International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Detecting Spam Vietnamese Email

Tisenko Victor Nikolaevich, Lai Van Duong, Ha Tuan Anh, Nguyen Quang Dam, Nguyen Quoc Hoang

Abstract- Spam messages have been causing much impact on email users. There are many different techniques and measures applied to detect and classify spam messages with other emails. Filters have been installed and configured to detect spam emails based on their identifying characteristics. However, relying only on filters is very easy to miss spam emails because spammers often try to use techniques to bypass the filter's control. The most effective approach in spam detection today is to find ways to analyze the content of emails. However, it is recognized that each natural language will have different analytical and identification characteristics, so it is not possible to use the same technique or method for all languages. In this paper, we will present a method to detect Vietnamese email based on the process of analyzing email content in both HTML and image formats.

Keywords: machine learning, spam, spam detection.

I. INTRODUCTION

1.1. About spam emails

There is currently no specific definition of spam email. It is hard to get the most accurate and complete definition because spam email is so personal. Many people argue that spam email is an "unwanted e-mail". This definition is not accurate, for example, an employee receives an email from their manager, which is an unwanted email but they are not spamming email. Other people said that spam email is "unsolicited commercial e-mail from the recipient" - these e-mails include advertising emails and phishing. This definition is not exactly accurate, which makes people believe spam email is junk mail. In 2005, at his work publishing, author Jonathan A. Zdziarski [1] stated: Spam email is a large number of unsolicited e-mail messages and most of them are an advertising and commercial e-mails. This can be considered as a clear and complete definition of spam email. Therefore, it is possible to define spam email as follows: Spam email is email in the form of pictures or text that are sent in bulk to many people. Spam email may contain meaningless content or advertising one which receivers feel troublesome. Even e-mail can be used to attach malicious code, viruses steal account e-mail, steal information from the recipient computer.

1.2. Some types of spam email

Phishing email: One of the most difficult types of spam email is phishing. These e-mails are designed to look like official e-mails from financial institutions or large companies, but they direct readers to phishing sites.

Revised Manuscript Received on March 30, 2020.
* Correspondence Author
Tisenko Victor Nikolaevich*, Peter the Great St. Petersburg Polytechnic University Russia, St.Petersburg, Polytechnicheskaya, 29 *v_tisenko@mail.ru;*
Lai Van Duong, FPT University Hanoi, Vietnam
Ha Tuan Anh, FPT University Hanoi, Vietnam
Nguyen Quang Dam, FPT University Hanoi, Vietnam
Nguyen Quoc Hoang, FPT University Hanoi, Vietnam

This type of email prompts the readers to enter their user names and passwords, which is then used by the site owner, then phishers use them to violate real accounts. Another form of spam email that is commonly used is Anti-Virus Spam email. This form of spam email usually sends messages to victims and informs them that their computers are infected with viruses. Some victims will be tricked into downloading antivirus software. Victims think they are downloading security software but are downloading programs that contain viruses. Another form of phishing spam email is Political or Terrorist Spam email. This type of spam email is one of the ways to steal personal information. The characteristics of this technique are fake from a famous politician or government office, claiming that the reader is in danger. To clarify the threat, e-mails require victims to provide personal information and sometimes cash. Email sent in large numbers: Emails are sent in large numbers, causing the reader to be upset because of the increased e-mail in the recipient's directory. Most spam email is not appreciated. Every day, hundreds of billions of advertising e-mails are sent, mostly selling advertising, online courses, cure Trojan Horse email: Trojan Horse email is considered as malicious e-mail messages that not only infect the victim's computer but also send to everyone in the victim's contact list. Once opened and downloaded, the attachment will damage the local computer and send it to the victim. Porn Spam email: Porn spam email is widely used and considered as a leading source of malicious content. Spam emails collect or buy people's email addresses, sending advertising emails that direct victims to pornographic sites

1.3. Contribution of the paper

In this paper, the author presents a method to detect Vietnamese spam emails using a machine learning method based on the email content analysis technique. Major contributions of the paper include:

- Developing a process to detect Vietnamese spam emails
- Apply the Vietnamese content analysis technique to the process of analyzing email content on photos and on HTML.
- Using machine learning techniques to classify spam messages.

II. RELATED WORKS

2.1. Some techniques for detecting Spam email

Today, in order to cope with the growing number and complexity of distributed forms of spam email, security experts recommend individuals and organizations to adopt some of the techniques below. These techniques are common techniques that contribute to preventing spam email.
Mail Blacklist/Whitelist: This method is based on an available list of email addresses including addresses of specialized email servers used to send spam email.

Setting a blacklist of black email addresses or e-mail servers will be confirmed by the user [3]. Some ISPs will use this type of blacklist and automatically refuse to receive email from addresses on that list. Users can install their Blacklist at their source email [2]. From the definition of spam email detection using Blacklist, it can be seen that this technique is sometimes not effective because the detection technique is based only on the given list. In addition, this detection technique often has a high false detection rate. Besides, this method may be overlooked if spam emails send back emails through a forwarding server that is not on the Blacklist list. In contrast to setting up a Blacklist, users can also set up a "Whitelist". The email addresses listed in this list have been automatically accepted by ISPs. By default, all other emails will be rejected. If the spam mailer sends an email with the "sender" of the email with the same whitelist domain, spam email emails can still be accepted.

Signature/Checksum: This is one of the methods of categorizing email based on content. When an email arrives, the "Signature / Checksum" value is calculated for each email and compares it to the calculated value from the spam email emails on the Internet. If the incoming "Signature / Checksum" value is the same as any value in the database, then the email is considered spam email [3]. Obviously, with this detection technique, it will not be effective because spam emailers can simply add a few letters or a nonsense word in spam email emails to make "signature" value different. Therefore, anti-spam email protection using the "signature/checksum" filter has never been a good way. However, this technique has the advantage of not properly classifying legitimate emails.

Heuristic Filtering: Heuristic Filtering uses a set of common rules to identify the quality of a particular spam email message. These attributes may be in the content or obtained by observing the specific structure of the spam email. Unlike primitive filters, heuristic filters have rules to detect both spam email and legitimate email. The simplest way to remove emails containing "bad" words (for example words that usually appear or only appear in spam email messages). This is also a weakness that spam emailers can use to pass filters by trying to avoid using the "bad" words and replacing them with "good" words (used in non-spam email emails). Meanwhile, legitimate emails can be removed if they contain some "bad" words. This makes the efficiency of this technique low. Another weakness is that these rules are static. When spam emailers find a new way to pass, filter builders will have to write new rules [3].

Challenge/Response: Challenge / Response is a similar approach to the Whitelist technique. Challenge / Response will automatically send a challenge message to the sender. In this message, the sender is asked to do a few actions (such as clicking on a link) to define the information before the email is sent to the recipient. Challenge / Response has pushed maintaining Whitelist's responsibility for the sender [2]. The advantage of this method is to leave out very little spam email. This method is suitable for users who want the sender to be authenticated before communicating and not interested in losing the mail from another source. The weakness of this approach is that it interferes with the sender. By using this technique, it is important to determine who sent the email. The system using this technique will have many non-spam email emails removed and delay time too long and consume bandwidth resources. For example, a person would like to invite a new friend to a party, but the friend will only see the email reply the next day and it's too late. The user will have to respond to the challenge messages leading to they will not want to communicate with the people requesting them to respond to the challenge letter.

Address Obfuscation: Address Obfuscation technique is a technique for disguising email addresses to hide that address from the spam emailer. This technique is used to against bots (a small computer program) that collects new e-mail addresses on websites to list as spam emailers. The feature of this technique is that instead of displaying an e-mail address like abc@fpt.edu.vn, users can see "abc [at] fpt dot edu dot vn". However, this approach will not work well because the bot crawlers are more and smarter, it can reassemble the above email address. In addition, spam emailers also receive email address information from users. For example, many ISPs and credit card companies sell lists of addresses for spam emailers.

Machine Learning: Machine Learning is a process having a computer program improves its performance at work through experience [4, 5]. Application of machine learning methods in classification problems, especially text classification into email classification problems, machine learning algorithms such as Naïve Bayes, Support Vector Machine [4], were used in the field of text classification. The main idea of this approach is to build a sector to classify a new sample by training the existing sample.

2.2. Some spam email detection tools

Apparently, spam email is causing more danger to organizations and individuals using email. Organizations and security vendors have been creating a number of technologies, including free and paid technology to help users and organizations detect spam email. Some open-source technologies are being used for spam email detection as follows:

Spam email fighter Pro: SPAM EMAIL fighter Pro is a dedicated tool to block spam email on your PC. SPAM EMAIL fighter Pro uses blacklist, whitelist, and heuristic filtering techniques to classify emails. It is compatible with email applications such as Outlook, Outlook Express, Windows Mail and Thunderbird. The program will integrate into the application and conduct an analysis of incoming emails to identify spam email. It also protects against phishing, identity theft, and email fraud. The latest version of the Spam email fighter tool is 7.6.144. Users can access the internet and download the latest version of Spam email fighter Pro [6].

Mail Washer Pro: Mail Washer Pro is a spam email filter with blacklist/whitelist and heuristic filtering. Mail Washer Pro also allows users to preview all emails and delete them before they access the email program. Besides, Mail Washer Pro also offers many filters, blacklist for users to choose from. Users can access the internet and download the latest versions of Mail Washer Pro [7]. In addition, users can learn about the new features of Mail Washer Pro on the website of this tool.

Choice Mail One: Choice Mail One is a powerful software to filter, remove and block spam email access to email addresses. Moreover, it also allows users to block receiving mail from anyone named in the email address list. This tool uses blacklist, whitelist, heuristic filtering and challenge/response techniques to classify emails [8].

Spam emailAssassin: Spam email Assassin is a software system that analyzes and evaluates incoming emails and concludes that it is spam email or legitimate email. The system is based on multiple detection techniques, where the main method is to compare different parts of an email with predefined codes. With each corresponding law, the email rating will be increased or decreased. An email that reaches a threshold. If it is high enough, it will be considered a spam email. Spam email Assassin rules consist of three parts: the header or body, the description and the score [8]. Spam email Assassin is an open-source tool that can handle many languages and it is supported by most email servers. Although there are many products that detect and block spam email, Spam email Assassin is one of the most popular.

III- Vietnamese spam email detection method

3.1. Model recommendation

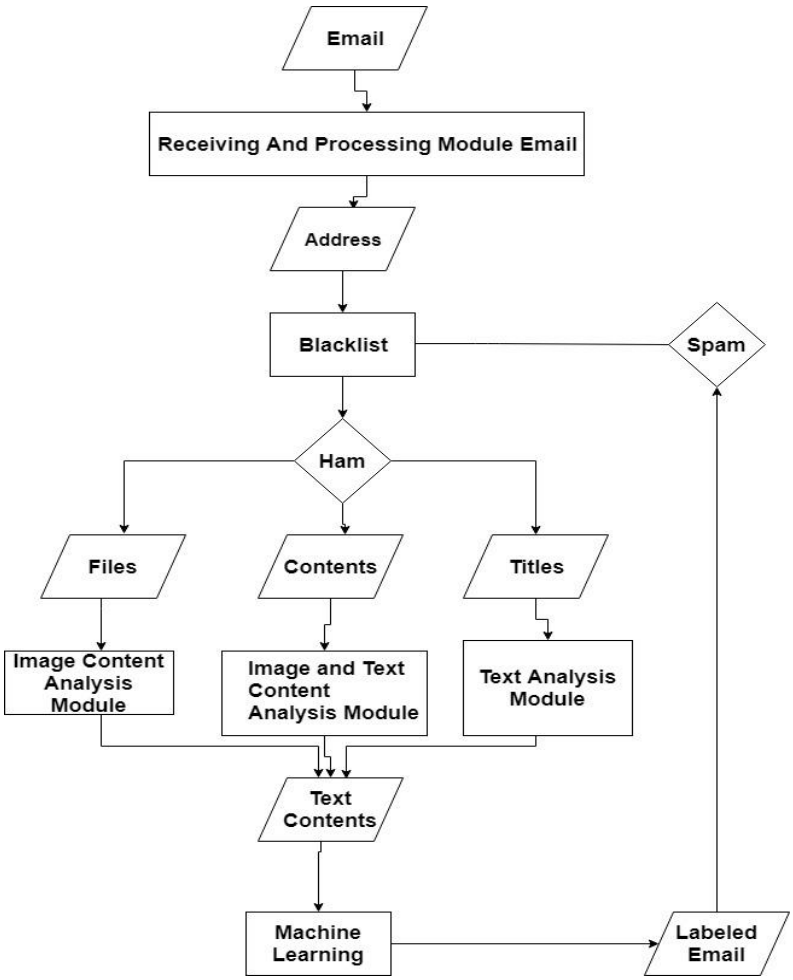

Figure 1. Model of Vietnamese spam email detection

When the system receives a new email, the process will read each part of the entire message package of the email, including headers, attachments, and content. The first classification module that can be used is Subject-based, IP-based filters (Blacklist / Whitelist). If spam emails are detected (for example, emails with addresses on The Blacklist), this filter will return the classification results to the program and may stop the classification process. If the email passes the header-based filter, the content-based classification module will classify the text content classification in the email, including the text in the inline-image.

Inline images are images embedded in the HTML code of an email, which can be displayed directly when the user reads the email. Inline images are often included by spammers in email content, usually product descriptions, brand logos, advertising banners.…

The recognition of characters in inline images requires special support tools. We will use the VietOCR tool - a tool to recognize Vietnamese and English words in images. VietOCR is built on the open-source library Tesseract of Goole. This tool supports the most popular printed fonts in Vietnamese, can reach 97% accuracy for Vietnamese content. The result of the process of separating the content in the image can be incorporated into the text of the email. Here, the spam filter with a trained model will classify this email as spam or ham and return results to the program. Details of the handling process in the Vietnamese spam email detection model using machine learning are as follows:

- Email pre-processing:

The content of the message after the HTML tags are removed, only the plain/text part will be included in the standardization process of the message by: Remove special characters; Eliminate numbered characters and hyphens; Spelling correction.

The preprocessing step minimizes noise in the data and reduces processing time.

- Analyzing the content of letters

From the above comparison table, content processing and bilingual mail training can be handled on the same model. The requirement is to be able to handle each specific part of both languages, including separating words in Vietnamese and returning them to their original forms in English.

.

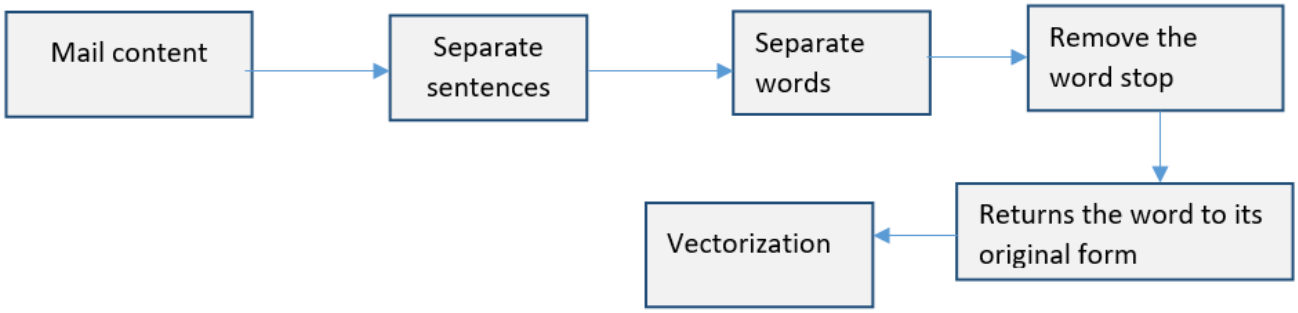

Figure 2. Process of processing Vietnamese mail content

The process of handling the content of bilingual English-Vietnamese letters is described in Figure 2. In which:

Step 1: Separate the sentences

Based on the sentence splitting symbols, the word processor can split the paragraph into separate, grammatical separated sentences. Common sentence separators:

- Dots (.)
- Comma (,)
- Semi-colon (;)
- Double quotes (“)
- The colon (:)
- Square brackets []
- Braces {}
- Parentheses ()
- Word operators + - / * = <>

Step 2: Separate words

To separate single and compound words in Vietnamese, we use the vnTokenizer tool [9]. This tool is based on the Maximum Matching method with a set of data used in Vietnamese syllables and Vietnamese vocabulary dictionary. Process of performing magnetic separation of vnTokenizer tool:

- The input of the tool separating from vnTokenizer is a sentence or a document saved as a file (encoded using Unicode UTF-8 encoding).
- The output is a sequence of separated word units, words in the dictionary with punctuation, strings of numbers, foreign characters.

Step 3: Remove the word stop

In English and Vietnamese, stop words only have a grammatical meaning and not a vocabulary meaning. When it comes to a stop word we have no knowledge of things or phenomena. Afterword separation is complete, stop words can be omitted and not affect the text content. Common stop words in English are articles (articles), prepositions (prepositions), conjunctions (conjunctions) and some pronouns (pronouns). Some typical examples are: "a", "about", "an", "are", "as", "at", "be" ... Vietnamese stop words include many different components including: including compound words. Some typical examples are: "suffer", "by", "all", "the", "sure", "sure" ... To eliminate stop words in emails, we use a dictionary. stop words 3240 words, which include both English and Vietnamese words. Vietnamese stop words of compound words have been separated from the word. The handler will in turn check and remove words in the email content if they are in the list of stop words, and the sentence separator, extra spaces will be removed.

Step 4: Return the word to its original form

To convert English words into their native form, we use the Stanford CoreNLP library [10] from Stanford University, USA. The Stanford CoreNLP library provides a set of natural language processing tools. This tool can find out basic forms of words, parts of words, whether it's the company name, person name, etc. by analyzing sentence structure in phrases and grammatical dependencies. , indicates terms that refer to the same entity.

Step 5: Vectorization

To vectorize text documents, the program needs a dictionary-like mapping. With the built-in word set, the processed text data will be converted into a number format (corresponding to the order in the dictionary). In this step, words that are not in the list of dictionaries and punctuation will be considered as meaningless and removed.

3.2. Algorithm for detecting spam email

There are many different algorithms that assist in the classification to detect spam emails and regular emails. However, in this study, we choose to use the Naïve Bayes algorithm.

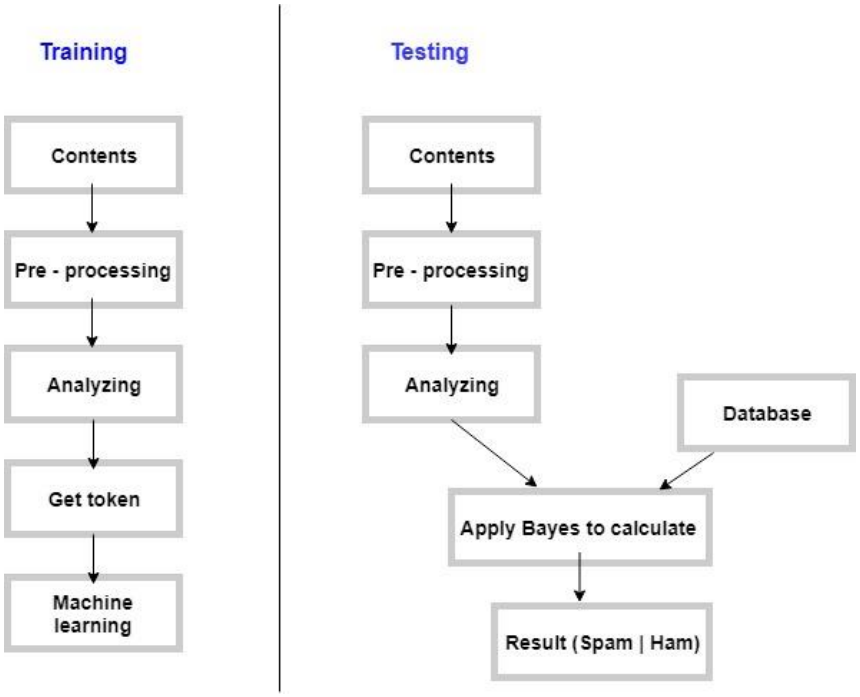

Figure 3. Vietnamese spam email detection process using the Naïve Bayes algorithm

The process of applying the Naïve Bayes algorithm in our research is as follows [4, 5]:

Assume that *D* is a training set consisting of the patterns in the form $X = \langle x_1,...,x_n \rangle$. $C_{i,D}$ is the set of samples of *D* which belong to class $C_i (i = \{1,...,m\})$. The attribute $x_1,...,x_n$ are independent in a condition of every pair. Classification algorithm using the Naïve Bayes algorithm:

Step 1: Train Naïve Bayes on the training data set. The calculation of values $P(C_i)$ and $P(X_k|C_i)$ according to the Naïve Bayes theorem is presented in [4, 5].

Suppose $X = \langle x_1,...,x_n \rangle$. In it, x_i receive discrete values. In that, $P(C_i)$ and $P(X_k|C_i)$ calculated by the formula:

$$P(C_i) \approx \frac{|C_{i,D}|}{D}$$

$$P(X_k|C_i) \approx \frac{C_{i,D}\{x_k\}}{|C_{i,D}|}$$

To avoid the case of value $P(X_k|C_i)$ = 0 because there is no sample in the training data, it should be smooth by adding some virtual samples. Apply Laplace smooth method [12], obtained:

$$P(C_i) \approx \frac{|C_{i,D}+1|}{D+m}$$

$$P(X_k|C_i) \approx \frac{C_{i,D}\{x_k\}+1}{|C_{i,D}|+r}$$

With m is the number of class need classified, r is the discrete value of the characteristic

Step 2: Newly unclassified data points will be labeled according to the class which has the largest formula value:

$$\arg\max_{C_k} P(C_i)\prod_{k=1}^{n} P(x_k|C_i)$$

In that, *n* is the number of words in the text to be classified

IV- Experimental and evaluated

4.1. Experimental data

4.1.1. English email data

English spam data is taken from the Enron dataset [11]. This is the personal data of more than 150 employees of Enron company, USA. This data includes a large number of personal emails, made public and used as a standard dataset to evaluate spam filters. Enron mail including spam and spam, a total of 21783 mail is divided into 6 parts simulating many different situations that users may encounter in reality. All messages have been preprocessed (removing information fields, HTML tags, and non-Latin characters). The paper will use Enron 1 for training and part of Enron 2 for testing. The total number of English letters used is 15,000 letters - equal to the number of emails of the Vietnamese set.

4.1.2. Vietnamese data

Vietnamese spam data is created from newsletters, recruitment, promotions, scams and partly taken from the mailbox of the Gmail account. In particular, the spam content often appears as:

- Advertising: discounted products, English courses, dance courses, study abroad advice, medicine, cosmetics, recruitment information, …
- Service brokerage: travel, healthcare, securities, virtual currency trading, pornography, online games, …
- Phishing: notice of winning, request for money transfer, …

4.1.3. Summary of data sets

All emails are separated from the body of text, correct spelling and numbered in ascending order. These emails contain both Vietnamese and English content. The data set after being processed is as follows.

- The following is data which is used to perform an experiment:
- 9000 Vietnamese clean letters
- 6000 spam messages in Vietnamese
- 9000 clean English letters
- 6000 English spam messages
- 500 images with spam email contents
- 500 images with ham contents

4.2. Experimental measurement

- TP (True Positive): Spam email detection rate correctly
- TN (True Negative): Ham detection rate correctly
- FP (False Positive): Incorrect detection rate of ham emails as spam email
- FN (False Negative): Incorrect detection rate of spam email mails as ham
- Accuracy (AC): The accuracy to detect correctly spam email mail
- Training time (TT): Time is required to make training data
- Prediction time (PT): Time required to get calculate the result

4.3. Experimental results

Experimental results of detecting Vietnamese spam emails are shown in Tables 1 and 2 below.

Table 1. Compare experimental results with text content

		Ham email	Spam email	AC (%)	TN (%)	TP (%)	FP (%)	FN (%)	TT (m)	PT (m)
	Total	**9,000**	**6,000**							
Protected	Ham	8,532	127	97.8	94.8	97.8	5.2	2.11	100	250
	Spam email	468	5,873							

From the above test results table, we can see that the tested bilingual mail filter model has quite accurate classification results for both languages. The highest filtering rate of Vietnamese spam reached 97.8%.

Table 2. Compare experimental results with images content

		Ham email	Spam email	AC (%)	TN (%)	TP (%)	FP (%)	FN (%)	TT (m)	PT (m)
	Total	**500**	**500**							
Protected	Ham	480	40	92	96	92	4	8	100	250
	Spam email	20	460							

From the test results table in Table 2, it can be seen that the bilingual mail filter model tested has quite accurate classification results for both languages. The highest filtering rate of Vietnamese spam reached 92%. This empirical result shows that the supporting tools for word separation and image processing applied in the paper have provided good results and can be applied in the detection models

III. CONCLUSIONS

In this paper, we have presented methods of detecting Vietnamese spam emails using machine learning algorithms, including Vietnamese word separation and analysis techniques. Experimental results in the paper showed that our proposed spam detection system quickly and accurately detected spam emails in English and Vietnamese not only in text form but also in image format. This proves that the process of separating and processing Vietnamese words applied in the detection system has brought high efficiency. In the future to optimize the detection process of Vietnamese spam emails we will incorporate more methods of word prediction and natural language processing.

REFERENCES

1. Zdziarski, Jonathan. (2005). Ending Spam: Bayesian Content Filtering and the Art of Statistical Language Classification.
2. P.Graham, Stopping Spam email, http://www.paulgraham.com/stopspam email.html, [Accessed February 15, 2020].
3. The Spam emailhaus Block List. https://www.spam emailhaus.org/sbl/. [Accessed February 15, 2020].

4. Smola, A.; Vishwanathan, S.V.N. Introduction to Machine Learning; Cambridge University Press: Cambridge, UK, 2008.
5. Stanford University, Text Classification and Naïve Bayes, 10/2011.
6. MAILWASHER PRO STOPS SPAM EMAIL. https://www.firetrust.com/products/mailwasher-pro[Accessed February 15, 2020].
7. ChoiceMail One 2010.1 Review. https://www.toptenreviews.com/software/security/best-spam email-filter/choicemail-one-review/[Accessed February 15, 2020].
8. A. Schwartz, Spam emailAssassin, Sebastopol, CA: O'Reilly, 2004. [Accessed February 15, 2020].
9. http://mim.hus.vnu.edu.vn/dsl/tools/tokenizer[Accessed February 15, 2020].
10. Zimbra Collaboration 8.7 Open Source Messaging and Collaboration. https://s3.amazonaws.com/files.zimbra.com/public/collateral/Zimbra %20Collaboration%20Datasheet%20-%20VT.pdf [Accessed February 15, 2020]
11. Enron Spam Dataset, http://www2.aueb.gr/users/ion/data/enron-spam/ [Accessed February 15, 2020]
12. Vartziotis, Dimitris & Himpel, Benjamin. (2014). Laplacian smoothing revisited.

AUTHOR PROFILE

Tisenko Victor Nikolaevich My position is the professor of Institute of computer sciences and technologies in Peter the Great Saint-Petersburg Polytechnic Ubiversity. I have received the degree Doctor of Technical Sciences in 1998 in accordance of scientific speciality "Systems of automatic Desing" in SPbPY. The area of scintific interest is use of new type of fuzzy logics in different applications. I think that we could cooperate intensively in future. E-mail: ***v_tisenko@mail.ru***

Nguyen Quang Dam, are fourth-year students majoring in information security at FPT University. These students have over 2 years of experience working with APT attack detection issues. E-mail: longdhse05220@fpt.edu.vn

Ha Anh Tuan, are fourth-year students majoring in information security at FPT University. These students have over 2 years of experience working with APT attack detection issues. E-mail: hiepnvtse05065@fpt.edu.vn

Le Van Duong, are fourth-year students majoring in information security at FPT University. These students have over 2 years of experience working with APT attack detection issues. E-mail: anglqse04676@fpt.edu.vn

Nguyen Quoc Hoang are fourth-year students majoring in information security at FPT University. These students have over 2 years of experience working with APT attack detection issues. E-mail: hoangnqse06012@fpt.edu.vn