KACST Arabic Text Classification Project: Overview and Preliminary Results

AbdulMohsen Al-Thubaity, Information Center, Riyadh, KSA, althubaity@ic.org.sa
Abdulrahman Almuhareb, KACST, Riyadh, KSA, muhareb@kacst.edu.sa
Sami Al-Harbi, Information Center, Riyadh, KSA, s.h.alharbi@ ic.org.sa
Abdullah Al-Rajeh, KACST, Riyadh, KSA, asrajeh@kacst.edu.sa
Mohammad Khorsheed, KACST, Riyadh, KSA, mkhorshd@kacst.edu.sa

Abstract

Electronically formatted Arabic free-texts can be found in abundance these days on the World Wide Web, often linked to commercial enterprises and/or government organizations. Vast tracts of knowledge and relations lie hidden within these texts, knowledge that can be exploited once the correct intelligent tools have been identified and applied. For example, text mining may help with text classification and categorization. Text classification aims to automatically assign text to a predefined category based on identifiable linguistic features. Such a process has different useful applications including, but not restricted to, E-Mail spam detection, web pages content filtering, and automatic message routing. In this paper an overview of King Abdulaziz City for Science and Technology (KACST) Arabic Text Classification Project will be illustrated along with some preliminary results. This project will contribute to the better understanding and elaboration of Arabic text classification techniques.

1. Introduction

In addition to WWW pages, government organizations and business enterprises have huge repositories of written texts in electronic format for different needs. Managing available knowledge to discover the hidden relations between texts in these repositories or to create a knowledge-base for organization expertise will leverage the knowledge in these organizations and improve their competitive advantage. Text classification as a process of automatically assigning the text to predefined categories can help in managing the content and thereby the knowledge they contain. There are three main consecutive phases in building a classification system. First, the training data must be compiled. Second, a set of features must be selected to represent defined classes. Third, a chosen classification algorithm must be trained and tested using the corpus compiled in the first stage.

To the best of our knowledge, it would appear that efforts to investigate Arabic text classification and categorization are rare. A search was conducted for research efforts on Arabic text classification. We found (and outline below) six full text papers on the subject using different data sets, different feature selection methods and different classification algorithms [24,8,12,13,26,17].

Sawaf et al. [24] used a statistical approach based on the maximum entropy technique to classify the Arabic NEWSWIRE corpus of the Linguistic Data Consortium (LDC) of the University of Pennsylvania which covers four classes: politics, economy, culture and sports. The main objective was to simplify Arabic classification difficulties by avoiding morphological analysis and to use sub-word units (character n-grams).They argued that "even with no morphological analysis, we gain satisfying results".

ElKourdi et al. [8] used Naïve Bayes algorithm to classify 300 web documents into five classes (health, business, culture, science and sport). The average accuracy achieved was 68.78% in cross validation and 62% in evaluation set experiments. They found that "the performance of NB algorithm in classifying Arabic documents is not sensitive to the Arabic root extraction algorithm[8].

Kanaan et al.[12] used Naïve Bayesian classifier to classify 600 texts distributed equally into 6 classes (architecture, economy, health and medicine, politics, science, and sports). They discovered that classification accuracy varied between classes (from 41% to 100%) and the overall average accuracy achieved was a relatively poor 57.19%.

N-gram text classification based on two distance measures, Manhattan measure and Dice’s measure, was used by Khreisat [13] to classify Arabic newspapers into four classes (sport, economy, technology and weather). The experiment showed poor results. Khreisat argued "The poor performance of the measure for Arabic in this study can be attributed to the nature of the Manhattan measure, and the complex morphological structure of Arabic, which is quite different than the structure for English"[13]. She concludes that "N-gram text classification using the Dice measure outperforms classification using the Manhattan measure"[13].

K-NN and Rocchio algorithms were used by Syiam et al. [26] to classify 1132 texts compiled from three main Egyptian newspapers into six classes (arts, economics, politics, sports, women and information technology). They used six feature selection methods and four term weighting criteria. Syiam et al. suggest the use of combining Document Frequency Thresholding and Information Gain for feature selection, normalized-TFiDF for term weighting and Rocchio algorithm as a classifier for Arabic text classification.

Mesleh [17] used three classification algorithms, namely SVM, K-NN and Naïve Bayes, to classify 1445 texts taken from online Arabic newspaper archives. The compiled texts were classified into nine classes: computer, economics, education, engineer, law, medicine, politics, religion and sports. Chi square statistics was used for feature selection. Mesleh argued that "Compared to other classification methods, our system shows a high classification effectiveness for Arabic data set in terms of F-measure (F=88.11)" [17].

According to the above presented research efforts, it is difficult to suggest which combination of feature selection method, term weighting and classifier is the optimal solution for Arabic text classification. The research undertaken so far has required a relatively large number of well defined experiments. The amount of such experiments in Arabic text classification has encouraged the KACST Arabic Text Classification Project team to start this project.

2. KACST Arabic Text Classification Project

The scope of the Arabic Text Classification project is to experiment with the well known feature selection methods and classification algorithms and suggest a text classification system for Arabic language texts based on that. In addition, the project aims to compile a benchmarking data set (Corpus) for Arabic text classification. Such a system requires: (i) different datasets to train and test the system, (ii) feature selection tools able to extract lexical features and weight the extracted features according to different features' selection methods and, (iii) classification tools to classify texts based on different classification algorithms. According to these requirements, Arabic Text Classification can be divided into three major parts namely; compiling different data sets, building features extraction and selection tools, and investigating the performance of different classification algorithms for Arabic language using the available tools and packages.

2.1 Compilation of training data set

For text classification, a representative corpus of labeled texts must be assembled. Each text in the data set must be assigned to one of the defined classes. Different training data sets are available for text classification in English. Reuters-21450 and Reuters-810000 collections of news stories are popular and typical examples. The Linguistic Data Consortium (LDC) provides two Arabic corpora, the Arabic NEWSWIRE and Arabic Gigaword corpus. Both corpora comprise newswire stories. One of the aims of this project is to compile representative training data sets for Arabic text classification that cover different text genres which can be used in the future as a benchmark. Up to this stage of the project, seven corpora were assembled comprising 17,658 texts with more than 11,500,000 words covering seven different written genres. The Internet was used to collect texts. An overview of compiled corpora (genres) statistics is shown in Table 1.

Table 1: An overview of compiled data sets (corpora) for KACST Arabic Text Classification Project

Genre	No. of Text	No. of Classes	Classes
Saudi Press Agency (SPA)	1,526	6	Cultural News, Sports News, Social News, Economic News, Political News, General News
Saudi News Papers (SNP)	4,842	7	Cultural News, Sports News, Social News, Economic News, IT News Political News, General News,
WEB Sites	2,170	7	IT, NEWS, Economics, Religion, Medical, Cultural, Scientific
Writers	821	10	Ten writers
Discussion Forums	4,107	7	IT, NEWS, Economics, Religion, Medical, Cultural, Scientific
Islamic Topics	2,243	5	Tafseer, Feqah, Aqeedah, Hadeeth, Linguistics
Arabic Poems	1,949	6	Gazal, Hega'a, Madeh, Retha'a, Wasf, Hekmah

2.2 Feature Extraction and Selection

Generally, there are two kinds of features related to text: *external features* which are not related to the content of the text such as writer, publisher, language, and number of pages, and *internal features* which are related to text content and are linguistic features such as lexical items, single or compounds, grammatical categories and semantic relations. Arabic language and hence Arabic text have different characteristics resulting in several difficulties and challenges. For feature extraction, these challenges can be summarized as follow: (i) Diacritics are used in some texts and neglected in many others, (ii) The use of Kashida, and (iii) Neglecting the use of *Hamza* and *Taa Marbutta*.

Much of the work in text classification treats a document as a kind of bag-of-words with the text represented as a vector of a weighted frequency for each distinct word or token. Such a simplified representation of text has been shown to be quite effective for a number of applications [7][25]. There have been several attempts to enhance text representation using concepts [4] or multi-word terms [20][27].

Text representation, in the bag-of-words technique, usually leads to high dimension input space affecting the efficiency of classification algorithms, which makes the feature set exceed the training set by orders of magnitude. To handle this problem, several techniques are used to reduce the input space by selecting a subset of features that may lead to better classification. Term frequency (TF), document frequency (DF), Chi-Squared (CHI) [2], Information Gain (IG)[19] for instance are the most commonly used metrics for feature selection. Numerous metrics are also used for feature selection/reduction like Gain Ratio, Odds Ratio, and Probability Ratio [9]

A tool was implemented (ATC Tool) for feature extraction and selection as part of the Arabic Text Classification Project. This tool is able to perform the following main functions:

(1) Automatically divide the data set into two partitions - training and testing - according to the user input of training and testing size.
(2) Extract the lexical features (single word) and generate the feature frequency profile for both the training set and testing set with options to explore the profile for each class and each file. Also, the user has the option to exclude stop words and remove the diacritics and Kashida from the texts or to generate the frequency profile of a certain list of words.
(3) Calculate the importance of each feature locally (for each class)and globally (for all classes) based on ten feature selection methods (Term Frequency, Document Frequency, Information Gain, Chi Square, NGL Coefficient, DIA Association Factor, Mutual Information, Odds Ratio, GSS Coefficient and Relevancy Score).
(4) Generate training and testing matrices and weight their elements (selected features) according to seven weighting methods (Boolean, Frequency, TFiDF, TFC, LTC, Entropy, Relative Frequency).

2.3 Classification Algorithms

Different approaches may be used for text classification such as rule based [14], machine learning and data mining techniques. Machine learning and data mining techniques have proved their ability to perform very well in many different cases. For machine learning, the text classification problem is a standard supervised learning, where the learning algorithm maps inputs to desired outputs. The algorithm is required to develop a function which maps the input vector into one of several classes by looking at several input-output examples of the function.

Numerous machine learning techniques can be used for the task of text classification such as Naive Bayes[23], Naïve Bayes (MultinomialUpdateable) [16], support vector machines[11], artificial neural network [3] and K-Nearest Neighbor [15]. Different data mining techniques are also used for text classification techniques namely Conceptual Structure [1][6], Decision Tree [10][5], C4.5 decision tree [21] and C5 decision tree[22]. Free software, RapidMiner[1] 4.0[18], and a commercial package, Clementine[2], were chosen as classification tools. Both of them are well known tools providing several classification algorithms which will allow a wide range of experiments.

3. Preliminary Results

In this section some preliminary results will be illustrated and discussed. The Saudi Press Agency (SPA) corpus provided data. Literature review shows considerable usage of Reuters data sets for text classification and categorization. Unfortunately, no Arabic data set exists for Reuters newswire. SPA was chosen to be the source of newswire for Arabic text classification in this project for different reasons: (i) to be able to compare the classification results with those who used Reuters data set, (ii) SPA have five explicit classes for its newswire and (iii) the availability of SPA news on the Web[3]. SPA corpus statistics are shown in Table 2.

ATC Tool was used to: (i) split SPA corpus to 70% training and 30% testing, (ii) remove stop words and diacritics from the SPA corpus and (iii) select features based on information gain (IG).

[1] http://rapid-i.com
[2] http://www.spss.com/clementine
[3] http://www.spa.gov.sa/

Table 2: SPA corpus details.

Source	Classes	No. of Texts	No. of Words	No. of unique Words	Average Text Length
Saudi Press Agency	Cultural News	258	44,713	12521	173
	Sports News	255	40006	8641	157
	Social News	258	51714	11818	200
	Economic News	250	38766	9158	155
	Political News	250	34940	9720	140
	General News	255	43333	12639	170
TOTAL		**1,526**	**253,472**	**36,497**	**166**

The information gain weighting formula measures the amount of information gained for class prediction as a function of the presence or absence of a term in a document. Equations 1 and 2 represent the local and global functions of IG respectively:

$$IG(t_j, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_j, \bar{t}_j\}} P(t,c) \log \frac{P(t,c)}{P(t)P(c)} \qquad (1)$$

$$\begin{aligned} IG(t) = & -\sum_{i=1}^{i=m} P(c_i) \log P(c_i) \\ & + P(t) \sum_{i=1}^{i=m} P(t, c_i) \log P(t, c_i) \\ & + P(\bar{t}) \sum_{i=1}^{i=m} P(\bar{t}, c_i) \log P(\bar{t}, c_i) \end{aligned} \qquad (2)$$

Where

m = Number of classes

$P(t_j)$ is the probability that the term t_j occurred in a document,

$P(\bar{t}_j)$ is the probability that the term t_j does not occur,

$P(c_i)$ is the probability of the class c_i,

$P(\bar{c}_i)$ is the probability that a random document does not occur in the class c_i,

$P(t_j, c_i)$ is the joint probability of the class c_i and the occurrence of the term $t_{j.}$

The ten highest ranked terms/features according to IG for SPA corpus are shown in Table 3 following exclusion of the stop words:

Boolean and TFiDF weighting methods were used to weight training and testing matrices elements. RapidMiner 4.0 and Clementine provides several well known learners (classification algorithms).

Table 3**:** The ten highest ranked terms/features according to IG for SPA corpus

Cultural News	بعنوان، محاضرة، المحاضرة، ينظم، الثقافية، اليوم، يلقيها، العلمية، تنظم، الندوة
Sports News	لكرة، القدم، الشباب، المنتخب، لرعاية، كأس، نواف، منتخب، البطولة، الاتحاد
Social News	الخيرية، جمعية، الاجتماعية، الجمعية، الاجتماعي، ريال، الايتام، الفطر، الخيري، المبارك
Econom ic News	ر، دولار، البنك، الاقتصادية، مؤشر، المال، التجارة، مليون، سوق، ارتفع
Political News	السياسية، الخارجية، اليوم، الشعب، الفلسطينية، فهد، عبدالعزيز، الرئيس، العراق، الوزراء
General News	الحجاج، الحج، ضيوف، حجاج، موسم، بيت، المقدسة، الرحمن، الحرام، حج

The scope of KACST Arabic Text Classification Project is to experiment with all the classification algorithms provided by RapidMiner 4.0 and Clementine. The results of four classification algorithms namely C5.0, Neural Network (back propagation), SVM (C-SVC linear), and Naïve Bayes (MultinomialUpdateable) are presented here as an example. For more information about the mentioned algorithms, the reader is kindly requested to consult the provided references in this paper. Classification accuracy and experiment environments of the above mentioned classification algorithms (C5.0, Neural Network, SVM, and Naïve Bayes) are shown in Table 4.

Table 4: Classification accuracy and experiment environments

Experiment environment	Classifier	Features weighting	
		Boolean	TFiDF
data set: SPA, No. of classes:6 training:70%, testing:30% stop words and diacritics: removed, feature selection: Information Gain No. of Feature: 1% of training data size	C5.0	84.43%	84.21%
	SVM	76.10%	75.22%
	Naïve Bayes	75.66%	75.00%
	Neural Networks	63.78%	56.00%

The data shown in Table 4 suggests that C5.0 is outperforming all the classifiers used in this experiment with 84% accuracy. SVM and Naïve Bayes come after C5.0 and have almost the same accuracy (75%). Neural Networks shows the worst classification performance with 63% and 65% accuracy for Boolean and TFiDF weighting respectively. The data suggest that Boolean weighting gives better accuracy than TFiDF. The data shows that the obtained accuracy in our experiment for Naïve Bayes classifier is greater than the accuracy obtained by ElKurdi et al. [8] and Kanaan et al. [12] - 62% and 57.19% respectively. This kind of comparison is, however, not applicable because none of us (Elkourdi, Kanaan et al and ATC Project team) used the same data set or feature selection methods. The necessity of such a comparison sheds light on the importance of KACST Arabic Text Classification Project.

4. Afterward

It appears that there is a lack of thorough investigation and experimentation in Arabic text classification. The results of properly researched experiments will give clear insight into the capabilities and performance of different feature selection methods and classifiers. The availability of benchmarking data will help to achieve that. We believe that the KACST Arabic Text Classification Project contributes to the efforts of Arabic text classification research. In the near future, the implemented feature selection methods and feature weighting and the available classification algorithms in RapidMiner and Clementine package will be applied to the compiled corpora.

References

[1] Agrawal, R., and Srikant, R. "Fast Algorithms for Mining Association Rules in Large Databases," *In Proceedings of the 20th international Conference on Very Large Data Bases*, September 12 - 15, 1994. J. B. Bocca, M. Jarke, and C. Zaniolo, Eds. Very Large Data Bases. Morgan Kaufmann Publishers, San Francisco, CA, pp487-499.

[2] Alexandrov, M., Gelbukh, A., and Lozovo, G. "Chi-square Classifier for Document Categorization," *2nd International Conference on Intelligent Text Processing and Computational Linguistics*, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science N 2004, Springer-Verlag, pp. 455-457.

[3] Bhaumik, H., and Chowdhury, N. "A Neural Network Based Method for Text Classification using Root Words to Form Pattern Vectors," *In Proceedings of the 2nd Indian International Conference on Artificial Intelligence*, Pune, India, December 20-22,2005 Bhanu P. (Ed.), pp 503-510.

[4] Bloehdorn, S., and Hotho, A. "Text classification by boosting weak learners based on terms and concepts," *In Proceedings of the 4th IEEE International Conference on Data Mining (ICDM),* 2004, pp. 331—334.

[5] Cohen, W. W. 1996. "Learning to classify English text with ILP methods," In L. De Raedt, ed. *Advances in Inductive Logic Programming*, IOS Press, Amsterdam, 1996, pp. 124–143.

[6] Cohen, W. W. 2003. "Improving a page classifier with anchor extraction and link analysis," In S. Becker, S. Thrun, and K. Obermayer, ed. *Advances in Neural Information Processing Systems, Cambridge*, MA, MIT Press, 2003, pp. 1481–1488.

[7] Diederich, J., Kindermann, J. L., Leopold, E., and Paaß, G. 2003. "Authorship attribution with support vector machines," *Applied Intelligence*, 19(1/2), 2003, pp. 109–123.

[8] ElKourdi, M., Bensaid, A., and Rachid, T. "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," *in Proc. of COLING 20th Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, 2004.

[9] Forman, G. "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.* 3, 2003, pp. 1289-1305.

[10] Ghani, R., and Fano, A. E. "Using Text Mining to Infer Semantic Attributes for Retail Data Mining," *In Proceedings of the 2002 IEEE international Conference on Data Mining (Icdm'02)* (December 09 - 12, 2002). ICDM. IEEE Computer Society, Washington, DC, pp. 195-202.

[11] Joachims, T. 2001. "A statistical learning model of text classification for support vector machines," *In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, Louisiana, United States). SIGIR '01. ACM Press, New York, NY, 2001, pp. 128-136.

[12] Kanaan, G., Al-Shalabi, R., and Al-Azzam, O. "Automatic Text Classification using naïve Bayesian Algorithm on Arabic language," *IBIMA 2005 Conference on the Internet & Information Technology in Modern Organization*, Cairo, Egypt, 2005, pp. 327-339.

[13] Khreisat, L. "Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study," *DMIN*, 2006, pp. 78-82.

[14] Li, H., and Yamanishi, K. "Text classification using ESC-based stochastic decision lists," *Information Processing and Management*, 38(3), 2002, pp. 343-362.

[15] Lim, H. "Improving kNN Based Text Classification with Well Estimated Parameters," *In 11th International Conference on Neural Information Processing,* 2004, pp. 516-523.

[16] McCallum, A., and Nigam, K. "A Comparison of Event Models for Naive Bayes Text Classification," *In: AAAI-98 Workshop on Learning for Text Categorization*,1998, pp 41-48

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

[18] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. "YALE: Rapid Prototyping for Complex Data Mining Tasks," *in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2006, pp. 935-940.

[19] Mladenić, D., Brank, J., Grobelnik, M., and Milic-Frayling, N. "Feature selection using linear classifier weights: interaction with classification models," *In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Sheffield, United Kingdom, July 25 - 29, 2004). SIGIR '04. ACM Press, New York, NY, pp. 234-241.

[20] Peng, F., Huang, X., Schuurmans, D., and Wang, S. "Text classification in Asian languages without word segmentation," *In Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages - Volume 11* (Sapporo, Japan, July 07 - 07, 2003). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 41-48.

[21] Quinlan, R. *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA. 1993.

[22] Quinlan, R. "Improved use of continuous attributes in c4.5," *Journal of Artificial Intelligence Research*, 1996, pp. 4:77-90.
[23] Rish, I. "An empirical study of the naive Bayes classifier," *In Proceedings of IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence,* 2001, pp. 41-46.

[24] Sawaf, H., Zaplo, J., and Ney, H. "Statistical classification methods for Arabic news articles," *In Processing of the Arabic Natural Language Workshop (ACL2001)*, Toulouse, France.

[25] Sebastiani, F. "Machine learning in automated text categorization," *ACM Comput. Surv.* 2002, pp. 34, 1-47.

[26] Syiam, M. M., Fayed, Z. T., and Habib, M. B. "An Intelligent System for Arabic Text

[27] Tan, C. M., Wang Y. F., and Lee C. D. "The use of bigrams to enhance text categorization," *Information Processing and Management*, 38(4), 2002, pp. 529–546.