ARABIC TWEETS CATEGORIZATION BASED ON ROUGH SET THEORY

Mohammed Bekkali and Abdelmonaime Lachkar

L.S.I.S, E.N.S.A,University Sidi Mohamed Ben Abdellah (USMBA),
Fez, Morocco
bekkalimohammed@gmail.com, abdelmonaime_lachkar@yahoo.fr

ABSTRACT

Twitter is a popular microblogging service where users create status messages (called "tweets"). These tweets sometimes express opinions about different topics; and are presented to the user in a chronological order. This format of presentation is useful to the user since the latest tweets from are rich on recent news which is generally more interesting than tweets about an event that occurred long time back. Merely, presenting tweets in a chronological order may be too embarrassing to the user, especially if he has many followers. Therefore, there is a need to separate the tweets into different categories and then present the categories to the user. Nowadays Text Categorization (TC) becomes more significant especially for the Arabic language which is one of the most complex languages.

In this paper, in order to improve the accuracy of tweets categorization a system based on Rough Set Theory is proposed for enrichment the document's representation. The effectiveness of our system was evaluated and compared in term of the F-measure of the Naïve Bayesian classifier and the Support Vector Machine classifier.

KEYWORDS

Arabic Language, Text Categorization, Rough Set Theory, Twitter, Tweets.

1. INTRODUCTION

Twitter is a popular micro-blogging service where users search for timely and social information. As in the rest of the world, users in Arab countries engage in social media applications for interacting and posting information, opinions, and ideas [1]. Users post short text messages called tweets, which are limited by 140 characters [2] [3] in length and can be viewed by user's followers. These tweets sometimes express opinions about different topics; and are presented to the user in a chronological order [4]. This format of presentation is useful to the user since the latest tweets are generally more interesting than tweets about an event that occurred long time back. Merely, presenting tweets in a chronological order may be too embarrassing to the user, especially if he has many followers [5] [6]. Therefore, there is a great need to separate the tweets into different categories and then present the categories to the user. Text Categorization (TC) is a good way to solve this problem.

Text Categorization Systems try to find a relation between a set of Texts and a set of category.ies (tags, classes). Machine learning is the tool that allows deciding whether a Text belongs to a set

David C. Wyld et al. (Eds) : SAI, CDKP, ICAITA, NeCoM, SEAS, CMCA, ASUC, Signal - 2014
pp. 83–96, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41109

of predefined categories [6]. Several Text Categorization Systems have been conducted for English and other European languages, yet very little researches have been done out for the Arabic Text Categorization [7]. Arabic language is a highly inflected language and it requires a set of pre-processing to be manipulated, it is a Semitic language that has a very complex morphology compared with English. In the process of Text Categorization the document must pass through a series of steps (Figure.1): transformation the different types of documents into brut text, removed the stop words which are considered irrelevant words (prepositions and particles); and finally all words must be stemmed. Stemming is the process consists to extract the root from the word by removing the affixes [8] [9] [10] [11] [12] [13] [14]. To represent the internal of each document, the document must passed by the indexing process after pre-processing. Indexing process consists of three phases [15]:

a) All the terms appear in the documents corpus has been stocked in the super vector.

b) Term selection is a kind of dimensionality reduction, it aims at proposing a new set of terms in the super vector to some criteria [16] [17] [18];

c) Term weighting in which, for each term selected in phase (b) and for every document, a weight is calculated by TF-IDF which combine the definitions of term frequency and inverse document frequency [19].

Finally, the classifier is built by learning the characteristics of each category from a training set of documents. After building of classifier, its effectiveness of is tested by applying it to the test set and verifies the degree of correspondence between the obtained results and those encoded in the corpus.

Not that, one of the major problems in Text Categorization is the document's representation where we still limited only by the terms or words that occur in the document. In our work, we believe that Arabic Tweets (which are Short Text Messages) representation is challenge and crucial stage. It may impact positively or negatively on the accuracy of any Tweets Categorization system, and therefore the improvement of the representation step will lead by necessity to the improvement of any Text Categorization system very greatly.

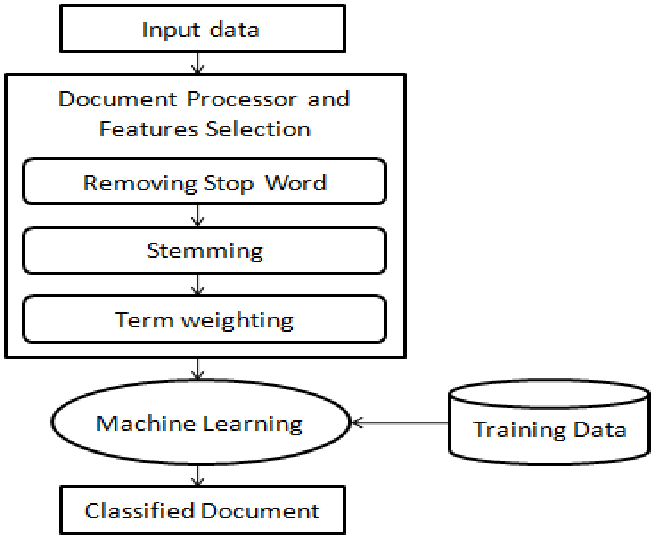


Figure .1 Architecture of TC System

To overcome this problem, in this paper we propose a system for Tweets Categorization based on Rough Set Theory (RST) [20] [21]. This latter is a mathematical tool to deal with vagueness and uncertainty. RST has been introduced by Pawlak in the early 1980s [20], it has been integrated in many Text mining applications such as for features selection [], in this work we proposed to use the Upper Approximation based RST to enrich the Tweet's Representation by using other terms in the corpus with which there is semantic links; it has been successful in many applications. In this theory each set in a universe is described by a pair of ordinary sets called Lower and Upper Approximations, determined by an equivalence relation in the Universe [20].

The remainder parts of this paper are organized as follows: we begin with a brief review on related work in Arabic Tweets Categorization in the next section. Section III presents introduction of the Rough Set Theory and his Tolerance Model; section IV presents two machine learning algorithms for Text Categorization (TC): Naïve Bayesian and Support Vector Machine classifiers used in our system; section V describes our proposed system for Arabic Tweets Categorization; section VI conducts the experiments results; finally, section VII concludes this paper and presents future work and some perspectives.

2. RELATED WORK

A number of recent papers have addressed the categorization of tweets most of them were tested against English Text [4] [30] [31]. Furthermore Categorization Systems that address Arabic Tweets are very rare in the literature [1]. This latter work realized by Rehab Nasser et al. presents a roadmap for understanding Arabic Tweets through two main objectives. The first is to predict tweet popularity in the Arab world. The second one is to analyze the use of Arabic proverbs in Tweets, The Arabic proverbs classification model was labeled "Category" with four class values sport, religious, political, and ideational.

On the other hand a wide range of Text Categorization based Rough Set Theory have been developed most of them were tested against English Text [39] [40]. Concerning Text Categorization Systems based on Rough Set that address Arabic Text is rare in the literature [41].

In Arabic Text Categorization we found Sawaf presented in [32] uses statistical methods such as maximum entropy to cluster Arabic news articles; the results derived by these methods were promising without morphological analysis. In [33], NB was applied to classify Arabic web data; the results showed that the average accuracy was 68.78%. The work of Duwairi [34] describes a distance-based classifier for Arabic text categorization. In [35] Laila et al. compared between Manhattan distance and Dice measures using N-gram frequency statistical technique against Arabic data sets collected from several online Arabic newspaper websites. The results showed that N-gram using Dice measure outperformed Manhattan distance.

Mesleh et al. [36] used three classification algorithms, namely SVM, KNN and NB, to classify 1445 texts taken from online Arabic newspaper archives. The compiled text Automated Arabic Text Categorization Using SVM and NB 125 were classified into nine classes: Computer, Economics, Education, Engineering, Law, Medicine, Politics, Religion and Sports. Chi Square statistics was used for features selection. [36] Discussed that "Compared to other classification methods, their system shows a high classification effectiveness for Arabic data set in terms of F measure (F=88.11)".

Thabtah et al. [37] investigate NB algorithm based on Chi Square feature selection method. The experimental results compared against different Arabic text categorization data sets provided evidence that features selection often increases classification accuracy by removing rare terms. In [38] NB and KNN were applied to classify Arabic text collected from online Arabic newspapers.

The results show that the NB classifier outperformed KNN base on Cosine coefficient with regards to macro F1, macro recall and macro precision measures.
Recently, Hadni et al. team [7] presents an Effective Arabic Stemmer Based Hybrid Approach for Arabic Text Categorization.
Note that, in any Text Categorization system the center point is the document and its representation that may impact positively or negatively on the accuracy of the system.

In the following section we present the Rough Set Theory, its mathematical background and also the Tolerance Rough Set Model which is proposed to deal with Text Representation.

3. ROUGH SET THEORY

3.1. Rough Set Theory

In this section we present Rough Set Theory that has been originally developed as a tool for data analysis and classification [20] [21]. It has been successfully applied in various tasks, such as features selection/extraction, rule synthesis and classification. The central point of Rough Set theory is the notion of set approximation: any set in U (a non-empty set of object called the universe) can be approximated by its lower and upper approximation. In order to define lower and upper approximation we need to introduce an indiscernibility relation that could be any equivalence relation R (reflexive, symmetric, transitive). For two objects x, y ϵ U, if xRy then we say that x and y are indiscernible from each other. The indiscernibility relation R induces a complete partition of universe U into equivalent classes $[x]_R$, x ϵ U [22].

We define lower and upper approximation of set X, with regards to an approximation space denoted by A = (U, R), respectively as:

$$L_R(X) = \{x \epsilon U: [x]_R \subseteq X\} \quad (1)$$

$$U_R(X) = \{x \epsilon U: [x]_R \cap X \neq \Phi\} \quad (2)$$

Approximations can also be defined by mean of rough membership function. Given rough membership function $\mu X: U \rightarrow [0, 1]$ of a set $X \subseteq U$, the rough approximation is defined as:

$$L_R(X) = \{x \epsilon U: \mu X(x, X) = 1\} \quad (3)$$

$$U_R(X) = \{x \epsilon U: \mu X(x, X) > 0\} \quad (4)$$

Note that, given rough membership function as:

$$\mu_X(x, X) = \frac{|\,[x]_R \cap X\,|}{|\,[x]_R\,|} \quad (5)$$

Rough Set Theory is dedicated to any data type but when it comes with Documents Representation we use its Tolerance Model described in the next section.

3.2. Tolerance Rough Set Model

Let D= $\{d_1, d_2\ldots, d_n\}$ be a set of document and T= $\{t_1, t_2\ldots, t_m\}$ set of index terms for D. with the adoption of the vector space model, each document d_i is represented by a weight vector $\{w_{i1}, w_{i2}\ldots, w_{im}\}$ where w_{ij} denotes the weight of index term j in document i. The tolerance space is defined over a universe of all index terms U= T= $\{t_1, t_2\ldots, t_m\}$ [23].

Let $f_{d_i}(t_i)$ denotes the number of index terms t_i in document d_i; $f_D(t_i, t_j)$ denotes the number of documents in D in which both index terms t_i an t_j occurs. The uncertainty function I with regards to threshold θ is defined as:

$$I_\theta = \{t_j \mid f_D(t_i, t_j) \geq \theta\} \cup \{t_i\} \quad (6)$$

Clearly, the above function satisfies conditions of being reflexive and symmetric. So $I_\theta(I_i)$ is the tolerance class of index term t_i. Thus we can define the membership function μ for $I_i \in T$, $X \subseteq T$ as [24]:

$$\mu_X(t_i, X) = v(I_\theta(t_i), X) = \frac{| I_\theta(t_i) \cap X |}{| I_\theta(t_i) |} \quad (7)$$

Finally, the lower and the upper approximation of any document $d_i \subseteq T$ can be determined as:

$$L_R(d_i) = \{t_i \in T: v(I_\theta(t_i), d_i) = 1\} \quad (8)$$
$$U_R(d_i) = \{t_i \in T: v(I_\theta(t_i), d_i) > 0\} \quad (9)$$

Once the documents handling is finished, the results will be the entry of any Text Categorization System. In the following section we present two of the most popular Machine Learning algorithms, Naïve Bayesian and the Vector Machine.

4. BASED MACHINE LEARNING

TC is the task of automatically sorting a set of documents into categories from a predefined set. This section covers two algorithms among the used known Machine Learning Algorithms for TC: Naïve Bayesian (NB) and Support Vector Machine (SVM).

4.1. Naïve Bayesian Classifier

The NB is a simple probabilistic classifier based on applying Baye's theorem, and its powerful, easy and language independent method. [25]

When the NB classifier is applied on the TC problem we use equation (10)

$$p(\text{class} \mid \text{document}) = \frac{p(\text{class}).p(\text{document} \mid \text{class})}{p(\text{document})} \quad (10)$$

where:

P (class | document): It’s the probability that a given document D belongs to a given class C
P (document): The probability of a document, it's a constant that can be ignored
P (class): The probability of a class, it’s calculated from the number of documents in the category divided by documents number in all categories
P (document | class): it’s the probability of document given class, and documents can be represented by a set of words:

$$p(\text{document} \mid \text{class}) = \prod_i p(\text{word}_i \mid \text{class}) \quad (11)$$

so:

$$p(\text{class} \mid \text{document}) = p(\text{class}).\prod_i p(\text{word}_i \mid \text{class}) \quad (12)$$

where:

$p(word_i \mid class)$: The probability that a given word occurs in all documents of class *C*, and this can be computed as follows:

$$p(word_i \mid class) = \frac{Tct + \lambda}{Nc + V} \quad (13)$$

where:

Tct: The number of times that the word occurs in that category C.
Nc: The number of words in category C.
V: The size of the vocabulary table.
λ: The positive constant, usually 1, or 0.5 to avoid zero probability.

4.2. Support Vector Machine Classifier

SVM introduced by Vapnik [26] and has been introduced in TC by Joachims [27]. Based on the structural risk minimization principle from the computational learning theory, SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set.

Given a set of N linearly separable points $N = \{x_i \in R^n \mid i = 1, 2\ldots N\}$, each point xi belongs to one of the two classes, labeled as $y_i \in \{-1, 1\}$. A separating hyper-plane divides S into 2 sides, each side containing points with the same class label only. The separating hyper-plane can be identified by the pair (w, b) that satisfies: $w.x + b = 0$

and:

$$\begin{cases} w.x_i + b \geq +1 \text{ if } y_i = +1 \\ w.x_i + b \leq -1 \text{ if } y_i = -1 \end{cases} \quad (14)$$

For i = 1, 2… N; where the dot product operation (.) is defined by:

$$w.x = \sum w_i.x_i$$

For vectors w and x, thus the goal of the SVM learning is to find the Optimal Separating Hyper plane (OSH) that has the maximal margin to both sides. This can be formularized as:

minimize:

$$\frac{1}{2} w.w$$

subject to

$$\begin{cases} w.x_i + b \geq +1 \text{ if } y_i = +1 \text{ for } i = 1, 2, \ldots, N \\ w.x_i + b \leq -1 \text{ if } yi = -1 \end{cases} \quad (15)$$

Figure 2 shows the optimal separating hyper-plane

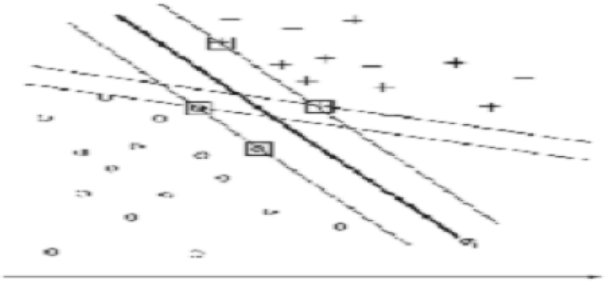


Figure .2 Learning Support Vector Classifier

The small crosses and circles represent positive and negative training examples, respectively, whereas lines represent decision surfaces. Decision surface σi (indicated by the thicker line) is, among those shown, the best possible one, as it is the middle element of the widest set of parallel decision surfaces (i.e., its minimum distance to any training example is maximum). Small boxes indicate the Support Vectors.

During classification, SVM makes decision based on the OSH instead of the whole training set. It simply finds out on which side of the OSH the test pattern is located. This property makes SVM highly competitive, compared with other traditional pattern recognition methods, in terms of computational efficiency and predictive accuracy [28].

The method described is applicable also to the case in which the positives and the negatives are not linearly separable. Yang and Liu [28] experimentally compared the linear case (namely, when the assumption is made that the categories are linearly separable) with the nonlinear case on a standard benchmark, and obtained slightly better results in the former case.

5. The Proposed System For Arabic Tweets Categorization

In this section we present in detail our proposed system for Arabic Tweets Categorization. The proposed system The proposed system contains two mains components, the first component generate the Upper Approximation for each Tweet to extend the Tweet's Representation by taking into account not just their words but also the words with which there is semantic links (Figure 3); the second one does the categorization using the Naïve Bayesian and the Support Vector Machine.

The treatment goes through the following steps: each Tweet in the corpus will be cleaned by removing Arabic stop words, Latin words and special characters like (/, #, $, ect…). After that for each word in corpus we make the following operations:

- Apply a stemmer algorithm to generate the root and eliminate the redundancy. too many algorithms have been proposed in this topic such as Khoja [11] and Light Stemmer [13] which are both the most known algorithms for Arabic text preprocessing; we used in this stage Khoja as a stemmer algorithm.

- Calculate the frequency in the document and also in the hole corpus.

- Determine the tolerance class of terms which contains all the words that occur with our word in the same document a number of times upper than θ. This tolerance class is defined using the formula (6).

After these operations we can calculate the approximations for each document by using the formula (8) for the lower and the formula (9) for the upper approximation. Then term weighting in which, for each term a weight is calculated by TF-IDF which combine the definitions of term frequency and inverse document frequency.

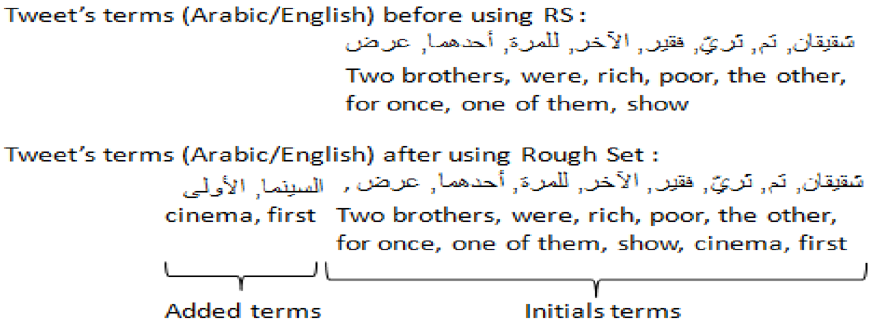

Tweet's terms (Arabic/English) before using RS :
شقيقان, تم, ثري, فقير, الآخر, للمرة, أحدهما, عرض
Two brothers, were, rich, poor, the other,
for once, one of them, show
Tweet's terms (Arabic/English) after using Rough Set :
شقيقان, تم, ثري, فقير, الآخر, للمرة, أحدهما, عرض ,
السينما, الأولى
cinema, first
Two brothers, were, rich, poor, the other,
for once, one of them, show, cinema, first
Added terms
Initials terms

Figure .3 Example of Rough Set Application

To illustrate the semantics links discovered by the Upper Approximation, Figure 3 presents an example of Arabic Tweet after the pre-processing; the initial tweet contains the word show / عرض we saw that the word cinema / السينما was added to the generated Upper Approximation because the two words are semantically related and often found with each other.

Finally, the classifier is built by learning the characteristics of each category from a training set of Tweets. After building of classifier, its effectiveness is tested by applying it to the test set and verifies the degree of correspondence between the obtained results and those encoded in the corpus.

6. EXPERIMENTS RESULTS

To illustrate that our proposed method can improve the Tweet's representation by using the Upper Approximation of RST and therefore can enhance the performance of our Arabic Tweets Categorization System. In this section a series of experiments has been conducted.

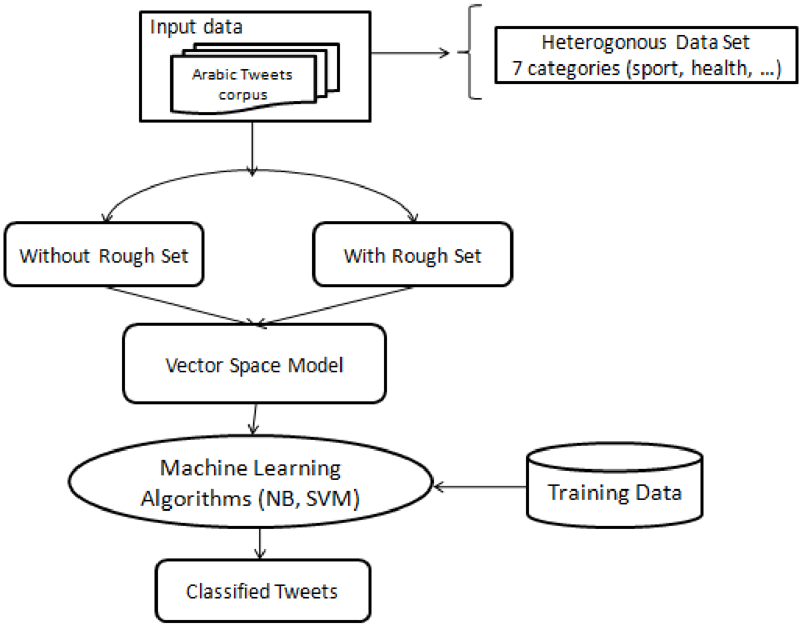


Figure .4 Descriptions of Our Experiments

Figure 4 describes our experiment with and without applying the Upper Approximation based Rough Set Theory.

The data set used in our experiments is collected from Twitter by using *NodeXL Excel Template* which is a freely Excel template that makes it super easier to collect Twitter network data [29]. This corpus is manually classified into six categories (Table 1). These categories are: Cinema/ السينما, News/الأخبار , Documentary/وثائقي , Health/الصحة , Tourism/السياحة and Economics/ الاقتصاد. Table 2 contains some examples of Tweets collected from Twitter.

The dataset has been divided into two parts: training and testing. The training data consist of 70% the documents per category. The testing data, on the other hand consist of 30% the documents of each category.

Table 1. The 6 categories and the number of Tweets in each one

Category	**Number of Tweets**
Cinema/السينما	79
Documentary/وثائقي	64
Economy/الاقتصاد	63
Health/الصحة	71
News/الأخبار	90
Tourism/السياحة	83
Total/المجموع	450

To assess the performance of the proposed system, a series of experiments has been conducted. The effectiveness of our system has been evaluated and compared in term of the F1-measure using the NB and the SVM classifiers used in our TC system.

F1-measure can be calculated using Recall and Precision measures as follow:

$$\text{F1-measure} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \tag{16}$$

Precision and Recall both are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP+FP})} \tag{17}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP+FN})} \tag{18}$$

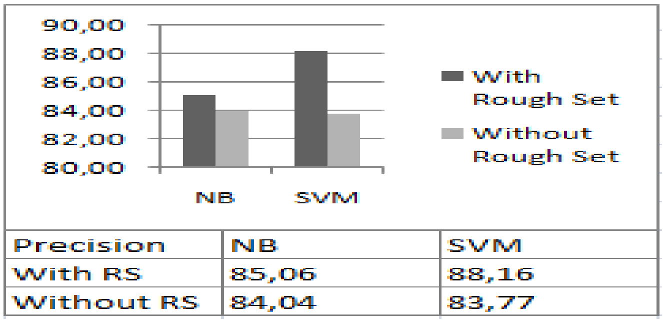
where:

- True Positive (TP) refers to the set of Documents which are correctly assigned to the given category.
- False Positive (FP) refers to the set of documents which are incorrectly assigned to the category.
- False Negative (FN) refers to the set of documents which are incorrectly not assigned to the category.

Table 2. Examples of the Tweets in the corpus

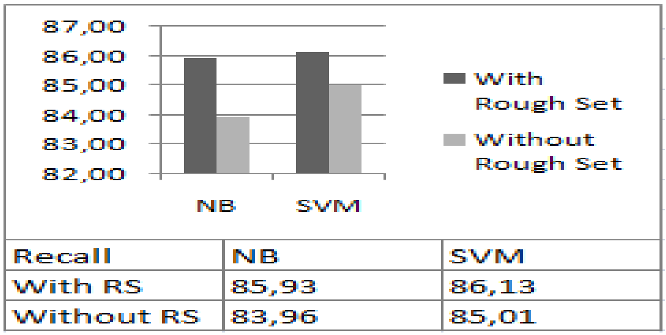
Category	Tweets in Arabic	Tweets in English
Cinema/السينما	تغرق في عالم يكتنفه الجنون و القتل و القوى الخارقة للطبيعة	Drowning in a world beset by madness, murder and supernatural powers
	القصة الحقيقية لمخطط اغتيال الرئيس ريتشارد نيكسون	The real story of the assassination's plot of the President Richard Nixon
Documentary/ وثائقي	تحت المجهر "حرب السدود": نهر النيل الآن على الجزيرة الوثائقية	Under the microscope "war dams": the Nile River is now on Al Jazeera Documentary
	بقي 6 أيام و يعرض البرنامج العالمي "الكون الكبير" على قناة ناشونال جيوغرافيك : أبو ظبي	6 days left and the global program the "big universe" will be displayed on the National Geographic Abu Dhabi Channel
Economy/ الاقتصاد	صعود الذهب إلى أعلى مستوياته ثلاثة أسابيع ونصف بعد موافقة واشنطن على توجيه جوية على العراق	The rise of gold to its highest level, three and a half weeks after the approval of Washington to direct flights on Iraq
	المبيعات الإجمالية لشركات الاسمنت تتراجع بنسبة 2% مقارنة بالشهر الماضي	Total sales of cement companies falls by 2% compared to last month
Health/الصحة	ذكرنا أعراض و علامات السكتة القلبية في تغريدتنا السابقة في تمام الساعة 10 صباحا	We mentioned symptoms and signs of heart attack in the previous Tweet at 10:00
	أخصائية تغذية لعمل نظام غذائي مرتفع السعرات ومناسب لك	Dietitian makes a high-calorie diet and suitable for you

News/الأخبار	سقوط قذيفة هاون افضى إلى إصابة محافظ الأنبار بجروح بليغة	A mortar shell fell led to the injury of the governor of Anbar by a seriously injuries
	حوارات و تقارير: في تونس سباق محموم للظفر بمنصب الرئيس	Dialogues and reports: in Tunisia a frantic race to win the post of president
Tourism/السياحة	الأماكن السياحية في تركيا: التفاصيل	Tourist places in Turkey: Details
	منتجع شاموني يعتبر أحد أكثر المنتجعات شهرة في فرنسا	Resort of Chamonix is one of the most famous resorts in France



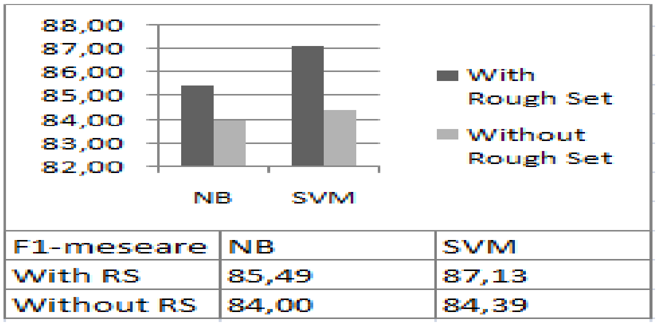
Precision	NB	SVM
With RS	85,06	88,16
Without RS	84,04	83,77

Figure .5 Representation of the Precision's Average



Recall	NB	SVM
With RS	85,93	86,13
Without RS	83,96	85,01

Figure .6 Representation of the Recall's Average



F1-meseare	NB	SVM
With RS	85,49	87,13
Without RS	84,00	84,39

Figure 7. Representation of the F1-measure's Average

Figure 7 shows the obtained F1-measure results with and without using the Rough Set Theory in our Arabic Tweets Categorization System. These results illustrate that using the Rough Set Theory enhances greatly the performance of Arabic Tweets Categorization.

Using the SVM classifier, applying RST performed better results for the five classes: Cinema (86, 4), Documentary (91, 9), Economy (90, 98), News (87, 52) and Tourism (88, 51) compared to the Categorization without using the RST: Cinema (85, 6), Documentary (84, 4), Economy (85, 35), News (85, 45) and Tourism (85, 54). But an average F1-measure the 87, 13 % with using the RST compared to 84, 39 % without using the RST.

To validate our results, we used another classifier which is the NB and the results was 85, 49 % with applying the RST compared to 84 % without using the RST.
The precision and the recall results showed in Figure 5 respectively in Figure 6 illustrate also that using the RST influence positively in the process of Arabic Tweets Categorization by enriching the Tweet representation with others terms that not occurred in and they have some semantic links with the Tweet's terms.

7. CONCLUSION AND FUTURE WORK

Tweets Categorization becomes an interest topic in recent years especially for the Arabic Language. Tweets Representation plays a vital role and may impact positively or negatively on the performance of any Tweets Categorization System. In this paper, we have proposed an effective method for Tweets Representation based on Rough Set Theory. This latter enriches and adds other terms which are semantically related with the original terms existing in the original Tweets. The proposed method has been integrated and tested for Arabic Tweets Categorization using NB and SVM classifiers.

The obtained results show that using the Upper Approximation of the Rough Set Theory increases significantly the F1-measure of the Tweets Categorization Systems.

In our future work, we will focus on using an external resource like Arabic WordNet or Arabic Wikipedia to add more semantic links between terms in Tweets Representation step.

REFERENCES

[1] Rehab Nasser, Al-Wehaibi*, Muhammad Badruddin, Khan "Understanding the Content of Arabic Tweets by Data and Text Mining Techniques", Symposium on Data Mining and Applications (SDMA2014)

[2] K. Lee, D. Palsetia, R. Narayanan, Md Patwary, A. Agrawal, A. Choudhary. "Twitter Trending Topic Classification", 11th IEEE International Conference on Data Mining Workshops 2011, 978-0-7695-4409-0/11

[3] A. Go, R. Bhayani, L. Huang. "Twitter Sentiment Classification using Distant Supervision", Processing (2009), S. 1—6

[4] Bharath Sriram. "Short Text Classification In Twitter To Improve Information Filtering". Computer Science and Engineering. Ohio State University, 2010

[5] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu. "Short Text Classification in Twitter to Improve Information Filtering", SIGIR'10, July 19–23, 2010, Geneva, Switzerland. ACM 978-1 60558-896-4/10/07.

[6] Sebastiani F. "Machine learning in automated text categorization". ACM Computing Surveys, volume 34 number 1. PP 1-47. 2002.

[7] M.Hadni, A.Lachkar, S. Alaoui Ouatik "Effective Arabic Stemmer Based Hybrid Approach for Arabic Text Categorization", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013

[8] Al-Fedaghi S. and F. Al-Anzi. “A new algorithm to generate Arabic root-pattern forms”. In proceedings of the 11th national Computer Conference and Exhibition. PP 391-400. March 1989.

[9] Al-Shalabi R. and M. Evens. “A computational morphology system for Arabic”. In Workshop on Computational Approaches to Semitic Languages, COLING-ACL98. August 1998.

[10] Aljlayl M. and O. Frieder. “On Arabic search: improving the retrieval effectiveness via a light stemming approach”. Proceedings of ACM CIKM 2002 International Conference on Information and Knowledge Management, McLean, VA, USA, 2002, pp. 340-347.

[11] Chen A. and F. Gey. “Building an Arabic Stemmer for Information Retrieval”. In Proceedings of the 11th Text Retrieval Conference (TREC 2002), National Institute of Standards and Technology. 2002.

[12] Khoja S.. “Stemming Arabic Text”. Lancaster, U.K., Computing Department, Lancaster University. 1999.

[13] Larkey L. and M. E. Connell. “Arabic information retrieval at UMass in TREC-10”. Proceedings of TREC 2001, Gaithersburg: NIST. 2001.

[14] Larkey L., L. Ballesteros, and M. E. Connell. “Improving Stemming for Arabic Information Retrieval: Light Stemming and Co occurrence Analysis”. Proceedings of SIGIR’02. PP 275–282. 2002.

[15]Sebastiani F. “A Tutorial on Automated Text Categorisation”. Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence. PP 7-35. 1999.

[16] Liu T., S. Liu, Z. Chen and Wei-Ying Ma. “An Evaluation on Feature Selection for Text Clustering”. Proceedings of the 12th International Conference (ICML 2003), Washington, DC, USA. PP 488-495. 2003.

[17] Rogati M. and Y. Yang. “High-Performing Feature Selection for Text classification”. CIKM'02, ACM. 2002.

[18] Yang Y., and J. O. Pedersen. “A comparative study on feature selection in text categorization”. Proceedings of ICML-97. PP 412-420. 1997.

[19] Aas K. and L. Eikvil. “Text categorisation: A survey”, Technical report, Norwegian Computing Center. 1999.

[20] Pawlak, Z. Rough sets: Theoretical aspects of reasoning about data. Kluwer Dordrecht, 1991.

[21] Jan Komorowski, Lech Polkowski, Andrzej Skowron “Rough Sets: A Tutorial”

[22] Ngo Chi Lang “A tolerance rough set approach to clustering web search results”

[23] Jin Zhang and Shuxuan Chen “A study on clustering algorithm of Web search results based on rough set”, Software Engineering and Service Science (ICSESS), 2013

[24] Ngo Chi Lang, “A tolerance rough set approach to clustering web search results”, Poland: Warsaw University, 2003.

[25] Saleh Alsaleem, “Automated Arabic Text Categorization Using SVM and NB”, International Arab Journal of e-Technology, Vol. 2, No.2, June 2011.

[26] Vapnik V. (1995). The Nature of Statistical Learning Theory, chapter 5. Springer-Verlag, New York.

[27] Joachims T. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," In Proceedings of the European Conference on Machine Learning (ECML), 1998, pp.173-142, Berlin

[28] Yang Y. and X. Liu, “A re-examination of text categorization methods,” in 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’99), pp. 42–49, 1999.

[29]http://social-dynamics.org/twitter-network-data/

[30]B. Sriam, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, “Short text classification in twitter to improve information filtering,” in Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010, pp. 841–842.

[31] D. Antenucci, G. Handy, A. Modi, M. Tinkerhess “Classification Of Tweets Via Clustering Of Hashtags”, EECS 545 FINAL PROJECT, FALL, 2011

[32] Sawaf, H. Zaplo,J. and Ney. H. "Statistical Classification Methods for Arabic News Articles". Arabic Natural Language Processing, Workshop on the ACL,2001. Toulouse, France.

[33] El-Kourdi, M., Bensaid, A., and Rachidi, T."Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," 20th International Conference on Computational Linguistics, 2004, Geneva

[34] Duwairi R., “Machine Learning for Arabic Text Categorization,” Journal of the American Society for Information Science and Technology (JASIST), vol. 57, no. 8, pp. 1005-1010, 2005.

[35] Laila K. “Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study,” DMIN, 2006, pp. 78-82.

[36] Mesleh, A. A. "Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System," Journal of Computer Science (3:6), 2007, pp. 430-435.

[37] Thabtah F., Eljinini M., Zamzeer M., Hadi W. (2009) Naïve Bayesian based on Chi Square to Categorize Arabic Data. In proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt 4 - 6 January. (pp. 930-935).

[38] Hadi W., Thabtah F., ALHawari S., Ababneh J. (2008b) Naive Bayesian and K-Nearest Neighbour to Categorize Arabic Text Data. Proceedings of the European Simulation and Modelling Conference. Le Havre, France, (pp. 196-200), 2008

[39] Y. Li, S.C.K. Shiu, S.K. Pal, J.N.K. Liu, "A rough set-based case-based reasoned for text categorization", International Journal of Approximate Reasoning 41 (2006) 229–255, 2005

[40] W. Zhao, Z. Zhang, "An Email Classification Model Based on Rough Set Theory", Active Media Technology 2005, IEEE

[41] Yahia, M.E. "Arabic text categorization based on rough set classification", Computer Systems and Applications (AICCSA), 2011 IEEE