Myanmar Spam Filtering using Bayesian Model

Thae Naw Naw Kyaw, Dr. Nyein Nyein Myo
University of Computer Studies, Mandalay
thaenawnawkyaw13@gmail.com

Abstract

Internet has become rich of information. This information should be properly managed. Electronic mail (E-mail) has become a major problem for internet users and providers. People are using e-mail services but most of the e-mails are irrelevant or junk called spams. Spam e-mail is well known problem for both corporate and personal users of e-mail. The volume of non-English language spam is increasing day by day. The motivation for this research is to find a solution for the internet users in the Myanmar language with Myanmar e-mail messages received every day in their mailboxes. So, a classification filter for these e-mails should be applied on e-mail servers. Bayesian approach is being popular for filtering spams. This approach is based on the bayes method. To filter the spam messages, this research applied Bayesian model for the framework of Myanmar e-mail spam filtering. In this paper, we are presenting e-mail mining and bayesian spam filtering method.

Keywords: *E-mail mining, Bayesian spam filtering*

1. Introduction

With the rapid development of internet, e-mail has become a powerful tool for information exchange in everyday life. As the popularity of e-mail increased, it becomes an important form of communication for many computer users, for both legitimate and illegitimate activities. Therefore, it is necessary to develop the technology of spam filtering [2]. Overall spam filtering task is divided into two steps. One of them is training of spam filter and other is classification of e-mails. In first step training of filtering is done by calculating probabilities and in classification step, an e-mail is classified based on the calculated probabilities.

Spams are undesired e-mails, which we don't to be in our e-mail account, so filtering of spam is becoming very necessary. E-mail servers offer a system for filtering e-mails and save our time and bandwidth. Spam filtering is the text classification technique which proved to be a great technique for dealing with spams. It refers to the automated assigning of e-mails to predefined classes as Spam or Legitimate e-mail based on their contents [10].

A variety of technical measures are: decision tree (DT), support vector machine (SVM), K-nearest neighbor algorithm (KNN), naïve bayes (NB), neural networks, etc. Most of the techniques above can be effective applied to the problem of spam, but among them, content-based filtering (Bayesian filtering) is playing a key role in reducing spam e-mail. And this approach is the statistical-based spam filtering method. The spam filtering is actually to classify the e-mails into legitimate and spam. This need to use the theory of Bayes to predict whether the received e-mail is spam or not, according to the correctly classified e-mails [14].

2. Related Work

There are some works in research compare different machine learning methods that filter anti-spam English e-mail messages. For example, work of [4] who presented an empirical evaluation of four machine learning methods which are Naïve Bayes (NB), Term Frequency – Inverse Term Frequency (TF-ITF), K-nearest Neighbors (KNN) and Support Vector Machine (SVM). Several solutions to the spam problem involve detection and filtering of the spam e-mails. Machine learning approaches have been used in the past for this purpose. Some examples of this are: Bayesian classifiers as Naïve Bayes [6], [7], Ripper [3] and Support Vector Machine (SVM) [5]. In many of these approaches, Bayesian classifiers were observed to give good results and so they have been widely used in several spam filtering software. In this paper, the Naïve Bayesian classification method is applied for analyses of spam and legitimate messages. In our case, the text documents are textual e-mails. In spite of the fact that there are many approaches to representation of text documents, the most widespread of them is the Naïve Bayesian classification method [6].

2. E-mail Mining

E-mail mining can be considered an application of the upcoming research area of text mining on e-mail data. Text mining is an emerging field that has attracted the interest of researchers from areas like

machine learning, data mining, natural language processing, and computational linguistics. But there are some specific characteristics of e-mail data that set a distinctive separating line between e-mail and text mining [13].

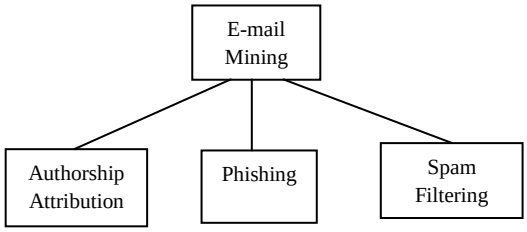

Figure 1. Categories of E-mail Mining

2.1. Authorship Attribution

E-mail authorship attribution means identify the most plausible author of an anonymous e-mail from a group of potential suspects. For author attribution various techniques used by various authors. The various topics on which work was done are gender, language, and various writing styles. Every person has unique identify, features and writing styles.

2.2. Phishing

Phishing can be defined as a scam by which e-mail users are duped into surrendering private information that will be used for identity theft. Phishing attacks use both social engineering and technical subterfuge to steal personal identity data and financial account credentials. It is one of the fastest growing scams on the Internet. The exclusive motivation of phishers is financial gain.

2.3. Spam Filtering

Spam is a big problem because of the large amount of shared resources it consumes. The time spent by people in reading and deleting the spam emails is a waste. Filtering is a simple and efficient way to combat against spam.

3. Spam E-mail Filtering

Filtering spam is a task of increased applicative interest that lies at the junction between filtering and classification. Several standard text classifications have been applied to spam filtering [12]. Bayesian approach is most popular of them. The major class of spam filter relies on information outside of the content of the individual e-mail messages. There are two types of spam filters: Reputation-Based and Content-Based. We will use the content-based spam filtering in our research. In content-based, there are three types of techniques such as heuristic (rule-based), fingerprint (honeypot, digest, signature/checksum schemes), and machine learing (zombie host detection, decision tree, artificial immune system, statistical methods, etc) [1].

3.1. Content-Based Filters

Content-based filters detect spam by examining the content of e-mail messages. These filters require the body of a message before they can classify messages as spam or ham. This content-based method is the most used method. Each message is searched for spam features like indicative words (e.g. “ဒီနေရာကိုနှိပ်ပ ”, “အောက်မှာဒေါင်းယူပါ”, etc), unusual distribution of punctuation marks (e.g. “!!!!!!!”,”$$$$$”), etc.

3.2. Machine Learning Filters

The training and testing is done using one machine learning classifier. There are various of machine learning methods such as Naïve Bayes (NB), Decision Trees (DT), Neural Network and etc.

Statistical filters rely on a corpus of spam e-mails and legitimate e-mails to conclude features which can be used to classify incoming e-mails. If the statistical properties are closer to corpus of spam e-mails, the e-mail is classified as a spam. Otherwise, e-mail is classified as a legitimate if the statistical properties are closer to legitimate e-mails corpus. A selection of statistical filters is: Bayesian, chi-squared, support vector machines (SVM)**,** boosting, maximum entropy models and memory-based learning techniques. Bayesian spam filters consider the historical probability of each word in the message occurring in either spam or legitimate messages [11]**.**

3.3. Bayesian Filtering Techniques

Content-based spam filters can be built manually, by hand-engineering the set of attributes that define spam messages. These are often called heuristic filters [5], and some popular filters like SpamAssasssin have been based on this idea for years. Content-based filters can also be built by using machine learning techniques applied to a set of pre-classified messages. These so-called Bayesian filters are very accurate according to recent statistics [9].

The Naïve Bayes (NB) learner is the simplest and most widely used filter that derives from Bayesian decision theory. NB classifier is the most employed in

spam filtering because of its simplicity and high performance [9]. From Baye's theorem and the theorem of total probability for a message with vector $\vec{x} = \langle x_1, ..., x_n \rangle$ belongs to a category $c_i \in \{c_s, c_l\}$ is

$$P(c_i \setminus \vec{x}) = \frac{P(c_i).P(\vec{x} \setminus c_i)}{P(\vec{x})} \qquad (1)$$

Since the denominator does not depend on the category, NB classifies each message in the category that maximizes $P(c_i).P(\vec{x}/c_i)$. In spam filtering domain it is equivalent to classify a message as spam (c_s) whenever

$$\frac{P(c_s).P(\vec{x} \setminus c_s)}{P(c_s).P(\vec{x} \setminus c_s) + P(c_l).P(\vec{x} \setminus c_l)} \rangle T \qquad (2)$$

with T = 0.5.

Bayesian filters automatically learn a spam classifier from a set of manually classified examples of spam and legitimate messages of the training collection. Bayes method is a probability-based approach. Five versions of Naïve Bayesian Classification are:

i. Multi-variate Bernoulli NB
ii. .Multinomial NB, TF attributes
iii. Multinomial NB, Boolean attributes
iv. Multi-variate Gauss NB and
v. Flexible Bayes

The Bayesian algorithm predicts the classification of new e-mail by identifying an e-mail as spam or legitimate. This is achieved by looking at the features using a 'training set' which has already been pre-classified correctly and then checking whether a particular word appears in the e-mail. High probability indicates the new e-mail as spam e-mail.

We have downloaded Myanmar e-mail messages from 10 personal mailboxes and various websites including Myanmar's news sites, Myanmar's political side, Adin garden website, and Myanmar's health website. The learning process takes as input the training collections, and consists of the following steps.

- *Normalization*, the text is converted to UTF-8 encoded.
- *HTML* and *XML* tags are striped out.
- *Tokenization* which divides the message into semantically coherent segments (e.g. words, other character string, etc.).
- *Representation* which converts a message into an attribute-value pairs vector, where the attributes are the previously defined tokens and their values can be binary, (relative) frequencies, etc.
- *Selection* which statistics detection of less predictive attributes (using e.g. quality metrics like Information Gain).

4. System Description

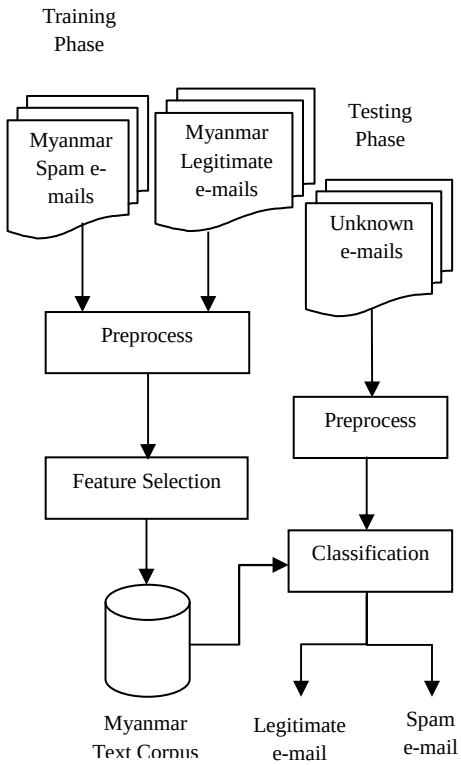

Figure 2. The System Structure for Training and Testing Phases

In this work, a system has been constructed to test the classifiers that classify e-mail messages. The structure of the system is depicted in Figure 2. The system consists of two subsystems; one for training and the other for testing.

In the above figure, the training phase takes the training e-mails such as Myanmar spam e-mail and Myanmar legitimate e-mails, and then preprocess these e-mails. And Feature selection trains the data for constructing the Myanmar text corpus.

In testing phase, the input of this part is unlabeled Myanmar e-mail messages after applying the same preprocessor. The output is classifying each new e-mail into legitimate e-mail message or spam e-mail message that uses the corpus.

4.1. Preprocessing

Before applying machine learning methods, for both training and testing datasets, some preprocessing text are performed in Myanmar text. All e-mails are

Myanmar text only. In the preprocessing, the text is converted to UTF-8 encoded. Then HTML and XML tags are striped out. Tokenization which divides the message into semantically coherent segments (e.g. words, other character string, etc.). For example. Input Myanmar messages စာတွေရော အဆင်ပြေလား □ မအား‌လို‌ပြန်‌မ‌ပြော‌ဖြစ်‌တာ‌ပါ After using the Myanmar Word Segmentation, စာတွေ့_ ရော_ အဆင်ပြေ_ လား_ ။ မအား_ လို့_ ပြန်_ မ_ ပြော_ ဖြစ်တာ_ ပါ_ ။ Representation which converts a message into an attribute-value pairs vector, where the attributes are the previously defined tokens and their values can be binary, (relative) frequencies, etc.

4.2. Feature Selection

In general, the size of the training corpus is large. To reduce the high dimensionality of the words, feature selection is performed. In this case the features are the words to be trained in e-mail messages. Feature selection is the attributes (words) dependent. Information Gain (IG) is computed to select the most appropriate feature in the text messages.

$$IG(X,C)=\sum_{x\in\{0,1\},c\in s\{pamlegitimat\}} P(X=x,C=c)\times\log\frac{P(X=x,C=c)}{P(X=x)\times P(C=c)} \quad (3)$$

In this formula IG is computed for words W and class C where C denoted the class (spam or legitimate). $P(W=w,C=c)$ is the probability that the word W occurs $(W=1)$ or does not occur $(W=0)$ in Spam $(C=spam)$ or legitimate $(C=legitimate)$ e-mail message and $P(W=w)$ is the probability that the word W occurs or not in all e-mail messages, $P(C=c)$ is the probability that an e-mail is spam or legitimate. Then, the features with the n highest IG score are selected.

4.3. Corpus

Corpus is not as large as public datasets. Myanmar e-mail corpus which is a collection of spam and legitimate messages from users' mail boxes. Corpus manually classified as spam or legitimate in the training phase. In Myanmar, there are 109 different languages and a plenty of text for these languages. In this system, we will use Myanmar corpus as training corpus for both spam and legitimate that is constructed by Information Gain (IG). Corpus structure changes according to the methods. Collection words have still not covered all the valid words in our corpus. So, we will collect words from the dictionary. We will use words such as (ထူးထူးခြားခြား□ ဟုတ်ကဲ့) in everyday life. The corpus will consist of 500 messages: 250 spam messages and other 250 messages are legitimate for training step.

4.4. Myanmar Spam Definition

- Message contains many special characters such as ။။။။, ။။။။, +++++, ---
- Message body contains phrases such as ဒီနေရာကိုနှိပ်ပါ, အောက်မှာဒေါင်းယူပါ
- The new message probability is equal to or greater than the pre-defined probability, the new message is spam
- Body message is plain text.
- Messages which occur less than 20 words are discarded (e.g. ဟုတ်ဘူး ၊ မေးလ်အခုမှတွေ့တာ၊အဆင်ပြေလား၊ဂိမ်းနှိပ်နေတယ် …).

Example:

Spam #1 for □□□:

from: nawnaw nawnaw thaenawnawkyaw13@gmail.com

to: Phyo ko ko

kyaw<bobokyaw7@gmail.com>, thenawnaw@gmail.com

date: Thu, Sep 12, 2013 at 1:39 PM

subject: မြန်မာစာ

mailed-by: gmail.comsigned-by: gmail.com: Important mainly because of the people in the conversation.

။။။ တစ်ခါတစ်ရံမှာ ။။။ နွေးထွေးမှုပေးတဲ့နေမင်းတစ်စင်းလို ။။။ တစ်ခါတစ်ရံမှာ ။။။ ၂ နှစ်သမီးလေးလောက်တောင် ။။။ မပါးနပ်သလို။။။

Spam #2 for [illegible]:

နိုင်ငံခြား သတင်းထောက်။ ။ ခင်ဗျားကို **လူဆိုးကြီး၊ သူဆိုးကြီး** လို့ ပြောကြတယ်။ ခင်ဗျား ဘယ်လို ထင်သလဲ။ အမေးခံရသူ။ ။ ဘာဖြစ်လို့ **လူဆိုးသူဆိုးကြီး**လို့ မေးတာလဲဗျ။ မြန်မာ ပြည်သားတွေကြားရင် ခင်ဗျား အရိုက်ခံရနိုင်တယ်။

နိုင်ငံခြား သတင်းထောက်။ ။သူကိုယ်တိုင်က ချမ်းသာပြီး မြန်မာပြည်မွဲရတာ၊ အဆင်းရဲဆုံး နိုင်ငံ ဖြစ်ရတာ သူ .လက်ထက်ကျမှ ပိုဆိုးတာလို့ ကြားရ ဖတ်ရလို့ပါဗျာ။

အမေးခံရသူ။ ။ အဲလို မမေးသင့်ဘူး။

နိုင်ငံခြား သတင်းထောက်။ ။ အော်ခင်ဗျားက အကောင်းမြင်သမားထင်တယ်။ ဘယ်လိုမေးသင့်လဲဗျ။ လုပ်စမ်းပါဦး။

အမေးခံရသူ။ ။အာဇာအလွဲသုံး **လူသတ်ကောင်၊ စစ်တပ်အသုံးချ ဓားပြအကြီးစား၊ တရုတ်အားကိုး သေချင်းဆိုး၊ ပြည်ဖျက်ကောင်၊ အယုတ်တမာကြီး**.....

4.5. Performance Measurement

Classifier needs to evaluate based on performance of information retrieval (recall, precision and derived measures) and decision theory (false positive and false negative). Accuracy, spam precision and spam recall are the most important performance parameters. Recall indicates the number of correctly classified spam against spam that is misclassified as legitimate and the number of spam recognized as spam.

Precision presents the ratio between the numbers of correctly classified spam to the number of all messages marked as spam. Accuracy represents the ratio between the number of correctly classified spam and legitimate mails to the total e-mails used for testing that is all e-mails are correctly classified by the classifier. These parameters can be measured using the following equations:

$$Accuracy(A) = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$\text{Pr } ecision \ (p) = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Re } call \ (r) = \frac{TP}{TP + FN} \quad (6)$$

$$F_i = \frac{2 * r * p}{r + p} * 100\,\% \quad (7)$$

Spam e-mails that are classified as legitimate e-mails are referred to as false negatives (FNs) where else legitimate e-mail classified as spam is referred to as false positives (FPs). True positive (TP) means spam e-mails that correctly predicted as spam; True negative (TN) is the number of e-mail that is legitimate and is truly predicted as legitimate.

5. Conclusion

E-mail is very popular and necessity of many users. The paper gives an overview of e-mail. E-mail mining raised new difficulties and challenges for the text-mining community. Spam filters used contents of e-mails body. This paper gives a framework of e-mail mining, and machine learning technique (Bayesian spam filtering).

References

[1] H.S. Alkahtani, P. Gardener-Stephen, and R. Goodwin, "A Taxonomy of E-Mail Spam Filters", King Faisal University, P.O, Kingdom of Saudi Arabia and Flinders University, Australia

[2] G. SUganya, S. KarPagavalli, and V. Christina, "A Study on Email Spam Filtering Techniques"..*Ingernational Jorunal of Computer Applications* (0975-8887), Volume 12-No.1.December 2010

[3] W.W. Cohen, "Learning Rules that Classify E-mail", in *Proceedings of the 1996 AAAI Spring Symposium on Machine Learning in Information Access*, pp.203-214, 1996

[4] C. Corts and V. Vapnik, "Support-Vector Networks", *Machine Learning*, 1995

[5] H. Druker, "Support Vector Machines for Spam Categorization", *in IEEE Transaction on Neural Networks*, pp 1048-1054, 1999

[6] P. Graham, "Better Bayesian Filtering", Internet.

[7] D. Heckerman, E. Horvitz M. Sahami, and S. Dumais, "A Bayesian Approach to Filtering Junk E-mail", in *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, pp.1048-1054, 1998

[8] G. Paliouras, I. Androutsopoulos, and V. Metsis, "Spam Filtering with Naïve Bayes – which Naïve Bayes?", *in Proceedings of the 3rd International Conference on E-mail and Anti-Spam, Mountain View*, CA, USA, July 2006, pp. 1-5

[9] G. Paliouras, G. Spyropoulos, I. Androutsopoulos, J. Koutsias, and K.V. Chandrinos, "An Evaluation of Naïve Bayesian Anti-spam filtering". *Proceedings of the Workshop on Machine Learning in the New Information Age*, 11th European Conference on Machine Learning (ECML 2000), pp.9-17, 2000

[10] A. Rajput, and D. Toshniwal, "Adaptive Spam Filtering based on Bayesian Algorithm". *Electronics and Computer Science Department*, IIT Rooorkee, India

[11] M. Sahami, and S. Dumais, "A Bayesian Approach fo Filtering Junk E-Mail", *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, Wisconsin, pp.1-8, 1998

[12] A.Y. Taqar, H.A. Jalab, and S. Thamarai "Overview of textual anti-spam filtering techniques", *Computer System and Technology*, Faulty of Computer Science and Information Technology, University Malaya, Malaysia, 2010

[13] G. Tsoumakas, I. Katakis, and I. Vlahavas, "E-mail Mining: Emerging Techniques for E-mail Management". Aristotle University of Thessaloniki, Greece, 2007

[14] L. Zhitang, W. Meizhen, and W. Hantao, "An improved Bayes algorithm for filtering spam e-mail". *Natural Science Edition* , Vol 37, No 8. Aug 2009.