Arabic Spam Filtering using Bayesian Model

Abdulkareem Al-Alwani
Department of Computer Science & Engineering
Yanbu University College
Yanbu, Saudi Arabia

Majdi Beseiso
Department of Computer Science & Engineering
Yanbu University College
Yanbu, Saudi Arabia

ABSTRACT

Many of us are concerned about an onslaught of SPAM email. Spam has become major problem for the email communications. The number of spam mails is increasing daily – studies show that over 45-50% of all current email communication is spam, it is an ever-increasing problem and will reach up to 70% in coming years. The volume of non-English language spam is increasing day by day. The motivation for this research is to find a solution for the millions of internet users in the Arabic language struggling with hundreds of SPAMS being received every day in their mailbox. To filter this kind of messages, this research applied Bayesian Model which provides the framework for building intelligent learning system.

General Terms

Spam, spam filtering, Bayesian model.

Keywords

Email, spam, spam filtering, machine learning techniques, Bayesian model.

1. INTRODUCTION

Email spam refers to sending irrelevant, inappropriate and unsolicited email messages to numerous people [1]. This is possible due to the low entrance barrier and low cost of sending emails, which makes it one of the most popular forms of spam [2]. The main motivation of email spam is advertising, promotion, and spreading backdoors or malicious programs. Currently phishing is also considered as one of the main goals of spammers when employing email spams [3].

The problem of spam e-mail has been increasing over the years. In earlier but recent statistics, 40% of all emails are spam, which total to about 12.4 billion email per day [4]. To handle the problem of email spam various methods are presented such as spam filtering which is a program that is used to find out unwanted email and prevent them from getting into the user's mailbox. Programs such as Bayesian models or heuristic filters attempt to identify spam mail through suspicious word patterns and word frequency.

Nowadays Arabic language spam is also increasing very fast, According to a latest internet study, 89% of emails are spam constituting to around 260 billion spams being sent every day [16]. Among such spam emails, around 90% are in English but the number of spam mails in other foreign languages are also on the rise and among them emails in Arabic language are also a lot [16]. Spam emails in English can be dealt with very efficiently using very innovative spam filtering techniques but filtering spam mails in foreign languages poses a problem as there is not much done in certain foreign languages such as Arabic. The spam emails received by users in Arabic and Arabic and English mixed in Arab speaking countries including Saudi Arabia is difficult to deal with owing to lack of efficient spam filtering systems. Machine learning methods such as NB, Term Frequency Inverse Document Frequency (TF-IDF), K-NN and SVM are found to be effective.

We have divided algorithm in two parts. One is about the training and the second is for filtering. We are finding the spam probability for each word and built the Bayesian model depending on the probabilities and trained the dataset. For filtering, we are using Bayesian model which is output of trained algorithm.

The rest of the paper is organized as follows: Section 2 is regarding related work done in the field of spam filtering in Arabic language. Section 3 will elaborate on motivation and challenges with this research. Section 4 is about methodology and proposed approach. Section 5 reports our experiments of the proposed method and compares the results of the different email. Finally we will close this paper with summary and conclusions.

2. RELATED WORK

According to Jaramh et al [12], six supervised machine learning approaches were conceived by El-Halees and a good detection rate were shown for Arabic, English and Arabic and English mixed. Jaramh et al [12], presented new features to improve the efficiency and accuracy of the web spam detection classifiers. Using the same set of classifiers, Decision Tree, K -Nearest Neighbour (K - NN) and Naïve Bayes, their study yielded Decision tree with best results.

Another method used by Goweder et al. utilizes term frequency Inverse Document Frequency Technique and good results were obtained for both English and Arabic emails.

According to Saad [13], a study analyses different criteria for analysis of content, types of content, units of analysis, different ethical issues faced and the software that aids the analysis. It explores certain fundamental issues faced in content analysis by referring to different studies conducted in last decade. This paper provides starting point for content analysis for future researchers.

Wahsheh, and Al-Kabi[14], have built the first ArabicWeb Spam corpus manually, that has only 400 spam ArabicWeb pages using three classifiers Decision Tree, K -Nearest Neighbour (K - NN) and Naïve Bayes. The result of their study showed that K -NN at K =1 is better than the other two classifiers in detecting Web spam pages in Arabic.

The literature study indicated that there have been many studies done on spam filtering for English language but only very little have been done for Arabic language. Our main goal is to develop a client based system for Arabic email spam filtering.

3. MOTIVATION AND CHALLENGES

The motivation for this paper is to find a solution for the millions of internet users in the Arabic language struggling with hundreds of SPAMS being received every day in their mailbox. There are 22 countries with Arab speaking people located in Asia and African continents. Arabic language is spoken by around 300 million people in this world and more and more of these people have started using internet and email daily for communication.

The challenge in this endeavor is the lack of previous works done in the Arabic language. Also lack of Arabic spam dataset makes it difficult for researchers. The language itself is much more complex than other languages. It is difficult to even transfer the language to internet as there is not much work done on morphological analyzers to analyze and identify structure and different features of the language.

Representation style of Arabic language is unique and it is difficult to identify the different character sets and codes. The unique way of writing and reading the language from right to left is another challenge to deal with.

4. THE METHODOLOGY

Methodology involves evaluating the behaviors of the spammers in the content-based spam Arabic Web pages.

4.1 Detecting scenarios for spams

This is done by observing various types of spams and start categorization and classification of email text. Text classification techniques are used in many applications, including e-mail filtering, mail routing, spam filtering etc.

4.2 Data Collection based on detecting scenarios

This step involves building a spam dataset of words mostly frequently used by Arab users, Using Site Content Analyser software, extract a set of content-based features from the collected Arabic Web pages. This is done only for the 10 Arabic words that are most frequently used keywords by Arabic Web searchers and evaluating the spam techniques using data mining tool that uses the Decision Tree Classification Algorithm. It has been found through these exercises is that spammers use the most popularly used Arabic key words in their spam webpages.

4.3 Arabic text classification using N-grams

According to Khreisat [5] Arabic language is complex and rich in nature for text categorization. The Arabic language consists of 28 letters. The language is written from right to left. It has very complex morphology, and the majority of words have a tri-letter root. The rest have either a quad letter root, penta-letter root or hexa-letter root.

An N-gram [6] is an N-character slice of a string. The Ngram method is language independent and works well in the case of noisy-text (text that contains typographical errors). We used tri-grams for text classification. The trigrams of a string or token is a set of continuous 3-letter slices of the string. For example, the tri-grams for the word المودعين are لمو, ا لم, , مود عين, دعي ودع in general, a word of length w has w-2 tri-grams. According to Zipf's law [7] : "The nth most common word in a human language text occurs with a frequency inversely proportional to n"

This has the implication that documents belonging to the same class or category will have similar N-gram frequency distributions.

We are using N-grams for frequency count of the multi words which increases the efficiency of algorithm. We are taking N=3, N=2 and N=1 which is single word. By using N-grams, we will be able to work on phrases of the sentences used in emails.

4.4 Frequency & Pattern analysis

The frequency and pattern of spam mails need to be analyzed to find the most frequently used Arabic words in such mails and how these are used. Pattern of these mails also needs to be analyzed thoroughly. Such an analysis is required to develop an effective tool for filtering spam e-mails.

4.5 Arabic spam filtering: Bayesian Model

Spam filtering is divided in two major steps. One is training of spam filter and the second is testing of spam filter. Training of filter is done by calculating probabilities of words/ phrases and the classification is done according to calculated probabilities.

4.5.1 Phase 1: Preprocessing

This phase involves four major steps.

4.5.1.1 Tokenization

First we split text into units called tokens; this process is called tokenization. As this text is being read, tokenizing it into tokens (words) actually takes place as using a blank space or any other character as a delimiter [9].

4.5.2 Clean Words

This step involves deletion of punctuation marks and other symbols from the text such as (أ , ,إ ,ه), (آ) ة .

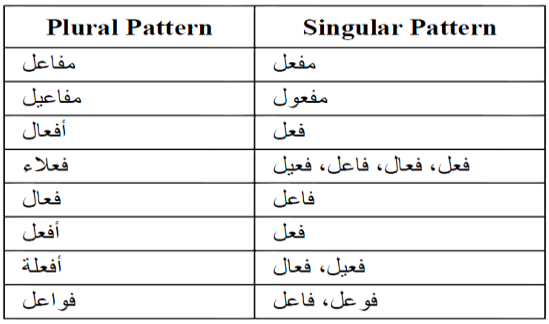
4.5.3 Stop Words Removal

The next step of e-mail text pre-processing involves the removal of some Arabic stop words; removal of those words that have little meaning and usually appear frequently in text documents eg: "or" (أو), "whose" (لمن), "on" (على) , "where" (حيث), "in" (في), "from "(من), "beyond" ( ), غِثثث and "all" ( ).كثثث Removing stop words leads to document with shorter length, which results in more effective processing, and enhances the efficiency of terms indexing procedure [10].

4.5.4 Stemming

Arabic language is highly derivative where tens or even hundreds of words could be formed using only one stem, furthermore, a single word may be derived from multiple stem. According to Ahmed A Elbery, working with Arabic document words without stemming results in an enormous number of words being input into the classification phase. This will definitely increase the classifier complexity and reduce its scalability.

Table 1. Plural and Singular Patterns



Plural Pattern	Singular Pattern
مفاعل	مفعل
مفاعيل	مفعول
أفعال	فعل
فعلاء	فعل، فعال، فاعل، فعيل
فعال	فاعل
أفعل	فعل
أفعلة	فعيل، فعال
فواعل	فوعل، فاعل

4.5.5 Phase 2: Bayesian Model

4.5.6 Calculating Probabilities in Bayesian Model

After phase 1, we calculate the probability of being part of spam message for each word by counting the number of frequency in spam texts and normal texts.

Example: For a word ‘حسابك’ the output is

$$\begin{bmatrix} rate\ of\ bad = \dfrac{\#\ of\ occurence\ in\ spams}{total\ number\ of\ words\ on\ spams} = \dfrac{2}{1062} = \ 0.0018832391 \\ rate\ of\ good = \dfrac{2\ X\ \ \#\ of\ occurence\ in\ normal\ msgs}{total\ number\ of\ words\ in\ normal\ msgs} = \dfrac{0}{1062} = \ 0.00000 \\ probability\ of\ spam = \ 0.99 \end{bmatrix}$$

Note: we multiply 2 in good rate case to help fight against false positives.

4.5.7 Calculating Benefits

When we want to check a message, first we select the useful features to make a decision. We have already calculated the probability from the previous step, now we will follow the following criteria to evaluate the benefit of using each feature:

$$Benifit(W) = \frac{\#\ of\ occurrence\ in\ good\ texts\ -\ \ \#\ of\ occurrence\ in\ bad\ texts}{(\#\ of\ occurrence\ in\ good\ texts +\ \ \#\ of\ occurrence\ in\ bad\ texts)}$$

We selected top 15 useful items for further study.

Example: For the message :
“Subject: 9- اعلان من مجمع دار السلام
يعلن مجمع دار السلام للابسة الجاهزة عن وجود زبائننا الكرام ..
تخفيضات هائله في الملابس قد تصل الى 20%
وللفترة من 24/ 7 /2011 ولغايه10 /8 /2011 وللمزيد
من المعلومات يرجى زياة احد فروعنا في ماليزيا.

Vector of useful criteria depending on the last benefit equation:

$$\begin{bmatrix} \text{سلام} \\ \text{يعلن} \\ \text{معلومات} \\ \text{يرجى} \\ \text{اعلا} \\ \text{زبائن} \\ . \\ . \\ . \end{bmatrix}$$

Each word in the last vector is supported by the probabilities as in the last stages with interesting value.

4.5.7.1 Make a decision

Depending on the last selected items we determine the kind of email using the following heuristics:

$$\frac{\prod_{selected\ items} spam_{prob(Wi)}}{\prod_{selected\ items} spam_{prob(Wi)} + \prod_{selected\ items} NotSpam_{prob(Wi)}}$$

We use the output of the last two steps and apply it to the last equation to determine the nature of message.
We are trying to explain main steps in block diagram of training algorithm and filtering algorithm which is shown in fig 1. This diagram will describe basic blocks of algorithm.

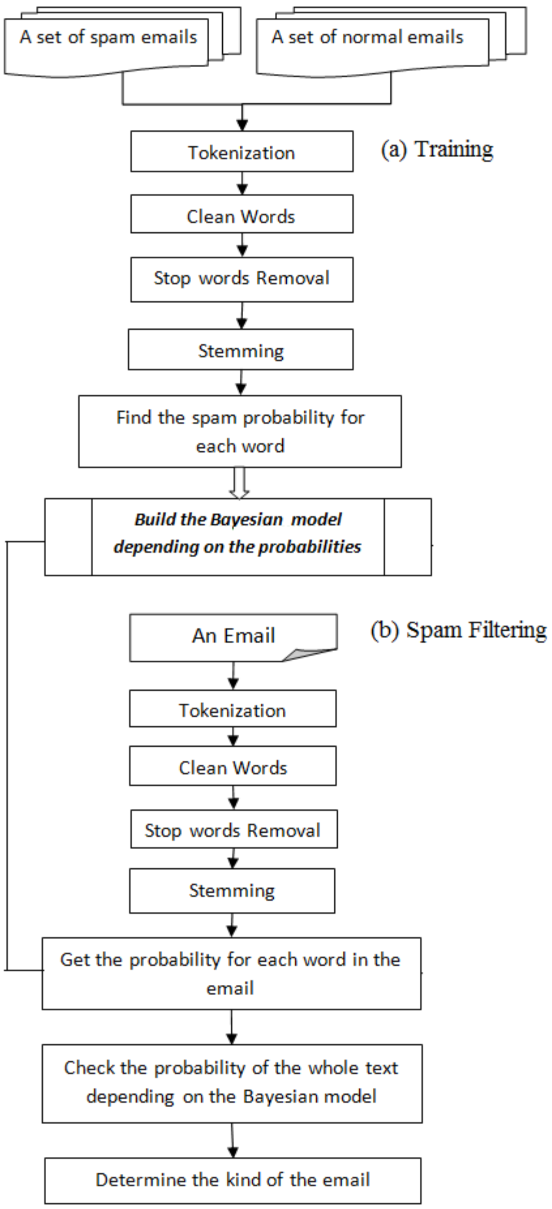
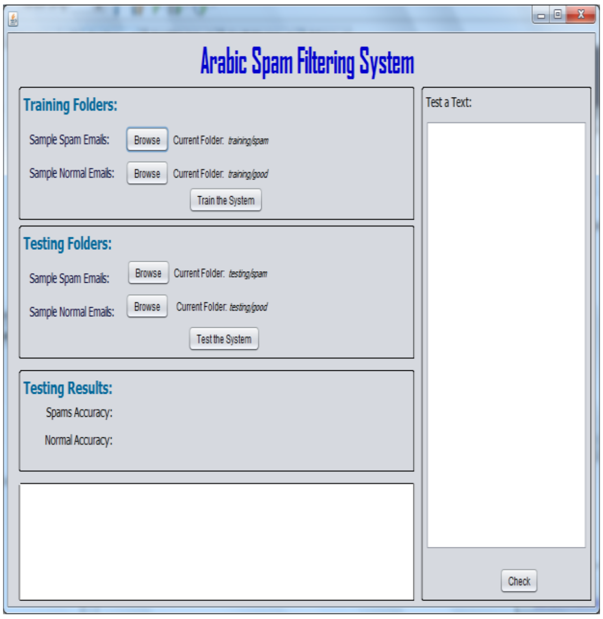

A set of spam emails
A set of normal emails
Tokenization
(a) Training
Clean Words
Stop words Removal
Stemming
Find the spam probability for each word
Build the Bayesian model depending on the probabilities
An Email
(b) Spam Filtering
Tokenization
Clean Words
Stop words Removal
Stemming
Get the probability for each word in the email
Check the probability of the whole text depending on the Bayesian model
Determine the kind of the email

Fig 1: Block diagram of training and filtering algorithm

Below we are showing main screen of the application with source code which is executable with NetBeans using JAVA version 1.7.



Sample Spam Emails: Current Folder: training/spam

Sample Normal Emails: Current Folder: training/good

Sample Spam Emails: Current Folder: testing/spam

Sample Normal Emails: Current Folder: testing/good

Fig 2: Screen shot of GUI of implemented system

5. EXPERIMENTS & RESULTS

We have used 100 Arabic emails for training from the dataset and tested the system with 81 emails and 189 normal texts.

Table 2. Result of evaluation of emails

Result	Is Spam	Is not Spam
The output of the system is spam	66	15
The output of the system is not spam	0	189

As we can see 66 emails are detected correctly from the 81 emails so the accuracy is 81%. If we take 189 normal texts, then it is showing the text is not spam. So the total accuracy is 94 %.

6. CONCLUSION

The research tackles the spam filtering issue for Arabic language emails along with mixed Arabic-English emails. Tackling Arabic language spams is a challenging task due to unique features of the language and very little research available on the subject for Arabic language based texts and emails. The Bayesian approach to spam filtering for Arabic emails, Arabic and English mixed emails provided significantly improved filtering system for Arabic language. The accuracy of spam filtering system for Arabic emails is more than 80% which is very good. For Arabic text, it is correctly predicting no spam.

7. REFERENCES

[1] Junod, John. 1997 Servers to spam: drop dead Computers and Security 16.7 , 623-623.

[2] Email spam statistic. 2007, http://spamnation.info/stats/

[3] Hayati, P., & Potdar, V. 2008. Evaluation of spam detection and prevention frameworks for email and image spam: a state of art. In Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services (pp. 520-527).

[4] Drucker, H., Wu, D., & Vapnik, V. N. 1999. Support vector machines for spam categorization. Neural Networks, IEEE Transactions on, 10(5), 1048-1054.

[5] Khreisat, L. 2006. Arabic text classification using N-gram frequency statistics a comparative study. In Conference on Data Mining| DMIN (Vol. 6, p. 79).

[6] Damashek, M. 1995. Gauging similarity with n-grams: Language-independent categorization of text. Science, 267(5199), 843-848.

[7] Getis, A. 2007. Reflections on spatial autocorrelation. Regional Science and Urban Economics, 37(4), 491-496

[8] Rajput, A., & Toshniwal, D. Adaptive Spam Filtering based on Bayesian Algorithm.

[9] Qasem A. Al-Radaideh & Ahmed F. AlEroud. Arabic alert E-mail detection using rule based filter.

[10] Abu El-Khair, I., 2006. Effects of Stop Words Elimination for Arabic Information Retrieval: A Comparative Study. International Journal of Computing & Information Sciences, Volume 4, Number 5, p.p. 119 – 133.

[11] Khorsi, A. 2007. An overview of content-based spam filtering techniques. Informatica (Slovenia), 31(3), 269-277.

[12] Jaramh, R., Saleh, T., Khattab, S. and Farag, I., 2012. Arabic Email Spam Detection Techniques and Related Arabic Text Preprocessing Options: A Survey. Detecting Arabic Spam Web.

[13] Saad, M., 2006. "The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification", Master of Science, Computer Engineering, The Islamic University Gaza.

[14] Wahsheh, H. A., Al-Kabi, M. N., & Alsmadi, I. M. 2012. Spam Detection Methods for Arabic Web Pages. In First Taibah University International Conference on Computing and Information Technology-Information Systems ICCIT (pp. 486-490).

[15] Farmer, James John 2003. "3.4 Specific Types of Spam" (FAQ). An FAQ for news.admin.net-abuse.email; Part 3: Understanding NANAE. Spam FAQ. Archived from the original.

[16] Sophos 2008. Sophos report reveals rising tide of spam in April–June 2008" (Press release).