Arabic Text Classification Algorithm using TFIDF and Chi Square Measurements

Aymen Abu-Errub
Assistant Professor
Department of Network and
Information Security, Faculty of IT,
Al-Ahliyya Amman University,
Amman, Jordan

ABSTRACT
Text categorization is the process of classifying documents into a predefined set of categories based on its contents of keywords. Text classification is an extended type of text categorization where the text is further categorized into sub-categories. Many algorithms have been proposed and implemented to solve the problem of English text categorization and classification. However, few studies have been carried out for categorizing and classifying Arabic text. Compared to English, the Arabic text classification is considered as a very challenging due to the Arabic language complex linguistic structure and its highly derivational nature where morphology plays a very important role. This paper proposes a new method for Arabic text classification in which a document is compared with pre-defined documents categories based on its contents using the TF.IDF method (Term Frequency times Inverse Document Frequency) measure, then the document is classified into the appropriate sub-category using Chi Square measure..

General Terms
Information Retrieval.

Keywords
Text Categorization, Text Classification, Term Frequency, Inverse Document Frequency, Chi Square.

1. INTRODUCTION
According to the Internet World Stats (http://www.internetworldstats.com/stats7.htm), Arabic is the fifth most used language in the world, spoken by almost 340 million people in 27 states. The Arabic language is one of the oldest known spoken languages as well as one of the official languages of the United Nations. It belongs to the Semitic language family originated in the Arabian Peninsula in pre-Islamic times, and spread rapidly across the Middle East [4]. The Arabic language is very interesting in terms of its history, the strategic value of its people and the region they live in, and its cultural legacy. Historically, for more than fifteen centuries, classical Arabic remained unaffected, comprehensible and functional. At the Strategic level, Arabic is the native language of almost 340 million speakers occupying a main region with vast oil reserves important to the world economy. Culturally, the Arabic language is closely associated with Islam in which 1.4 billion Muslims perform their prayers five times daily [6].

Text categorization (TC) refers to the process of classifying free text, to predefined categories, assigning to it one or more category label. TC has been widely employed in many areas such as email classification, information retrieval and junk email filtering. There are two main approaches for text categorization: the knowledge engineering approach and the supervised learning approach. In the knowledge engineering approach, the classification rules are manually created by domain experts. While in the supervised learning approach, machine learning techniques automatically build the classifiers based on a set of labeled documents. Technically, for each input document *d* and category *c*, text classification involves two steps: (1) estimating the extent to which *d* shares semantics with *c*, and (2) based on the estimation, deciding whether *d* may be classified into *c*. In the first step, classifiers need to assess the similarity score of *d* with respect to *c*. In the second step, classifiers need to associate a threshold to *c*. If the similarity score of *d* with respect to *c* is higher than or equal to the threshold of *c*, *d* is classified into *c* , otherwise it is not classified into *c* [8, 10, 16].

Nowadays, due to the increasing amount of valuable textual information, it becomes a problem or a challenge for humans to manually manipulate this huge amount of information and identify the most relevant information or knowledge. Therefore, Automatic text categorization plays an important role in helping information users overcome such a challenge by reducing the time needed to classify thousands of daily arrived documents, without the need for experts. Thus, Automatic TC can significantly reduce the cost and effort of manual categorization [3]. For example, it has been reported in the Internet World Stats (http://www.internetworldstats.com/stats7.htm) that the number of Arabic speaking Internet users has grown 2,501.2 % in the last eleven years (2000-2011), which is the highest growth rate among other languages. Consequently, with the increasing amount of Arabic textual information, there is a need to propose, develop and employ Arabic TC algorithms in order to store and divide textual information into categories. Thus, assisting the Arabic users to navigate to the information they would like to obtain.

Compared to other languages, there is still a limited research which has been carried out for the Arabic text categorization and classification due to the complex and rich nature of the Arabic language and its highly derivational nature where morphology plays a very important role [6, 14]. Additionally, most of such research includes supervised machine learning techniques in which most of these techniques have complex mathematical models and do not usually lead to accurate results for Arabic TC [14]. Accordingly, much more research is needed to further develop and refine the area of Arabic TC. In this study, the researcher will apply both a vector classification method and Chi square measurement for Arabic

text classification in which similar documents are grouped into categories and sub-categories based on their contents. The rest of this paper is organized as follows: In Section 2, a brief overview of the related studies in which a number of research papers that deal with the area of Arabic TC and Arabic root extraction are reviewed. Section 3 shows the proposed algorithm. Section 4 presents the experimental results. Finally, conclusions and directions for future study are provided in Section 5.

2. RELATED WORKS

Many researches have been carried out on text categorization in English. However, researches on text categorization for Arabic language are quite limited [6, 14]. Among the successful approaches for Arabic Text categorization, a number of recent studies have been proposed [1, 2, 5, 7, 9, 11-15, 17].

In the paper of Syiam et al. [15], an intelligent Arabic text categorization model is presented. For Arabic text categorization, the proposed model uses: 1) statistical n-gram stemmer for document pre-processing, 2) a hybrid approach of Document Frequency Thresholding and Information Gain for feature selection, 3) normalized TF-IDF for term weighting, and 4) Rocchio classifier for classification. Experimental results demonstrate the effectiveness of the proposed model and gives generalization accuracy of about 98%.

Mesleh [11] implemented a text classification system for Arabic language articles. The implemented system uses 1) CHI statistics as a feature extraction method in the pre-processing step of the text classification system design procedure, and 2) Support Vector Machines (SVMs) classification model for TC tasks for Arabic language articles. The author collected corpus from online Arabic newspaper archives, including *Al-Jazeera*, *Al-Nahar*, *Al-hayat*, *Al-Ahram*, and *Al-Dostor* in addition to a few other specialized websites. The collected corpus contains 1445 documents which were falling into 9 classification categories. Experimental results show a high classification effectiveness for Arabic data set in term of F-measure (F=88.11) compared to other classification methods.

Al-Harbi et al. [2] presented the results of experiments of document classification performed on seven different Arabic corpora using statistical methodology. A tool was implemented for feature extraction and selection and the performance of two popular classification algorithms (SVM and C5.0) in classifying the seven Arabic corpora has been evaluated. Generally, C5.0 classifier shows better classification accuracy than SVM.

Harrag et al. [9] enhanced Arabic text classification by feature selection based on hybrid approach of Document Frequency Thresholding using an embedded information gain criterion of the decision tree algorithm. Experiments are conducted over two self collected data corpus. The first corpus is a set of Arabic texts from the Arabian scientific encyclopedia *"Do You Know"*. It contains 373 documents fitting in 8 categories. The second corpus is a set of prophetic traditions collected from Prophetic encyclopedia *"The Nine Book"*. It contains 435 documents fitting in 14 categories. The study demonstrated the effectiveness of proposed classifier and gives classification accuracy of 0.93% for the scientific corpus and 0.91% for the literary corpus.

Noaman et al. [13] introduced the use of rooting algorithm with Naïve Bayes classifier to resolve the problem of Arabic document categorization. To validate the proposed classification algorithm, the authors created a corpus of 300 documents belong to 10 categories which were selected based on the most popular categories from many newspaper articles collected from different online newspaper websites. The experimental study shows the success of the proposed classifier in terms of error rate, accuracy, and micro-average recall measures, and achieves 62.23% of classification accuracy.

Alsaleem [5] discussed the problem of automatically classifying Arabic text documents and used Naïve Bayesian method (NB) and Support Vector Machine algorithm (SVM) on different Arabic data sets to handle the Arabic text classification problem. The Experimental results against different SNP Arabic text categorization data sets confirm that SVM algorithm outperforms the NB with regards to F1, Recall and Precision measures.

Molijy et al. [12] proposed and implemented an automatic Arabic document indexing method to automatically create and Index Arabic books. The proposed method depends mainly on text summarization and abstraction processes to gather main topics and statements in the book. It is start by the pre-processing step which removes irrelevant text (e.g. punctuation marks, diacritics, non-letters, etc.). Then it computes the frequency of every term in the document and reorders them in a descending order. After that, a ranking algorithm is used to remove all terms with highest and lowest frequency. Finally, the system matches between the term and the page number where the term occurs in the document and automatically adds the index to the end of the document. Experimental results in terms of accuracy and performance show that the proposed method can effectively replace the human time consuming effort for indexing a large number of documents or books.

Al-Diabat [1] investigated the problem of Arabic text categorization using different rule-based classification data mining algorithms. These algorithms, which have been contrasted on the problem of Arabic text classification, are: One Rule, rule induction (RIPPER), decision trees (C4.5), and hybrid (PART). Inclusive experiments have been carried out against known Arabic text collection called SPA with respect to different evaluation measures such as error rate, number of rules, etc, to determine the best performing algorithm in regards to the Arabic text classification problem. The results show that the most applicable algorithm is the hybrid approach of PART in which it achieved better performance that the rest of the algorithms.

Zaki et al. [17] proposed a hybrid approach based on n-grams and the OKAPI model for the indexing and classification of an Arabic corpus. The hybrid approach takes into account the concept of the semantic vicinity of terms and the use of a radial basis modelling. The use of semantic terminological resources such as semantic graphs and semantic dictionaries significantly improves the process of indexing and classification. The hybridization of NGRAMs-OKAPI statistical measures and a kernel function is used to calculate the similarity between terms in order to identify the relevant concepts which represent best a document.

Goweder et al. [7] developed a Centroid-based technique for Arabic text classification. The proposed algorithm is evaluated using a corpus containing a set of 1400 Arabic text documents covering seven distinct categories. The experimental results show that the adapted Centroid-based algorithm is applicable to classify Arabic documents. The

performance criteria of the implemented Arabic classifier achieved approximately figures of 90.7%, 87.1%, 88.9%, 94.8%, and 5.2% of Micro-averaging recall, precision, F-measure, accuracy, and error rates respectively.

Sharef et al. [14] introduced a new Frequency Ratio Accumulation Method (FRAM) approach for the Arabic TC. The proposed approach has a simple mathematical model and it combines the categorization task with the feature selection task. The combination leads to reduce the computational operations of Arabic TC system unlike the other methods which deal with feature selection and classification as a major process of automated TC. The performance of FRAM classifier is compared with three classifiers based on the Bayesian theorem, namely the Simple NB, Multi-variant Bernoulli Naïve Bayes (MNB) and Multinomial Naïve Bayes models (MBNB). Experimental results show that the FRAM has outperformed the simple NB, MNB and MBNB which are the major methods of supervised Machine Learning. The FRAM achieved 95.1% macro-F1 value by using unigram word-level representation method.

Yousef et al. [14] introduced a new technique to extract Arabic word's roots using N-gram method. The proposed algorithm consists of several steps; it starts by normalizing the word, then dividing it into bi-grams and calculate the similarity between the original words and candidate roots select from the roots list. The researchers tested their algorithm on a 141 roots corpus. The results show that the proposed algorithm is capable of extracts the most possible roots of nearly 80% of the strong roots.

3. PROPOSED ALGORITHM

The proposed algorithm is consists of two stages; categorization stage and classification stage. Following is the description of each stage.

Categorization Stage: in this stage the tested document is categorized into one of 10 categories. The categorization process is done by comparing the key words of the test document with the key words of each category, by using TF.IDF measurement, and then the most related category is chosen. Steps 1 to 6 of the proposed algorithm represent this stage.

Classification Stage: in this stage a further comparing process is done, this time between the tested document and the documents in sub-categories of the chosen main category. The comparing process is done by using Chi square measurement to find the index words of each sub-category of the main category. This stage is represented by steps 7 and 8 of the proposed algorithm. Following are the steps of the proposed algorithm:

Step1 (Categorization Stage): Delete stop words from the tested document. Table (1) shows some Arabic stop words:

Table 1. Arabic Stop Words

أنت	أنتما	أنتم	أنتن	أنا	نحن
هو	هي	هما	هم	هن	هؤلاء
هذا	هذه	هذان	هاتان	أولئك	أولئكم
إياك	إياكما	إياكم	إياكن	إياي	إيانا
إياه	إياها	إياهما	إياهم	إياهن	الذي
التي	اللذان	اللتان	الذين	اللاتي	اللواتي
له	لها	لهما	لهم	لهن	لا
لم	لن	حتى	بلى	في	على
عن	من	أو	ثم	أي	مثل
كي	نعم	أجل	أين	كم	متى
هاهنا	هيهات	مذ	منذ	قط	إيان
لدى	أنى	لدن	حين	حينما	رب
ربما	حسب	حسبما	سيما	ريثما	بين
بينما	بيننا	كيف	كيفما	عند	عندما
ذلك	تلك	تلكما	تلكم	حيث	حيثما
دون	دونما	ما	كما	نحو	فوق
أعلى	تحت	أمام	وراء	خلف	قبل
قدام	جوار	هنا	هناك	مجاور	محاذي
بعد	بعدما	خلال	الآن	إذ	إذا
لما	بلا	خلا	عدا	حاشا	إلا
قد	قال	مقابل	عوض	بدل	كل
جزء	معظم	أغلب	بعض	قليل	كثير
إدبار	أول	آخر	طلوع	شروق	غروب
طويل	قصير	كان	يكون	كن	صار
أصبح	أمسى	ظل	بات	ليس	انفك
زال	برح	دام	إن	ليت	كأن
لعل	لكن	يمكن	صفر	واحد	احد
إحدى	اثنان	اثنتان	ثلاث	أربع	خامس
ست	سبع	ثمان	تسع	عشر	مائة
ألف	مليون	بليون	مليار	بداية	نهاية
برهة	صباح	مساء	بكرة	عشيا	ظهر
ساعة	سنة	عام	البارحة	أمس	اليوم
غدا	ليلة	ليال	ليل	نهار	نصف
ثلث	ربع	خامس	سدس	ثامن	حقا

Step2: Normalize the rest of the tested document: this step consists of several processes such as:

- Removing punctuation.
- Deleting numbers, spaces and single letters.
- Converting the letters (ء), (ؤ), (ئـ), (أ), (إ) to (ا) and (ة) to (ه).

Step3: Apply stemming process to the tested document's words to delete affixes (suffixes and prefixes) letters and extract the root of each word in the document.

Step4: Find the index terms of both the testing documents and the tested document by calculating the weight of each word using TFIDF - Term Frequency (tf_{ij}) and the Inverse Document Frequency ($\log (N/df_j)$)- measurement as shown in the following equation:

$$Wij = tf_{ij} * \log (N/df_j)$$

Step5: Choose the top three words that have the largest weight in the tested document (index words).

Step6: Compare the index words of the tested document with the index words of each testing category to find the most suitable main category.

Step7 (Classification Stage): Calculate the weight of each word in the main category chosen in step 6 using Chi square measurement to select the index words of each sub-category. This step is done by applying the following equation:

$$\chi^2 (w, s_i) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$$

Where:

w: The word to be weighted.

s_i: The i^{th} sub-category.

N: Total number of documents in the main category.

A: Number of documents in sub-category s_i that containing word w.

B: Number of documents in sub-category s_i that do not containing word w.

C: Number of documents not in sub-category s_i but containing word w.

D: Number of documents neither in sub-category s_i nor containing word w

Step8: Calculate the similarity between the index words of the tested document and the index words of each sub-category of the chosen main category.

Step9: Return the name of main and sub category the has highest matching percentage

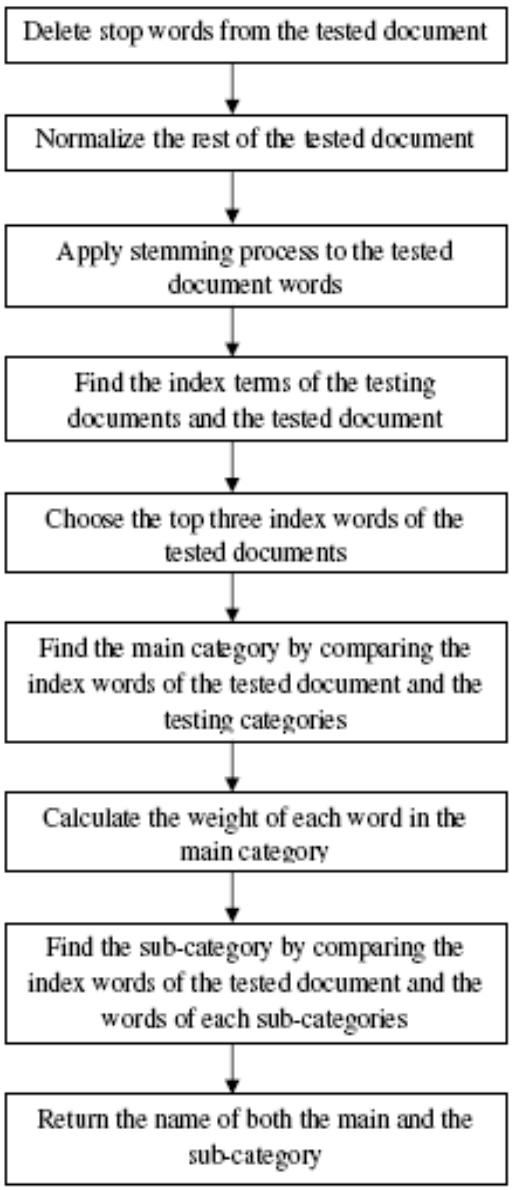

Figure 1: Arabic Text Classification Algorithm

4. EXPERIMENTAL RESULT

To examine the proposed algorithm the researcher chose a training set of Arabic articles covering different topics. These documents are categorized into ten main categories and 50 subcategories containing 1090 documents with variant size and content. These documents are used as learning document set to apply the proposed algorithm. Table 2 shows both the main and sub-categories of the training set and the number of documents in each category:

Table2: Main and Sub Categories of the Training Documenting Set.

Main Category	Sub Categories	# Doc.	Main Category	Sub Categories	# Doc.
Agriculture	Aquatic Species	20	Environment	Ecology	20
	Animal Feed	28		Physics	28
	Spices	23		Chemistry	23
	Poultry Farming	20		Soil Science	19
	Forestry	20		Geography	26
Astronomy	Moons	24	Biology	Cell Theory	21
	Planets	26		Evolution	18
	Stars	22		Genetics	24
	Nebulae	16		Homeostasis	16
	Galaxies	18		Systematic	19
Business	Electronic Commerce	27	Chemistry	Atom	21
	Industrial Automation Business	23		Element	23
	Social Business	24		Compound	18
	Marketing	19		Ions & Salts	27
	Healthcare	17		Acidity	22
Computer	Network	23	Sport	Swimming	21
	Hardware	25		Water Polo	18
	Software	20		Tennis	23
	CIS	21		Baseball	20
	Multimedia	19		Football	28
Economics	Markets	25	Tourism	Healthy Tourism	21
	Production & Efficiency	24		Cultural Tourism	25
	Supply & Demand	21		Nature Tourism	19
	Industrial Organization	19		Religion Tourism	20
	International Trade	23		Accommodation	23

The Proposed algorithm was implemented at a set of tested documents consist of 1100 document. The results of test show that the proposed algorithm is capable of categorize the tested documents to a main category and then classify these tested documents into a suitable sub-category. Table (3) shows the results of categorizing selected tested documents to main category. Table (4) shows the results of classifying tested documents into sub-category of the main category

Table 3: Sample of Proposed Algorithm Percentage Results for Categorization Stage

Doc. # / Cat.	79	894	440	126	969	1050	281	709	673	64
1	86.87	26.05	1.77	**91.53**	19.35	21.12	16.17	6.43	2.12	30.82
2	75.20	32.78	73.89	62.31	0.86	33.27	48.17	0.00	2.10	11.93
3	**90.48**	73.93	47.80	0.00	54.54	**90.26**	0.11	38.08	0.99	31.08
4	22.71	11.65	0.00	18.02	55.09	5.84	**82.60**	7.61	10.27	40.47
5	49.61	14.91	0.42	1.41	58.40	42.91	7.29	7.55	7.64	29.64
6	3.07	0.93	1.68	83.48	**96.39**	1.04	29.68	**98.93**	**92.31**	57.20
7	18.47	40.73	**94.10**	17.22	13.75	26.54	9.68	0.00	1.12	95.88
8	61.70	3.98%	37.67	26.39	25.29	75.42	19.20	27.48	37.46	72.65
9	13.36	**79.19**	34.43	23.36	1.31	9.11	61.21	79.04	75.56	70.71
10	6.29	5.50	88.33	26.03	49.24	42.50	43.10	16.39	87.54	6.67

Table 4: Sample of Proposed Algorithm Percentage Results for Classification Stage

Doc. # / Sub- Cat.	969	709	673
Ecology	20.33	22.43	17.06
Physics	0.90	35.00	48.80
Chemistry	59.55	**98.93**	0.12
Soil Science	**96.39**	6.53	**92.31**
Geography	66.09	45.42	7.33

5. CONCLUSION

This research introduced a dual-stages Arabic text classification algorithm using TFIDF measurement for categorization stage and Chi square measurement for classification stage. The researcher examines the proposed algorithm using 1090 testing (training) documents categorized into ten main categories and 50 sub categories. The tested documents set was consists of 1100 different documents. The experimental results show that the proposed algorithm is capable of classifying the tested documents to its appropriate sub category.

6. REFERENCES

[1] A. Alatabbi, and C. S. Iliopoulos, "Morphological analysis and generation for Arabic language." pp. 1-9.

[2] A. Farghaly, and K. Shaalan, "Arabic Natural Language Processing: Challenges and Solutions," ACM Transactions on Asian Language Information Processing, vol. 8, no. 4, pp. 1-22, 2009.

[3] R. Guzmán-Cabrera, M. Montes-y-Gómez, P. Rosso et al., "Using the Web as corpus for self-training text categorization," Information Retrieval, vol. 12, no. 3, pp. 400-415, 2009.

[4] A. H. Wahbeh, and M. Al-Kabi, "Comparative Assessment of the Performance of Three WEKA Text Classifiers Applied to Arabic Text," Abhath Al-Yarmouk: Basic Sci. & Eng., vol. 21, no. 1, pp. 15-28, 2012.

[5] R. L. Liu, "Context recognition for hierarchical text classification," Journal of the American society for information science and technology, vol. 60, no. 4, pp. 803-813, 2009.

[6] R. Al-Shalabi, G. Kanaan, and M. Gharaibeh, "Arabic text categorization using kNN algorithm." pp. 1-9.

[7] B. Sharef, N. Omar, and Z. Sharef, "An Automated Arabic Text Categorization Based on the Frequency Ratio Accumulation," International Arab Journal of Information Technology (IAJIT), vol. 11, no. 2, pp. 213-221, 2014.

[8] A. Goweder, M. Elboashi, and A. Elbekai, "Centroid-Based Arabic Classifier." pp. 1-8.

[9] A. A. Molijy, I. Hmeidi, and I. Alsmadi, "Indexing of Arabic documents automatically based on lexical analysis," International Journal on Natural Language Computing, vol. 1, no. 1, pp. 1-8, 2012.

[10] M. Al-diabat, "Arabic Text Categorization Using Classification Rule Mining," Applied Mathematical Sciences, vol. 6, no. 81, pp. 4033-4046, 2012.

[11] S. Alsaleem, "Automated Arabic Text Categorization Using SVM and NB," Int. Arab J. e-Technol., vol. 2, no. 2, pp. 124-128, 2011.

[12] T. Zaki, D. Mammass, A. Ennaji et al., "Arabic Documents Classification by a Radial Basis Hybridization." pp. 37-44.

[13] M. M. Syiam, Z. T. Fayed, and M. Habib, "An intelligent system for Arabic text categorization," International Journal of Intelligent Computing and Information Sciences, vol. 6, no. 1, pp. 1-19, 2006.

[14] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity et al., "Automatic Arabic text classification." pp. 77-83.

[15] A. M. d. A. Mesleh, "Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System," Journal of Computer Science, vol. 3, no. 6, 2007.

[16] F. Harrag, E. El-Qawasmeh, and P. Pichappan, "Improving arabic text categorization using decision trees." pp. 110-115.

[17] H. M. Noaman, S. Elmougy, A. Ghoneim et al., "Naive Bayes Classifier based Arabic document categorization." pp. 1-5.