1	0	0
سمعت	سمع	فَعِلَت	0	1	0
سماع	سمع	فَعَال	0	1	0
استمعت	سمع	افْتَعَلْت	0	0	1
الدرس	درس	فَعْل	1	0	0
دراسة	درس	فَعَالَة	0	1	0
ودرست	درس	فَعَلْتُ	0	0	1
دراسته	درس	فَعَلْتُ	0	0	1

With the current representation of descriptors no similarity between each couple of documents will be detected, because they have no term in common. We will regroup the descriptors that have a dependency $\Re \in]0,1]$ and for each class we selected a representative which is the root of words of each cluster.

TABLE V
GROUPING DESCRIPTORS THAT HAVE A DEPENDENCY $\Re \in]0,1]$

representative	Document 1	Document 2	Document 3

rep1 = سمع	ℜ(سمع, الاستماع)	ℜ(سمع, سماع)
		ℜ(سمع, اسمعت)
			ℜ(سمع, استمعت)

rep1 = در	ℜ(درس, الدرس)	ℜ(درس, دراسة)	ℜ(درس, دراسته)
			ℜ(درست, درس)

We then reduced the descriptor space since each cluster is represented by the "root". Measuring dependencies between descriptors allow us to consider the dependence measure as a dimension. We divide measure of dependence in B intervals (for example, if B = 4 there will be 4 measurement intervals dependence $\Re \in]0,0.25]$, $\Re \in]0.25,0.5]$, $\Re \in]0.5,0.75]$, and $\Re \in]0.75,1]$), then we calculated the value of the component b by counting occurrences of desc with dependency with the representative of the class between $\frac{b-1}{B}$ et $\frac{b}{B}$.

TABLE VI
SIGNAL REPRESENTATION OF DOCUMENTS

Rep	Doc1	Doc2	Doc3
سمع	0.5	0.9 – 0.4	0.7
درس	1	0.7	0.6 – 1

signal	Doc1	Doc2	Doc3
s_{doc} سمع	[0,1,0,0]	[0,1,0,1]	[0,0,1,0]
s_{doc} درس	[0,0,0,1]	[0,0,1,0]	[0,0,1,1]

In the tables above, we presented only the occurrence of descriptors in each component of the signal b.

B. *Corpus Representation*

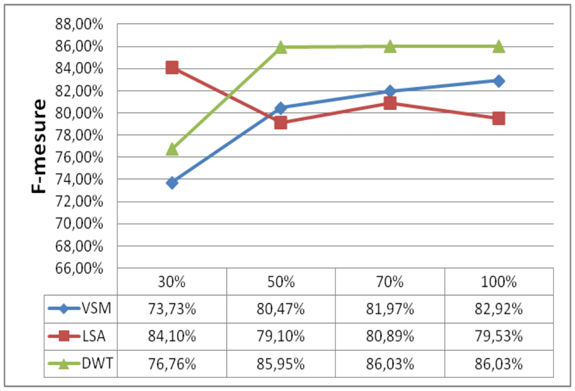
We build an Arabic corpus from Wikipedia we named WikipediaArabia2012. The size of the extracted corpus is 4856 documents, distributed over six topics, in this case:

Class A : Engineering هندسة تطبيقية ;
Class B : Philosophy of science فلسفة العلوم ;
Class C : Sociology علم الاجتماع ;
Class D : Mathematics رياضيات ;
Class E : Artificial intelligence ذكاء اصطناعي ;
Class F : Economy اقتصاد .

The corpus was divided into two subsets of documents. Where 90% of the corpus was dedicated to training and 10% of the overall documents formed the evaluation corpus.

Text pre–processing is the first step in a Text Classification. It aims to reduce the noise in documents by removing all the unnecessary terms and mistyped words along with transforming documents content from a plain text to a suitable form that can be easily handled by automated programs. The most important text pre-processing operations are:

1) Documents encoding unification: the encoding unification process ensures the same encoding for all the documents in the document collection. In our work we adopted the UTF–8 character set, which supports the characters of the Arabic language.



	30%	50%	70%	100%
VSM	73,73%	80,47%	81,97%	82,92%
LSA	84,10%	79,10%	80,89%	79,53%
DWT	76,76%	85,95%	86,03%	86,03%

2) Documents normalisation: suppression of symbols, numbers, markers, special characters, etc.

3) Normalization of certain Arabic characters: a/ Removal of diacritics : We have removed the following vowels: *Fatha, Damma, Kasra, Sukun, Shadda, double Fatha, double Damma*, and double *Kasra*. b/ Removal of *Tatweel* (Elongation of letters). c/ Normalization of *Hamza*: The following letters are converted to *Alef* by systematically removing the *Hamza* (*Alef Madda, Alef Hamza Aabove, Below Alef Hamza, Hamza Aabove, and Below Hamza*).

4) Stems extraction: In our work we used the Alkhalil [11] morphological analyser to generate a list of stems for each document. Alkhalil analysis each word in the documents and returns among other morphological information the word's possibly related stems, roots and patterns. We have also implemented a Viterbi algorithm to select, exclusively the stems that are relevant to the context.

5) Stop words elimination: elimination of noisy words by comparing each word with the elements of a handmade list of noisy words including: prepositions, demonstrative pronouns, identifiers, logical connectors, etc. Stop words do not carry any useful information and therefore their removal will not affect our process.

In order to evaluate our model based on Discrete Wavelet Transform we compare our model to the Vector Space Model and Latent Semantic Analysis in the task of text classification in Arabic context. the results are shown in Figure 4 and 5.

C. Results

We use three standard indicators: precision, recall and F–score.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (12)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (13)$$

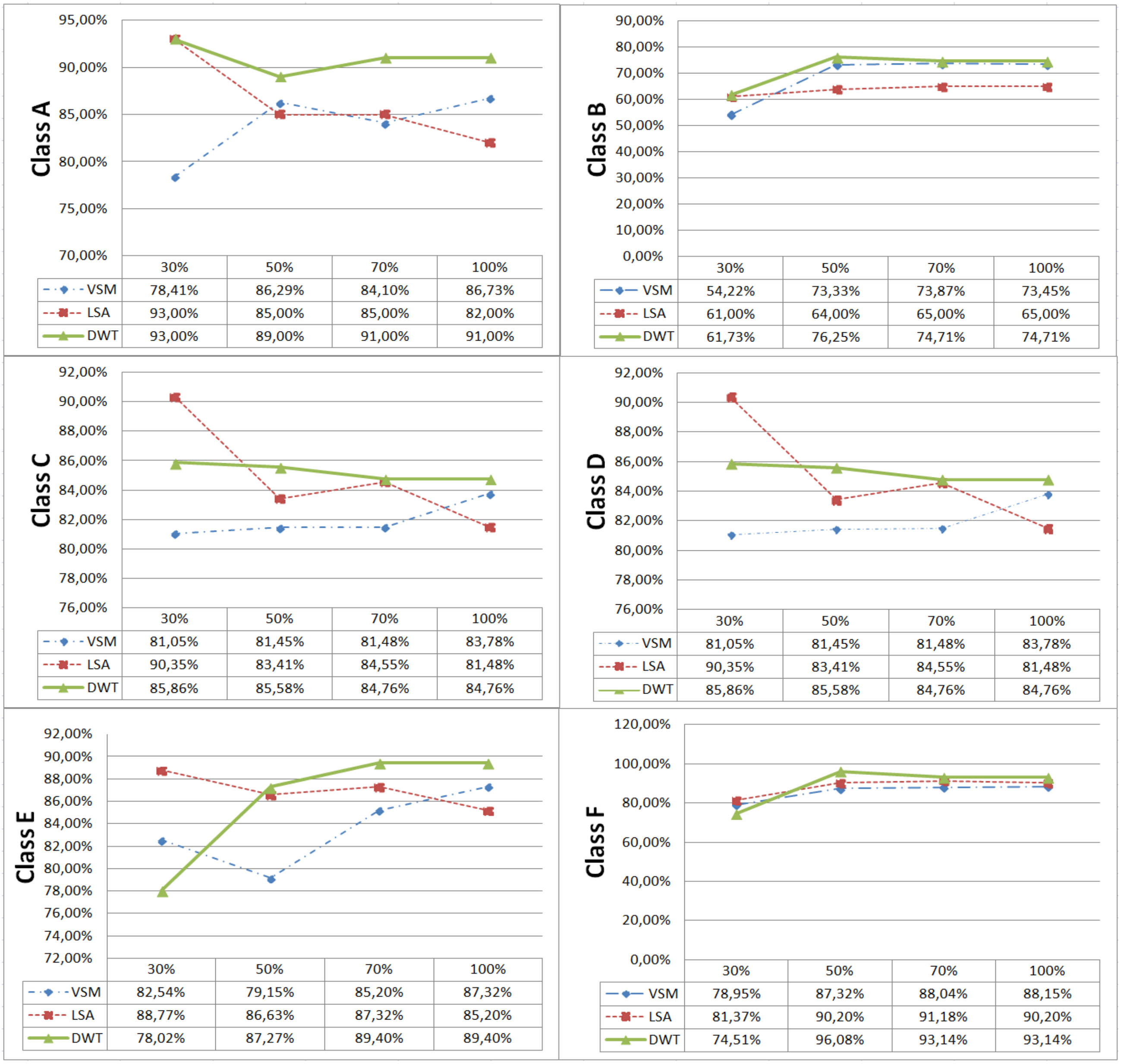
$$\text{F} - \text{score} = \frac{(1+\beta^2)\text{Precision} \times \text{Recall}}{\beta^2\text{Precision} + \text{Recall}} \quad (14)$$

The parameter β is set to 1 to provide the same importance to the recall and precision. The following table illustrates the four categories TP, TN, FP and FN.

TABLE VII
EVALUATION OF TEXT CLASSIFICATION

	Real class	Other classes
Predicted class	TP	FN
Other classes predicted	FP	TN

Fig. 4 Model performance comparison based on average of F-measure



Class A

	30%	50%	70%	100%
VSM	78,41%	86,29%	84,10%	86,73%
LSA	93,00%	85,00%	85,00%	82,00%
DWT	93,00%	89,00%	91,00%	91,00%

Class B

	30%	50%	70%	100%
VSM	54,22%	73,33%	73,87%	73,45%
LSA	61,00%	64,00%	65,00%	65,00%
DWT	61,73%	76,25%	74,71%	74,71%

Class C

	30%	50%	70%	100%
VSM	81,05%	81,45%	81,48%	83,78%
LSA	90,35%	83,41%	84,55%	81,48%
DWT	85,86%	85,58%	84,76%	84,76%

Class D

	30%	50%	70%	100%
VSM	81,05%	81,45%	81,48%	83,78%
LSA	90,35%	83,41%	84,55%	81,48%
DWT	85,86%	85,58%	84,76%	84,76%

Class E

	30%	50%	70%	100%
VSM	82,54%	79,15%	85,20%	87,32%
LSA	88,77%	86,63%	87,32%	85,20%
DWT	78,02%	87,27%	89,40%	89,40%

Class F

	30%	50%	70%	100%
VSM	78,95%	87,32%	88,04%	88,15%
LSA	81,37%	90,20%	91,18%	90,20%
DWT	74,51%	96,08%	93,14%	93,14%

Fig. 5 Model performance comparison on WikipediaArabia2012 data set based on F-measure

VI. Conclusion

Experiments conducted on Arabic scientific corpus shows that the spectral method based on wavelet analysis gives an average F-measure score higher in the task of document classification, namely 83.05%, when compared to Latent Semantic Analysis (82.46%) and Vector Space Model (79.77%). It is an alternative approach giving a sufficiently rich representation to capture the relationship between the objects described in the documents in addition of the occurrence.

However, our method could be computationally expensive. Indeed we use, in Arabic context, a measure based on the results of the automatic processing of natural language yet Natural Language Process operates in the order of a few words per second. It remains a challenge to see how the spectral semantic representation can be made much more efficient for very large text corpora.

For future work, our proposed technique could possibly be applied to other languages by defining a quantitative measure of similarity between two descriptors. It could also be applied to other types of Text Mining tasks such as text clustering and selection of concepts taking into account the descriptors dependency.

REFERENCES

[1] L. A. Park, K. Ramamohanarao and M. Palaniswami, " A novel document retrieval method using the discrete wavelet transform," *ACM Transactions on Information Systems (TOIS), 23(3),* pp. 267-298, 2005.

[2] S. Thaicharoen, T. Altman and K. J. Cios, "Structure-Based Document Model with Discrete Wavelet Transforms and Its Application to Document Classification," 2008.

[3] N. E. Miller, P. C. Wong, M. Brewster and H. Foote, " (). a wavelet-based text visualization system," in *In Visualization'98. Proceedings,* 1998, October.

[4] I. El Hassani, A. Kriouile and Y. BenGhabrit, "Measure of fuzzy presence of descriptors on Arabic Text Mining," *IEEE In Information Science and Technology (CIST),* pp. 58-63, 2012, October.

[5] L. Park, M. Palaniswami and R. Kotagiri, "Internet document filtering using fourier domain scoring," *Principles of Data Mining and Knowledge Discovery,* pp. 362-373, 2001.

[6] L. Park, M. Palaniswami and K. Ramamohanarao, "A novel web text mining method using the discrete cosine transform," *Principles of Data Mining and Knowledge Discovery,* pp. 385-397, 2002.

[7] L. A. Park, M. Palaniswami and K. Ramamohanarao, " A new implementation technique for fast spectral based document retrieval systems," in *In Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference,* 2002.

[8] L. A. Park, K. Ramamohanarao and M. & Palaniswami, "Fourier domain scoring: A novel document ranking method," *Knowledge and Data Engineering IEEE Transactions,* pp. 529-539, 2004.

[9] S. Mallat, A wavelet tour of signal processing, Academic press, 1999.

[10] H. A, "Zur theorie dur othogonalen funktionensysteme.," *Mathemat. Annalen 69,.HARMAN ,* p. 331–371, 1910.

[11] I. Daubechies, "Othonormal bases of compactly supported wavelets," *Pure Appl.Math. 41,* no. 1988, p. 909–996.

[12] A. Boudlal, A. Lakhouaja, A. Mazroui, A. Meziane, M. O. a. o. bebah and a. M. Shoul., "Alkhalil morphosys: Morphosyntactic analysis system for non ocalized arabic," in *Seventh International Computing Conference in Arabic,* 2011.

BIOGRAPHY

Ibtissam El HASSANI received her Master's Degree in Industrial Engineering in ENSAM (Ecole Nationale Supérieure d'Art et Métiers) My Ismail University, Meknes Morocco in 2006. She is currently pursuing his Ph.D. at Doctoral Studies Center in My Ismail University, Meknes Morocco, with interests in Mathematical Modeling for Analysis and Decision and Data Mining.

Dr Tawfik Masrour is working since 1998 as a Professor, Mathematical Modeling and Computing Research Laboratory (M2.I) , Mathematical Modeling for Analysis and Decision Making Research team (M2APD), Ecole Nationale Supérieure d'Arts et Métiers (ENSAM), My Ismail University, Meknes, Morocco. He received his Ph.D. in Applied Mathematics and Informatics, Ecole Nationale des Ponts et Chaussées ENPC-Paris-France (1995). He has done his M.phil in Modeling, Numerical Analysis and Scientific Calculus (D.E.A. Paris 6 and Ecole Polytechnique- 1992).He was Temporary Assistant Professor (ATER) in Paris 7 University-Paris-France (1995-1997) and also as invited Temporary Assistant Professor (ATER) in Franche Comté University - Besançon- France (2001- 2003). He Has about 20 years of experience of Teaching and research.