# Arabic Email Spam Filter Using Probabilistic Model

Basel AlTamimi[1], Ibrahim Qdemat[1], Mohammad Thwaib[1] and Hashem Tamimi[2]

College of Information Technology and Computer Engineering, Palestine Polytechnic University,
Hebron, Palestine

E-mail: [1]{basel_tamimi, qusaymusa, mohammadThwaib } @webmail.student.ppu.edu
[2]htamimi@ppu.edu

## Abstract

A spam filter is a software that is used to detect unsolicited and unwanted emails. It prevents those messages from getting to a user's inbox. The aim of this project is to develop an intelligent Arabic spam filter based on one aspect of Artificial Intelligence. This aspect focuses mainly on using a Content-Based probabilistic model decision making system.

### Keywords

Arabic probabilistic model, spam filter, content based, dictionaries , spam email.

### INTRODUCTION

Emails are important in our life , therefore, they must be supported and protected by sufficient techniques to be useful, reliable, and to prevent any undesirable messages. We depend on the content of the massage, not on the massage source in our classification. So any massage from any source can be classified as spam or not-spam based on the massage contents. This work is done on Arabic messages.

In order to avoid crisp classification, we follow a machine learning approach. This approach depends on a probabilistic model that uses Bayes' and Laplace smoothing rules which make the classifier more practical.

Our model is able to take the correct decision under the shortage of information which usually the case in real massages. It is not necessary that the whole massage contents exist in the dictionary to be able to classify. This feature comes as a result of using probabilistic model combined with Laplace smoothing that prevents the zero and one hundred probabilities any more.

*Laplace smoothing:*

Laplace smoothing is an approximation function, which attempts to capture important patterns in the data. The probability of the word $w_i$ is its count $c_i$ normalized by the total number of word tokens N:

$$P_{Laplace}(w_i) = \frac{c_i + K}{N+(K.|w_i|)} \qquad (1)$$

Where: K is a non-zero number, and $|w_i|$ is the cardinality of input words.

*Bayes' theorem:*

Bayes' theorem is a probabilistic model used for calculating conditional probabilities. The simplest statement of Bayes' theorem:

$$P(B|A) = \frac{P(A|B).P(B)}{P(A)} \qquad (2)$$

## METHODOLOGY

First we are concerned in the creation of dictionaries. In other words, we have to find a good source for spam and inbox dictionaries. Spam dictionary contains words that come from a real spam massages, Some of repeated topics in forums, sub-words and single letters. Inbox dictionary consists of words that come from a real inbox massages, everyday speaking language and some of important words included in spam messages could be found in any inbox messages.

The second phase is to calculate probabilities of spam and ham which evaluated based on Spam-inbox-dictionaries and Equation (1).

After that, the system can make the classification by evaluating the probability of each word in this massage based on Equation (2). Finally, the system will classify the massage to be a spam or ham depending on the higher probability.

## EXPERIMENTS

We used 20 different messages – 10 of them are labeled to be spam messages - through two main tests:

The first test aims to determine the smoother (k) value, which maximizes the total of successful spam messages detection, and minimizes the number of messages to be marked as spam when they are not.
The second test aims to determine the possibility of reducing the size of spam and inbox dictionaries, while maintaining the results and their accuracy.
The outcomes of both experiments are listed next:

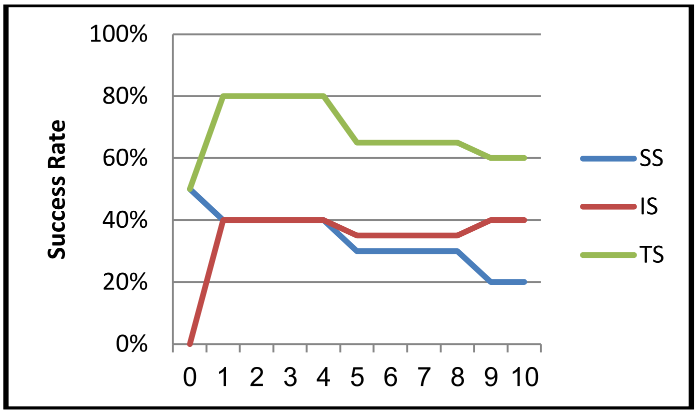

Figure 1: Relationship Between (K) Value and Success Rate

In Figure (1), the importance of using the smoother (k) is clear at the beginning of the curve, as the results otherwise would be so inaccurate, by using a smoother (k = 1), the accuracy of determining inbox messages was decreased by 20%, but the accuracy of determining spam messages was increased greatly, which has positively influenced the total successful attempts of filtering messages.
The results became more stable till (k = 4), after which both curves started decreasing at the same time, which decreases the total success curve. After (k = 8), results become more inaccurate as the system fails to detect spam messages.

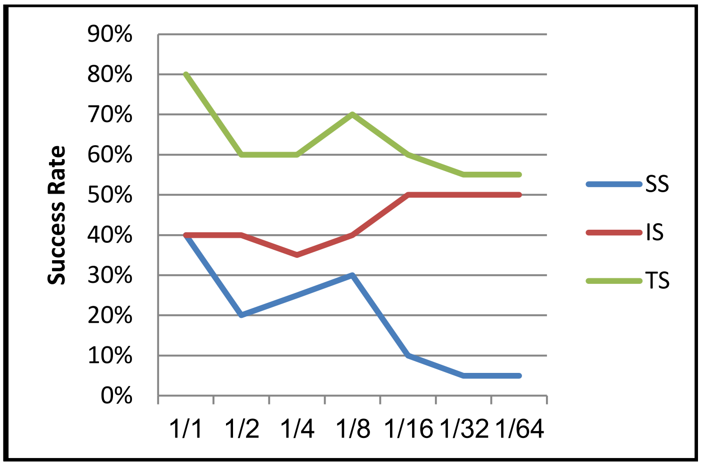

**Figure 2: Relationship Between Dictionary Resize Ratio and Success Rate**

In Figure (2)**:** After stating (k) value as (3) – at the preceding conclusion from previous experiments - the changes were made on the dictionaries' size. By taking the original dictionaries and comparing them with the result of applying the same test on the same messages using different dictionaries' size, the rate of success was dynamically changing between case (1) - the original size- and case (4) – where the size of the dictionaries became (1/8) of the original size, due to the content of the dictionaries at these points. After this point the system failed to detect spam messages from other messages, as the dictionaries became too small, which is considered as a drawback of the system.

## CONCLUSION

In this system we were able to classify Arabic incoming messages into spam or not using content based probabilistic model with a high success rate.

It can be concluded from Figure (1) that the maximum success rate is in the range of 1 to 4. By considering the values one and four as critical points and the left side of this curve as much more stable than the right side. The best (k) value would be (3).

From Figure (2), we note that the larger the dictionaries are the more accurate the result is. When the dictionaries' size is reduced, the content of those dictionaries plays an important role in detecting spam messages.

This emphasizes that reducing the size of the dictionaries would make a trade off with accuracy: gaining smaller sizes does not always maintain or enhance the total accuracy rate, since the total accuracy rate is the basic success measure, it is better to consider the dictionaries' sizes as a secondary element in this context, and not to make it a critical base for achievement. So within all cases; we found that it is not a good choice to reduce the dictionaries' size.

## REFERENCES

[1] Ahmed Khorsi, "An Overview of Content-based Spam Filtering Techniques", *Informatica*, vol. 31, no. 3, October 2007, pp 269-277.

*[2] Yizhou Sun, Hongbo Deng and Jiawei Han,"Probabilistic Models For Text Mining", pp 260-290*

*[3] C. Andrieu, N. De Freitas, A. Doucet, and M. Jordan. An introduction to mcmc for machine learning. Machine learning, 50(1):5 –43, 2003.*