Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
13/2 (2020), 97-107. DOI: http://dx:doi:org/10:21609/jiki:v13i2.871

SPAMMER DETECTION BASED ON ACCOUNT, TWEET, AND COMMUNITY ACTIVITY ON TWITTER

Arif Mudi Priyatno, Agus Zainal Arifin, Rizka Wakhidatus Sholikah

Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia

E-mail: arifmudi11@gmail.com

Abstract

Spammers are the activities of users who abuse Twitter to spread spam. Spammers imitate legitimate user behavior patterns to avoid being detected by spam detectors. Spammers create lots of fake accounts and collaborate with each other to form communities. The collaboration makes it difficult to detect spammers' accounts. This research proposed the development of feature extraction based on hashtags and community activities for the detection of spammer accounts on Twitter. Hashtags are used by spammers to increase popularity. Community activities are used as features for the detection of spammers so as to give weight to the activities of spammers contained in a community. The experimental result shows that the proposed method got the best performance in accuracy, recall, precision and g-means with are 90.55%, 88.04%, 3.18%, and 16.74%, respectively. The accuracy and g-mean of the proposed method can surpassed previous method with 4.23% and 14.43%. This shows that the proposed method can overcome the problem of detecting spammer on Twitter with better performance compared to state of the art.

Keywords: *Spammer detection, account feature, tweet feature, community feature, twitter, hashtag*

Abstrak

Spammer adalah aktivitas pengguna yang menyalahgunakan Twitter untuk menyebarkan spam. Spammer meniru pola perilaku pengguna yang sah untuk menghindari terdeteksi oleh pendeteksi spam. Spammer membuat banyak akun palsu dan berkolaborasi satu sama lain untuk membentuk komunitas. Kolaborasi ini membuat sulit dalam mendeteksi akun spammer. Penelitian ini mengusulkan pengembangan ekstraksi fitur berdasarkan hashtag dan aktivitas komunitas untuk mendeteksi akun spammer di Twitter. Hashtag digunakan oleh spammer untuk meningkatkan popularitas. Aktivitas komunitas digunakan sebagai fitur untuk mendeteksi spammer, sehingga memberi bobot pada kegiatan spammer yang terkandung dalam suatu komunitas. Hasil percobaan menunjukkan bahwa metode yang diusulkan mendapatkan kinerja terbaik dalam akurasi, recall, presisi, dan g-means dengan masing-masing 90,55%, 88,04%, 3.18%, dan 16.74%. Akurasi dan g-mean dari metode yang diusulkan dapat melampaui metode sebelumnya dengan 4,23% dan 14,43%. Ini menunjukkan bahwa metode yang diusulkan dapat mengatasi masalah mendeteksi spammer di Twitter dengan kinerja yang lebih baik.

Kata Kunci: *Spammer detection, account feature, tweet feature, community feature, twitter, hashtag*

1. Introduction

Twitter is one of online social media which develops rapidly. Established in 2006, Twitter has appeared as the most popular microblogging platform in which the users can share news, media, meme, point of view, and update in the form of tweet. Tweet is the writing containing text and limited URL HTTP until 280 characters [1]. Unfortunately, the growth of Twitter social interaction has attracted the cyberspace criminals who exploit the trust relationship among the users to distribute evil content to big number of victims in the network. The most well known spamming type in Twitter is catching hot recent topics [2]. Whenever the event occurs, the users try to express the opinion or information about the event, by using hashtag or same keywords. If the topic is the most tweeted topic in that day, then it will be seen by all Twitter users in their home as the hot recent topic. Spammer uses the same hashtag to be seen by users basis in big scale after certain trend event but with URL that is not asked and led to unrelated web site. Because of 280 characters limitation on Twitter, spammer usually shares URL using URL shortener service.

Spammer usually imitate the behavior pattern of official user to avoid detected by spam detection technique. Spammer develops the device and technique to avoid the existed detection technique. Besides that, the research trend nowadays about spam detection has complexity obstacle or owns some warnings that can be passed by spammer. In this case, it is extremely necessary to detect and block spammer from social network such as Twitter to save resources and human efforts from unwanted users. Included the stronger feature and more difficult to be imitated. And the usage of user interaction in and out of the community structure which can be used to build spam classification model which will make the spammer difficult. Spammer makes many fake account, and collaborate one and another forming tight community to increase their credibility. Therefore, spammer account tends to connect socially to the highest classification coefficient [3].

Various methods have been conducted to detect spammer in Twitter. There are 5 characteristics of bot spammer according to [4] such as spam containing active link, spam containing certain product, owning the same similarity between the tweet before and after, new account and spam frequently uses hashtag. The research from aditya et. al. [5] conducted bot spammer detection by looking at the characteristics of posting time and the sentiment of the tweet done. Another research from Inuwa-Dutse et. al. [6] conducted spammer detection by optimizing a series of feature from tweet history and information of users' account. From the analysis result conducted, it can be seen that spammer tends to be selective in following other users, until forming spammer connection. Beside that, mostly spam account automatically posted at least 12 tweets per day at the period which is well determined. Bhat and Abulaish [7] conducted spammer identification in Facebook by using community feature. The community feature used in this research are total out-degree, total reciprocity, total in/out ratio, community memberships, foreign out-degree and foreign in/out ratio. From that research obtained conclusion that by combining the community and non-community feature can increase significant result of spammer detection. Sarlati et. al [8] adopted community feature to detect spammer and uses the feature selection of Principal Component Analysis for decreasing the feature volume used. Chen et. al. [9] found that the coordination of spammers makes detection difficult. Bindu et. al. [10] found that there is spammer community which works collectively for spreading the spam and avoid spammer detection technique in Twitter. Spammers collaborate and coordinate with the hashtag information on the tweet. Therefore, detecting spammers using hashtag and community activities features will increase success.

This research proposed the development of feature extraction based on hashtag and community activity for detecting spammer account on Twitter. Hashtag is used by spammer members for improving popularity. The community activity is used as the feature for spammer detection, until it can give weight towards spammer activity obtained in certain community. The community activity done such as tweet with hashtag usage, URL, and others.

2. Related Work

Perdana et. al. [11] conduct spammer detection by using consideration of tweet similarity done and interval time of doing the tweet. The level of tweet similarity is considered because spammer does sufficiently high tweet similarity, until disturbing the information spread in Twitter. However, spammer is getting smarter in doing his action until they make the tweet which is different from one and another. Spammer will string certain words in their action until making tweets that look good. Time interval entropy is considered because spammer tends to conduct their action in the time which approaches togetherness, or its interval is almost the same. But there is also spammer doing their action without managing the interval time, until it seems like a natural tweet.

Priyatno et. al. [12] conducted spammer detection by using Time Interval Entropy feature and global vector for word representation (Glove). The classification process uses convolution neural network. The tweet feature without omitting hashtag used as the input because spammer makes the tweet with hashtag for achieving certain purpose. Time Interval Entropy Feature is used because spammer does tweet with managed time until the range is not too far. However, there is also spammer who does rarely spam until it complicates the detection.

Aditya et. al. [5] conduct spammer detection by using sentiment analysis feature and time interval entropy (TIE). Sentiment analysis is used to detect the expression or opinion contained in the tweet. Sentiment analysis used combination of knowledge method-based and machine learning-based to obtain neutral tweet or the one which does not have social sentiment in which frequently appear at spam tweets. TIE was used to catch the regularity of posting the tweet which shows the tweet is posted automatically.

Inuwa-Dutse et. al. [6] conducted spammer detection by utilizing User Profile Feature (UPF), Account Information Feature (AIF), and Tweet

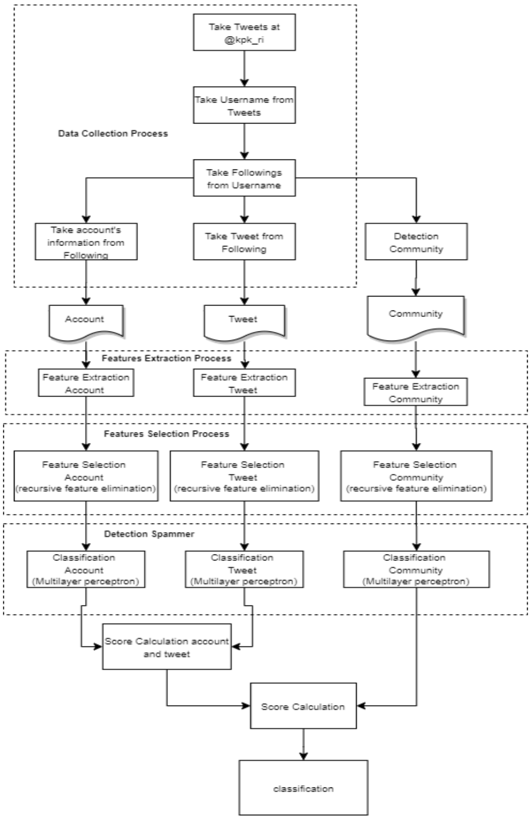

Data Collection Process
Take Tweets at @kpk_ri
Take Username from Tweets
Take Followings from Username
Take account's information from Following
Take Tweet from Following
Detection Community
Account
Tweet
Community
Features Extraction Process
Feature Extraction Account
Feature Extraction Tweet
Feature Extraction Community
Features Selection Process
Feature Selection Account (recursive feature elimination)
Feature Selection Tweet (recursive feature elimination)
Feature Selection Community (recursive feature elimination)
Detection Spammer
Classification Account (Multilayer perceptron)
Classification Tweet (Multilayer perceptron)
Classification Community (Multilayer perceptron)
Score Calculation account and tweet
Score Calculation
classification

Figure 1. Proposed Method

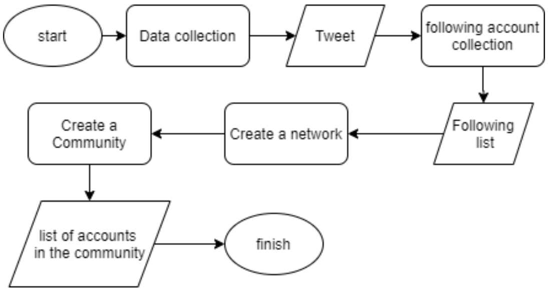


Figure 2. Community Detection Process

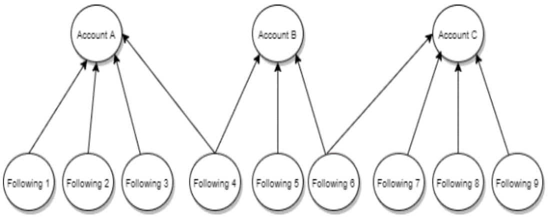


Figure 3. Following on The Account

Feature. User Profile Feature (UPF) included information about users such as username, screen name, location, and user description. Account Information Feature (AIF) consists of information such as time of creating the account (account age), and account verification sign (verified or not verified).

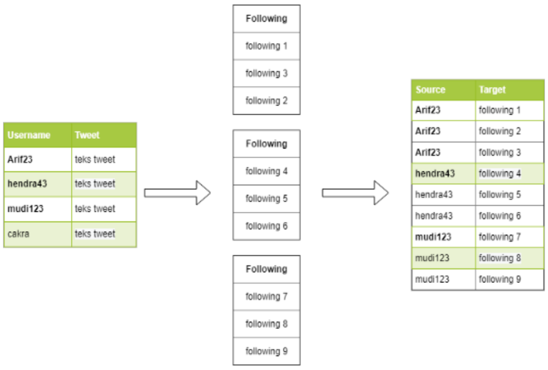
Chen et. al. [9] explains about three spammer intelligences in doing spam such as coordinated spam, machine base spam template or passive spam. The behavior of coordinated spam complicates spammer detection process.

3. Proposed Method

The proposed method consists of several steps, namely: community detection on Twitter, feature extraction, feature selection using recursive feature elimination (RFE), and classification using multi-layer perceptron. These stages can be seen in Figure 1.

Community Detection on Twitter

The process of community detection as showed in Figure 2 is started from data collection on Twitter on August 1st until September 10th, 2019. This process obtained tweets at home of Corruption Eradication Commission (KPK) @kpk_ri. This tweet is not only from KPK account only, but also tweet from the account that does mention to @kpk_ri. After the process of tweet collection with certain time interval, the next stage is collecting username that interrelated with the tweet. Username is obtained then do the process of taking the following account from each username as seen in Figure 3 and the example of taking following process is showed in Figure 4. The result of taking the following from each username saved in csv format. The document of csv has 2 headers such as source and target. After the process of obtaining the following, then the process done was uniting the data at one csv list containing the source of username in which its following is taken, and the target contained obtained following. This one csv list is called as edge list. The process of community detection aims to know existed community in the account. After the process of obtaining the following list at all accounts, then the process of forming the community by using louvain method from this research was used [13].



Username	Tweet
Arif23	teks tweet
hendra43	teks tweet
mudi123	teks tweet
cakra	teks tweet

Following
following 1
following 3
following 2

Following
following 4
following 5
following 6

Following
following 7
following 8
following 9

Source	Target
Arif23	following 1
Arif23	following 2
Arif23	following 3
hendra43	following 4
hendra43	following 5
hendra43	following 6
mudi123	following 7
mudi123	following 8
mudi123	following 9

Figure 4. Example of Taking the Following

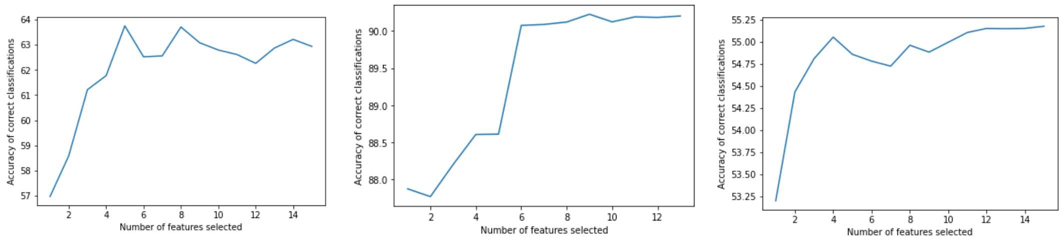
TABLE 1
FEATURES EXTRACTION

Features name	Features break down	Equation
Account	Account age	[6], [14], [10]
	Follower	[6], [14]
	Following	[6]
	Number of statuses	[6], [14]
	Name digit	[6]
	Username length	[6]
	Screen name length	[6]
	Similarity username and screenname	[6]
	Following ratio	[6]
	Follower ratio	[6]
	Interestingness	[6]
	Account activity	[6]
	Name ratio	[6]
	Indegree	[6]
Tweet	Average length of tweets	[6]
	Similarity tweet	[6]
	URL ratio	[10], [16]
	Mention ratio	[16]
	Lexrichwithuu	[6]
	Lexrichoutuu	[6]
	Unique URL ratio	[10], [16]
	Total account hashtag	$HA = n(H)$
	Account hashtag ratio	$RHA = \frac{n(kH)}{JKT}$
	Unique ration of account hashtag	$RUHA = \frac{Jumlah\ unique\ hastag}{HA}$
	Words ratio of spam account	$RKSA = \frac{n(kS)}{JKT}$
	Total words of spam account	$JKSA = n(kata\ spam)$
	Ratio of unique words of spam account	$RUKSA = \frac{n(unique\ kata\ spam)}{JKSA}$
Community	Total of indegree	[7], [8]
	Total community members	[7], [8]
	Total Hashtag Unique Community	$JHUK = \{\sum_{j \in K_i} unique\ hashtag_j \mid K_i \in K\}$
	Total community hashtag	$JHK = \{\sum_{i \in K_i} HA_j \mid K_i \in K\}$
	Community hashtag ratio	$RHK = \frac{\sum_{i=1}^{n} RHA_i}{JAK}$
	Unique Ratio of Community hashtag	$RUHK = \frac{\sum_{i=1}^{n} RUHA_i}{JAK}$
	Total URL of unique community	$JUUK = \frac{\sum_{i=1}^{n} UUA_i}{JAK}$
	Total community URL	$JUK = \{\sum_{j \in K_i} URL(A)_j \mid K_i \in K\}$
	Community URL ratio	$RUK = \{\sum_{j \in K_i} rasio\ url\ akun_j \mid K_i \in K\}$
	Community URL unique ratio	$RUUK = \frac{JUUK}{JUK}$
	Total community eigenvector	$JEK = \{\sum_{j \in K_i} eigenvector_j \mid K_i \in K\}$
	Community eigen ratio	$REK = \frac{JEK}{JAK}$
	Community spam word ratio	$RKK = \{\sum_{j \in K_i} RKSA(A)_j \mid K_i \in K\}$
	Total of words of community spam	$JKK = \{\sum_{j \in K_i} JKSA(A)_j \mid K_i \in K\}$
	The unique ratio of community spam	$RUSK = \{\sum_{j \in K_i} RUKSA\ (A)_j \mid K_i \in K\}$

Feature Extraction

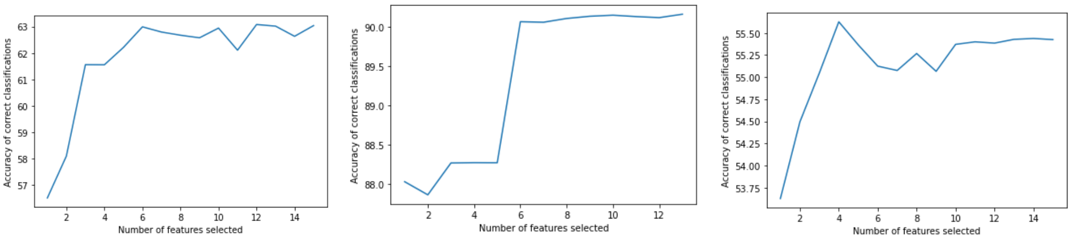
Feature extraction process is conducted by obtaining three big groups consisting of account feature, tweet feature, and community feature which can be seen in Table 1. Account feature is the feature which gives the description about the account information and activity information of users [6], [14], [10]. Tweet feature is the feature which gives information about tweet activities done [15], [16], [6], [14], [10]. Community feature [14], [7], [8] is the feature which gives information related to joint activities done by Twitter users such as total hashtag unique community (*JHUK*) which is total hashtag unique in one community (*K*), total account hashtag (*HA*) is total hashtag (*H*) at one account. Total community hashtag (*JHK*) is total hashtag used by all community members. Community hashtag ratio (*RHK*) is the quotient between ratio of account hashtag of one community and total community members (*JAK*). Account hashtag ratio (*RHA*) is the quotient of hashtag character length (*kH*) with total tweet character (*JKT*). Unique ration of account hashtag (*RUHA*) is the quotient of total unique hashtag and total account hashtag. Unique Ratio of Community hashtag (*RUHK*) is the quotient between total unique ratio of account hashtag and total community members. Total URL of unique community (*JUUK*) is the quotient between total URL of unique account and total community members. Total community URL (*JUK*) is the number of URL in the community. Ratio of community URL (RUK) is total ratio in the community. Unique ratio of community URL (RUUK) is the quotient between total unique of community URL and total community URL. Total community eigenvector (*JEK*) is the quotient of total community eigenvector and total community members. Words ratio of spam account (*RKSA*) is the quotient of total spam character (*kS*) and total tweet character. Total words of spam account (*JKSA*) is total spam words obtained in an account. Ratio of unique words of spam account (*RUKSA*) is the quotient between spam unique words and total words of spam account.

The next step is data cleaning on all features. Cleaning process is conducted to omit empty data feature and less complete one. After cleaning data process, normalization is conducted towards the data. Normalization process is conducted to equalize the feature range owned becomes range



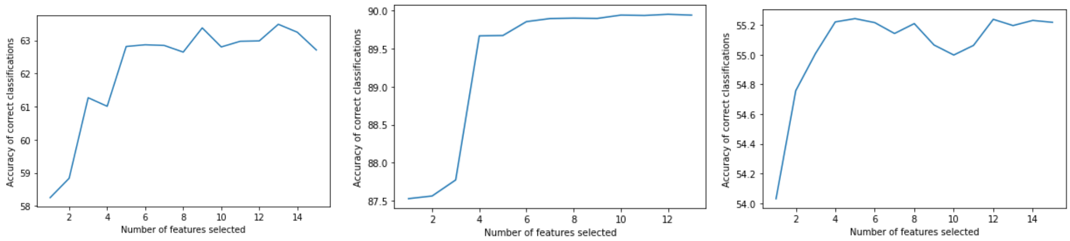
(A) (B) (C)

Figure 5. The results of feature selection the dataset 70:30 use recursive feature elimination: (a) feature accounts, (b) feature tweets and (c) feature communities.



(A) (B) (C)

Figure 6. The results of feature selection the dataset 80:20 use recursive feature elimination: (a) feature accounts, (b) feature tweets and (c) feature communities.



(A) (B) (C)

Figure 7. The results of feature selection the dataset 90:10 use recursive feature elimination: (a) feature accounts, (b) feature tweets and (c) feature communities.

TABLE 2
OPTIMAL FEATURES

Features name	Features break down
Account	Eigenvector Account age Length of username Length of screen name Username similarity and screen name Following ratio Interesting Account activeness Name ratio Indegree
Tweet	Average of tweet length URL ratio Mention ratio Lexrichoutuu Unique ratio of URL Total account hashtag Ratio of account hashtag Unique ratio of account hashtag Total words of spam account
Community	Total indegree Total community members Ratio of community URL Unique ratio of community URL Unique ratio of community hashtag Total community eigen

0.1 until 0.9. After normalization, recursive feature elimination process is conducted towards the data to obtain optimal feature.

Feature Selection using Recursive Feature Elimination (RFE)

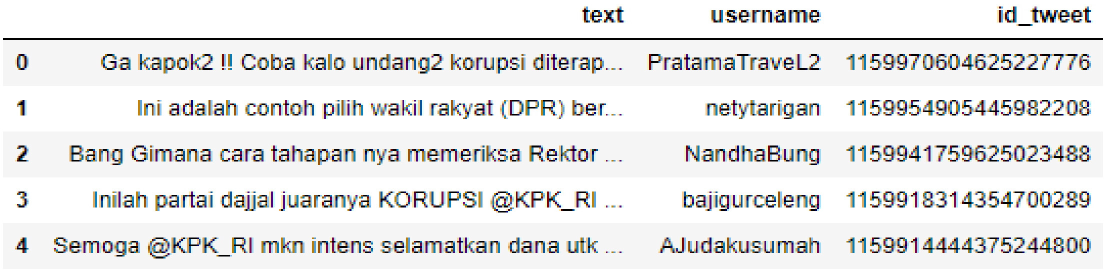
Feature selection process is conducted to obtain optimal features. Feature selection process uses support vector machine-recursive feature elimination which is adopted from the research [17], [18]. Support vector machine-recursive feature elimination (SVM-RFE) does feature selection in a backward way.

The process of SVM-RFE is started by conducting the training process of support vector machine, until the training result gains training weight. Then, weight calculation is conducted on the training result such as towards the length of dataset dimension. Then, we find the smallest criteria, then the result is used for feature improvement process. If feature improvement process has been done, then the process is continued with conducting update of rank orders of the existed features. Then, the process is continued by deleting the feature which has smallest criteria until the best features obtained. The feature is stated optimal if the value change is insignificant.

The results of feature selection as seen in Figure 5 by using the percentage of training data distribution and test data with the ration 70:30. The feature consist of account age, length of screen name, username and screen name similarity, following ratio and account activeness features from optimal account features. Average tweet length, URL ratio, mention ratio, lexirichoutuu, URL unique ratio, total account hashtag, account hashtag ratio, hashtag unique ratio and total words of spam account features from optimal tweet features. Total indegree, unique ratio of community hashtag, ratio of URL community, and unique ratio of URL community features from optimal community features.

The result of feature selection in Figure 6 uses training data percentage and test data with the ration 80:20. The features consist of eigen vector, account age, length of screen name, username similarity and screen name, following ratio, and account activeness features from optimal account features. Average length, unique URL ratio, total account hashtag, account hashtag ratio, unique ratio of account hashtag, and total words of spam account features from optimal tweet features. Total community member, ratio of community URL, unique ratio of community URL, and total eigen community features from optimal community features.

The result of feature selection as seen in Figure 7 with percentage of training data and test data with the ration 90:10. The features consist of eigenvector, account age, length of screen name, username similarity and screen name, follower ratio, interestingness, account activeness, name ratio and indegree features from optimal account features. Average length, URL unique ratio, total account hashtag, ratio of account hashtag, unique ratio of account hashtag and total words of spam account features from optimal tweet features. Total indegree, total community members, ratio of community URL, and unique ratio of community URL features from optimal community features.



	text	username	id_tweet
0	Ga kapok2 !! Coba kalo undang2 korupsi diterap...	PratamaTraveL2	1159970604625227776
1	Ini adalah contoh pilih wakil rakyat (DPR) ber...	netytarigan	1159954905445982208
2	Bang Gimana cara tahapan nya memeriksa Rektor ...	NandhaBung	1159941759625023488
3	Inilah partai dajjal juaranya KORUPSI @KPK_RI ...	bajigurceleng	1159918314354700289
4	Semoga @KPK_RI mkn intens selamatkan dana utk ...	AJudakusumah	1159914444375244800

Figure 8. Example of Tweet From @KPK_RI

The result of feature selection using recursive feature elimination thoroughly is eigenvector, account age, length of username, length of screen name, username similarity and screen name, following ratio, interesting, account activeness, name ratio and indegree features from optimal account features. The average of tweet length, URL ratio, mention ratio, lexrichoutuu, unique ratio of URL, total account hashtag, ratio of account hashtag, unique ratio of account hashtag and total words of spam account features from optimal tweet features. Total indegree, total community members, ratio of community URL, unique ratio of community URL, unique ratio of community hashtag and total community eigen features from optimal community feature. The list of optimal features can be seen in Table 2.

Detection Spammer

Spammer detection process is conducted by using multi-layer perceptron (MLP). We adopted research from Hans et.al. [19] that use MLP as classifier. The process of multi-layer perceptron has three big stages such as forward process, backward process, and process of weight change. Multi-layer perceptron uses some inputs in line with total features the result of feature selection process. Total hidden layers are 2 hidden layers with node hidden (15,15). Learning rate 0.1, 0.01 and 0.001. Maximum epoch used is 1000. The lowest error level is 0.0001. The process multi-layer perceptron uses input from features obtained from feature selection process. Then forward process was done towards input to hidden layer until output layer. The result of forward process is conducted activation function by using activation function of sigmoid biner. Then the next process is backpropagation. Backpropagation is conducted to count the error value obtained from the difference of output layer and ground truth. Backpropagation process is conducted for all layers, backward is started by finding the error in the layer. After backward obtains error value on all layers, MLP process is conducted, the process of weight change which is counted based on mistake value in each layer. This process is conducted continuously until stop value point is determined, either error minimal value or maximum iteration. If the training process has been done, then multi-layer perceptron obtains the model from the training result. The model is used for testing. Testing data are the data resulted from the distribution of main data divided to be two parts such as training data and testing data. The testing process of multi-layer perceptron is conducted at forward propagation phase.

TABLE 3
CONFUSION MATRIX

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

The merging process is done by adding the multiplication result from multi-layer perceptron output with each weight. Those multiplication are such as account feature weight (α) * the result of multilayer perceptron of account feature (A), tweet feature weight (β) * the result of multilayer perceptron of tweet feature (B), and weight of community feature (γ) * the result of multilayer perceptron of community feature (C). Total weight of and γ is one. Total weight of α weight and β weight is δ. The result of merging process then conducted classification by using threshold to obtain the classification result. The result of merging process is considered as spammer if score smaller from threshold and not spam if score bigger than threshold.

$$\boldsymbol{Account\ score} = \boldsymbol{\delta} * (\boldsymbol{AB}) + (\mathbf{1} - \boldsymbol{\gamma}) * \boldsymbol{C} \quad (1)$$

$$\boldsymbol{Tweet\ account\ score}\ \ (\boldsymbol{AB}) = \boldsymbol{\alpha} * \boldsymbol{A} + (\boldsymbol{\delta} - \boldsymbol{\beta}) * \boldsymbol{B} \quad (2)$$

Examples of merging process are β = 0.45 and γ = 0.1. δ obtained from 1 - 0.1 = 0.9. α is 0.9 – 0.45 = 0.45. This is according to the rules of δ and γ is 1. Total weight of α weight and β weight is δ. The value of the MLP results of the account is 0.59, the value of the results of the MLP tweet is 0.7, and the value of the MLP community is 0.61. The results of the tweet account score are 0.5816. The

TABLE 4
RESULTS OF THE SPAMMER DETECTION

Percentage of Data	Learning Rate	Feature	Accuracy	Recall	Precision	G-mean
70:30	0,1	Account	62,61%	89,48%	0,95%	9,24%
		Community	43,40%	67,87%	0,48%	5,71%
		Tweet	87,18%	88,04%	2,69%	15,40%
		Proposed	87,32%	**91,21%**	2,82%	16,03%
	0,01	Account	67,43%	82,85%	1,01%	9,16%
		Community	67,99%	45,68%	0.57%	5,12%
		Tweet	88,01%	86,31%	2,82%	15,61%
		Proposed	**90,55%**	88.04%	**3,18%**	**16,74%**
	0,001	Account	70,11%	79,11%	1,06%	9,14%
		Community	65%	48,99%	0,56%	5,25%
		Tweet	87,28%	88,04%	2,72%	15,46%
		Proposed	89,18%	89,19%	3,13%	16,94%
		Comparison [11]	86,32%	13,40%	0,40%	2,32%
80:20	0,1	Account	72,72%	78,79%	1,16%	9,55%
		Community	60,94%	47,62%	0,49%	4,83%
		Tweet	85,29%	85,93%	2,30%	14,05%
		Proposed	**89,35%**	85,50%	**3,14%**	**16,37%**
	0,01	Account	69,73%	82,90%	1,09%	9,50%
		Community	58,96%	52,16%	1,01%	5,16%
		Tweet	85,30%	87,23%	2,33%	14,27%
		Proposed	88,31%	87,23%	3,05%	16,30%
	0,001	Account	65,68%	86,80%	1,01%	9,34%
		Community	64,38%	45,67%	0,51%	4,85%
		Tweet	85,20%	87,23%	2,32%	14,22%
		Proposed	87,77%	**88,96%**	2,85%	15,91%
		Comparison [11]	83,09%	16,02%	0,39%	2,49%
90:10	0,1	Account	68,82%	87,01%	1,11%	9,82%
		Community	64,98%	42,86%	0,49%	4,59%
		Tweet	85,82%	81,39%	2,26%	13,57%
		Proposed	**89,24%**	84,85%	**3,08%**	**16,17%**
	0,01	Account	66,25%	**88,74%**	1,05%	9,36%
		Community	65,22%	41,99%	0,49%	4,51%
		Tweet	86,06%	80,25%	2,28%	13,56%
		Proposed	88,36%	86,15%	2,90%	15,82%
	0,001	Account	69,89%	86,15%	1,14%	9,90%
		Community	58,95%	46,75%	0,46%	4,62%
		Tweet	85,60%	83,12%	2,27%	13,74%
		Proposed	88,70%	85,28%	2,87%	15,66%
		Comparison [11]	80,68%	19,05%	0,40%	2,76%

account score results are 0.5840. account score results are carried out with a value of 0.5 then the results obtained legitimate accounts.

4. Experiment and Analysis

This research used Twitter data which were collected from the account of Corruption Eradication Commission (KPK) @kpk_ri with tweet interest target is about "corruption". Started data collection on Twitter on August 1st until September 10th, 2019. Data collection from Twitter did not use official API from Twitter but used python library GetOldTweet3 because if the process of taking tweet used official API from Twitter, data obtained will be only the last 7 days. Total of tweets obtained is 22.281 tweets. an example of a tweet is Figure 8. After the process of tweed data collection about corruption at KPK account, then the next process is taking username involved in the tweet interest of "corruption". The total username is 10.961 usernames. From the username obtained then conducted taking the following at each username. The example of taking following process in username can be seen in Figure 4. The Process of taking the following on username by using python twint library. The total username from the following is 4.995.357 usernames. The total unique username is 1.392.841 usernames. The total unique username is done by the process of retrieving tweets, account information and the process of getting the community. Account information attributes that will be taken are name, username, bio, join date, total tweets, total following, total followers and verified. The tweet attributes that will be taken are username, date, time, tweet, mentions, URLs, hashtags, and retweet. Tweet used is Indonesian. so that accounts using tweets other than Indonesian are deleted. Total accounts obtained are 575.851 accounts. The total spammer accounts are 2.312 accounts and the total legitimate accounts are 573.539 accounts.

The evaluation of success level from the proposed strategy is by using accuracy, recall, precision, and g-mean [20]. The calculation of accuracy, recall, precision, and g-mean used confusion matrix as showed in Table 3. Accuracy is the measurement of success level in detecting spammer (True Positive) and legitimate (True Negative) in all data. The accuracy calculation is done by using Equation 3. Recall is the measurement of success level in detecting spammer (True Positive) in all spammer data (actual positive). Recall is counted by using Equation 4. Precision is the accuracy level of information obtained. The precision calculation is conducted by using Equation 5. G-mean [21] conducts the calculation for the relative balance from the classification performance in positive and negative class. G-mean uses recall and precision. G-mean is counted by using Equation 6.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+FP+TN+FN}} \tag{3}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{4}$$

$$\boldsymbol{Presition} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \tag{5}$$

$$\boldsymbol{G-mean} = \sqrt{\boldsymbol{Recall * Presition}} \tag{6}$$

Table 4 is the evaluation result obtained. The data percentage of 70:30 gains the best results in accuracy, recall, precision, g-mean respectively 90,55%, 91,21%, 3,14%, and 16,74%. All best result obtained by proposed strategy. This shows that the success level in recall, precision, g-mean, and accuracy of proposal can improve spammer detection. In data distribution with percentage 80:20 obtained the result of accuracy, recall, precision, and g-mean respectively are 89,35%, 88,96%, 3,14%, and 16,37%. Proposed has the success for detecting spammer account and legitimate account based on accuracy, recall, precision and g-mean. At percentage 90:10 obtains the best result of accuracy, recall, precision and g-mean respectively 89,24%, 88,74%, 3,08%, and 16,17%. The best recall is at percentage 90:10 obtained by account feature. This shows that account feature also can detect spammer in overall spammer data. However, account feature decreases in g-mean, precision, and accuracy ability. For the success of spammer and legitimate detection, proposed is the best based on accuracy and g-mean. This also prevails for precision and recall obtained. The result of experiment shows that the method proposed obtains the best performance in accuracy, recall, precision, and g-means and the value for each respectively are 90,55%, 88,04%, 3.18%, and 16.74%. Accuracy and g-mean from the proposed method can exceed the previous method with 4.23% and 14,43%. This shows that the method proposed can overcome spammer detection problem on Twitter with better performance.

The best account feature in spammer detection based on g-mean is 9,90%. The evaluation result of accuracy, recall, and precision are 69,89%, 86,15%, and 1,14%. The features used are account age, length of screen name, username similarity and screen name, and following ratio, account activeness, eigenvector, follower ratio, interestingness, name ratio, and indegree. All those features appear in each data distribution. Account feature at all data distribution are account age, length of screen name, username similarity, and screen name, following ratio, account activeness, and eigenvector. This shows that the account feature selected is the precise feature to be used. Tweet feature successfully detect spammer based on g-mean is 15,61%. The evaluation result of accuracy, recall, and precision are 88,01%, 86,31%, and 2,82%. The features used are tweet length, URL ratio, mention ratio, lexrichoutuu, URL unique ratio, total account hashtag, account hashtag ratio, unique ratio of account hashtag, and total words of spam account. The average feature of tweet length, URL unique ratio, total account hashtag, account hashtag ratio, unique ratio of account hashtag, and total words of spam account will appear in each data distribution. In tweet feature appears three features related to hashtag such as feature of total account hashtag, account hashtag ratio, and unique ratio of account hashtag. This shows that feature based on hashtag has effect in detecting the spammer. Community feature succeeds in detecting spammer with g-mean measurement is 5,71%. The evaluation result of accuracy, recall, and precision are 43,40%, 67,87%, and 0,48%. Optimal features used were indegree, unique ratio of community hashtag, ratio of community URL and unique ratio of community URL. Community ratio for all data distributions are ratio feature of community URL and unique ratio of community URL. Optimal feature in another data distribution is total community members and total eigen communities. Community feature has one optimal hashtag aspect. This fact strengthen more and more that hashtag has effect in spammer detection. Therefore, development of feature extraction based on hashtag and community activity for spammer account detection on Twitter with this detection strategy can increase the success and accuracy.

5. Conclusion

This research proposes the development of feature extraction based on hashtag and community activities for detecting spammer account on Twitter. Hashtag is used by spammer members to increase their popularity. Community activity is used as the feature for spammer detection until it can give weight towards spammer activity obtained in certain community. The experimental result shows that the proposed method got the best performance in accuracy, recall, precision and g-means with are 90,55%, 88,04%, 3.18%, and 16.74%, respectively. The accuracy and g-mean of the proposed method can surpassed previous method with 4.23% and 14,43%. This shows that the proposed method can overcome the problem of detecting spammer on Twitter with better performance compared to state of the art.

References

[1] A. Rozen, "Giving you more characters to express yourself," *blog.twitter.com*, 2017. https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html (accessed Aug. 27, 2019).

[2] J. Martinez-Romo and L. Araujo, "Detecting malicious tweets in trending topics using a statistical analysis of language," *Expert Systems with Applications*, vol. 40(8), pp. 2992–3000, 2013, doi: 10.1016/j.eswa. 2012.12.015.

[3] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammers' social networks for fun and profit," in *Proceedings of the 21st international conference on World Wide Web - WWW '12*, 2012, doi: 10.1145/2187836.2187847.

[4] C. Yang, R. C. Harkreader, and G. Gu, "Die Free or Live Hard? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers," *Springer Berlin Heidelb.*, vol. 11(I), pp. 318–337, 2011, doi: 10.1007/978-3-642-23644-0_17

[5] C. S. K. Aditya, M. Hani'ah, A. A. Fitrawan, A. Z. Arifin, and D. Purwitasari, "Deteksi Bot Spammer pada Twitter Berbasis Sentiment Analysis dan Time Interval Entropy," *Jurnal Buana Informatika*, vol. 7(3), pp. 179–186, 2016, doi: 10.24002/jbi.v7i3.656.

[6] I. Inuwa-Dutse, M. Liptrott, and I. Korkontzelos, "Detection of spam-posting accounts on Twitter," *Neurocomputing*, vol. 315, pp. 496–511, 2018, doi: 10.1016/ j.neucom.2018.07.044.

[7] S. Y. Bhat and M. Abulaish, "Community-based features for identifying spammers in online social networks," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, 2013, pp. 100–107, doi: 10.1145/2492517.2492567.

[8] Y. Sarlati, S. Hashemi, and N. Mozaffari, "Adopting Community Features to Detect Social Spammers," in *2015 European Intelligence and Security Informatics Conference*, 2015, pp. 153–156, doi: 10.1109/EISIC.2015.44.

[9] C. Chen, J. Zhang, Y. Xiang, W. Zhou, and J. Oliver, "Spammers Are Becoming 'Smarter' on Twitter," *IT Professional*, vol. 18(2), pp. 66–70, 2016.

[10] P. V. Bindu, R. Mishra, and P. S. Thilagam, "Discovering spammer communities in twitter," *Journal of Intelligent Information Systems*, vol. 51(3), pp. 503–527, 2018, doi: 10.1007/s10844-017-0494-z.

[11] R. S. Perdana, T. H. Muliawati, and R. Alexandro, "Bot Spammer Detection in Twitter Using Tweet Similarity and Time Interval Entropy," *Jurnal Ilmu Komputer dan Informasi*, vol. 8(1), pp. 19-25, 2015, doi: 10.21609/jiki.v8i1.280.

[12] A. M. Priyatno, M. M. Muttaqi, F. Syuhada, and A. Z. Arifin, "Deteksi Bot Spammer Twitter Berbasis Time Interval Entropy dan Global Vectors for Word Representations Tweet's Hashtag," *Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 5(1), pp. 37-46, 2019, doi: 10.26594/register.v5i1.1382.

[13] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008(10), 2008, doi: 10.1088/1742-5468/2008/10/P10008.

[14] F. Masood *et al.*, "Spammer Detection and Fake User Identification on Social Networks," *IEEE Access*, vol. 7(3), pp. 68140–68152, 2019, doi: 10.1109/ACCESS.2019.2918196.

[15] M. Washha, A. Qaroush, M. Mezghani, and F. Sedes, "Unsupervised collective-based framework for dynamic retraining of supervised real-time spam tweets detection model," *Expert Systems with Applications*, vol. 135, pp. 129–152, 2019, doi: 10.1016/j.eswa. 2019.05.052.

[16] J. Rodríguez-Ruiz, J. I. Mata-Sánchez, R. Monroy, O. Loyola-González, and A. López-Cuevas, "A one-class classification approach

for bot detection on Twitter," *Computers & Security*, vol. 91, 2020, doi: 10.1016/j.cose. 2020.101715.

[17] D. Park, M. Lee, S. Park, J.-K. Seong, and I. Youn, "Determination of Optimal Heart Rate Variability Features Based on SVM-Recursive Feature Elimination for Cumulative Stress Monitoring Using ECG Sensor," *Sensors*, vol. 18(7), 2018, doi: 10.3390/s18072387.

[18] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46(1), pp. 389–422, 2002, doi: 10.1023/A:101248 7302797.

[19] K. Hans, L. Ahuja, and S. K. Muttoo, "Detecting redirection spam using multilayer perceptron neural network," *Soft Computing*, vol. 21(13), pp. 3803–3814, 2017, doi: 10.1007/s00500-017-2531-9.

[20] N. Japkowicz, "Assessment Metrics for Imbalanced Learning," in *Imbalanced Learning*, Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 187–206, 2013, doi: 10.1002/9781118646106.ch8

[21] M. Kubat, R. C. Holte, and S. Matwin, "Machine Learning for the Detection of Oil Spills in Satellite Radar Images," in *Machine Learning*, vol. 30, pp. 195–215, 1998, doi: 10.1023/A: 1007452223027.