IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. IV (Mar-Apr. 2014), PP 24-35
www.iosrjournals.org

Improvement of Arabic Spam Web pages Detection Using New Robust Features

Mohammed A. Saleh[1], Hesham N. El mahdy[1],Talal Saleh[2]
[1](*Information Technology Dept , Faculty of Computers and Information, Cairo University, Egypt)*
[2]*(Computer Science Dept., Faculty of Computer Science and Engineering, Hodeida University, Yemen)*

Abstract: *In this paper we proposed new set of features to improve the detection ofthe Arabic spam web pages. These Features include: Global Popular Keywords (GPK) features, Character N-Gram Graph (CNGG)features and Sentence Level Frequent Words (SLFW) features. We denoted the new proposed set of features as features B in contrast to the state-of-art featured which denoted by features A. We combined our features (B) with the state-of-art features (A) to get features(AB) and then fed the resulting features AB into different classification algorithms include Ensemble Decision Tree with Bagging and Boosting ensemble methods, Decision Tree J48 , and Random Forest classifiers to achieve our results. In our results we achieved an F-measure of about 99.54% with the Random Forest classifier. We applied our new features on a dataset of about 15962 Arabic web pages which include spam and non-spam pages. We also compared our results with results of a previous study in the field of Arabic spam web pages and we found that, our results (F-measure of 99.64%) have outperformed their results (98%) with the same dataset they used in their study (Dataset 2010).*

Keywords: *Apriori Algorithme ,Arabic web spam,Detecting Arabic Spam, Machine learning algorithms, N-Gram graph, Popular keywords.*

I. INTRODUCTION

With the explosive growth of the World Wide Web, there are a vast amount of web pages populated on every imaginable subject from human rights to Daily News to Sport Games as presented in online news articles, forums, and blogs. These pages may contain also a mixture of computer data such as voice, graphics, pictures, videos and multimedia. Web involves a huge number of users from different geographical regions. Users turn to Web search engines such as Google, Yahoo and Bing for searching about useful information. Many millions of web pages maybe retrieved for each query, but the user only looks for a few and specific web pages . In fact, according to [1] 85 percent of search engines users only look at the first page of the returned Search Engine Results Page (SERP) and only three to five links of them are clicked [2]. As a result of that, the competition for a high ranking for user queries became necessary. Therefore, inclusion of the most important and the highest ranked websites pages in the first SERP has a great economic necessity due to the increase in website traffic. However, some website owners attempt to use spam pages to manipulate search engine rankings.

Web spam is defined as a behavior with the effect of manipulating and deceiving search engines ranking algorithms, and it is usually aimed to obtain a high rank for pages more than they deserve [3]. Web spam is a serious problem, since it harms the search engine ranking and weakens the trust of the users in the search results. Spammer manipulate search engines by using multiple Globally Popular Keywords (GPK) unrelated to the contents of the web pages such as , العاب(Games), بنات(Girls), شات(Chat), فايس بوك(Facebook), ياهو(yahoo)[4]. They also use duplicated, near duplicated phrases or sentences and sentences from different topics that have no relations to each other . The purpose of this manipulation is to make their pages more relevant to user queries, thus manipulating the search engines to increase the spam webpage rank to be included in the first ten links appear in the first page of SERP. An Example of Arabic spam web pages that contain duplicated and near duplicated sentences shown in fig.1.


مشاهدة مباراة برشلونة وريال مدريد بث مباشر اليوم 26-2-2013,اونلاين الدوري الأسباني 2013,مشاهدة مباراة برشلونة وريال مدريد بث مباشر 26-2-2013
بث مباشر برشلونة وريال مدريد اونلاين مباشر , نقل حي برشلونة وريال مدريد , اونلاين برشلونة وريال مدريد ,توقيت مباراة برشلونة وريال مدريد ,مشاهدة برشلونة وريال مدريد , شاهد برشلونة وريال مدريد , رابط برشلونة وريال مدريد ,لينكات برشلونة وريال مدريد , شاهد مباراة برشلونة وريال مدريد , القنوات الناقله لمباراة برشلونة وريال مدريد , الجزيره +2, الجزيرة HD2 , موعد مباراة برشلونة وريال مدريد , معلق مباراة برشلونة وريال مدريد يوسيف سيف , أهداف برشلونة وريال مدريد , اعداه مباراة برشلونة وريال مدريد مشاهده مباراه برشلونة وريال مدريد بث مباشربرشلونة وريال مدريد الدوري الأسباني , 2013,رابط مباراه برشلونة وريال مدريد مباشرنتيجة مباراه برشلونة وريال مدريد لايف الدوري الأسباني 2013 , شاهدة مباراة برشلونة و ريال مدريد شاهد مباراة برشلونة وريال مدريد يوم الاربعاء 26 فبراير , مشاهدة مباراة برشلونة وريال مدريد بث مباشر اليوم 26-2-2013,2013 موعد المباراة في تمام الساعة بتوقيت جرينتشبث مباشر برشلونة
source http://www.harfkids.com/upload/showthread.php?t=34722

Figure 1: An Example of Arabic spam web page that contain duplicated and near duplicated sentences.

There are a lot of benefits of having a high ranking, but the superior benefit of having a high ranking is the internet marketing and online advertisements [5]. For example, owners of commercial sites are very interested in ranking their sites on search engines, because appearing in the first page of SERP means getting a larger number of visits, and thus, getting a higher opportunity to promote and sell their goods and products[6].

According to the statistics of US Census Bureau [7] they found that the number of Internet users in the Arab world is 86 million in Des-2011 rising at a rate of 42% over 60 million in Des-2009, which representing 3.8% of the total internet users around the world and 23.9% of the Arabic speaking regions population.

It's hard to know the exact number of Arabic web pages on the Internet, but according to the estimation proposed by MDR [8] there are about 2 billion Arabic web pages which represents 2% of the total web pages on the internet, with growing rate about 50% average annual rate, while the Arabic blogs have reached around 750,000 by August 2012 and the Wikipedia's Arabic content totaled as 1.34% of all published content. However, those contents represent a considerably low percentage when compared to the number of the Arabic speakers.

A few studies discuss the problem of the Arabic spam web pages compared to the English language, as a result of lacking of data set in this field.

In this paper we built a new Arabic Spam Web Page corpus contains 15962 web pages , we also introduce new features to improve the detection of the Arabic spam web pages.

Our contributions in this paper can be summarized as fellow. First, we built a new Arabic web spam corpus, containing 15962 pages collected manually and relied on international and Arabic search engines. Second, we compared three groups of features A, B and AB, where A represent features used in [4] with Arabic spam web pages corpus which contain 12742 pages, and B including Our features introduced in this paper, and AB features (combination of A and B features).We also compared A, B and AB using Decision Tree (J48)[9], Bagging[10], Boosting(Adaboost)[11], and Random Forest[12]. Third, we introduced new features based on contents analysis to improve the detection of Arabic spam web pages, including three types, namely Globally Popular keywords (GPK), Character N-Gram Graphs (CNGG) and Sentence Level Frequent Words (SLFW) features. Fourth, we showed that ensemble classifiers such as bagging , Boosting and Random Forest could improve accuracy. Fifth, we tested our method on two available data sets, Our data set , and data set-2010 which collected by researchers as proposed in [4].

The rest of this paper is organized as follows: In Section 2, we discussed the related work. In Section 3, we presented our Arabic Web spam corpus. In section 4 we explained three groups of features A,B and combination(AB).In Section 5, we described our experimental results for A,B , And AB features with four classifier, also we described in details the experimental results of our features (B) that include three types, namely GFK, CNGG, and SLFWfeatures and compared them among four classifier. Finally, in Section 6, we presented the conclusions and the future work.

II. RELATED WORK

In this section, first, the studies of Non-Arabic web spam detection will be described. Then the previous studies of Arabic web spam detection will be described as well.

1. Non-Arabic Web Spam Detection Studies:

Non Arabic web pages spam detection studies mainly focus on the content-base analysis and link-base connection. In Content-base analysis studies, Ntoulas et al. [13] introduce several anti-spam heuristics according to the content of the web page such as: the number of words, the amount of anchor text, the fraction of globally popular words, and the likelihood of independent n-grams. These heuristics are combined and treated as features in classifying spam Web documents. They used Decision Tree classifier with 2,364 web pages and achieved an F-measure of 0.862 and a false positive rate of 1.3%. Bencz´ur et al. [14] proposed spam web pages detection classifier by extracting features based on the occurrence of the keywords that have a high spam or high advertisements value. The experiments showed that those features improved the accuracy of classification by 3% on the publicly available WEBSPAM-UK2006 features.

Wang et al. [15] considered the problem of spam detection as a ranking problem. In addition to traditional text features, the information quality based evidence is proposed to define the trust feature of spam information, and according to these evidences the authors developed a novel content trust learning algorithm. The experiments with real data showed that the proposed method works well in finding spam Web pages. Sumit et al. [16] discussed the features which are responsible for web pages ranking. They proposed a feature that computes the matching score between title and URL of the page based on N-Gram algorithm. Experimental results with WEBSPAM-UK2006 dataset showed that the introduced feature is helpful in detecting spam problem. Pera et al. [17] presented a novel approach which relies on the actual word-semantic measure in the content and markup of a given Web page to identify spam effectively. Experimental results applied with two well-known Web spam-detection datasets, i.e., WEBSPAM-UK2006 and WEBSPAM-UK2007 showed the

reliability of the introduced approach with 84.5% accuracy on the average.Egele et al. [18] introduced an approach to detect and to remove web spam pages that are returned by a search engine. At the first, they have conducted many experiments to find out the most important features, and then based on these features they develop a system to identify web spam pages in search engine results.

There are some studies detect spam based on link analysis. Becchetti et al. [19] introduced dumping function that uses the incoming and outgoing links for web spam detection. The concept of spam mass has introduced in [20], it reflects the effect of link spamming on the ranking of the page and used for detecting spam pages. Cheng -et al. [21]used information from Search engine optimization (SEO) forums to determine spam sites and thus link farms.Other well-known anti-spam techniques that rely on link-based connectionor content-based analysis are described in [22].

2. Arabic Web Spam Detection Studies:

H. Wahsheh et al. [23] have introduced an approach for detecting Arabic spam web pages. New features included such as the complexity factor of Web page within lexical density and the number of words in the page. They collected 402 Arabic web pages, and applied classification with Decision Tree, Naïve Bayes, and K-nearest neighbors. The best accuracy achieved is 96.8% with the K-nearest neighbors. In [4] Jaramahet al.proposed a set of new features and combined them with the public content features in [13] to improve the detection of Arabic spam Web pages, their features include cosine similarity among components of a Web page, number of excessively repeated words, ..,etc. They used the Decision Tree, Naïve Bayes, and LogitBoost classifiers with a collection of 12747 pages. The best result achieved were an F-measure of 98% and a false positive rate of 1.9% using the Decision tree. H. Wahsheh et al. [24] Proposed a method based on the rule-based of the Decision Tree for detection the content of Arabic spam web pages. The experimental results with 15,000 pages collected manually showed that the efficiency of the proposed system to detect Arabic spam web page with an accuracy of 83%. They also studied[25] the Machine Learning algorithms used in detecting Arabic spam web pages. They compared Decision tree with Naive Base algorithms and they conducted tests showed that the Decision Tree outperforms the Naïve Bayes.

H. Wahsheh et al. [26] have conducted a study to analyze the behaviors of the spammers based on the weights of the most 10 Arabic popular words used in the content of the HTML tags. The Decision Tree classifier has been applied on 3500 web-pages for evaluation and achieved and accuracy of 90% in detecting Arabic spam web page.In the study [27] H. Wahsheh et al. have proposed a link and content hybrid approach for detecting Arabic spam web pages, many content-based and link-based features are extracted and the rules of Decision Tree classifier have been used to build their system. The proposed system was tested with 5000 web pages and has achieved an accuracy of 93.1% for Arabic link-based, 90.1% for Arabic content-based, and 89% in detecting both Arabic content and link Web spam.H. Wahsheh et al. [28] have analyzed and evaluated the behavior of the top 100 popular Arabic websites. They have relied on the top 10 most popular Arabic keywords to collect those sites. The results showed that 37 of Arabic websites are using unethical techniques in order to improve their ranks in SERP.

III. DATA COLLECTION

1. Our Data Set

The main challenges that meet researchers in the field of detecting Arabic spam web pages is the lack of dataset, there is no public and standard dataset published over the internet for this type of problem,so we are enforced to collect our dataset manually .We relied on the collection methodology used in [4] to collect our corpus, as a result 15962 Arabic web pages were collected from June 2012 to March 2013. Some of the spam web pages were collected using search engines (e.g., Google, Yahoo, Bing, Maktoob, and Ayna) with a spamming query such as pornography contents, we also relied on commercial and marketing web sites to collect Arabic spam web pages that were designed in order to get a high ranking in the search engines. Spam web pages were also collected from Forums, Sites and Blogs that are so many and spread widely and contain some illegitimate pages. For labeling the spam web pages we look in the source content of the web page, spam web pages contain repetition, copy and paste contents, and contents from different topics. Non-spam pages were collected from several trusted sites, such as educational and governmental,human rights and news sites. In the end, 7423 spam and 8539 non-spam web pages were evaluated by two assessors before using them in our experiments.

2. Data Set -2010

This data was used by researchers in [4] and containing 12747 pages, 5645 spam and 6602 non-spam pages. They collected them from February 2010 to July 2010, so it is called 2010 Data Set.

IV. FEATURES EXTRACTION

1. State-of-Art Features (Features A)

There are many features that are always used to detect spammed web pages either in the Arabic language[4] or in the English language[13]. These state-of-art features can be summarized as follows: Amount of Anchor Text, Independent N-Gram, Compressibility, Average Length of Words, Number of Words in the page, Number of Words in the Title, Fraction of Visible Content … etc. These state-of-art features have been extracted from the Arabic web pages to be used to detect either this page spam or non-spam.

There are many new features that have been used to analyze the Arabic web pages, these features that have been proposed by other researchers include: First, Globally Popular Arabic Words that are the most repetitive Arabic words used by users of the search engines. Second, Counting that includes the number of words with length >15 characters, number of words repeated >=10 times in the web page. Third, Cosine Similarity that finds the similarity between the components of the web page such as finding the similarity between title anddocument body, similarity between document and description of the web page and so on so forth. Fourth, the N-Gram syntax with n = 2, 3 and 4.[4]

2. The Proposed Features (Features B)

We presented new features in this paper to analyze the Arabic web pages and then to detect them to differentiate between spammed and non-spammed web pages. These new proposed features have been proven to be very efficient in web pages spam detection problems. They have enhanced the ability of classifying the web pages into spammed and non-spammed ones. These features can be listed as follows:

a. Global Popular Keywords (GPK) Features

Globally popular keywords are those words that mostly and often used by visitors to search for their needs in the search engines .A spammer deliberately introducing such words to their web pages in order to get the largest possible number of visits, and as a result of this their pages get a high rank in the search engines and appear within the first page of the SERP. We have collected 150 GPK based on statistics of Google Adwordstools[29]. We used Google tools to identify the frequency of each GPK. GPK contain (pornographic, Sports, Games, personalities ...etc) content, where pornographic words constitute about thirty percent of them.

Some of these words could exist in more than one type, as shown in Table (1). Type 1 contains words in modern standard Arabic language. Type 2 is written in English as a translation of Arabic words. Type 3 is written in Franco-Arabic forms while the pronunciation of words is in Arabic, some words of this type contain numbers embedded in it like (3, 7) which mapping the Arabic characters (ع, ح) respectively, this type used by spammers to attack users whomused to searchin Franco-Arabic. Type 4 which is useless to the users, they are written in Arabic while the language of operating system is English, words of this type may containsymbols like ([,], l, ; , . , /) mapping Arabic characters (ظ, ز ,ك, ط , د , ج) respectively. Type 5 is written in English while the language of operating system is Arabic. Examples of different type are shown in table.1

Type1	Type2	Type3	Type4	Type5
ألعاب	Games	Al3ab	Hguhf	لشةثس
دردشة	Chat	Dardasha	]v]am	ؤاشف
مباشر	Online	Mubasher	Lfhav	خنمهبث
بنات	Girls	Banat	Fkhj	لهقمس

Table.1 Examples For GPK.

In our work, we derived the following features using the GPK:

- Fraction of GPK in the page components, including (1) Main fields of the web page, namely URL, title, Meta keyword and Meta description. (2) Highlight Tags (<a>, , <i>, <b> and). (3) Structure Tags (text of header, table or list, and this includes <h1> . . . <h6>, <dd>, <table>, and <li>). (4) Attributes of Tags (text that is invisible to the user including image title, alt and src tags attributes).
- Difference: The difference between the density of the GPK in anchor text[1] and other contents of the page. Spammers put too many GPK in anchor text for tow reason: first, words in anchor text are an important factor in ranking process [6]. Second, to redirect the users who surfs the page to other spam pages.
- Percentage of the GPK within the repeated words.
- Percentage of the GPK within the unique words.
- Types: spammers insert GPK with more than one type to get large number of visits. Pages which contain GPK written in more than one form give us a clear signal that it is designed for spamming

[1] a visible caption for a hyperlink.

purposes. We extract two important features namely: (1) the number of types in the page. (2) The number of GPK that written in more than one type.

- Placement/Positions: where the GPK are close to each other. How these words are distributed through the Page? We represent the Page as an array of words with a length equal to the number of words in the page. For example, suppose that the words in the page are [**M**, B, R, S, S, N, **K**, T], and (**M,K**)are GPKwith indexes (1,7) respectively, then the minimum number of GPK between **M** and **K** based on the indexes values is the distance between the positions of these two words and that is six in this case. Therefore, we derive four features as follows. (1) The longest series of GPK that appear sequentially. (2) The Number of similar GPK that repeated sequentially. (3) Minimum distance between similar GPK. (4)The number of Structure Tags that contain at least one.

b. Character N-Gram Graph(CNGG) Features

Due to the different types of page contents in the web, i.e. news articles web pages which have lower noise than forums or blogs web pages, and as reported in [30] Forums and blogs contents represent 35% of the Arabic content in the internet, weusedthe CNGG model[31][32] which is robust to the noise, non-standard expressions, spelling mistakes and grammatical mistakes. CNGG model works efficiently in many applications such as text classification [32] and summary evaluation [33]. CNGG model takes in account the neighboring pairs of n-gram and the edges between nodes which represents the number of co-occurrences.An example is the character tri-gram graph for the phrase"لعب أطفال" shown in the Fig.1.(More explanation about CNNG shown in[32])

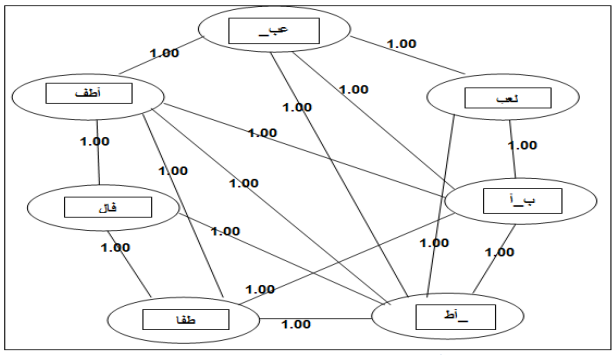


Figure 2 : Tri-gram graph of " لعب أطفال" string.

Definition (N-Gram graph) [32]: *"An n-gram graph isan undirected graph* $G = \{V^G, E^G, W\}$*, where* V^G*is the setof vertices that are labeled by the corresponding n-grams,* E^G*is the set of edges that are labeled by the concatenation of thelabels of the adjacent vertices (in alphabetic order), and* W*is a function that assigns a weight to every edge".*

In this work, we used character tri-gram graph, At First, we built a CNGG model for the main components of the pages including title, description anddocument body of the web pages. Then the similarities between title and document , title and description, description and document have been estimated by applying the following similarity metrics[32]:

1. **Containment Similarity (CS)**, which represents the proportion of edges of a graph G^i that are shared with a second graph G^j.

$$CS(G^i, G^j) = \frac{\sum_{e \in G^i} \mu(e, G^j)}{min(|G^i|, |G^j|)}$$

G is an n-gram graph, e is an n-gram graph edge, μ (e , G)is a function with value=1 if and only if ($e \in G$) and zero otherwise, $|G|$ denotes to the number of edges in the graph *G*.

2. **Value Similarity (VS)**, which indicates how many of the edges included in graph $G^i$are contained in graph G^j, as well, taking in to account the weights of the matching edges.

$$VS(G^i, G^j) = \frac{\sum_{e \in G^i} \frac{min(w_e^i, w_e^j)}{\max(w_e^i, w_e^j)}}{max(|G^i|, |G^j|)}$$

where *w* represents the weight of edges.

3. **Normalized Value Similarity (NVS)**, which is an important metric that takesintoaccountthesize of comparisons graphs.

$$NVS(G^i, G^j) = \frac{VS(G^i, G^j)}{SS(G^i, G^j)}$$

The denominator is defined as the **Size Similarity (SS)** that denotes to the ratio of size between two graphs.

$$SS(G^i, G^j) = min(G^i, G^j)/(max(G^i, G^j))$$

In addition to the similarity features, we extracted the following features from the document body graph. (1) Maximum number of weight edges. (2) Minimum number of weight edges. (3) The difference between maximum number and minimum number of weight edges. (3) Fraction of edges with weights large than ten. (4) Fraction of edges with weights less than five.

c. Sentence Level Frequent Words (SLFW) Features

In this subsection, we focus on the frequent words at the sentences level, the frequentwordsarethose wordswhichappears together inmanysentences in a frequency that is enough to equal at least the suggested threshold. Rather than comparing all words of the document with the main contents of the page (Title, description and Keywords), we only compare the frequent words with the main content of the page. Apriori Algorithm[34]is used for mining the frequent words.

In this work, we extracted the sentences from the document body based on the punctuation marks. Each sentence is considered as a Transaction (T), and every word in the sentence is considered as an (item). Stop words such as (to/الى ,from/من) [35] were removed. Sentences that contain unique words that are not repeated in any other sentence were eliminated.

Before going through the Apriori Algorithm we should define some terms:

Itemset (I) **-** Itemset define as a collection of items in a database which is denoted by $I = \{i1, i2, \dots, in\}$, where i is the number of items.
Transaction (T) **-** Database entry which contains collection of items and denoted by T, and $T \subseteq I$. A transaction contains set of items T={i1,i2,..,in}.
Minimum support (Min_sup) **-** Minimum support considered as a condition or suggested threshold which helps in removal of the in-frequent items in any database. Usually it's given in terms of percentage.
Frequent itemset **-** The itemsets which satisfies the minimum support condition. It is denoted by Li where i indicate the i-itemset.
Candidate itemset **-** Items which are only to be consider for the processing. Candidate itemset are all the possible combination of itemset. It is usually denoted by Ci where i indicate the i-itemset.
Support **-** Support defines as the number of transactions that contain the itemset.

The Apriori Algorithm

Apriori algorithm [34] is an influential algorithm which uses level-wise search that works in iterative approach to explore itemsets efficiently, with the property that "*All nonempty subsets of a frequent itemset must also be frequent*". Apriori property used to reduce search space and improve efficiency. A priori algorithm depends on two actions called self join and prune. In self join action k-itemset is joined with itself to produce k+1-itemset through joining process, while prune process used to remove infrequent itemsets.

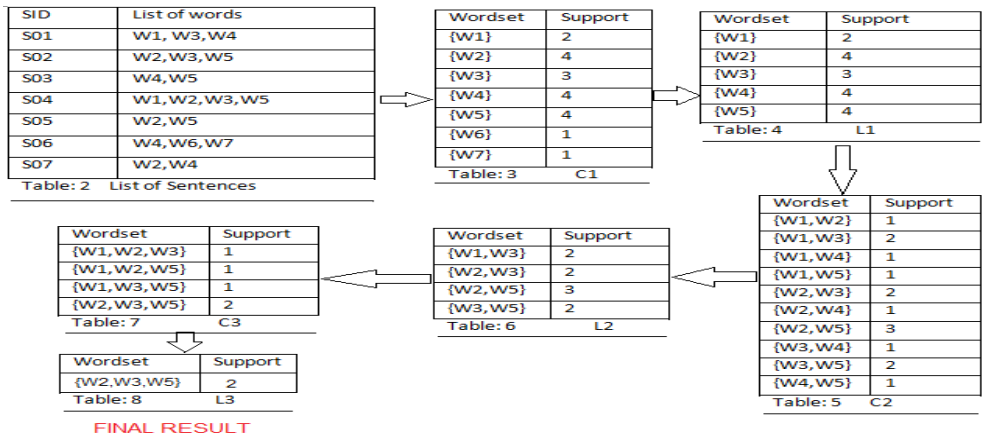
The Apriori Algorithm works as follows:
Step 1: Scan the database to find the 1-itemsets, the collection of 1-itemsets is denoted by C1.
Step 2: Scan the database to find the support of each item to put it as the support of each 1-itemset, in which the support is the frequency of each item.
Step 3: Apply Pruning on C1, in which the support of every 1-itemset will be compared with the minimum support, items with Support less than Minimum Support will be removed. This step results in a table called L1, which consists of 1-itemsets with their support greater than or equal to the Minimum Support.
Step 4: Finding 2-itemsets collection depending on L1, this collection is denoted by C2.
Step 5: Scan the database to calculate the Support of each 2-itemset then compare it with the Minimum Support, 2-itemsets with Support less than the Minimum Support will be removed. This step results in a table called L2 which consists of 2-itemsets with their support greater than or equal to the Minimum Support.
Step 6: Repeat the above step till there is no more frequent itemsets.

***Pseudo Code* [34]-**

Ck: Candidate itemset of size k
Lk : frequent itemset of size k
L1 = {frequent items};
for(k = 1; Lk != Φ; k++) **do begin**
Ck+1 = candidates generated from Lk;
for each transaction t in database do
increment the count of all candidates in Ck+1
that are contained in t
Lk+1 = candidates in Ck+1 with min_sup
end
return∪kLk;

We can easily understand the concepts used by the Apriori Algorithm with the help of a sample example, Let's look in table-2 that contains seven sentences, every sentence contains many words Wi where i = 1,… N, where Wi corresponds to words like "اخبار", "اليوم", and SID is a unique identification given to the each sentence. the Apriori Algorithm steps that used to find out frequent words works as follow:

- Scan all sentences in order to count the number of occurrences of each word (1-wordset); we define it as the support. Each word is a member of the set of candidate 1-wordsets,C1, as shown in table.3.
- Let minimum support equal to 2, that is, Min_sup=2. The frequent 1-wordsets, L1, can be determined from candidate 1-wordsets satisfying min_sup. Prune all Wordsets that have support lesser than min_sup. Table.4 shows the result of pruning Step.
- Generate all possible, but unique candidate 2-wordsets by join L1 with itself. This table will be denoted by C2.. Table.5 shows all the possible combination that can be made from Table.4 1-Wordset.
- The pruning step has to be done on the basis of min_sup criteria. From Table.5 six Wordsets will be removed. After pruning we get the result of frequent 2-wordsets as shown in Table.6.
- The same procedure gets continued till there are no frequent wordsets or candidate set that can be generated. The further processing is described in Table.7 and Table.8. Here we called the final table (**FINAL RESULT**) Table, which is Table.8 in this example.



Here, we derive the following features:

- Number of sentences in the List of Sentences table that at least 50% of their words appear in the Final Result table (i.e. table 8).
- Matching Score: Matching score between Words appear in Final result table with title, Final result with description, Final result with keywords, Frequent 1-wordsets(i.e. All words in Table.4) with title, Frequent 1-wordsets with description, Frequent 1-wordsets with keywords. We used the following formula to calculate the similarity factor :$\boldsymbol{S = \frac{2X}{Y+Z}}$

  Where $\boldsymbol{S}$ is matching score, $\boldsymbol{X}$ number of common words between two string of words $(\boldsymbol{Y},\boldsymbol{Z})$, $\boldsymbol{y}$ Number of unique words in first string, $\boldsymbol{Z}$ Number of unique words in second string.
- Number of words in the Final result table.
- Number of words in the frequent 1-wordsets.
- Number of words in the frequent 2-wordsets.

- Maximum value of the support of frequent 1-wordsets.
- Maximum value of the support of frequent 2-wordsets.

We also extracting some features based on the sentences level as listed below :

- Maximum cosine similarity between the sentences with title.
- Maximum cosine similarity between the sentences with keywords.
- Maximum cosine similarity between the sentences with description.
- Fraction of sentences that express unique words not repeated in any other sentences.
- Fraction of sentences that match at least one words of the title.
- Fraction of sentences that match at least one words of the description.
- Fraction of sentences that match at least one words of the keywords.

3. The Combined Features (AB Features)

The combined Features are the resulted Features from the combination of Features A and B. These features are fed into the classifiers to produce our results. The combined features (AB features) have proved a great effectiveness in the web page spam classifications. Using these features with Random Forest classifier have given the best results in the spam web page classification.

V. THE EXPERIMENTAL RESULTS

The experiments were performed using the Weka data mining toolkit [36], we evaluated our experiments using F-measure described in [37] which is a measure of accuracy.We used four classifiers, namely Decision Tree-J48,(ie. the WEKA [36] implementation of the C4.5 algorithm [9]). Ensemble Decision Tree with Bagging[10] and Boosting[11] and finally Random Forestclassifier [12]. We used J48 as weak classifier for bagging and boosting. In all experiments We used different train-test split percentages. Every experiment applied ten times and the average results of F-measure are reported with 95% confidence interval as error bars in the Figures.

In this study, we conducted several experiments. First, we conducted experiments using Combined Features (AB) including Features that introduced in studies [4][13], We called them Features (A) and they also called state-of-art features, and Features proposed in this paper, we called them Features(B). We compared AB Features using four classifier techniques namely J48, Bagging, Boosting, and Random Forest. Second, we compared A, B, and AB features among four classifiers in details to measure the improvements gain we obtained from the introduced Features. Third, we analyzed and compared three groups of features introduced in this paper including Globally Popular Keywords (GPK), Character N-GRAM Graph (CNGG), and Sentence Level Frequent Words (SLFW) Features using the four previously mentioned classifiers. Four, we Applied the combined Features (AB) to the Arabic Spam Web Page Corpus which researchers used in [4] to measure the improvements gain over F-measure and compare them against the winner of Arabic spam web page corpus.

1. AB Features results with J48, Bagging, Boosting and Random Forest Classifiers

Here, we use all combined features (AB) with four classifiers, and the results are reported in Fig.3. We see that the Random Forest classifier produces the best results, followed by the Boosting, Bagging, and J48 classifiers respectively. The best result achieved is 99.57% when the training set size is 75% for the Random Forest classifier, which is higher than 99.25%, 99.05%, and 98.4962 F-measure value obtained by Boosting, Bagging, and J48 classifier respectively.

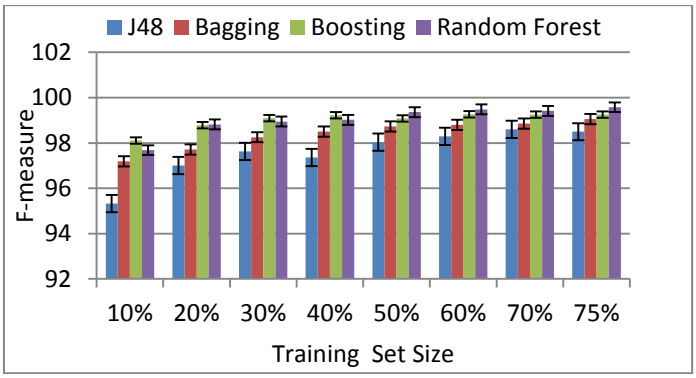


Figure 3: Classification Result with different Training Set Sizes for AB features by J48,Bagging, Boosting, and Random Forest

2. A, B and AB Features results comparison using J48, Bagging, Boosting and Random Forest Classifiers

In order to measure the effects of our proposed features (B) with the used classifiers in improving the detection of Arabic spam pages in our data set we applied several experiments. We experimentally evaluated A, B and AB features using J48, Bagging, Boosting, and Random forest classifiers in details. In Fig 4, we plot theclassification results with different training set sizes for A, B, and AB features by Random Forest classifier. In Fig 5, we plot the classification result with different training set sizes for A, B, and AB features by Boosting classifier. In Fig 6, we plot the classification result with different training set sizes for A, B, and AB features by Bagging classifier. In Fig 7, we plot the classification result with different training set sizes for A, B, and AB features by J48 classifier.

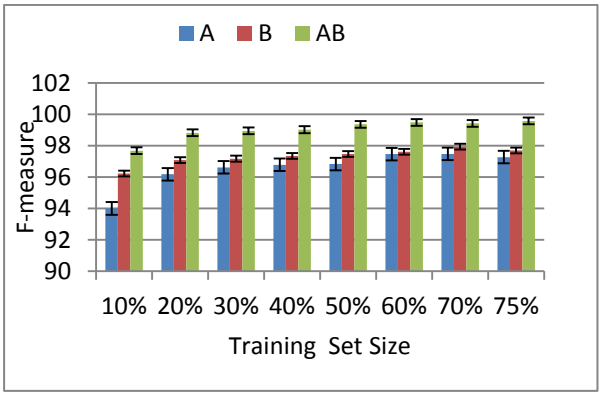


Figure 4: Classification Result with different Training Set Sizes for A, B, and AB features by Random Forest.

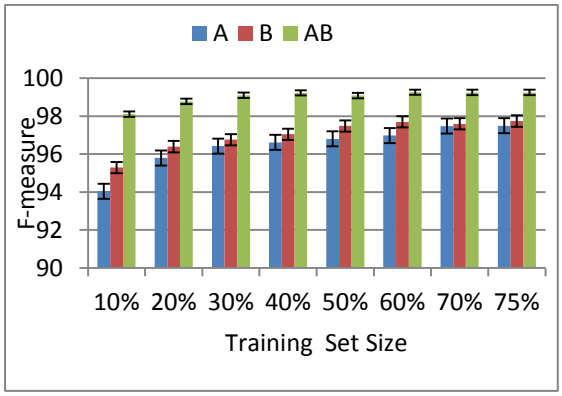


Figure 5: Classification Result with different Training Set Sizes for A, B, and AB features by Boosting.

From Fig.4 we note that the B features slightly outperformed A Features, However we gain an improvements close to 1.5% when we used AB Features by Random Forest with training data size 75%, which reflects the effectively of our introduced features in improving the Arabic spam web pages detection problem .

3. GPK, CNGG and SLFW analysis and comparison using J48, Bagging, Boosting and Random Forest Classifiers

In the following experiments, we analyzed and compared three groups of features introduced in this paper including Globally Popular Word (GPK), Character N-GRAM Graph (CNGG), and Sentences Level Frequent words (SLFW).We performed experiments using only one type of features at a time. In Fig 8, we plot the classification results using GPK features. In Fig 9, we plot the classification results using CNGG features while in Fig 10, we plot the classification results using FWS features.

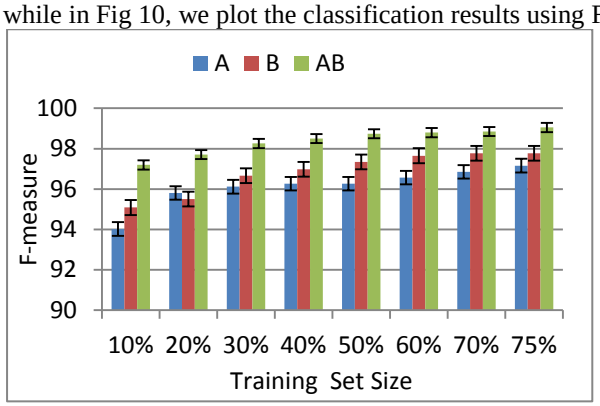


Figure 6: Classification Result with different Training Set Sizes for A, B, and AB features by Bagging.

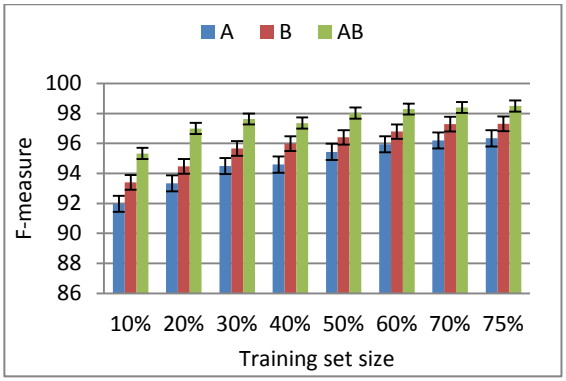


Figure 7: Classification Result with different Training Set Sizes for A, B, and AB features by J48

Comparing Fig 8 with Figs 9 & 10, one can clearly note that the SLFW features achieved the highest performance Compared to CNGG and GPK features alone. CNGG features also had relatively good performance which is close to SLFW feature, and the lowest individual performance is achieved by GPK features. We expect the SLFW to be a very strong indicator of the nature of the page to be spam or non-spam and hence, the high results of SLFW features obtained because they don't only take into account relationship of the frequent words with the words in the meta-tags contents, but they also take the relationship between sentences in the document to their consideration. In addition to that the good results of CNGG features reflect their efficiency in detecting spam pages. The GPK features are most helpful in detecting spam pages that include GPK, the low performance of GPK is not surprising because the collected corpus contain a large portion of data set labeling to spam and non-spam due to the duplication and repetition content.

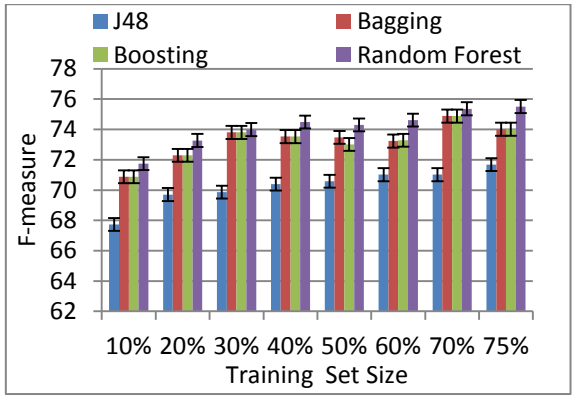


Figure 8: Classification Result with different Training Set Sizes for PGK features by J48,Bagging, Boosting, and Random Forest

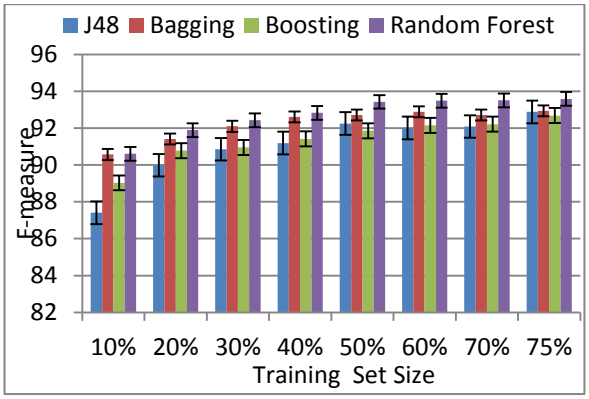


Figure 9: Classification Result with different Training Set Sizes for CNGG features by J48,Bagging, Boosting, and Random Forest

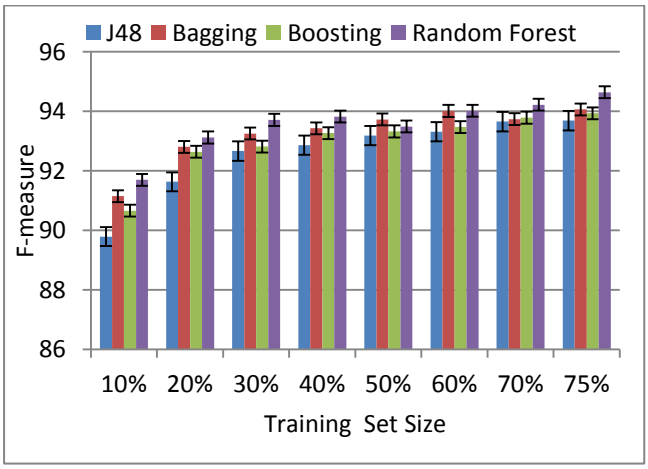


Figure 10: Classification Result with different Training Set Sizes for SLFW features by J48,Bagging, Boosting, and Random Forest

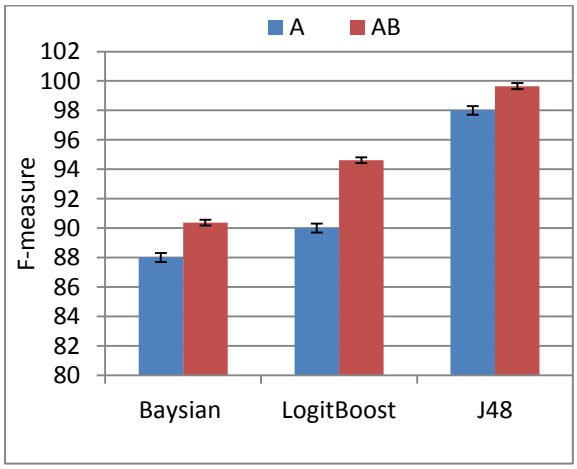


Figure 11: Comparing A and AB Features using Data set-2010 proposed in[4].

4. Experimental Results with Data Set-2010

In this experiment we followed the evaluation protocol of the Web pages Spam corpus in [4].Using this evaluation procedure we could compare our results with other studies. The main purpose of this experiment is to measure the improvement that we could gain by applying ALL features (AB Features) over the winner of the Arabic Spam Web pages corpus who use Features (A) alone.We compared our results against the best results obtained on this dataset which was obtained when the training data size was 66% with the J48 decision tree classifier. The winners of this corpus have achieved a classification result of 98% with J48. As shown in Fig.11, the classification resultshave been raised from 98.00% to 99.64 % when applying all features with J48.

VI. CONCLUSIONS AND THE FUTURE WORK

Most People turn to search engines for searching about helpful information. However, some of the spam web pages used to manipulate search engines to get a high ranking. Such pages cause multiple negative

effects to the user and search engines. Several studies have been conducted in the field of web spam detection as described in the related work.

For finding Arabic spam web pages, features and classification methods play a very important role .To enhance the detection of Arabic spam web pages we have proposed a new features. Specifically, we have introduced three types of features namely GPK, CNGG, and SLFW. We have also collected a corpus of Arabic Web pages contain 15942 Arabic web pages(spam and non-spam). We used combined features AB (A features with B features). We experimented with the Decision Tree (J48) and ensemble methods include bagging, boosting and Random Forest. We also made a comparison between them. We observed that the ensemble methods could improve the classification results. And the best results achieved were with the combined Features AB using the Random Forest with an F-measure value 99.54%.

We plan to collect large Corpus of Arabic web pages up to millions Arabic spam Web pages. We also plan to apply our features to detect spam in social media environments, Natural language processing techniques also could be used to recognize artificial text.

References

[1] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, "Analysis of a very large web search engine query log," in ACM SIGIR Forum, vol. 33, no. 1. ACM, 1999, pp. 6–12.

[2] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthroughdata as implicit feedback," in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005, pp. 154–161.

[3] Z. Gyongi and H. Garcia-Molina, "Spam: It's not just for inboxes anymore," Computer, vol. 38, no. 10, pp. 28–34, 2005.

[4] R. Jaramh, T. Saleh, S. Khattab, and I. Farag, "Detecting arabic spam web pages using content analysis," International Journal of Reviews in Computing, vol. 6, pp. 1–8, 2011.

[5] Y.-M.Wang, M. Ma, Y. Niu, and H. Chen, "Spam double-funnel: Connecting web spammers with advertisers," in Proceedings of the 16th international conference on World Wide Web. ACM, 2007, pp. 291–300.

[6] A. Bifet, C. Castillo, P.-A.Chirita, and I. Weber, "An analysis of factors used in search engine ranking," 2005.

[7] US Census Bureau, "Arabic Speaking Internet Users Statistics," in Usage and Population Statistics of Internet Coaching Library ,Last visited on march 2014 from (http://www.internetworldstats.com/stats19.htm).

[8] Madar Research and Development Center "Ict use and social network adoption in the arab world,".WAMDA, 2012.from (http://www.wamda.com/2012/12/ict-use-and-social-network-adoption-in-the-arab-region-report)

[9] J. R. Quinlan, C4. 5: programs for machine learning. Morgan kaufmann, 1993.

[10] L. Breiman, "Bagging predictors," Machine learning, vol. 24, no. 2, pp. 123–140, 1996.

[11] Y. Freund, R. E. Schapire et al., "Experiments with a new boosting algorithm," in ICML, vol. 96, 1996, pp. 148–156.

[12] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[13] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in Proceedings of the 15th international conference on World Wide Web. ACM, 2006, pp. 83–92.

[14] A. Benczúr, I. Bíró, K. Csalogány, and T. Sarlós, "Web spam detection via commercial intent analysis," in Proceedings of the 3rd international workshop on Adversarial information retrieval on the web. ACM, 2007, pp. 89–92.

[15] W. Wang, G. Zeng, and D. Tang, "Using evidence based content trust model for spam detection," Expert Systems with Applications, vol. 37, no. 8, pp. 5599–5606, 2010.

[16] R. V. Sumit Sahu, Bharti Dongre, "Web spam detection using different features," in International Journal of Soft Computing and Engineering, July, vol. 1, no. 3. IJSCE, 2011.

[17] M. S. Pera and Y.-K. Ng, "A structural, content-similarity measure for detecting spam documents on the web," International Journal of Web Information Systems, vol. 5, no. 4, pp. 431–464, 2009.

[18] M. Egele, C. Kolbitsch, and C. Platzer, "Removing web spam links from search engine results," Journal in Computer Virology, vol. 7, no. 1, pp. 51–62, 2011.

[19] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates, "Using rank propagation and probabilistic counting for link-based spam detection," in Proc. of WebKDD, vol. 6, 2006.

[20] Z. Gyongyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen, "Link spam detection based on mass estimation," in Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment, 2006, pp. 439–450.

[21] Z. Cheng, B. Gao, C. Sun, Y. Jiang, and T.-Y. Liu, "Let web spammers expose themselves," in Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011, pp. 525–534.

[22] N. Spirin and J. Han, "Survey on web spam detection: principles and algorithms," ACM SIGKDD Explorations Newsletter, vol. 13, no. 2, pp. 50–64, 2012.

[23] H. A. Wahsheh and M. N. Al-Kabi, "Detecting arabic web spam," in The 5th International Conference on Information Technology, ICIT, vol. 2011, 2011, pp.1–8.

[24] H. A. Wahsheh, M. N. Al-Kabi, and I. M. Alsmadi, "Spam detection methods for arabic web pages," in First Taibah University International Conference on Computing and Information Technology-Information Systems ICCIT, 2012, pp.486–490.

[25] H. Wahsheh, I. A. Doush, M. Al-Kabi, I. Alsmadi, and E. Al-Shawakfa, "Using machine learning algorithms to detect content-based arabic web spam." Journal of Information Assurance & Security, vol. 7, no. 1, 2012.

[26] H. Wahsheh, I. Alsmadi, and M. Al-Kabi, "Analyzing the popular words to evaluate spam in arabic web pages," IJJ: The Research Bulletin of JORDAN ACM–ISWSA, vol. 2, no. 2, pp. 22–26, 2012.

[27] H. A. Wahsheh, M. N. Al-Kabi, and I. M. Alsmadi, "A link and content hybrid approach for arabic web spam detection," International Journal of Intelligent Systems and Applications (IJISA), vol. 5, no. 1, pp. 30–43, 2013.

[28] H. A. Wahsheh, I. M. Alsmadi, and M. N. Al-Kabi, "Evaluation of web spam behaviour on arabic websites popularity," 2013, in the 6th International Conference on Information Technology, ICIT, 2013

[29] Globally popular Keywords, googleadwords, Last Visited on march 2013 from (https://adwords.google.com) .

[30] Tarabaouni. MENA Online Advertising Industry. Retrieved October, 28, 2011 from (http://www.slideshare.net/aitmit/mena-online-advertising-industry).

[31] F. Aisopos, G. Papadakis, K. Tserpes, and T. Varvarigou, "Textual and contextual patterns for sentiment analysis over microblogs," in Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012, pp. 453–454.
[32] G. Giannakopoulos, P. Mavridi, G. Paliouras, G. Papadakis, and K. Tserpes, "Representation models for text classification: a comparative analysis over three web document types," in Proceedings of the 2nd International Conference onWeb Intelligence, Mining and Semantics. ACM, 2012, p. 13.
[33] G. Giannakopoulos, V. Karkaletsis, G. Vouros, and P. Stamatopoulos, "Summarization system evaluation revisited: N-gram graphs," ACM Transactions on Speech and Language Processing (TSLP), vol. 5, no. 3, p. 5, 2008.
[34] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques. Morgan kaufmann, 2006.
[35] Arabic Stop Words, Last Visited on march 2013 from (http://sourceforge.net/projects/arabicstopwords).
[36] S. R. Garner et al., "Weka: The waikato environment for knowledge analysis," in Proceedings of the New Zealand computer science research students conference. Citeseer, 1995, pp. 57–64.
[37] J. Rennie, "Derivation of the f-measure, 2004," URL http://people.csail. mit.edu/jrennie/writing/fmeasure. pdf, 2008.