In Proceedings of IADIS European Conference on Data Mining 2013 (ECDM'13), 22 – 24 July, Prague, Czech Republic

MINING SOCIAL NETWORKS' ARABIC SLANG COMMENTS

Taysir Hassan A. Soliman[1], M. Ali M.[2]
Information Systems Dept. Faculty of Computers & Information, Assiut University[1], Fayoum University[2]
Assiut[1], Fayoum2, Egypt
taysirhs@ieee.org[1], mostafa_elmasry2006@yahoo.com[2]

Abdel Rahman Hedar[3], M. M. Doss[4]
Computer Sciences Dept.[3], Electrical Engineering Dept.[4] Faculty of Computers & Information[3], Faculty of Engineering[4], Assiut University, Assiut, Egypt
hedar@aun.edu.com[3], magdy@aun.edu.eg[4]

ABSTRACT

Social networks have affected the way the new generation think all over the world, specifically in the Middle East. Their effectiveness appears in the revolutions of Tunisia, Egypt, and Syria. Therefore, social networks opinion mining for Arabic slang language has become an essential since it is widely used between the youth generation. Arabic slang language suffers from two main problems, which are the new expressive (opinion) words and idioms as well as the unstructured format. Mining Arabic slang language requires efficient techniques to extract youth opinions on various issues, such as news websites. In this paper, we propose a SVM-based classifier for Arabic slang language, applying sentiment analysis, to classify youth news' comments on Facebook. This classifier consists of three main phases: 1) Arabic comments' data preparation, 2) Data preprocessing, and 3) data classification. In addition, a Slang Sentimental Words and Idioms Lexicon (SSWIL) of opinion words is built, used by Arab youth in their comments on news topics, Facebook1 posts and comments, twits in Twitter2 and reviews. This paper works on users' comments and SSWIL enhances the classification task to be 86.86% of classified comments instead of 75.35% when using classical opinion words lexicon with precision 88.63 and recall 78 instead of 82.4 and 59.33 respectively.

KEYWORDS

Opinion mining, Social Network, sentiment analysis, support vector machines, Arabic Classification.

1. INTRODUCTION

In recent years, the use of Internet interest in interaction between the websites and its users. Social networking is the grouping of individuals into specific groups, like small rural communities or a neighborhood subdivision, if you will. Although social networking is possible in person, especially in the workplace, universities, and high schools, it is most popular online [1]. Most (Social Network Sites) SNSs also provide a mechanism for users to leave messages on their Friends' profiles. This feature typically involves leaving "comments," although sites employ various labels for this feature. In addition, SNSs often have a private messaging feature similar to webmail. While both private messages and comments are popular on most of the major SNSs, they are not universally available [2]. Reviews and comments play a vital role in the interaction between source of events and destination which are web users. Opinion mining (OM) includes several subtasks, such as subjectivity detection, polarity classification, review summarization, humor detection, emotion classification, and sentiment transfer [3]. Opinion mining, which is called sentiment analysis, can be viewed as a classification process that aims to determine whether a certain document or text is written to express a positive or a negative opinion about a certain object (e.g., a topic, product, or person). The process has been referred to as 'document level-sentiment classification', where the document is seen as

[1] www.Facebook.com
[2] www.Twitter.com

an opinionated artifact. The more fine-grained problem of identifying the sentiment of every sentence has also been studied [4]. Sentiment analysis is typically performed using one of two basic approaches: rule-based classifiers, in which rules derived from linguistic study of language are applied to sentiment analysis and machine learning classifiers [4]. Currently, most of the systems built for sentiment analysis are tailored for the English language [5] but there has been some work on other languages. In this paper, an opining mining methodology is proposed to classify Arabic slang comments, based on Support Vector Machine (SVM) for comments' classification to conclude the summary of the web users' opinions. The paper is organized as follows: Section two illustrates previous work. Section three explains the characteristics of slang Arabic language. Section four shows proposed slang sentimental words and idioms. Section five explains SSWIL phases. Section six discusses the results. Finally, section seven concludes our work and introduces future work in this track.

2. RELATED WORK

There are many researches in Arabic text mining. T. helmy and A. Daud [6] prove that Bayes Point Machine (BPM) is obvious than SVM in classification task. The authors apply the two methods in trustworthy and untrustworthy classification of Islamic Hadith Narration. A. El-Halees [7] combine classification methods to classify Arabic documents because the accuracy of most methods is law. S. AbdelRahman *et al.* [8] present a novel solution for Arabic Named Entity Recognition (ANER) problem, which aims to boost the identification of extracted named entities. They utilize a machine learning technique use pattern recognition to classify name entities (NE). M. Elarnaoty *et al.* [9] based their feature analysis on a semi-supervised pattern recognition technique to extract opinion holder in Arabic news. M. R.S and M. Teresa [5] apply SVM and NB classifiers to identify the polarity of web users' comments. They translate collected Opinion Corpus for Arabic (OCA) into English version OCA (EVOCA) and apply classification task. M. Elhawary and M. Elfeky [10] apply sentiment analysis on Arabic reviews to extract features by using Arabic lexicon words to identify reviews' polarity (positive, negative or neutral). Omar Zaidan and Chris Callison-Burch [11] work on Arabic Commentary Dataset (AOC) to apply dialect labeling task. They partition news' comments into dialectal sentences which is more accurate than Modern Standard Arabic sentences. Motaz K. Saad and W. Ashour [12] evaluate text preprocessing on Arabic text mining, stemming and pruning, document normalization and term weighting and enhance text representation. They study the impact of text-preprocessing and different term weighting schemes on Arabic text classification. They apply Boolean model, TF, IDF and TF-IDF for term weighting and apply C4.5 decision tree in classification task. M. Hijjawi and Z. Bander [13] represent an approach of identification of opinions based on ontological exploration of texts. Lexicon of emotions is used in extracting Elementary Opinions Units (EOUs) and the authors use supervised classification techniques to identify opinions polarity. Most of these researches work on structured text or modern Arabic text but not work on unstructured, ungrammatically Arabic. Our research focuses on free text Arabic written by web users who comment on Facebook and Twitter or news websites. Comments on those networks are written with new sentiment words and idioms. So, they need new lexicon to facilitate feature extraction and sentiment analysis. In the next section, formal and slang Arabic language will be discussed.

3. FORMAL AND SLANG ARABIC LANGUAGE

As the official language of 22 countries, Arabic is spoken by more than 300 million people, and is the fastest-growing language on the web (with an annual growth rate of 2,501.2% in the number of Internet users as of 2010, compared to 1,825.8% for Russian, 1,478.7% for Chinese and 301.4% for English) [14]. There are about 65 million Arabic-speaking users online, or about 18.8% of the global Internet population [14]. Arabic is divided into three types: Classical Arabic, Modern Arabic, and Colloquial Arabic [13]. Classical Arabic is the language used during the period before and during the Islam era and in which the holy Quran was subsequently written. It contains a rich vocabulary and sophisticated grammar. Modern Arabic is derived from the classical type and nowadays it became the formal language for literature, media, and education. It has less sophisticated grammar and it is the language type that this paper is targeting. Finally, Colloquial Arabic is the language used between people in every day communication and it varies between countries, each country has its own dialect [13]. Modern Arabic, is different from Latin based languages because Arabic

is highly inflectional and derivational language. Our work based on Colloquial (Slang) Arabic which written in free text with new sentiment words and idioms.

3.1 New Tools exiles Analysis

Tools exiles in formal Arabic language are " لا، لن، لم، ما، ليس، ليست", which are "la, ln, lm, ma, lays, laysat" mean "No, Not" and negative suffixes and prefixes, which are exiles in English. In slang Arabic language, there is an exile tool "ش، شي" added as a suffix of the Arabic verbs like "يحكم" which is "Govern" to get the opposite meaning, which is "يحكمش" or "يحكمشي", where it is "لا يحكم" in formal Arabic language that means "Not Govern" in English. Sometimes, Egyptians use this exile tool another way, they add character "م" at the beginning of the verb (replace the formal exile tool "ما") then they add the previous suffix, the previous example will be "ميحكمش" or "ميحكمشي" that give the same meaning. To clarify the previous exile tool, for example " مرسي الاخواني ميحكمش مصر" means "Brotherhood Morsi not govern Egypt" since the word "ميحكمش" contains prefix "م" and suffix "ش". Another exile tool is a suffix "مش" which mean "لا" as formal Arabic exile tool and mean "No" in English. The example will be "مرسي الاخواني مش يحكم مصر".

4. PROPOSED SLANG SENTIMENTAL WORDS AND IDIOMS (SSWIL)

SSWIL is new sentiment descriptive words, which are spread among Egyptian internet users. Egyptians comment on topics with new slang language sometimes they called it "Franco-Arab" [15]. Comments perform a binary classification problem, having people "satisfied" and unsatisfied with particular news. After surveying comments of 112 news pages and comments of 20 famous Arabic Facebook pages, new 43 words and new 27 idioms are collected, having a "Satisfied" class, and about new 91 words and new 31 idioms means, having an "unsatisfied" which do not exist in Arabic lexicon. This way of writing has new opinion words that describe the satisfaction or dissatisfaction of comments like " حلو، كويس، تظبيط، جامد" , which are "Helw, kowyes, tazbeet, gamed" which means "Satisfy" and "وحش، فاكس، تعبان، زفت" which are "wehesh, fakes, ta'ban, zeft" which means "Dissatisfy". There are also new idioms in this way of writing that describe the satisfaction or dissatisfaction of comments like "روش طحن، زي الفل" which are "Rewesh tahn, zay elfol" which means "Satisfy" and " يخبط فالحلل، هبل فالجبل" which are "yehkabat felhelal, habal felgabal", which mean "Dissatisfy". SSWIL is organized in XML format as shown in Fig. 1.

5. PROPOSED OPINION MINING APPROACH

In this section, the proposed opinion mining approach phases, based on SVM, is illustrated in Fig. 2. The approach is organized into three phases: data preparation (comments collection, XML View), data preprocessing (remove stop words, data auto correction, stemming) and data classification. In last phase, the system classifies the data in three types: classifying comments based on classical opinion words lexicon, classifying comments based on classic lexicon and SSWIL and classifying using SSWIL only.

3.2 Data preparation

The first phase of the proposed system is the data preparation, which consists of two sub phases: comments collection and converting collected comments into XML format.

3.2.1 Collecting Comments

To collect comments, the dataset is collected from news websites like: Aljazera[3], bbcarabic[4], Alyoum Alsabe[5] and Alarabia[6], where these portals have the most readers all over the Arabic countries. We collect more than 1350 comments from previous websites as in Table 1 as the following:

[3] http://www.aljazeera.net/portal
[4] http://www.facebook.com/bbcarabicnews?ref=ts
[5] www.youm7.com

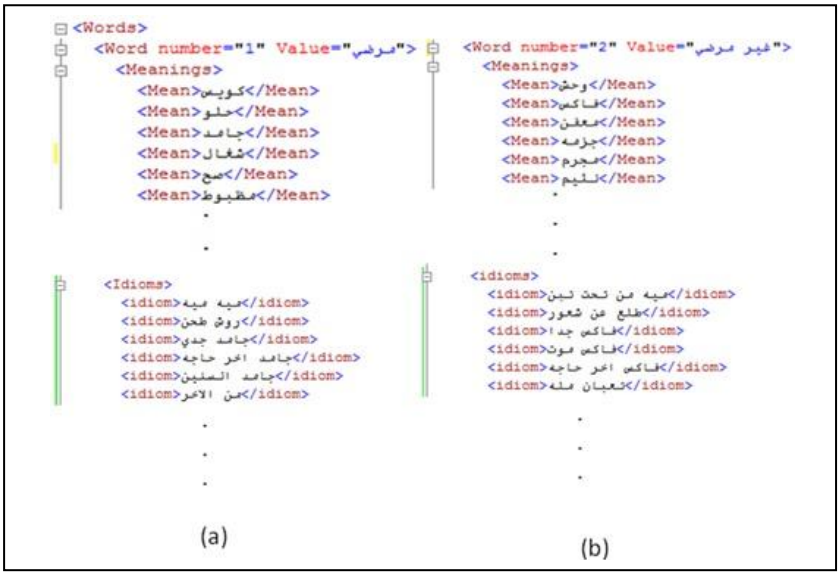

Figure 1. Example of SSWIL structured as XML: (a) SSWIL for satisfaction words and idioms. (b) SSWIL for dissatisfaction words and idioms

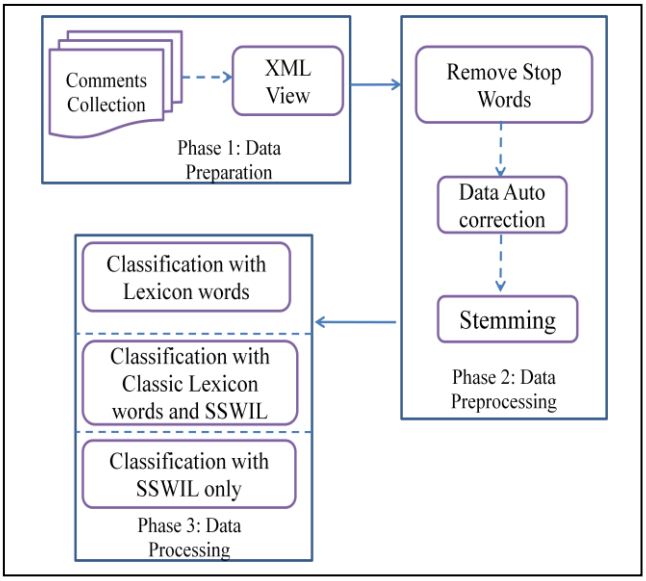

Figure 2. Proposed system phases

Table 1 Comments Dataset and its sources

News Website	Num. of comments
Aljazera.net	370
Alarabia.com	472
BBCArabic.com	23
Youm7.com	490

The comments are written in Arabic language with free text format. Many readers from many countries comment on the news to show their satisfaction or dissatisfaction for different topics. Taking Egyptian presidential elections as an example, the percent of satisfaction and dissatisfaction on new Egypt's president

[6] www.alarabiya.net/

'Morsi' are taken. Taking the comments of the same news from various news websites, comments are written in free text slang Arabic language. The comments are written freely, without grammar rules, unstructured and with slang language that have non lexicon words and idioms. Example is:

يعم مرسي وطي صوتك حرام عليك وديتنا في داهيه

This example indicates that there are new idioms not included in Arabic lexicon; "حرام عليك" and "داهية" which shows that the commenter is not satisfied with Morsi as a president. The comments are written in many forms.

- *Direct Comments:* are comments that are related with the topic and it is useful in our study. These comments are written directly with expressive words so it helps researcher in their automatic analysis.
- *Direct Modern Comments:* Some commenters may use non lexicon idioms and words to create their comments; they write comments in slang Arabic language so the sentiment of the sentence will be understand by human analysis. For instance, instead of comment an item, the user may replay to another commenter, another comments may written sentimentally and subjectively in Arabic but with English characters, for instance the following comment:

Morsy da ragel koyes ya gedaan w nsebo yakhod forsetoh

Comments are written sentimentally and subjectively in Arabic but with English characters and numbers which is called FrankoArabic Language. Many Egyptians' web users use numbers instead of Arabic characters like "5" to replace "خ". For instance, comment may be like this:

Morsi mn ele5wan wmosh hayenfa3 yo7kom masr

The system focuses on the comments written in Arabic language and ignores comments written in Franco-Arab language [16].

- *Indirect Comments:* Many comments in different web pages are not related to the topic. People attempt to comment on anything, even with unrelated words or nonsense [5].

5.1.1 5XML View

Comments are collected manually from four different news websites. The comments are opinions of a specific topic, which is "Egyptian presidential election". The system extracts the comments from the web pages and organizes the data in XML view to facilitate the manual analysis and system process. The system builds an XML, as shown in Fig. 3, schema for the data which contains specific data about the commenter (Name, Title, and Comment).

5.2 Data preprocessing

5.2.1 Remove stop words:

A list of Arabic stop words is used to remove unneeded words to facilitate the data processing. A list of more than 600 words is used for removing unneeded words (من، الى، عن، فوق، الخ). Removing unneeded words is a basic operation when mining the unstructured data because the data will be converted to numbers as input for statistical equations. For example, after removing stop words, the comment "ربنا يقدرك يا محمد مرسي على خدمة مصر والف مبروووك" will be "ربنا يقدرك محمد مرسي خدمة مصر والف مبروووك".

5.2.2 Data Auto correction:

As the web users write comments in free text, ungrammatically and unstructured Arabic language, comments always contains many syntax errors that make the mining process very difficult. As example of errors, web users may repeat a character in a word like "مبروووووووك" instead of "مبروك", which means "Congratulation". Our system tries to solve like these problems to repack words to its correct syntax. Another syntax error in users' comment; the user may use " ه " instead of using " ة " at the end of the word. The system tries to make all the final characters of such words to be "ه" to make the system work well and avoid mismatch through comparison or matching. After making data auto correction on the previous example, it will be " ربنا يقدرك محمد مرسي خدمه مصر والف مبروك".

5.2.3 Stemming:

Arabic stemmer is used to stem the yield words of each comment to get the words' root. We use stemmer of [16] to stem the words to return its root to easily extract features and identify satisfaction classification. The system tokenizes the comments into words and stems the words. The stemmer does not stems unknown words which may be new modern words or not stems the words not written in syntax well. After stemming the previous example it will be "رب قدر محمد مرسي خدم مصر والف مبروك".

```
<Commenter Number="245">
  <Name>said -</Name>
  <Time></Time>
  <Comment>الله اكبر ولله الحمد</Comment>
</Commenter>
<Commenter Number="246">
  <Name>محمود من سوهاج - lugl</Name>
  <Time>lugl</Time>
  <Comment>اللعب يا معلم لعبتها مع كسبتها مع</Comment>
</Commenter>
<Commenter Number="247">
  <Name>سمير لندن - London</Name>
  <Time>London</Time>
  <Comment>مبروك للامة العربية ولكن يد وحدة ما تصفق
  انشالله العاقبة لدول الأخرة خاصة الجزائر</Comment>
</Commenter>
<Commenter Number="248">
  <Name>مصري - بلاد الغربة</Name>
  <Time>بلاد الغربة</Time>
  <Comment>مرسي اخد الكرسي ، اما نشوفك هاتعمل اية</Comment>
</Commenter>
```

Figure 3. Samples of collected comments in XML view

5.3 Data Classification:

In the classification phase, three techniques are performed. The first one is used to classify comments without applying SSWIL, the second one is used to classify comments after the creation of SSWIL and finally, classifying comments using SSWIL only.

5.3.1 Comments classification Using Classical Lexicon without SSWIL

The system classifies the comments using SVM technique into two classes: satisfaction class and dissatisfaction class. The system uses a list of more than 600 satisfaction words to classify the comments as a satisfaction comment and use a list of 700 dissatisfaction words to classify the comments as dissatisfaction comments. The comments of the most four famous news websites described previously and in Table 1, are used as a dataset to test the classification method. All the words in the two lists are lexicon words, which are formal Arabic language words. The result of the classification process is tabulated in Table 2.

Table 2 Classic Classification Results

Comments Source	#of comments	satisfaction	dissatisfaction	Outliers
Alarabia.com	472	298	92	82
Algazira.net	370	252	39	79
Youm7.com	490	240	81	169
Bbcarabic.com	23	15	4	4

5.3.2 Comments' Classification using Classic Lexicon and SSWIL

As seen that the number of outliers' comments is more than any of the two target classes as the web users use new modern sentimental words not belong to Arabic lexicon. After using the SSWI as lists of sentimental

words with SVM, we get better results than the first classification process. The new results of the classification are illustrated in Table 3 as follows:

Table 3 Results of using SSWIL and classical lexicon classification

Comments Source	#of comments	satisfaction	dissatisfaction	Outliers
Alarabia.com	472	312	86	74
Algazira.net	370	324	19	27
Youm7.com	490	343	72	75
Bbcarabic.com	23	17	4	2

5.3.3 Classification Using SSWIL only

When applying SSWIL only in the classification process, it is noticed that the SSWIL affect positively in the classification process. The affection is obvious in the satisfaction comments, this mean that the good opinion new words and idioms are used more in the users' comments. In Table 4, it is the results of classification process.

Table 4 SSWIL only classification results

Comments Source	#of comments	satisfaction	dissatisfaction	Outliers
Alarabia.com	472	68	0	404
Algazira.net	370	231	4	135
Youm7.com	490	261	11	218
Bbcarabic.com	23	8	0	15

6. RESULTS AND DISCUSSION

As a step of evaluation, the system results, we get at test set of 150 random comments and applying the three types of classification, the output results as in Table 5.

Table 5 System Evaluation Results

Classification Type	Precision %	Recall %	F-Measure%	Specificity %
Classic Classification	82.4	59.33	68.98	68.85
SSWIL with Classic Classification	88.63	78	82.97	54.54
SSWIL only Classification	83.07	36	50.23	88.54

As seen in the three types of classification, web users write the comments with a new syntax, which affect the results of classification processes. Extraction techniques fail to extract the opinion words at the first classification type but it performs well at the second classification type after adding the SSWI. Applying the system in all comments, the percent of classified comments in the first type (using classic lexicon) produce 75.35%, while the second classification type (using SSWIL with classic lexicon) produce 86.86%. Applying the system using SSWIL only, it gives 43.02% as a percent of comments classification and 56.98% not classified. As seen at Table 5 the results are enhanced in the second type after applying SSWI lists. According to the outlier comments, it seems that 24.65% of comments are not classified, but in the second classification it enhanced to be 13.14%. The results comparison is shown in Table 6.

Table 6 Percent of classification processes

Classification Type	# Of comments	#Of classified comments	Classified comments (%)	# of Outlier comments	Outlier comments (%)
Classic Classification	1355	1021	75.35	334	24.65
SSWI with Classic Classification		1177	86.86	178	13.14
SSWI only Classification		583	43.02	772	56.98

7. CONCLUSION AND FUTURE WORK

In the current work, an opinion mining approach was proposed to mine unstructured and ungrammatical customers' Arabic comments based in new slang sentiment words and idioms lexicon (SSWIL). The new lexicon collected manually from news websites, Facebook and Twitter pages, which were used as interaction and communication pages between web users. SVM technique was applied with SSWIL to classify comments to satisfy or dissatisfy comments. In a future study, we will concentrate on the new methodologies to improve the results of web users slang comments. In addition, SSWIL will be updated with new sentiment words and idioms and working on Franco-Arab comments.

REFERENCES

1. Internet world stats. http://www.internetworldstats.com/stats7.htm.
2. M. Abdul-Mageed and M. Diab, 2012 AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA).
3. Danah m. boyd and Nicole B. Ellison, October 2007 "Social Network Sites: Definition, History, and Scholarship" in Journal of Computer-Mediated Communication Volume 13, Issue 1, pages 210–230.
4. M. Korayem , David Crandall and M. Abdul-Mageed, 2012 "Subjectivity and Sentiment Analysis of Arabic: A Survey" in Advanced Machine Learning Technologies and Applications, Communications in Computer and Information Science series 322, (Springer). AMLTA
5. M. R.S and M. Teresa, 12-14 September 2011 "Bilingual Experiments with an Arabic-English Corpus for Opinion Mining" in Proceedings of Recent Advances in Natural Language Processing, pages 740–745, Hissar, Bulgaria,
6. T. helmy and A. Daud, November 7th, 2010 "Intelligent Agent for Information Extraction from Arabic Text without Machine Translation" in [C3LSW2010] Workshop on Cross-Cultural and Cross-Lingual Aspects of the Semantic Web Shanghai, China
7. A. El-Halees, December 11-14, 2011, "ARABIC OPINION MINING USING COMBINED CLASSIFICATION APPROACH" in 2011 International Arab Conference on Information Technology (ACIT'2011) in Riyadh, Saudi Arabia.
8. S. AbdelRahman, M. Elarnaoty, M. Magdy and A. Fahmy , July 2010. "Integrated Machine Learning Techniques for Arabic Named Entity Recognition" in IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 3
9. M. Elarnaoty, S. AbdelRahman, and A. Fahmy, March 2012 "A MACHINE LEARNING APPROACH FOR OPINION HOLDER EXTRACTION IN ARABIC LANGUAGE" in International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2
10. M. Elhawary and M. Elfeky "Mining Arabic Business Reviews" in 2010 IEEE International Conference on Data Mining Workshops DOI 10.1109/ICDMW.2010.24
11. Omar Zaidan and Chris Callison-Burch "The Arabic Online Commentary Dataset" an Annotated Dataset of Informal Arabic with High Dialectal Content. ACL (Short Papers) 2011: 37-41
12. Motaz K. Saad and W. Ashour, 2010 "Arabic Text Classification Using Decision Trees" in Workshop on computer science and information technologies CSIT'2010, Moscow – Saint-Petersburg, Russia
13. M. Hijjawi and Z. Bander, 12-14 April 2011 "An Arabic Stemming Approach using Machine Learning with Arabic Dialogue System" in ICGST AIML-11 Conference, Dubai, UAE.
14. Internet world stats. http://www.internetworldstats.com/stats7.htm.
15. A. Hofert and A. Salvatore "Shaping Modernity In A Tran cultural Space between Europe and Islam" Chapter 5, 2000.
16. S. R. Elbeltagy and Ahmed Reafea , February 2011, An accuracy-enhanced light stemmer for Arabic text, in ACM Transactions on Speech and Language Processing (TSLP) Volume 7 Issue 2 Article No. 2