Who are Tweeting in the 2014 Indonesia's Legislative Elections?

Rischan Mafrur[1], M Fiqri Muthohar[2], Gi Hyun Bang[3], Do Kyeong Lee[4], Deokjai Choi[5]
School of Electrical Engineering Chonnam National University
Gwangju, South Korea
Email : {rischanlab,fiqri.muthohar}@gmail.com[1,2],badmanner@naver.com[3], ldk7175@nate.com[4], dchoi@jnu.ac.kr[5]

***Abstract*—In April 2014 Indonesia has held legislative elections. Fifteen political parties have been participated to this election. Each parties has different strategic for campaign including social media campaign. In this paper we interested with one of political party which very active in social media campaign especially in Twitter, Prosperous Justice Party (PKS). Both of supporters and haters are active tweeting on Twitter about the goodness and badness of this party. This thing begs the question that "Who they are? It is really the voice of Indonesia, It represent Indonesian public opinion or just tweets from twitter campaign accounts ?".**

This paper tackles the above question by presenting the result of analysis with empirical data. We collected all tweets which related with this party, total more than 250 thousand tweets. We extract the data and classify to two types of twitter accounts: real accounts and campaign accounts. We use some features and Naive Bayes as method for classification. Finally we can determine who are really tweeting it.

Keywords*—Twitter for political campaign, social media analytics, web mining.

I. INTRODUCTION

This year (2014) is the politics year for Indonesia. In this year, Indonesia will have new president with the democratic elections. Beside presidential elections Indonesia also has held legislative election. Indonesia has many of political parties. Each political party has different strategy for campaign. Most of them use online media such as Facebook, Twitter, Youtube video for campaign. In this paper we interested about social media campaign especially on Twitter. As we know Twitter is not only micro-blogging service but also provides some features like real time trending topics and other features. Twitter provides "#" called "hashtag" it can used by user for giving some topics of their tweets. When many of people use the same hashtag, it will raise the possibility of the hashtag become a trending topics. The campaign schedule for the legislative election was from March 16 until April 5 2014. On that time each party has strategy for campaign include in social media campaign but we interested with one of parties, Partai Keadilan Sejahtera (PKS) or Prosperous Justice Party. The reasons why we interested with this party are as follows:

1) This party very active in social media campaign especially in Twitter campaign.
2) This party has many opposition or haters that they always tweeting about weakness of this party.
3) This party also has many supporters that they always tweeting about the goodness of this party.

The opposition or haters of this party use the hashtag #TolakPartaiPoligami. It means "We refuse party which has polygamy leader". This hashtag like a sarcasm because the leader of this party has more than one wife. This incident was published in several Indonesia online media such as Liputan 6, Tribun, and Republika. Finally this hashtag became worldwide trending topic at 9:30 PM, 20 Mar 2014 GMT+7. On the other side, the followers or supporters of this party use Twitter hashtag #SayaPilihPKS. It is mean "I choose PKS". This hashtag also became trending topic at 8:36 AM-22 Mar 2014 GMT+7 but only in Indonesia region not worldwide trending topics.

Our objective is to know who are tweeting both of hashtags. We want to classify to two types: real account and campaign account. In this case, real account means the account created by user for using Twitter such as for communication or tweeting something but not for spamming, promotion or campaign. We can determine which is real account by some features such as: creation date, tweet contents, period of tweeting, followers and friends, etc. On the other side, we found account which always tweeting the same content for specific purpose, so we think this is not real twitter account. Actually there are some differences between real account and campaign accounts such as the age of twitter accounts, the number of followers and following, the ratio tweeting and etc.

II. BACKGROUND AND RELATED WORK

A. Twitter

Twitter is one of most visited site now, if we look up on Alexa.com top site, twitter on top ten positions for the popular site in the world. Twitter is social networking site founded in 2006. People can share what is happening, they can post something that we call tweets. Twitter has some unique rules, Twitter only allow 140 characters for post tweets including HTTP link. We also can use hashtags ("#") for identifying the topic of our tweets. When many people use same hashtag then the hasthtag will be trending topics or current trend on Twitter at that time. The hashtag is like a keyword so when we click the hashtag we will find all tweets that use the same hashtag.

B. Various topics research on Twitter

Research on Twitter has been commonly with various topics. Jansen et al [1] mentioned that Twitter is an important tool for communication in marketing. Thelwall et al [2] research about reaction and public sentiment of popular events. Becker at al. [3] observed about real world event identification based on twitter trending topics.

As we know in 2014 Turkey government restricted access to Twitter because political issues reason. It is the indication that Twitter can affect the political situation in a country. There are many papers also about twitter in political issues. Small [4] mentioned in their research about Twitter in political campaigning and election. Wigand [5] presents some positive findings from the use of Twitter in terms of overcoming the limits of traditional communications between people with government stakeholders. They found that USA federal and local governments adopt Twitter faster than state agencies. Cho and Park [6] conducted in social networking and semantic content analysis of the Twitter account of a large South Korean Ministry. They mentioned that Twitter in government could function as an effective information distribution because Twitter can make mutual communication and direct conversation although with some limitations.

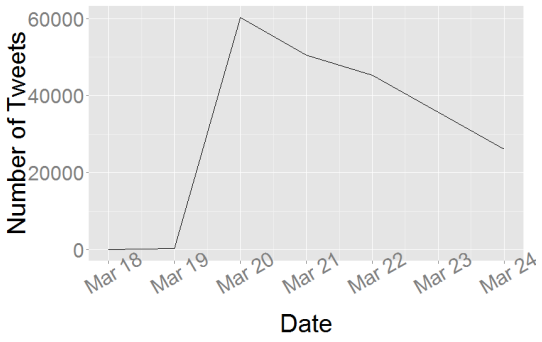
C. Twitter users classification

According to the introduction, we already explain about our purpose to classify twitter accounts. We found many of papers related with twitter accounts classification but most of them concern on spam and non-spam twitter account classification. Kwak et al. [7] filtered tweets from users who have been on Twitter for less than a day as well as tweets that contain three or more trending topics. They made classification between spam and non-spam account and then reported spam on the twitter data they collected. Yard et al. [8] studied the behavior of a small group of spammers. They found that the spammers have different behavior with non-spammers user such as replying tweets, followers, and friends. Wang [9] collected thousands users on Twitter and used classification to distinguish the suspicious behaviors from normal user.

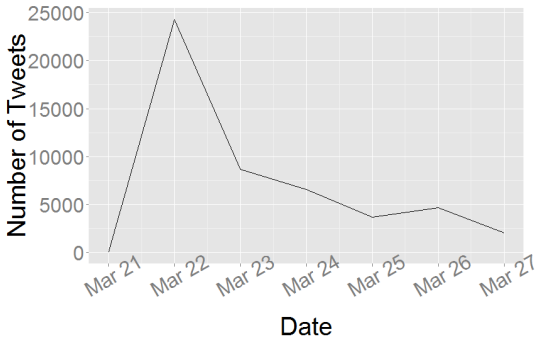
Zi Chu et al. [10] collected thousands Twitter users, They proposed features and techniques to classify Twitter users to three types : bot, human, or cyborg (human and bot). J. Song at al. [11] proposed new approach for classification between spam and non-spam Twitter users using sender and receiver relationship. Benevenuto et al. [12]. In their work, they collecting a large dataset and then they classify spam and non-spam users. They also provide some features, evaluate it using X2 statistic. C Yang [13] analyzing evasion tactics of twitter spammers and then they provide robustness features for solve it. They also evaluated 24 features for twitter users classification then make rank from low until high robustness.

D. Trending topics on Twitter

Trending topics are valuable to informs user what is the current trend in Twitter. We already mention about Thelwall et al [2] and Becker at al [3] researches. They use twitter trending topics for their researches. G. Stafford et al [13] gathered over 9 million tweets in Twitter trending topics over a 7 day period. They want to know effect of spammers in Twitter trending topics. They use Bayes classifier method to classify spam tweets. They found that spammers not drive the trending topics in Twitter. This research similar with our work, the different is Grant Safford et al [13] concern on question ”whether spammers can manipulate and drive twitter trending topics?” but in our work we concern to classify who are tweeting the hashtag. We want to know who they are and how many real accounts or campaign accounts.



(a) #TolakPartaiPoligami hashtag



(b) #SayaPilihPKS hashtag

Fig. 1: Tweets distribution

III. EXPERIMENT

In this part we will describe about how we get and extract the dataset, the ground truth creation, and the list of features that we used for accounts classification.

A. Data Collection

We collected the dataset for the 7 days. Figure 1 shows that number of tweets distribution per days, we can see the #TolakPartaiPoligami hashtag on March 20 the number of tweets almost 60,000 tweets and on that day this hashtag became trending topics. As well as the #SayaPilihPKS hashtag, the highest number of tweets is on March 22, almost 25,000 number of tweets. Total number of tweets are 222,444 tweets from #TolakPartaiPoligami hashtag and 48,135 tweets from #SayaPilihPKS hashtag. The total all of tweets are 270,579. From the all of dataset we select randomly 7,000 tweets from #TolakPartaiPoligami hashtag and 3,000 tweets from #SayaPilihPKS hashtag, we called this dataset is dataset I and the other dataset is dataset II. The dataset I is for ground truth creation purpose.

B. Data Extraction and Ground Truth Creation

The dataset that we gathered from twitter contains many of things such as username, tweets and other variables. Our purpose is to know who are tweeting the hashtag so we need to pick the twitter username from the dataset. We proposed algorithm for picking and counting twitter username from dataset. Actually this Algorithm 1 came from MapReduce, we

real	campaign	total
1,479	201	1,680

TABLE I: Hand labeled dataset I overview

modifying it according to our goal. In previous we already mentioned about ground truth dataset. We have 10,000 tweets (dataset I). To get the all of twitter username or twitter account from this dataset we need to apply the Algorithm 1 to this dataset. After we applied it we got impressive result, from the 10,000 tweets (dataset I) only came from 1,680 twitter accounts. It means many of twitter accounts they are tweeting the hashtags more than one times. From 1,680 twitter accounts we gave the label real and campaign accounts. We classified and gave hand-labeled to the twitter accounts manually one by one, the result can be seen on Table I.

Algorithm 1 Pick twitter username from dataset

```
Input dataseta : raw data (username, ..., tweet)
AccountMapping (String key, String rawdata)
for all username in rawdata do
  EmitIntermediate (username, "1")
end for

AccountReducing (String key, String Value[1..m])
int acc_count = 0
for all v in Value[1..m] do
  account_count += ParseInt(v)
end for
Emit(key, AsString(acc_count))
```

C. Features and Classification Methods

For classification features we use previous work that have been purposed by Benevenuto at al. [12] and C. Yang at al. [13]. They identified and provided the following features as being useful for detecting spam in Twitter. Benevenuto at al. [12] provide 10 features and C. Yang at al. [13] also provide 24 features but some of their features is same. Because of our purpose is not to classify between spam and not-spam so we need to determine which of the features were the most relevant to our task and dataset. We use 14 features such as : 1) average number of hashtags per tweet; 2) location data; 3) the age of twitter account; 4) hashtags ratio(day); 5) tweet ratio(day); 6) protected twitter account; 7) reputation; 8) number of all tweets; 9) API ratio(day); 10) URL ratio(day); 11) number of followings; 12) number of followers; 13) mention ratio(day); 14) characters length of description profile. We applied the Information Gain, Chi Square and ReliefF to our dataset (dataset I) then we make ranked the effectiveness and the last we only choose the top ten most important features. To classify we employed the popular machine learning algorithms, Naive Bayes. To evaluate the effectiveness of the classifiers we use standard information retrieval metrics: precision, recall, and accuracy with k-Fold cross validation, k=10.

IV. RESULT AND DISCUSSION

A. Features Evaluation

We analyze 14 features from previous research which related with our goal and whether it could be employed to our dataset. Table II shows the result of the rank top ten features evaluation. The result from Information Gain, Chi Squared, and ReliefF obtained the same features for the first rank, the age of twitter accounts. The result from Chi Squared and Information Gain is very similar only different on the third and fifth rank but not for the result from ReliefF. Why we show this result?, actually we use 14 features for this classification but the last four features did not have good value and did not affect the accuracy when we remove it. The four features that we removed are : 1) average number of hashtags per tweet; 2) location data 3) protected twitter account; 4) characters length of description profile. From this result now we know the most important features in our dataset for the classification.

B. Classifier Performance Evaluation

Table III shows the confusion matrix obtained from our Naive Bayes classifier on the dataset I. From 1,680 twitter accounts on dataset I, Naive Bayes has 12 classification error for classifying real accounts and 15 error for classifying campaign accounts. Table IV shows the information retrieval metrics for the classifier.

		Predicted	
		real	campaign
True	real	1467	12
	campaign	15	186

TABLE III: Confusion Matrix

	real	campaign
precision	0.99	0.92
recall	0.98	0.94
accuracy	0.98	0.98

TABLE IV: Classifier Performance

According to Table IV, the accuracy using Naive Bayes is pretty good, 98 %. Actually we also employed SMO (SVM) to this dataset but SVM did not perform better than Naive Bayes, so we decided to use Naive Bayes as the method.

C. Who are Tweeting

Before we talking about the result, to know who are tweeting the hashtag we need to pick the twitter username from dataset II using Algorithm 1. From The dataset II #TolakPartaiPoligami hashtag the total tweets are 215,444 came from 9,651 twitter accounts. The second hashtag #SayaPilihPKS, total tweets are 45,135 came from 5,639 twitter accounts. The total tweets in dataset II only came from 15,290 twitter accounts.

The results of Naive Bayes classification can be seen in Figure 2. #TolakPartaiPoligami hashtag has been classified to 6,621 (69%) campaign accounts and 3,030 (31%) real accounts. #SayaPilihPKS hashtag has 2,334 (41%) campaign accounts and 3,305 (59%) real accounts. Actually in this research we did not consider to divide between tweet and re-tweets data. As me mentioned in section III-B that most of twitter accounts

Value	Rank
1538.0	the age of twitter account
1433.8	number of followings
1321.3	number of followers
1238.5	mention ratio(day)
1214.8	number of all tweets
1206.5	hashtag ratio(day)
1150.8	tweet ratio(day)
481.7	reputation
84.6	API ratio(day)
0	URL ratio(day)

(a) Chi Squared

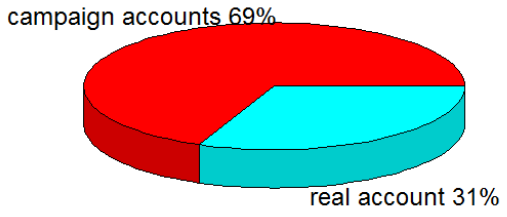
Value	Rank
0.45	the age of twitter account
0.41	number of followings
0.40	number of all tweets
0.38	mention ratio(day)
0.37	number of followers
0.35	hashtag ratio(day)
0.33	tweet ratio(day)
0.18	reputation
0.06	API ratio(day)
0	URL ratio(day)

(b) Information Gain

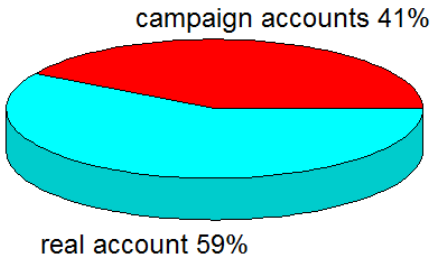
Value	Rank
0.50	the age of twitter account
0.25	mention ratio(day)
0.23	number of all tweets
0.13	API ratio(day)
0.12	hashtag ratio(day)
0.11	tweet ratio(day)
0.09	URL ratio(day)
0.08	number of followings
0.01	reputation
0.008	number of followers

(c) ReliefF

TABLE II: Features Evaluation



(a) #TolakPartaiPoligami hashtag



(b) #SayaPilihPKS hashtag

Fig. 2: Percentage of campaign and real accounts

tweeting more than one time and maybe many twitter accounts they only re-tweet tweets from their friends, so it could be our next future work.

In this research we also tried to analyze the distribution of number twitter accounts with the features. As we mentioned in IV-A the most important features is the age of twitter accounts. It is understandable, when we make a little observation with the twitter campaign accounts most of them created on January or February 2014, two or three months before campaign schedule. We thought this accounts will active tweeting about politics until the Indonesia presidential elections finished. Figure 3 shows the plotting of distribution twitter accounts with the age of twitter accounts. The x-axis is the number of days and y-axis is the density(the number of twitter account). Most of twitter campaign accounts (red line) from the Fig. 3a and Fig. 3b the age around of 100 days or three months. On the other side, most of real accounts they has average age around 700 days (almost 2 years).

V. CONCLUSION

Twitter is one of tool which not only for communication with others but twitter can be used for business, promotion, administration, or political campaign. In political terms a person can easily use twitter or other social media for campaign.

When there are many newspapers or online media were reported that there are many people who hate or love with one of party because many people tweeting about it, It could not be used as a basis of truth. Not all tweets on the Twitter derived from the real accounts, it could be from a bot, cyborg or campaign accounts. This paper descibed about it, we collected all tweets from the two kinds of hashtags that total all of them are more than 250 thousand tweets which only came from 15 thousand twitter accounts. The #TolakPartaiPoligami hashtag that became worldwide trending topics has almost 70% tweets came from the campaign accounts as well as the #SayaPilihPKS hashtag which became a indonesian regional trending topics has 40% tweets from campaign accounts.

VI. ACKNOWLEDGEMENT

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2014-H0301-14-1014) supervised by the NIPA(National IT Industry Promotion Agency).

Basic Science Research program through the National Research Fund of Korea (NRF) funded by the Ministry of Education, Science, and Technology (MEST), Korea (2012-035454).

REFERENCES

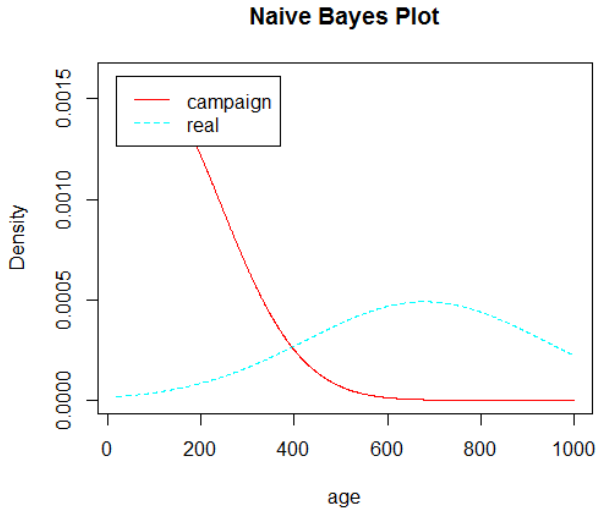
[1] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *Journal of the American Society for Information Science and Technology*, vol. 62(11), pp. 2169–2188, 2009.

[2] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in twitter events," *Journal of the American Society for Information Science and Technology*, vol. 62(2), pp. 406–418, 2011.

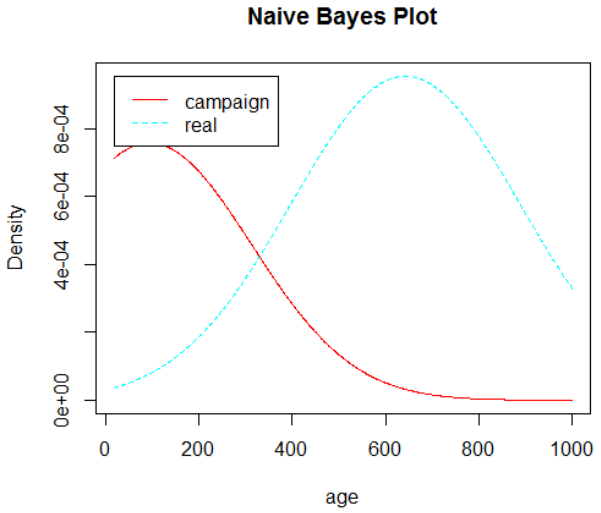
[3] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on twitter." in *ICWSM*, Barcelona, July 17-21 2011.

[4] T. A. Small, "What the hashtag? a content analysis of canadian politics on twitter," *Journal Information,Communication and Society*, vol. 14(6), pp. 872–895, 2011.

[5] R. D. L. Wigand, "Tweets and retweets: Twitter takes wing in government," *Journal Information Polity*, vol. 16(3), pp. 215–224, August 2011.



(a) Age of twitter accounts from #TolakPartaiPoligami hashtag



(b) Age of twitter accounts from #SayaPilihPKS hashtag

Fig. 3: x-axis: the age of twitter accounts (days), y-axis: twitter accounts distribution

[6] S. E. Cho and H. W. Park, "Government organizations' innovative use of the internet: The case of the twitter activity of south korea's ministry for food, agriculture, forestry and fisheries," *Journal Scientometrics*, vol. 90(1), pp. 9–23, January 2012.

[7] H. Kwak, G. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 591–600.

[8] S. Yardi, D. Romero, G. Schoenebeck, and danah boyd, "Detecting spam in a twitter network," *First Monday*, vol. 15, pp. 1–4, January 2010.

[9] A. H. Wang, "Dont follow me: Spam detection in twitter," in *International Conference on Security and Cryptography (SECRYPT)*, July 26-28 2010.

[10] Z. Chu, S. GIanvecchio, H. Wang, and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?" *Journal IEEE Transactions on Dependable and Secure Computing*, vol. 9(6), pp. 811–824, November 2012.

[11] J. Song, S. Lee, and J. Kim, "Dont follow me: Spam detection in twitter," in *Proceedings of the 14th international conference on Recent Advances in Intrusion Detection*, September 20-21 2011.

[12] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Collaboration, Electronic messaging, AntiAbuse and Spam Conference (CEAS)*, July 2010.

[13] C. Yang, R. C. Harkreader, and G. Gui, "Die free or live hard?empirical evaluation and new design for fighting evolving twitter spammers," in *Proceedings of the 14th international conference on Recent Advances in Intrusion Detection*, 2011.