Adaptive Classification for Spam Detection on Twitter with Specific Data

Thayakorn Dangkesee
Department of Computer Engineering, Faculty of Engineering
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
Email: napaa_sky@hotmail.com

Sutheera Puntheeranurak
Department of Computer Engineering, Faculty of Engineering
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
sutheera.pu@kmitl.ac.th

***Abstract*—At present, the popularity of Twitter is more and more increasing. Many users can find information that tweets to Twitter. Some information is beneficial, and at the meantime, some information is created from spammers who would like to promote their websites or services. They are harmful to normal users by using the Twitter channel to exploit common user interests, such as sharing malicious links and ads. In order to stop spammers, many researchers have proposed tools to filter those junk. However, the focus of recent works is how to create streaming spam detection methods. In this paper, we proposed the adaptive data classification for spam detection by using spam word lists and a commercial URL-based security tool. We analyzed data by Naïve Bayes algorithm with both data types including all data and specific data. It can help to improve the performance of the spam detector is better than usual. We can show our proposed methods fulfillment in the experiments result.**

Keywords— spam detection; adaptive classification; streaming spam

I. INTRODUCTION

Nowadays, social networks applications widely used in all corners of the world, such as Facebook, Twitter, Instagram. As the statistics show, Instagram announces 700 million users in its blog post. Twitter has updated its active user numbers over a long time to 328 million [1] and is expected to increase by several million per month. Social networks are becoming a way for spammers to spread malicious or annoying messages to normal users.

Developers and researchers are challenging to create a system to detect streaming spam for the benefit of storage and general users because the population on the social network tends to increase hugely. Therefore, streaming spam detection is a useful tool in filtering valuable data to reduce processing resources and also reduces errors in other types of evidence analysis. Moreover, spammers will find the new ways to avoid detection of spam detector, especially the release of malicious links. Many recently tools use spam word lists to filter all spam, but they always get errors. Because spammers always find the ways to aim their goal. Therefore, spam detection is a task that should be adopted consistently. Alternatively, it can develop itself.

We were interested in developing an adaptive classification using both the spam word list and the commercial URLs to improve the filter rule and analyze data as a whole detection or accurate information discovery.

The remainder of the paper is organized in the following: Section II deals with related work. Section III provides a detailed our proposed methodology. Section IV describes experiments and results. Finally, we discussed conclusions and future work in section V.

II. RELATED WORK

According to a study, many researchers have proposed a number of mechanisms to solve spam detection problems. SJ. Soman [2] explored the behavior of spammers. They studied Honeypot and other research to classify these behaviors into five types: Text Based Spam, Comment Based Spam, Spams found in Bookmarking Systems, Spams in SMS and Email and Spams in Streaming Media. Therefore, we can divide spam detection processes into two parts: data preparation and data analysis. Data preparation is classified into data collection and data extraction. PC. Lin and PM. Huang [3] used URL classification in data collection part. They invented classification using URL rate and Interaction rate from some URLs and calculated interactions on tweet per total tweets. K. Kandasamy and P. Koroth [4] used Natural language and URL in their proposed method. They used the blacklist URLs for the initial classification and discovered the words from these URL and recategorized those words. Then they used Natural Language to find the spam word sentences from the above classification and used the stemming techniques to locate the spam word datasets. H. Xu, W. Sun and A. Javaid [5] used data from a spammer account. They extracted data from the Twitter API and discovered the spam pattern from those data to do the classification.

P. Kaur, A. Singhal, and J. Kaur [6] divided the data extraction into four main sections: User-based has taken advantage of basic user information such as account, age, followers count, and the following count. Content-based used data from those messages, which are filled with hashtags count, words count, and URLs count. Hybrid-based used additional techniques to find other types of information, such as URLs shorting services. Relation-based located the relationship

978-1-5386-0787-9/17/$31.00 ©2017 IEEE

between the spammer account was created, and the spammer account was detected. [6] and [7] used both user-based and content-based that use twitter to user-based from account_age and number of followers, following/friends, favorites, lists, and tweets. Moreover, they got content-based from some retweets, hashtags, mentions, URLs, characters, and digits. Y. Zhu and Y. Tan [9] used hybrid-based and relational-based techniques for Their local concentration model with Information Gain, Term Frequency Variance and Document Frequency. F. Yu, M. Moh, and TS. Moh [10] used some Filter rule including only English tweet, remove the similar tweet and remove all tweets published by users with over 10,000 followers. Then, they use Top 15 Drugs data to Filter again for specialized information to optimize their tools.

III. Our Proposed Methodology

A. Adaptive Classification

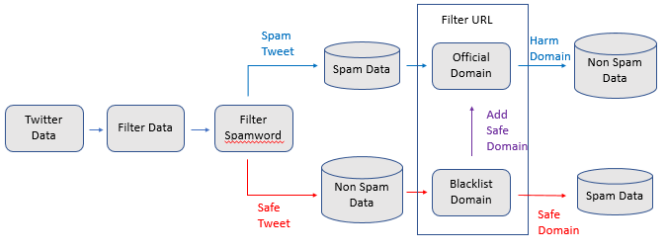

Fig. 1. Adaptive Classification Methodology

TABLE 1. Category Data Filter Rules

Category	Word Rules
Ads	ads, images, banners, Hedberg, RealMedia, img, announcer, popup, offer, adserver, sales, gifs, media, exit, out, adv, splash, pub, pop, graphics
Books	catalog, book, patterns, weaving, product, academic, news, ebook, educator, library, store, wileycda
E-commerce	shop, store, catalog, tickets, art, users, business
Games	Juegos, Jeux, category, game, Xbox, jeunesse, pc, online, Comunidad, consoles, flash, PSP, arcade, Wii, emuladores, gratis, Nintendo, PlayStation
Medical	health, conditions, article, content, diseases, meds, group
News	news, newspapers, media, publications, section, feed, opinion, business, community, archive, papers, profile
Porn	sex, free, porn, pics, pictures, nude, lesbian, pussy, hardcore, amateur, teen, avs, gals, black, gay, xxx, Asian, sexy, galleries, naked, hot, mature, girls, babes
Sport	sport, athletics, team, basketball, football, college, women, track, tennis, soccer, baseball, golf, mens

We proposed the adaptive classification methodology as shown in Fig.1. Twitter provides the Streaming API for developers and researchers to access public tweets in real time. We, therefore, start by collecting Twitter data from the Twitter API and then going through the Category Filter for specific data. We classify all data by using the Spam Word Filter. So we can separate the result as spam and non-spam datasets. For non-spam data, we pass it directly to the commercial URL Filter by using the Blacklist Domain to detect malicious links. The datasets that have elapsed will be moved to the spam set, and the remaining domain name, especially the ".com" domain will be updated in the official domain list for further filtering. Spam packets will be filtered through the Official Domain to find a possible share of potential news. After that, it will be changed to Non-Spam and will be used for data analysis later.

We can explain in detail in the steps of designing the rules for our proposed filtering as follows.

1) Category Data Filter: In our proposed, we choose the seven category which has ads, book, e-commerce, games, medical, news, porn, and sport. Each category will consist of words that derived from the link in each category of URLblacklist [3] because we would like to focus on the adaptive URL Filter so we try to find category rules from the relationship of those URLs. We truncate the domain name and the special symbol, which contains / and then take the remaining words to count as shown in Fig.2. After that, we select the word from the top 30 words to be the rules in the Category Data Filter as shown in Table 1.

Example urls of porn category from URLBlacklist

```
xxxfantasyland.com/teenleg
xxxfantasystories.com/free/slavedave
xxxfeature.com/sex-pictures-free-movies
```

When cut domain and special symbol out

```
teenleg
free slavedave
sex pictures free movies
```

Then count it

```
1 teenleg
2 free    1 slavedave
1 sex     1 pictures    1 movies
```

Fig. 2. Category Data Count Method

2) Spam Word Filter: we choose a series from [11], for example, Billion dollars, Dear friend, Free sample, Brand new pager, Dear somebody, Free trial and more to be the rules in the Spam Word Filter.

3) URL Filter: We take the link to check that it is a shortened URL [12]. Then we find the destination link and apply only domain name to the rules of the Blacklist domain that will be taken from URLBlacklist.com. [16] After that, we will update the Official domain to the URL Filter.

B. Feature Extraction

Feature extraction involves reducing the number of resources needed to explain the large datasets with creating a combination of variables to solve many analytical problems. It requires much memory and computational power while it still is explaining the data with sufficient accuracy.

We choose user-based and content-based twelve features [7] [8] as shown in lists:

– The consisting of account_age is days of an account since its creation until the time of sending a recent tweet.

- The no_of followers are some followers of this user.
- The no_of followings are the number following/friends of this user.
- The number user favorites are the number of favorites that received.
- The no_lists is the number of lists that added.
- The no_tweets is the number of a tweet that sent.
- The no_retweets is the number of retweets.
- The no_hashtags is the number of hashtags.
- The no_usermentions is the number of mentions.
- The no_urls is the number of URLs.
- The no_chars is the number of characters.
- The no_digits is the number of digits.

IV. Experiments and Results

A. Experiment Setup

We collected data from the Twitter API via tweepy [15] by using Python 2.7, which collected a total of 50,000 data during the May-August 2017 period. We divided into three series, 5,000, 10,000, and 50,000. We applied all data with the category data filter. Therefore the dataset has remained as shown in Table 2 will be included in each category as the complete rules.

TABLE 2. Amount of Data that through the category filter

Category	Amount of data		
	5,000	***10,000***	***50,000***
All	3,634	7,336	36,230
Ads	1,208	2,501	11,489
Books	288	609	3,076
Ecommerce	532	1,064	4,829
Games	445	956	3,634
Medical	200	435	1,935
News	458	930	4,273
Porn	1,112	2,246	13,646
Sports	481	972	4,666

The rest of the data is divided into 70% for training dataset and 30% for the testing dataset. Then, we use the twelve features as mentioned above to apply with the Naïve Bayes algorithm [13] for data analysis. Moreover, we evaluated by calculating precision, recall, and F1-Score [14]. We divide the experiments into two cases to compare between classification by using only spam word list and our proposed adaptive classification.

B. Experimental Results

We defined that Process1 is classification by using only spam word list, and Procress2 is our proposed adaptive classification. We evaluated from the trial data which show an overall output in Fig. 3, Fig.4, and Fig. 5.

We can show that the precision of the adaptive classification is higher than the traditional one for any volume of data. In the meantime, Recall and F1-score of the adaptive classification are less than the traditional one. In process2, the highest precision is 0.984502 in the 10,000 datasets. On the other hand, the accuracy of process1 has reduced when the number of data is increased.

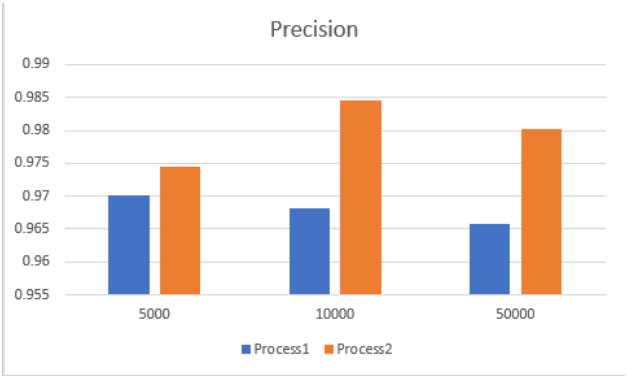

Fig. 3. Precision of All data

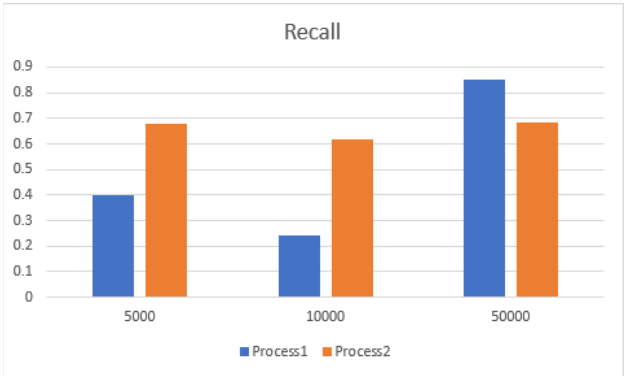

Fig. 4. Recall of All data

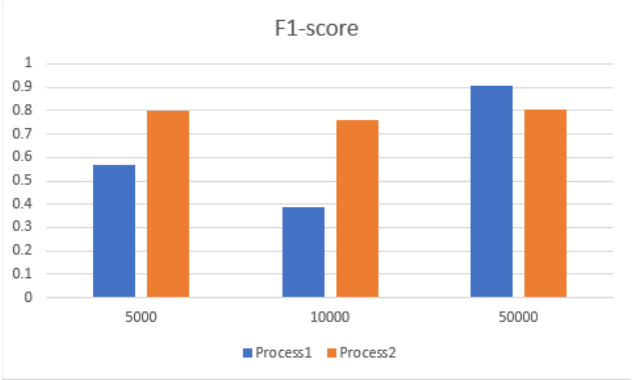

Fig. 5. F1-Score of All data

Based on the results of the specific data experiments have shown in Fig. 6, Fig. 7, and Fig. 8, the precision of our proposed method is still higher than the traditional one. In

comparison for 50,000 datasets, the E-commerce, News, and Porn, their precision values are 0.975887, 0.981848 and 0.984331, respectively. The only E-commerce is less than All Data. From Table 1, the data for the three analyses were 4,829, 4,273, and 13,646, respectively. It can show that our proposed is good for the amount of data to obtain good performance.

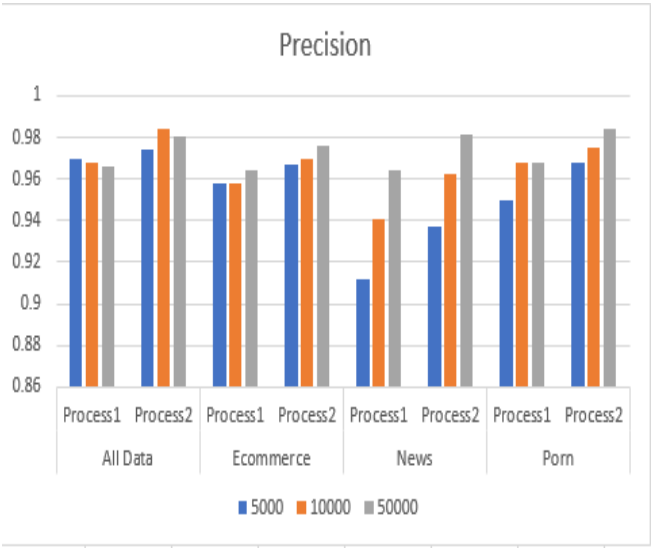

Fig. 6. Precision of All Data and Some Specific Data

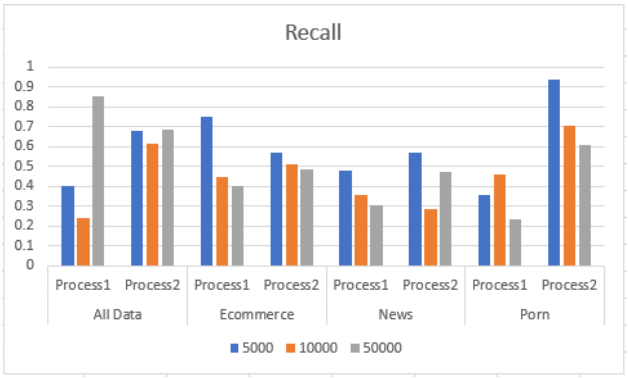

Fig. 7. Recall of All Data and Some Specific Data

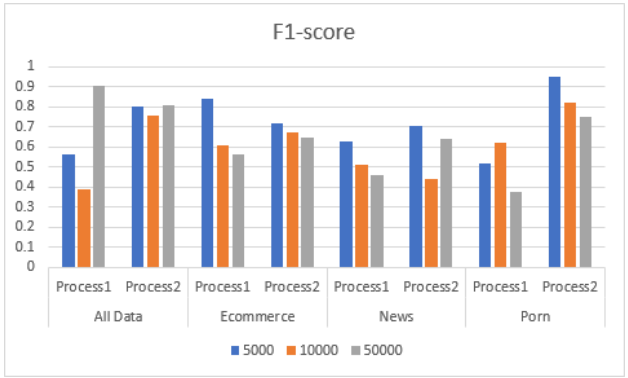

Fig. 8. F1-Score of all Data and Some Specific Data

V. Conclusion and Future Work

This paper proposed the adaptive classification by using spam word list and Blacklist URL to apply to some specific data. It can show that our proposed method has more efficient than traditional classification for all dataset. In the future work. We are interested in how to apply for using Safe Browsing instead of URL blacklist to detect potentially dangerous links, as well as changing some features. In the data analysis section we succeeded in developing an adaptive classification for Naïve Bayes. Then, we will refine algorithms and compare with others for data analysis that may improve stability and performance better.

References

[1] (August 20, 2017) Top 15 Most Popular Social Networking Sites and Apps [Online]. Available: https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/

[2] SJ. Soman, "A survey on behaviors exhibited by spammers in popular social media networks, "In Proceedings of International Conference on Circuit, Power and Computing Technologies (ICCPCT), 2016, pp. 1-6.

[3] PC. Lin, PM. Huang," A study of effective features for detecting long-surviving Twitter spam accounts," In Proceedings of 15th International Conference on Advanced Communications Technology (ICACT), 2013, pp. 841-846.

[4] K. Kandasamy, P. Koroth, " An integrated approach to spam classification on Twitter using URL analysis, natural language processing, and machine learning techniques,"In Proceedings of IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 2014, pp. 1-5.

[5] H. Xu, W. Sun, A. Javaid, "Efficient spam detection across Online Social Networks," In Proceedings of IEEE International Conference on Big Data Analysis (ICBDA), 2016, pp. 1-6.

[6] P. Kaur, A. Singhal, J. Kaur, "Spam detection on Twitter: A survey," In Proceedings of 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 2570-2573.

[7] C. Chen, J. Zhang, Y. Xie, Y. Xiang, W. Zhou, MM. Hassan, A, AlElaiwi, M. Alrubaian, "A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection," In Proceedings of IEEE Transactions on Computational Social Systems, Volume 2, Issue 3, 2015, pp. 65-76.

[8] C. Chen, J. Zhang, X. Chen, Y. Xiang, W. Zhou, "6 million spam tweets: A large ground truth for timely Twitter spam detection," In Proceedings of IEEE International Conference on Communications (ICC),2015, pp.7065-7070.

[9] Y. Zhu, Y. Tan, "A Local-Concentration-Based Feature Extraction Approach for Spam Filtering," In Proceedings of IEEE Transactions on Information Forensics and Security Volume 6, Issue 2, 2011, pp. 486-497.

[10] F. Yu, M. Moh, TS. Moh, "Towards Extracting Drug-Effect Relation from Twitter: A Supervised Learning Approach," In Proceedings of IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), 2016, pp. 339-344.

[11] (August 20, 2017). A List of Common Spam Words [Online]. Available: https://emailmarketing.comm100.com/email-marketing-ebook/spam-words.aspx

[12] (August 20, 2017) List of URL Shorteners [Online]. Available https://bit.do/list-of-url-shorteners.php

[13] (August 20, 2017) Naïve Bayes [Online]. Available: http://scikit-learn.org/stable/modules/naive_bayes.html

[14] (August 20, 2017) Precision-Recall [Online]. Available: http://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

[15] (August 20, 2017) Tweepy [Online]. Available: http://tweepy.readthedocs.io/en/v3.5.0/

[16] (May 14, 2017) URLBlacklist.com [Online]. Available: http://urlblacklist.com/cgi-bin/commercialdownload.pl?type=download&file=bigblacklist