Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). 8/1 (2015), 19-25
DOI: http://dx.doi.org/10.21609/jiki.v8i1.280

BOT SPAMMER DETECTION IN TWITTER USING TWEET SIMILARITY AND TIME INTERVAL ENTROPY

Rizal Setya Perdana, Tri Hadiah Muliawati, and Reddy Alexandro

Informatics Engineering Departement, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Jalan Raya ITS, Surabaya, 60111, Indonesia

E-mail: rizal13@mhs.if.its.ac.id, muliawati.trihadiah@gmail.com, reddy.alexandro@gmail.com

Abstract

The popularity of Twitter has attracted spammers to disseminate large amount of spam messages. Preliminary studies had shown that most spam messages were produced automatically by bot. Therefore bot spammer detection can reduce the number of spam messages in Twitter significantly. However, to the best of our knowledge, few researches have focused in detecting Twitter bot spammer. Thus, this paper proposes a novel approach to differentiate between bot spammer and legitimate user accounts using time interval entropy and tweet similarity. Timestamp collections are utilized to calculate the time interval entropy of each user. Uni-gram matching-based similarity will be used to calculate tweet similarity. Datasets are crawled from Twitter containing both normal and spammer accounts. Experimental results showed that legitimate user may exhibit regular behavior in posting tweet as bot spammer. Several legitimate users are also detected to post similar tweets. Therefore it is less optimal to detect bot spammer using one of those features only. However, combination of both features gives better classification result. Precision, recall, and f-measure of the proposed method reached 85.71%, 94.74% and 90% respectively. It outperforms precision, recall, and f-measure of method which only uses either time interval entropy or tweet similarity.

Keywords: *spam, Twitter, automation, bot spammer, entropy, tweet similarity*

Abstrak

Ketenaran Twitter mengundang *spammer* untuk menggunakannya dalam penyebarluasan pesan *spam*. Penelitian terdahulu menunjukkan bahwa kebanyakan pesan *spam* dihasilkan secara otomatis oleh *bot*. Deteksi *bot spammer* akan dapat mengurangi jumlah pesan spam pada Twitter secara signifikan. Akan tetapi, sejauh yang penulis ketahui, masih sedikit penelitian yang fokus dalam deteksi *bot spammer* pada Twitter. Sehingga, paper ini mengusulkan pendekatan baru untuk membedakan antara *bot spammer* dan pengguna sah menggunakan *time interval entropy* dan kemiripan antar *tweet*. Kumpulan *timestamp* digunakan untuk menghitung *time interval entropy* dari tiap akun pengguna. *Uni-gram matching-based similarity* akan digunakan untuk menghitung kemiripan antar *tweet*. Dataset diambil dari Twitter yang terdiri atas kumpulan akun normal dan akun yang terindikasi sebagai *bot spammer*. Hasil percobaan menunjukkan beberapa pengguna sah Twitter juga memiliki kebiasaan yang teratur dalam menghasilkan *tweet* sebagaimana *bot spammer*. Beberapa pengguna sah juga terdeteksi menghasilkan *tweet* yang mirip. Oleh karena itu, deteksi *bot spammer* menggunakan satu fitur saja akan kurang optimal. Akan tetapi, kombinasi atas kedua fitur tersebut memberikan hasil klasifikasi yang lebih baik. Presisi, *recall*, dan *f-measure* dari metode yang diusulkan mencapai 85.71%, 94.74% dan 90%. Nilai ini melampaui presisi, *recall*, dan *f-measure* dari metode yang hanya menggunakan baik *time interval entropy* maupun kemiripan antar *tweet* saja.

Kata Kunci: *spam, Twitter, otomatis, bot spammer, entropy, kemiripan antar tweet*

1. Introduction

Due to rapid development in internet connection, the number of user in Online Social Networking (OSN) websites are also increasing. Nowadays, OSN has been part of many people's daily routine. People may spend significant amount of time on popular OSN where they store and share personal information. Among various types of OSN, Twitter is considered as one of the most popular OSN. In last quarter of 2012, Twitter has been reported by Global Web Index as the fastest-growing website with a growth rate in active users of 714% since July 2009 [1]. Moreover, Twitter belongs to top 10 most viewed websites in Nov-ember 2014 [2]. Twitter is micro-blogging service that was founded in 2006. Twitter users are facilitated to communicate with each other by produc-

ing text-based post better known as tweet. The tweet size is limited to 140 characters. In total, there are 500 million tweets published by Twitter users per day. Its simplicity has attracted huge amount of people to join. Currently it has up to 284 million monthly active users [3].

However, the popularity of Twitter has also attracted many spammers to use it for disseminating large amount of spam messages. They try to exploit the network of trust among Twitter users for their own benefit, which are promoting personal blogs, spreading advertisements, phishing, and scam. The number of Twitter misuse can be worsening since the use of automated programs.

Automated program or better known as bot, short for robot, do not require human operator to execute its job. Preliminary studies had indicated that most of spam messages in Twitter are generated automatically by bot [4] and only very few of them are manually posted by humans [5]. Bot spammer can automatically generate spam message at given interval time using job scheduler [6]. Bot usage can reduce high cost of manually managing spam accounts, thus it is easier for spammer to generate more spam messages in Twitter.

The increasing number of spam message can deteriorate legitimate user experience in Twitter. It can pollute real time sharing information in Twitter and waste extra resource of legitimate user [7]. Therefore more rigorous efforts are required to stop further development of spammer in Twitter. Twitter itself has provided mechanism to stop spam development by inviting user to actively report spam message and account. However, it takes much time and resources due to several fake reports. Mistakenly labeling legitimate user account as spam can harm user's reliance toward Twitter [8]. Several researches have been conducted regarding automation (bot) and spam detection, to help fighting spam particularly in Twitter.

This paper proposes novel approach which combines entropy and tweet similarity to identify bot spammer. Time interval entropy is used to capture regularity of tweeting behavior which indicates automation. Entropy has been widely used to detect automation. Therefore several researches [5,10] utilize it to distinguish between bot and human behavior. In addition, tweet similarity is used to show the likelihood of Twitter account to be considered as bot spammer. Since many spammers tend to repeatedly tweet the same or similar post in order to increase the probability of successfully alluring legitimate users' visits. Their tweets used to have high homogeneous characteristics [4,9]. Instead of using cosine similarity as presented in [4], in this paper we prefer to use unigram matching-based similarity to overcome shortage of cosine similarity in short text as [11].

The rest of the paper is organized as follows. Related work is briefly reviewed in Section 2. The proposed method is elaborated in Section 3. Section 4 covers experiment section which includes not only data collection, but also result and discussion. Whereas, last section presents conclusion and future work.

Related Work

Several researches have been conducted regarding automation (bot) and spam detection, to help fighting spam particularly in Twitter.

Chu et al. [5] propose to classify Twitter users into several categories, which are human, cyborg, and bot. Entropy, spam detection, and account properties are used to identify bot and other categories. Among those features, the use of entropy produces the highest accuracy in classification. Entropy effectively captures timing behavior which distinguishes each category.

Zhang and Paxson [6] utilize Pearson x^2 algorithm to identify automation in Twitter using timestamp collection of users. Among observed users, 16% of them exhibit highly automation behavior. In addition, keywords which are associated with spam generally have higher automation rates than other keywords.

Amleshwaram et al. [4] introduce CATS which stands for Characterizing Automation of Twitter Spammers. They use various features to detect spam account, including tweet similarity by using cosine similarity.

Rather than detecting spam account, Stringhini et al. [9] create honey-profile in three popular OSN websites (Facebook, MySpace, and Twitter) to lure spam accounts and analyze their behavior. In the end, various features are utilized to identify spam account in aforementioned OSN, including message similarity. Since observed spammers tweet very similar messages, both in size and content as well as advertised websites.

2. Methods

In this paper, we propose novel approach to distinguish between bot spammer and legitimate user account. Due to its importance, spam detection has been widely researched, however few researches have focused in bot spammer detection. Even though preliminary studies have indicated that most spam messages are produced by bot.

Our proposed method utilizes not only behavior-based feature (time interval entropy) but also content-based feature (tweet similarity). For each user-k, its time interval entropy (H_k) and tweet similarity (Sim_k) will be calculated and combined to determine class which represents each user ac-

count. Flow mechanism of overall system is presented in Figure 1.

First, we collect timestamp of each user account which shows time interval needed by an account to post tweet. Time interval entropy (H) is calculated using equation(1) and equation(2) as used in [5].

$$H_{\Delta T}(T_i) = -\sum_{i=1}^{nT} P\Delta T(\Delta t_i)\log(P\Delta T(\Delta t_i)) \quad (1)$$

$$P\Delta T(\Delta t_i) = \frac{n\Delta t_i}{\sum_{k=1}^{n_T} n\Delta\, t_k} \quad (2)$$

Time interval between tweet is represented by ΔT, whereas $P\Delta T(\Delta t_i)$ denotes the probability of observing time interval Δt_i. The entropy component can detects periodic or regular timing which is strong indication of automation. Lower entropy value indicates regular behavior.

Since spammers tend to tweet similar message, we calculate tweet similarity using uni-gram matching-based similarity as presented in equation(3) and equation(4).

$$Sim\left(S_i, S_j\right) = \frac{(2 * |S_i \cap S_j|)}{(|S_i| + |S_j|)} \quad (3)$$

$$Sim_k = \frac{\sum_{i=1}^{m}\sum_{j=i+1}^{m} Sim\,(S_i, S_j)}{{}^1/_2\; m(m-1)} \quad (4)$$

For each user-k (Sim_k), we calculate its tweet similarity in pairwise. $Sim\left(S_i, S_j\right)$ calculates tweet similarity between tweet-i (S_i) and tweet-j (S_j). Whereas $|S_i \cap S_j|$ represents matching words between tweets. $|S_i|$ is defined as the number of words in tweet-i. Thus, Sim_k equals to the average value of pairwise tweet similarity within user-k. The number of tweet for each user-k is represented by m.

Before being calculated, each tweet has to be preprocessed. Preprocessing covers 4 steps, which are cleaning, stop-word removal, tokenizing, and stemming. Cleaning step aims to omit several parts of tweet including URL, mentioned user account, hash-tag, and RT. Moreover, stop-word will be removed by using stop-list which is implemented from [13]. Afterwards, each tweet will be tokenized and turned into root words using stemming algorithm which is proposed by Arifin and Setiono in [14].

Last, both values are combined using equation(5) to classify each user account into its designated class.

$$Score_k = \frac{\alpha(1 - H_k) + \beta\,(Sim_k)}{\alpha(\max(1 - H_k)) + \beta\,(\max(Sim_k))} \quad (5)$$

For each user-k, its time interval entropy (H_k) and tweet similarity (Sim_k) value should be multiplied by weighting factor to retrieve final value. Variable α and β denote weighting factor for time interval entropy and tweet similarity, respectively. Sum of both weighting factors should be equal to 1. Final score of user-k ($Score_k$) equals to sum of weighted time interval entropy and tweet similarity divided by sum of weighted maximum time interval entropy and tweet similarity.

3. Results and Analysis

In this section, we first describe the data collection. Detailed experiment is presented afterwards.

Data Collection

Datasets are crawled from Twitter using Twitter Streaming API. It facilitates third party to access Twitter's global stream of tweet data [12]. In total, there are 56 accounts which are written in Bahasa Indonesia to be observed containing both normal and spam accounts. Approximately 2000 tweets are collected from each account. Due to lack of ground-truth, we manually check each profile account and classify them into bot spammer or legitimate user. User is classified as spammer after checking its tweet content. Tweet which contains unsolicited advertisement is considered as spam. In addition, we also check following and follower ratio of each user profile account. According to preliminary studies in [5,7,9], spammer tends to follow many user accounts and have few number of follower. In total, dataset consists of 38 bot spammers and 18 legitimate user accounts.

Discussion

In order to quantitatively evaluate performance of the proposed method, precision, recall, and f-measure are utilized. Precision or positive predictive value is the fraction of retrieved instances which are relevant. Recall or sensitivity is the fraction of relevant instances which are successfully retrieved. F-measure is an accuracy measurement which considers both precision and recall value. Precision, recall, and f-measure are presented in equation(6), equation(7), and equation(8).

$$Precision = \frac{true\ positive}{true\ positive\ +\ false\ positive} \quad (6)$$

$$Recall = \frac{true\ positive}{true\ positive\ +\ false\ negative} \quad (7)$$

$$f - measure = \frac{2.(Precision\,.\,Recall)}{Precision + Recall} \quad (8)$$

According to equation(6), equation(7), and equation(8), we calculate precision, recall, and f-measure using combination of true positive, false negative, and false positive. In this paper, true positive refers to the number of correctly classified bot spammer. False positive represents the number of legitimate user which is incorrectly classified as bot spammer. Whereas, false negative is bot spammer which is incorrectly classified as legitimate user.

In the first experiment, we try to classify each user account using time interval entropy. Low entropy value indicates regular behavior. Therefore, user which has lower entropy than threshold will be classified as bot spammer. In this experiment we use 0,2 as threshold. Threshold is determined using exhaustive search algo-rithm by maximizing the f-measure which is not reported here. The initial value of threshold is 1 and being increased in steps of 0,5.

In the second experiment, instead of using time interval entropy, we utilize tweet similarity of each user account for classification. If user has higher value than threshold, it will be classified as bot spammer. In this experiment we use 0,6 as threshold. The same exhaustive search algorithm is utilized to determine threshold.

In the last experiment, we implement the proposed method which combines time interval entropy and tweet similarity to classify Twitter user account. Series of experiments are conducted beforehand to determine ratio of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ which are not reported here. According to aforementioned experiment, the best ratio of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is 1:1. Optimum threshold value for this experiment is derived using exhaustive search algorithm, which is 0,75. User account will be classified as bot spammer if its combined value is higher than threshold. The classification result of all methods is presented in Table 1.

According to classification result which is presented in Table I, several legitimate users are misclassified as bot spammer since they have lower entropy value than threshold. It can be inferred that legitimate user can also exhibit regular behavior in posting tweet. Thus, the use of time interval entropy is inadequate to distinguish bot spammer and legitimate user. Precision, recall, and f-measure for classification using time interval entropy are 86.49%, 84.21%, and 85.33% respecttively.

As presented in Table 1, even though most bot spammers are correctly classified, several legitimate users are misclassified as bot spammer. Those legitimate users tend to post tweet with similar topic, thus they have high value of tweet similarity. On contrary, several bot spammers are found to publish tweets which are quite heterogeneous. Even though they promote similar link, they use different wording. Therefore, those aforementioned bot spammers cannot be detected. Precision, recall, and f-measure for classification using tweet similarity are 75%, 63.16% and 68.57%, respectively.

The proposed method can produce better precision, recall, and f-measure which are 85.71%, 94.74% and 90%, respectively. Comparison among overall experiments is presented in Figure 2.

TIE, *TS*, and *PM* are abbreviation of Time Interval Entropy, Tweet Similarity, and Proposed Method, respectively. As presented in Figure 2, the proposed method has better performance than classification using tweet similarity only. However, it has slight lower precision than classification method which uses time interval entropy. The proposed method can increase the number of true positive and decrease the numbe of false negative. Thus, in general the proposed method still outperforms classification method which uses either time interval entropy or tweet similarity.

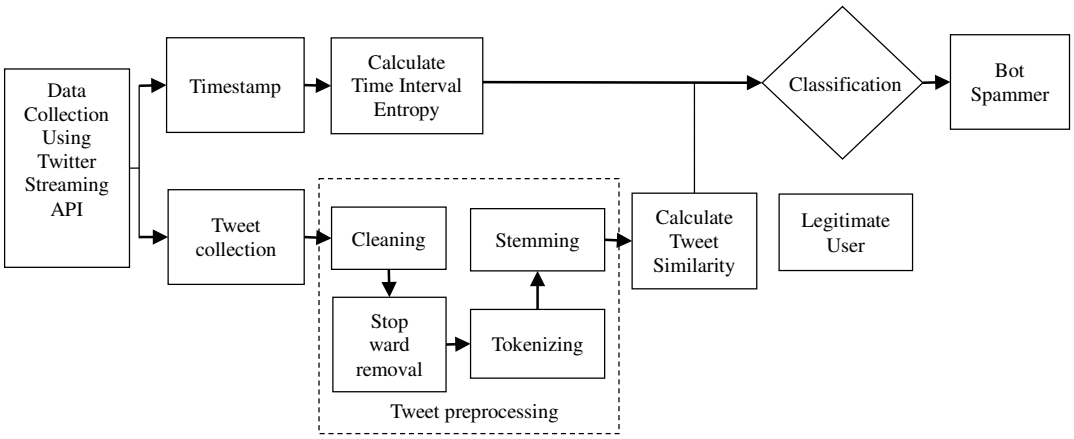
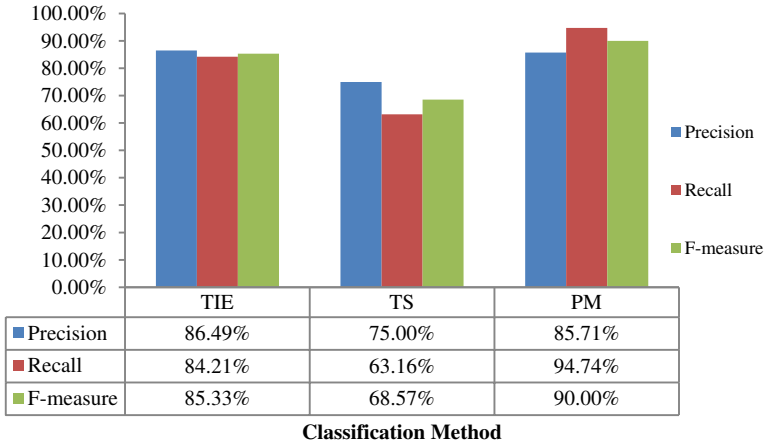


Figure 1. Flow mechanism of proposed method.

TABLE 1
CLASSIFICATION RESULT

User id	Value				Class		
	Time Interval Entropy	Tweet Similarity	Proposed Method	Ground truth	Time Interval Entropy	Tweet Similarity	Proposed Method
1	0,0567	0,036284	0,919687	Spam	Spam	Legitimate	Spam
2	0,0132	0,675997	0,98599	Spam	Spam	Spam	Spam
3	0,0126	0,647474	0,985506	Spam	Spam	Spam	Spam
4	0,0144	0,697067	0,985611	Spam	Spam	Spam	Spam
5	0,0115	0,783736	0,99168	Spam	Spam	Spam	Spam
6	0,0109	0,673837	0,988148	Spam	Spam	Spam	Spam
7	0,0156	0,474838	0,976121	Spam	Spam	Legitimate	Spam
8	0,0366	0,037363	0,939295	Spam	Spam	Legitimate	Spam
9	0,6156	0,123686	0,378855	Legitimate	Legitimate	Legitimate	Legitimate
10	0,0142	0,646494	0,983912	Spam	Spam	Spam	Spam
11	0,7083	0,183784	0,29086	Legitimate	Legitimate	Legitimate	Legitimate
12	0,2026	0,673939	0,801527	Spam	Legitimate	Spam	Spam
13	0,0169	0,686611	0,982785	Spam	Spam	Spam	Spam
14	0,0134	0,789979	0,990064	Spam	Spam	Spam	Spam
15	0,0181	0,047283	0,957677	Spam	Spam	Legitimate	Spam
16	0,6578	0,839294	0,36457	Legitimate	Legitimate	Spam	Legitimate
17	0,0132	0,072827	0,963403	Spam	Spam	Legitimate	Spam
18	0,0168	0,024637	0,958094	Spam	Spam	Legitimate	Spam
19	0,0132	0,026365	0,961664	Spam	Spam	Legitimate	Spam
20	0,0053	0,726082	0,995556	Spam	Spam	Spam	Spam
21	0,0823	0,30304	0,904753	Legitimate	Spam	Legitimate	Spam
22	0,0168	0,045373	0,958871	Spam	Spam	Legitimate	Spam
23	0,0132	0,838395	0,992071	Spam	Spam	Spam	Spam
24	0,0366	0,653434	0,962365	Spam	Spam	Spam	Spam
25	0,0201	0,024637	0,954882	Spam	Spam	Legitimate	Spam
26	0,7909	0,838395	0,234959	Legitimate	Legitimate	Spam	Legitimate
27	0,2017	0,134748	0,782212	Legitimate	Legitimate	Legitimate	Spam
28	0,7088	0,844746	0,315124	Legitimate	Legitimate	Spam	Legitimate
29	0,0183	0,654675	0,980227	Spam	Spam	Spam	Spam
30	0,7892	0,839294	0,236648	Legitimate	Legitimate	Spam	Legitimate
31	0,8689	0,637393	0,151498	Spam	Legitimate	Spam	Legitimate
32	0,6899	0,683933	0,327502	Spam	Legitimate	Spam	Legitimate
33	0,0286	0,738393	0,973334	Spam	Spam	Spam	Spam
34	0,8291	0,636827	0,190223	Legitimate	Legitimate	Spam	Legitimate
35	0,0181	0,639398	0,979849	Spam	Spam	Spam	Spam
36	0,0201	0,738382	0,981609	Spam	Spam	Spam	Spam
37	0,8976	0,738382	0,127339	Legitimate	Legitimate	Spam	Legitimate
38	0,1383	0,342334	0,851707	Legitimate	Spam	Legitimate	Spam
39	0,0793	0,342223	0,909141	Legitimate	Spam	Legitimate	Spam
40	0,2026	0,037228	0,777684	Spam	Legitimate	Legitimate	Spam
41	0,2165	0,636827	0,786605	Spam	Legitimate	Spam	Spam
42	0,0346	0,037228	0,941237	Spam	Spam	Legitimate	Spam
43	0,0443	0,593939	0,952641	Legitimate	Spam	Legitimate	Spam
44	0,8898	0,636827	0,13113	Legitimate	Legitimate	Spam	Legitimate
45	0,0689	0,037363	0,90785	Spam	Spam	Legitimate	Spam
46	0,2059	0,839294	0,804506	Spam	Legitimate	Spam	Spam
47	0,5019	0,128747	0,489735	Legitimate	Legitimate	Legitimate	Legitimate
48	0,8765	0,639398	0,144174	Legitimate	Legitimate	Spam	Legitimate
49	0,016	0,637393	0,981819	Spam	Spam	Spam	Spam
50	0,0132	0,683933	0,986287	Spam	Spam	Spam	Spam
51	0,0168	0,844746	0,988804	Spam	Spam	Spam	Spam
52	0,0983	0,493838	0,896322	Legitimate	Spam	Legitimate	Spam
53	0,0168	0,678384	0,982575	Spam	Spam	Spam	Spam
54	0,0183	0,036284	0,95707	Spam	Spam	Legitimate	Spam
55	0,3017	0,103586	0,683693	Legitimate	Legitimate	Legitimate	Legitimate
56	0,0053	0,023837	0,96926	Spam	Spam	Legitimate	Spam



	TIE	TS	PM
Precision	86.49%	75.00%	85.71%
Recall	84.21%	63.16%	94.74%
F-measure	85.33%	68.57%	90.00%

Classification Method

Figure 2. Performance evaluation of proposed method in comparison with other methods.

4. Conclusion

In this paper, a novel approach to detect bot spammer using combination of time interval entropy and tweet similarity has been proposed. Series of experiments has been conducted to evaluate performance of the proposed method.

It can be inferred from experimental results that the use of time interval entropy as behavioral feature is not sufficient to identify bot spammer. Even though entropy can capture automation behavior of Twitter account, however it cannot differentiate between bot spammer and legitimate user account.

Therefore, tweet similarity as content-based feature could be good match to complement it. The use of both features improves the overall system performance.

Further researches are needed to investigate the use of URL and URL shortening in spammer detection. Since Twitter limit each tweet to no more than 140 characters, spammer may use shorten website URL to lure legitimate user. In addition, several bot spammer also found to utilize trending topic in twitter to spread spam messages. They put trending topic into their published tweet, even though their tweet has no relation with trending topic.

References

[1] T.J. McCue, Twitter Ranked Fastest Growing Social Platform in the World, http://www.forbes.com/sites/tjmccue/2013/01/29/Twitter-ranked-fastest-growing-social-platform-in-the-world/, 2013, retrieved December 1, 2014.

[2] Alexa, The Top 500 Sites on The Web, http://www.alexa.com/topsites, 2014, retrieved December 1, 2014.

[3] About Twitter, https://about.Twitter.com/company, 2014, retrieved December 1, 2014.

[4] A.A. Amleshwaram, N. Reddy, S. Yadav, G. Gu, & C. Yang, "CATS: Characterizing Automation of Twitter Spammers" *In Communication Systems and Networks (COMSNETS), Fifth International Conference on IEEE*, pp. 1-10, 2013.

[5] Z. Chu, S. Gianvecchio, H. Wang, & S. Jajodia, "Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?" *Dependable and Secure Computing, IEEE Transactions on*, vol. 9, pp. 811-824, 2012.

[6] C.M. Zhang, & V. Paxson, "Detecting and analyzing automated activity on Twitter". In *Passive and Active Measurement*, *Springer*, pp. 102-111, 2011.

[7] F. Benevenuto, G. Magno, T. Rodrigues, & V. Almeida, "Detecting spammers on Twitter". *In Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, vol. 6, pp. 12, 2010.

[8] N. Cohen, Twitter Mistakenly Labels F.S.U. Article as Spam, for A Time, http://www.nytimes.com/2014/11/15/business/media/fsu-football-article-is-mistakenly-labeled-as-spam-on-Twitter.html?_r=0, 2014, retrieved December 1, 2014.

[9] G. Stringhini. C. Kruegel. & G. Vigna, "Detecting spammers on social networks" *In Proceedings of the 26th Annual Computer Security Applications Conference* ACM, pp. 1-9,2010.

[10] S. Gianvecchio, M. Xie, Z. Wu, & H. Wang, "Measurement and Classification of Humans and Bots in Internet Chat" *In USENIX security symposium*, pp. 155-170, 2008.

[11] K. Sarkar, K. "Sentence Clustering-based Summarization of Multiple Text Documents" *TECHNIA–International Journal of Co-*

mputing Science and Communication Technologies, vol. 2, 2009.

[12] The Streaming APIs, https://dev.Twitter.com /streaming/overview, retrieved December 1, 2014.

[13] F. Tala, "A study of stemming effects on information retrieval in Bahasa Indonesia", 2003.

[14] A.Z. Arifin. & A.N. Setiono, "Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering" *In Prosiding Seminar on Intelligent Technology and its Applications (SITIA), Teknik Elektro, Institut Teknologi Sepuluh Nopember Surabaya,* 2002.