Social-spam Profile Detection based on Content Classification and User Behavior

Thi-Hong Vuong[1], Van-Hien Tran[1], Minh-Duc Nguyen[1], Cam-Van Thi Nguyen[1],
Thanh-Huyen Pham[1,2], Mai-Vu Tran[1]

[1] Knowledge Technology Laboratory, University of Engineering and Technology, Vietnam National University Hanoi

[2] Ha Long University, Halong, Vietnam

{hongvt_57, hientv_55, ducnm_57, vanntc_58, vutm}@vnu.edu.vn

phamthanhhuyen34@gmail.com

***Abstract*—Web-based social system enables new community-based opportunities for participants to engage, share and interact. The rapid growth of Facebook has triggered a dramatic increase in spam volume and sophistication. Spammers post their status or comment in Page to send spam content to their friends or other users in the network. In this paper, we consider the problem of detecting spam accounts on Facebook based on comment content and user social behavior. We will propose a hybrid approach using Maximum Entropy (Maxent) model for classifying user comments as either spam or non-spam. We carefully conducted an empirical evaluation for our model on a large collection of comments in Vietnamese Facebook Pages and achieved promising results with an average accuracy of more than 90%.**

Keywords— social networks, spam accounts detection.

I. INTRODUCTION

The leading social networks are usually available for multiple languages and enable user to connect with friends across the globe. Approximately 2 billion internet users are using social networks and this figure is still expected to grow due to an increasingly prevalent trend of using mobile devices. According to Statista report[1], until April 2016, Facebook became the first social network to surpass 1 billion registered accounts with 1.59 billion monthly active users. In recent years, social networks have increasingly relied on social data to provide suitable information to their users. Many researches, Facebook users discover content based on what their friends and networking community like and comment [15]. However, how can we trust content generated by other users? Since social network has become more familiar with daily life, from work to shopping to socializing, it leads to a focus of spammers attempting to make off with money from Internet users by taking suspicious behavior. Unfortunately, the relative openness and reliance on users along with the growth of these social systems has also made them prime targets of social spammers.

On Facebook, Pages are used by organizations to interact with users. Users can comment on a Page to let their friends know about their interests and to receive content from the Page in their News Feed, the primary distribution channel on Facebook. At Facebook, there are many ways that attackers have attempted to spam content on Pages through a variety of deceitful methods, including malware, credentials stealing, social engineering, and fake accounts. However, Facebook already has many anti-phishing, anti-malware mechanisms making it difficult for real accounts to be compromised, and many algorithms to detect fake accounts. As a result, it is hard for an adversary to control many accounts and instead they need to use the same few to Like or Comment many Pages.

In our research, we focus on the problem of detecting spam accounts on Facebook. Features for detection of spammers could be content-based and user social behavior-based methods. We tried to collect a large range of user comments from selling Facebook pages in Vietnam, combining with user social behavior to build a dataset of users. We also have applied Maximum Entropy (Maxent) method on the dataset to create a model which detects spam accounts. The results of experiments conducted show the improvement in detecting spam account accuracy.

This paper is organized as follows. Related works are introduced in the second section. The framework of detecting spam accounts with more details in the third section. Experiments and results are presented in the fourth section. Conclusions are showed in the last section.

II. RELATED WORK

In this section, we review several major approaches to detect spam accounts in social networks. With the rapid development of social networks, social spam has attracted a lot of attention from both industry and academia. In industry, Facebook proposes EdgeRank algorithm[2] that assigns each post with a score generated from a few feature (e.g. number of likes, number of comments, number of reposts, etc.). Therefore, the higher EdgeRank score, the less possibility to be a spammer. This disadvantage of this approach is that spammers could join their networks and continuously like and comment each other in order to achieve a high EdgeRank score.

[1] http://www.statista.com

[2] http://techcrunch.com/2010/04/22/facebook-edgerank

978-1-4673-8929-7/16/$31.00 ©2016 IEEE

In academia, most detecting spam accounts are based on contents which exploit input of user profile, their comment and user social behavior. They have used machine learning to classification spam accounts or non-spam accounts. A few years ago, many researches around the world to solve the problem detecting spam accounts on social networks, such as [2, 5, 7, 9, 20] .The related works are indicated to be solutions and features to differentiate spam accounts on a particular social network to prevent or remove them. There are two common approaches to identify spam accounts: content-based method, such as [1, 3, 4, 10, 11, 16, 19] and social graph method, such as [12, 13, 17, 18].

Stringhini et al. [14] further investigates spammer feature via creating a number of profiles in three large social network sites (Facebook, Twitter and Myspace) and identifies five common potential features for spammer detection. Lee et.al [8] deployed social honeypots consisting of genuine profiles that detected suspicious users and its bot collected evidence of the spam by crawling the profile of the user sending the unwanted friend requests and hyperlinks in social networks (MySpace and Twitter) . Significant work has been done by Alex Hai Wang [16] in 2010 which used user-based as well as content-based features for detection of spam profiles in Twitter with Bayesian classification algorithm and presented user's behavior by social graph. Classic evaluation metrics have been used to compare the performance of various traditional classification methods like Decision Tree, Support Vector Machine (SVM), Naïve Bayesian and Neural Networks and amongst all Bayesian classifier has been judged as the best terms of performance. Grier et al [6] identified features related to tweet content and community features in Tweet. These characteristics are regarded as attributes in a machine learning process for classifying users as either spammers or non-spammers. However, these approaches are based on a large amount of selected features that might consume heavy computing capability and spend much time in model training. Beutel et.al proposes COPYCATCH which detects lockstep Page Like patterns on Facebook by analyzing only the social graph among users, pages and the times at which the edge in the graph were created [2].

From above analysis, we developed a spam account detecting model based on content and user social behavior. By using Maximum Entropy Modeling, it determines users as either spammers or non-spammers and makes the following contributions:

- A feature selection solution is applied to improve performance of obtained model.

- Two versions of data features are determined. Experiments show the feature combination of comment content and user behavior are better than the one of them.

III. Our Approach

In this section, we present the conceptual framework of the proposed approach and outline the research questions motivating our examination of this framework.

A. Problem Statement

The paper focuses to propose a frame of spam accounts detecting modeling in Facebook pages. In our research, the problem is described as follows.

In Facebook social network, there are a set of *n* users **U** = $\{u_1, u_2 \ldots u_n\}$ and a set of *m* pages **P** = $\{p_1, p_2, \ldots, p_m\}$. Each user u_i makes a comment on page p_j at the time *t*.

The spam account detection problem is to predict whether u_i is a spammer through a binary classifier **c**: $u_i \rightarrow$ {spammer, non-spammer}. To build **c**, we need to select a set of *l* features **F** = $\{f_1, f_2 \ldots fl\}$ from user comment dataset and user social behavior on Page p_j at the time *t*.

In this study, we consider two kinds of predictive models, related with two cases of data presentations. The first case, we only used user's comment on pages to build a predictive model for spam accounts. However, we utilized both comment contents and user social behavior for constructing the model in the second case and compared with the original model.

B. Solution Approach

We solved the problem of detecting spam accounts based on comments content and user social behavior on Facebook pages. **Fig.1** describes our frame of spam accounts detecting model, including three parts: Crawling and Preprocessing Data, Modeling and Using Outcomes.

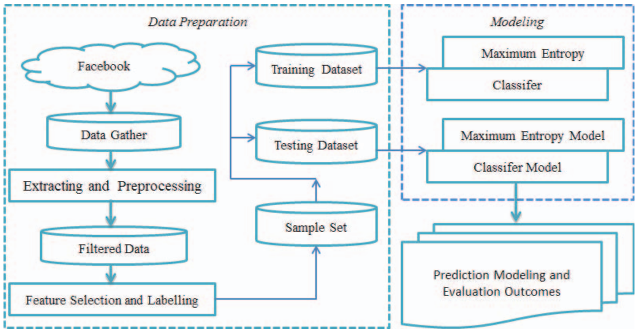

Fig.1. A social spam framework

1) Data Preparation Phase:

Firstly, we use Facebook's API methods and the Restfb library to gather a large of data from public Vietnam Facebook pages through searching related keywords, including page's information, user's comments and user's information. Each page will only get a post along with all its comments. After that, we extract valuable information like userID, pageID and comment; preprocess them by removing stop-word, error as well as splitting sentence.

Secondly, we select only a part of the data set to label for building the sample set. A group of students are asked to label it. They read all comments and assign labels (either SPAM or NON-SPAM) to them based on the agreement among them. For building the classification with MaxEnt, we need to define our features, including n-gram and user behavior feature. Depending on experiment, we obtained the best result when combining 1-gram and 2-gram. By observing user's behavior feature for detecting spammers, we calculated the number of comments of a user on one minute to make a user behavior feature because a normal person will not normally reply more 5 comments per 1 minute.

2) Modeling Phase: In this phase, Maximum Entropy Classification Model is applied because the model not only has the highest entropy but also satisfy constraints observed from empirical data. We selected features based on user comments and user behavior in pages. Then training data is used for modeling and the testing data is used for evaluation of the result model. As above description, the first case of building modeling is exploited user comments to select features. However, we supplement user behavior-based feature along with content-based features in the second case. Finally, comparing the effect of them help understand better about the importance of features.

IV. EXPERIMENTS AND RESULTS

In order to evaluate our approach above, we build an empirical model which detects spamming comments on the field of selling Facebook pages. Depending on the effective of the model, we deeply understand valuable content-based and user behavior-based features to distinguish spam accounts.

A. Experimental Data

We collected a large amount of data from Vietnam Facebook pages on the same day, based on domains keywords search. The data set contains 941,038 comments by 478,496 users from 23,461 Vietnam Facebook pages. We also selected a part of the data set to label manually for sample set. Depending on the survey and analysis the collected data, there are some signals to identify a comment by a spammer such as: the number of tagged people (>=1), containing links (https://; www), the length of comments and so on. The labeled data set is shown as follows:

TABLE I. **The information of the sample set**

Label	The number
SPAM	4,864
NON-SPAM	4,692

The sample set was divided randomly into four parts. We in turn took three parts for training and the one left for test to perform 4-fold cross-validation tests. The experimental results will be reported in the next subsection.

B. Experimental Results and Analysis

TABLE II. **The precision, recall and F-score of two classes of the best fold**

Class	Labeled	Predict	True positive	P	R	F1
SPAM	1197	1164	1093	93.90	91.31	92.59
NON-SPAM	1192	1225	1121	91.51	94.04	92.76
$F1_{macro}$						92.69
$F1_{micro}$						92.67

while the second used both content-based features and behavior-based features. As we can see, classification using behavior features give a better performance. Behavior features can improve the F1-score measure for more than 2% on average. From the results of 4-fold cross-validation tests, the results are quite stable over the four folds. This shows that the classification model works well on this data set.

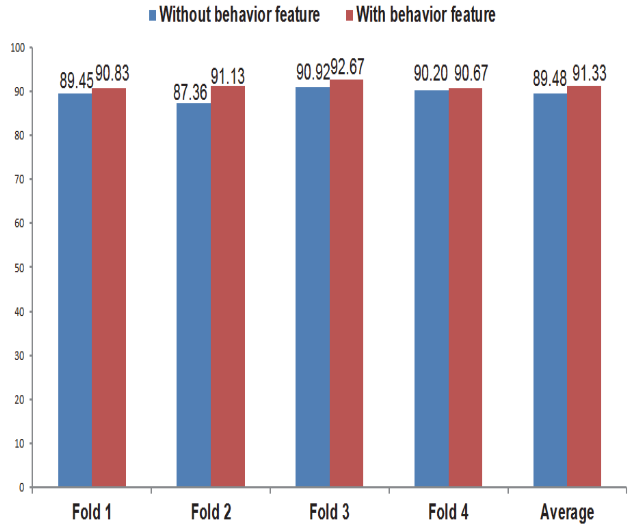

Fig.2 The F1-measure of the 4-folds cross-validation tests.

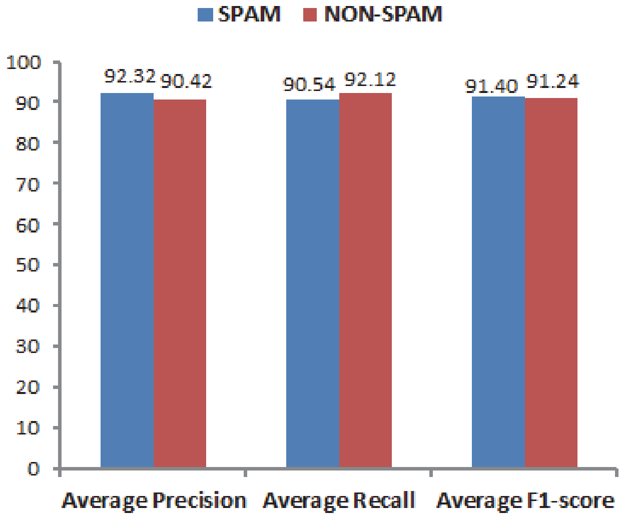

Fig.3 The average precision, recall and F_1-score of Spam and Non-Spam over the 4 folds (with behavior feature)

The high results indicate that spam signals are very useful in building the labeled set. They are highly distinguished features between spam and non-spam comments, such as the number of tagged users (≥1), links, the number of characters (≤5 or ≥100), …In fact, users who give comments too short or too long will be suspected as spammers.

We also calculated the average precision, recall, and F1 measure of the two classes: SPAM and NON-SPAM over the four folds. The results are shown in Figure 3. As we can see, the performance of NON-SPAM class is approximate to that of SPAM. This is in part because the number of comments carrying NON-SPAM is nearly equal (4,692 versus 4,864).

There are several hard comments for classification. Some comments which only contain a tagged person with a few abbreviated characters will be difficult for classifying. Spammers sometime tag users to advertise or pay attention while some real users want to tag their friends with short characters. A survey on collected comments shows that tagging attached by quite long text or without text will be labeled spam, whereas tagging with some additional information like address and phone number will be labeled non-spam. Finally, we find it difficult to label comments which carry sentiment or emotion such as: compliment, disparagement. To deal with these difficult cases, we need to integrate more high-level features to capture syntax, etc.

V. Conclusions

In this work, we have built a classification model based on the Maximum Entropy method to classify comments from Vietnam Facebook pages into SPAM or NON-SPAM. By combing content-based features and behavior-based ones, they help considerably get the best result. We have achieved an average F1-score of more than 90%, a promising result for further work on this work. The result also shows our right approach by using reasonable signals to detect spam comments. We also realized that we need to add better and higher level features as well as improve the quality of the sample set to the model in order to effectively distinguish ambiguous comments. This will be our focus in the future work.

Acknowledgment

This work was supported by the project QG.16.34 from Vietnam National University, Hanoi (VNU).

References

[1] F. Benevenuto, T. Rodrigues, V. A. F. Almeida, J. M. Almeida, and M. A. Gonçalves, "Detecting spammers and content promoters in online video social networks," in SIGIR 2009, pp. 620-627.

[2] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos, "CopyCatch: stopping group attacks by spotting lockstep behavior in social networks," WWW 2013, pp. 119-130.

[3] M. Fazeen, R. Dantu, and P. Guturu, "Identification of leaders, lurkers, associates and spammers in a social network: context-dependent and context-independent approaches," Social Netw. Analys. Mining, vol. 1, no. 3, pp. 241-254, 2011.

[4] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," ACM Conference on Computer and Communications Security, pp. 681-683, 2010.

[5] M. K. Girish Khurana, "Review: Efficient Spam Detection on Social Network," ISSN, vol. 3, no. 6, pp. 2321-9653, 2015.

[6] C. Grier, K. Thomas, V. Paxson, and C. M. Zhang, "@spam: the underground on 140 characters or less," in ACM Conference on Computer and Communications Security, 2010, pp. 27-37.

[7] K. Lee, J. Caverlee, and S. Webb, "The social honeypot project: protecting online communities from spammers," in WWW 2010, pp. 1139-1140.

[8] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots + machine learning," in SIGIR 2010, pp. 435-442.

[9] L. Liu, Y. Lu, Y. Luo, R. Zhang, L. Itti, and J. Lu, "Detecting "Smart" Spammers On Social Network: A Topic Model Approach," CoRR, vol. abs/1604.08504, 2016.

[10] M. McCord, and M. Chuah, "Spam Detection on Twitter Using Traditional Classifiers," in ATC 2011, pp. 175-186.

[11] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," Inf. Sci., vol. 260, pp. 64-73, 2014.

[12] S. Rayana, and L. Akoglu, "Collective Opinion Spam Detection: Bridging Review Networks and Metadata," in KDD 2015, pp. 985-994.

[13] B. Stone-Gross, T. Holz, G. Stringhini, and G. Vigna, "The Underground Economy of Spam: A Botmaster's Perspective of Coordinating Large-Scale Spam Campaigns," in LEET 2011.

[14] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in ACSAC 2010, pp. 1-9.

[15] Thi-Ngan Pham, Thi-Hong Vuong, Thi-Hoai Thai, Mai-Vu Tran, Quang-Thuy Ha, "Sentiment Analysis and User Similarity for Social Recommender System: An Experimental Study," ICISA vol. 376, pp. 1147-1156, 2016.

[16] A. H. Wang, "Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach," in DBSec, 2010, pp. 335-342.

[17] A. H. Wang, "Don't Follow Me - Spam Detection in Twitter," in SECRYPT 2010, pp. 142-151.

[18] C. Wilson, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao, "Beyond Social Graphs: User Interactions in Online Social Networks and their Implications," TWEB, vol. 6, no. 4, pp. 17, 2012.

[19] S. Yardi, D. M. Romero, G. Schoenebeck, and D. Boyd, "Detecting Spam in a Twitter Network," First Monday, vol. 15, no. 1, 2010.

[20] X. Zhang, and X. Zheng, "A novel method for spammer detection in social networks," in ICSDM 2015, pp. 115-118.