Opinion Extraction of Public Figure Based on Sentiment Analysis in Twitter

Nur Hayatin[1], Mustika Mentari[2], and Abidatul Izzah[3]

Abstract— **Twitter is a microblog that can generate an information from users such as sentiment about public figures. Sentiment analysis of public figure interpret the positive or negative response. This study aims to create system that automatically can extract the opinion about public figure based on sentiment analysis in twitter using two novel features, they are specific term and number of followers public figures lover and hater. Several step to determine the sentiment of public figure are preprocessing, weighting, classifying, and determining sentiment response. In this paper we use six public figures to be observed. This research resulting precision 99%, recall 75%, and accuracy 76,67%.**

Keywords—**Opinion Extraction, Sentiment Analysis, Twitter, Public Figure, Naive Bayes**

I. INTRODUCTION

Today, social media becomes a popular society in many activities, including goverment, science, and business activities. Moreover it cause the writing habits, that don't have clear purpose for important activites. Some social media continues to be developed by developer gradually. One example, twitter has practical and lighter appearance to be installed in various devices. Twitter is an online microblogging service used for read or send short messages that called tweets. Text message on twitter is limited to 140 characters. The text messages consist of some features which available at twitter such as hashtag, mention, url, picture, video, etc. People prefers to use twitter than other social media because it is more lighter to access. Moreover business executive or public figures choose twitter to promote biography or to publish important activity that many people want to know.

Twitter became one of the popular social media today. Twitter contains a set of messages about emotion and opinion as community responses. It can be used as dimention data to analyze sentiment of public figure become positive and negative responses from society which determined automatically with some step, until we conclude the tendency of society respons.

Research about sentiment analysis has been done by researcher using various types of data and methods. Some analysis of the public figures, made by Ismail Sunni [1]. The sentiment analysis resulting of twitter is user's keywords (tweets) and emoticons. Classification method that used in this research is Naive Bayes.

Once classified, the sentiment changes are presented in the form of curves. In addition, TF - IDF method used to search topics which causes sentiment changes. The final results of this study indicate that the preprocessing methods tested did not show any significant effect, but the decay factor for the search topic has increased the number of topic extraction.

Frameworks development about analysis sentiment of public response observed by Davidoff [2]. Sentiment analysis, obtained through the tendency of sentiment, such as Likely, focused, etc. Features uses for determines sentiment is hashtags and smileys. Then the classification method used is a supervised classification method k-nearest neighbor (kNN).

Other sentiment analysis performed by Montejo-Ráez in [3], which use the method of extracting vector graphics WordNet. Features that used in this research are tweets and emoticons. Classification method performed by comparing methods that already widely studied. Those methods are SVM and SentiWordNet Rank WordNet (SWN-RW). The results of comparing the two methods of classification, showed the proposed method has precision and recall that are not much different from SVM, but the proposed method has advantages to avoid the ambiguous meaning of word. A suggestion is adding a further study in certain specific tweet in a specific time that will be address sentiment analysis to the final result of classification.

Through the users of some social media, political analysis based on twitter users uses Conover content network analysis [4]. Classification methods used to test the accuracy of the SVM with 10-fold cross-validation-. The features used in this study is mention, hashtag and retweet. The research resulted high accuracy and strong relationships because of retweet feature.

Twitter as a media for brand sentiment analysis use integration method of n-gram analysis and dynamic artificial neural network which observed by Ghiassi [5]. Features that used in this research are tweet corpus,

[1]Nur Hayatin is with Departement of Informatics Engineering, Faculty of Engineering, University of Muhammadiyah Malang 65114, Indonesia. Email : noorhayatin@umm.ac.id

[2]Mustika Mentari is with Departement of Information System, Faculty of Computer Science, Narotama University Surabaya 60117, Indonesia. Email : must.mentari@gmail.com

[3]Abidatul Izzah is with Departement of Informatics, Faculty of Information Technology, ITS Surabaya 60111, Indonesia. Email : aza.syaifa@gmail.com

emoticons and twitter accounts that contain of love and hate words along with their synonyms. To determine the user's sentiment, two classification methods have been compared (SVM and DAN2). The results shows that DAN2 has better result than SVM from precision and recall computation.

Data tweet on twitter collected from message collections. This data has methods adapted to particular text data in Indonesian text, as research conducted by Arifin [6]. This study discusses about the method for identifying topics from an Indonesian news dataset based on a main topic query input by user. Topics are searched with extracted the keywords out of the query that relevant to the topic.

There is some studies that discuss about sentiment analysis. Many features that must adapted to the case to be resolved. There are classification methods to sentiment analysis for various cases have been implemented. Research [4] takes an additional feature. It covered term to be more specific which support the classification conclude good decision. In this paper, we proposed using two novel features as contribution to opinion mining against public figures based on sentiment analysis of twitter data. They are number of follower lover and hater account of public figures and a number of spesific terms that lead to the election. The data tweets will be classified using Naïve Bayes Classification (NBC). Thus, the study aims to create system that automatically can extract the opinion about public figure based on twitter using two novel features to determine the negative and positive response of public figures in Indonesia.

II. Sentiment Analysis using NBC

Sentiment analysis of public figures is searched through several processes ranging from preprocessing to classification using NBC. NBC uses statistical concepts in term weighting. The weights for each category will be calculated and categorized on negative, positive or neutral sentiment.

A. Naive Bayes Classification

NBC [7] is a simple probabilistic classifier based on Bayesian theorem with strong (naive) independence assumptions. A widely used framework for classification is provided by a simple theorem of probability known as Bayesian Rule on Eq (1)

$$P(C|F_1,\dots,F_n) = \frac{P(F_1,\dots,F_n|C)P(C)}{P(F_1,\dots,F_n)} \quad (1)$$

Where $P(C|F_1,\dots,F_n)$ is probability dependent category C conditional on several feature $F_1,\dots,F_n$ whereas $P(F_1,\dots,F_n|C)$ is probability of feature $F_1,\dots,F_n$ conditional on category C. We know that $P(F_1,\dots,F_n)$ has a constant value so a common strategy to assume that the distribution of C conditional on $F_1,\dots,F_n$ can be decomposed in Eq (2)

$$P(C|F_1,\dots,F_n) = P(C)\prod_k P(F_k|C) \quad (2)$$

Where $\prod_k P(F_k|C)$ is probability each feature on category C.

Classification apply by determining category c_i that have maximum value of $P(C = c_i|F_k)$ calculated by Eq (3) :

$$c^* = \arg\max_{c\in C} P(C = c_i)\prod_k P(F_k|C = c_i) \quad (3)$$

where c^* is category that has maximum value $P(C = c_i|F_k)$. In sentiment analysis, NBC is used to calculate probability of word in a tweet and determine probability of tweet is positive sentiment or negative one. Assume that tweet is a sets of words w_k so we can categorized based on Eq (4) :

$$c^* = \arg\max_{c\in C} P(c_i)\prod_k P(w_k|c_i) \quad (4)$$

where $P(c_i)$ is probability of tweet on category c_i and $P(w_k|c_i)$ is probability of w_k conditional on category c_i. Further, $P(c_i)$ and $P(w_k|c_i)$ are calculated by formula (5) dan (6):

$$P(c_i) = \frac{|T_i|}{|T|} \quad (5)$$

$$P(w_k|c_i) = \frac{n_k+1}{n+|vocabulary|} \quad (6)$$

where $|T_i|$ is a number of tweet categorized on c_i, $|T|$ is a number of all tweets in training process, n_k is frequency of word w_k conditional on category c_i, n is a number of word on category c_i, and $|vocabulary|$ is a number of unique word. Figure 1 shows the process of sentiment analysis using NBC.

B. Dataset

In this research, dataset (tweet) was collected from Twitter. The tweets were crawled based on keyword from name list of public figure that refers to Litbang Kompas survey [8]. The public figures were related to Indonesian election for 2014. There are six names used in this research from popular public figure. They are Aburizal Bakrie, Dahlan Iskan, Gita Wirjawan, Joko Widodo, Prabowo Subianto, and Rhoma Irama.

Tweet was collected by twitter API Streaming [9] started from November 18th, 2013 until November 21th, 2013 for training data and started from December 1th, 2013 until December 2th 2013 for testing data. There are 200 tweets for training and 100 tweets for testing for each public figure, so that we have 1800 tweets for six public figures. The examples of tweets based on keyword "Jokowi" and "Rhoma Irama" are shown in Table 1. Each tweet is labeled by positive, neutral, or negative sentiment manually.

We also use alternative keyword based on nickname, slang name or popular name. For example, "ical" is Aburizal Bakrie's nickname and "Bang haji Rhoma" is Rhoma Irama's popular name.

We also use number of follower lover and hater account to know who love or hate the public figure. For example, Prabowo Subianto has lover account @Vote_Prabowo and hater account @DosaBowo. The others lover and hater account for him are shown in Figure 2.

C. Sentiment Analysis

Sentiment analysis procedure has three main steps, which is training, testing, and determining public response. Figure 3 shows sentiment analysis procedure.

Training step consist of three phase, they are preprocessing, feature extraction, and weighting. Tweets are preprocessed to be a basic form by applying case-folding, tokenizing the term, normalizing tokenizes emoticon, normalization, and Nazief-Andriani stemming. The tweet will be separated between words and emoticons. Then prefix, infix, and sufix will be eliminated. Preprocessing aims to clean tweet from noise and slang words. In Bahasa Indonesia, there are slang words which are usually used in microblog conversation. However, those slang words are needed to be normalized into standard form. For example a tweep stated: “akyu gk pilih Gita”. It should be normalized be: “aku tidak pilih Gita“(I’m not vote for Gita).

It result will be extract to get unique-terms. These terms will be feature in classification. Over all, we have three term features. They are general term, specific term, and emoticons. For both specific term and emoticon will be change into two “negative” or “positive”. Stop list filtering used to reduce term dimension that not meaningful and have high frequency. The next step is calculation weight of unique-term and save it in the database. Testing step is classification phase using NBC algorithm that was described in section II.A. Finally, we combine between probability of sentiment tweet and number of follower to determine sentiment response. The followers get from lover and hater account of each public figure.

We evaluated the system by calculating precision and recall. This measurement can be calculated by Eq 7 and 8.

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

Probability theory is used to get sentiment tweet for every public figure. This probability calculated with Eq 9.

$$P(A) = \frac{n(A)}{n(S)} \quad (9)$$

For this formula $n(A)$ is a total of tweet that have positive sentiment. $n(S)$ is a total of retrieved tweet. This formula also used to calculated follower from lover and hater account.

D. Novel Feature

Novel features used are specific term and a number of followers from lover and hater account. Specific terms are words that commonly used in election and can represent negative and positive sentiment. For example, "pilih" is specific terms for positive sentiment and "tidak-pilih" for negative sentiment. They also used in a query when we collected tweets. Specific terms used are stated in Table 2.

Other feature used is number of follower from lover and hater account. Firstly we list lover and hater accounts. After that, we calculated number of follower from them for each public figure. So, the probability of lover can be calculated from number of follower from lover account divided by number of follower from both lover and hater account. It’s a same way to calculate probability of hater.

The testing process is done by giving a frequency of 60 Hz and then seen the maximum motor speed response. Then look at the maximum voltage of the inverter with a voltmeter. At this stage, making test program to try hardware induction motor in which the program will be included in the PIN FR and FC in the inverter. After that is done by replacing the value of the incoming voltage into the inverter which uses a scale of 0.1 volts.

III. Results and Discussion

A. Experiment Result

In preprocessing step, we must normalize terms to be formal one. In Bahasa Indonesia, we have *bahasa alay* (slang languages) likes “aku“ has a various term such as “akyu", "akuwh", "akku", "q", "aq", "aquwh", "gue", "gw", "q", and "ak”. Moreover, tweet contains of many symbol (emoticons) such as point, comma, slash, also parentheses and character combinations that are used to express sentiment on the service. Emoticon tokenizing aims to translate tweet sentiment that express how tweep feels. Emoticons are represented in the model by the features happy emoticon (positive sentiment), sad emoticon (negative sentiment), and neutral emoticon. The commonly employed happy, sad, and neutral emoticons are presented in Table 3. Here they are how to preprocess a tweet:

pak JOKO WIDODO haha :D http://t.co/ MO2oeVYrEi

For each tweet, we apply case folding, normalizing, and stemming process. In this process, we get terms without capital and URL as:

bapak joko widodo haha positifi t co mo2oevyrei

Then this result is extracted from unimportant and unique term, so we have only terms:

bapak positif

Finally, we must calculate term frequency to get weight for each term. Once the tweets are fully preprocessed and feature extraction techniques have been applied, classification process can begin. The Sentiment analysis output about political tweet using NBC shown in Table 3.

Finally, 600 tweets public for 6 political figures have been classified into 3 categories (neutral, positive, and negative). Performance evaluation in typical information retrieval tasks, precision and recall are common performance metrics. The precision is the number of relevant events detected over the total number of events detected, while the recall is the number of relevant events detected over the total number of relevant events that exist in the data streams. In this case, we get the value of precision for positive category is 0.98. Then the value of

recall for positive categories is 0.75. Calculation result about sentiment analysis shown in confusion matrix on Table 4.

By considering all features, we assume positive respond is mean from positive probability both of public tweet and account. In same way, negative respond is mean from negative probability both of public tweet and account. For example, we have value 0.66 for Aburizal Bakrie on positive probability of public tweet and 0.34 on positive probability of account. So we get mean both of them is 0.83 on positive respond whereas 0.17 on negative respond. As we have two conditions like positive respond and negative one, we choose 0.5 as threshold to determine dominant probability. In the Aburizal Bakrie case, he got 0.83 as final respond, so we conclude that he was preferred by tweeps. Further, calculation final respond for each political figure is shown on Table 6.

B. Discussion

The system which automatically extract opinion about public figure based on twitter has been created. But, there are some problems on several processes extractions. One of them is when tweets crawling process, manually many of them are positive sentiment and negative one. But, most tweets crawled are neutral because it contains news or link to such URL. It decreased the number of tweets sentiment that had been explored.

In this paper, we consider use a number of followers from lover and hater account. By using this feature, we would know how many tweeps have similar characteristics. But in fact, all of them have lover account and 3 of 6 political figures have no hater account. It's caused the positive probability of account be 1 and increase respond value.

Sometimes the specific features cause errors in certain cases. For example, negative tweet about Wirjawan Gita written by HV_nes :

> *RT @ibnux: Hayoloh, pilih pak gita wirjawan, ga bisa nikmatin smartphone beli dr luar negeri, kasian yg roaming [pic]*

Had been positive detected by the system. Weighting the specific term "pilih" made a positive outcome regardless of the classification into different meanings.

Based on the result, we obtained the value which represented respond about each political figure. It turn out that each political figure got more than 0.5 in positive respond, so we can conclude that all of them were preferred by tweeps. The systems that have been evaluated have precision for neutral, positive, and negatives categories are 0, 0.98, and 0.28. Then the value of recall for neutral, positive, and negatives categories is 0, 0.75, and 0.95. In contrast, when we check the manual label (Table 7), Gita and Rhoma have a small number in positive sentiment, the system instead identified many positive tweets than negative. It shown that system have a poor classification (76.67%). A poor classification is caused by weighting in training process.

Some misclassification caused by incomplete dictionary. If we are not found match term in dictionary, it will not be processed to the next step. This case occurred in Gita Wirjawan's tweet, because we have no term 'ilfeel' so a tweet likes

> *RT @GuyonWaton: Gita wiryawan ki sopo ? Pengen banget dadi presiden kok le ngoyo, kabeh selebtwit di jak kampanye. Malah njuk ilfeel aku.*

will be extracted and term that will be processed are

> *gita ki ingin banget dadi presiden kok ngoyo di kampanye malah aku*

Because there are positive term such as 'ingin' and 'presiden' processed and negative term such 'ilfeel' did not processed, so this tweet detected as positive sentiment.

IV. CONLUSION

This paper proposed opinion mining of public figure based on sentiment analysis from twitter using two additional features as the contribution. That features are number of followers lover and hater account and several specific terms. This research begins from preprocessing, weighting, until sentiment analysis using NBC. This method has been tested and resulted in the evaluation of specific and unspecific terms. By sentiment analysis of public tweet we obtained that all of public figures have good responds. This result is validated using precision and recall. We obtained 99% positive, and 28% negatives for precision values and 75% positive, and 95% negatives for recall values. Those respond combined by the number of follower lover or hater account. Finally, six public figures got probability value above 0.5 for positive sentiment. It means that all of them got a positive response from the public.

Future research that can be extended from this research is addition of automatic filtering method for tweets selection, so the crawled tweets have no duplication. Enhanced slang word on dictionary words also can be added appropriate with updated slang. As an effort to improve the accuracy of the preprocessing methods, it may be used reliable method and considering error correction letters or writing word. It also has an opportunity to change manual labeling with automated methods. In this paper, sentiment analysis using value 0.5 as threshold because there are two types of sentiment, positive and negative. This can be developed by using another threshold with particular consideration. The final results of this study concluded that every public figure is resulted in positive and negative sentiment, but it may be predicted with the best public figure. This could be an opportunity to develop this research in the future with certain consideration.

REFERENCES

[1]. I Sunni and DH Widyantoro, "Analisis Sentimen dan Ekstraksi Topik Penentu Sentimen pada Opini terhadap Tokoh Publik," *Jurnal Sarjana ITB*, vol. 1, pp. 200-206, July 2012.

[2]. D Davidov, O Tsur, and A Rappoport, "Enhanced Sentimen Learning Using Twitter Hashtags and Smileys," in *Coling 2010: Poster Volume*, Beijing, 2010, pp. 241-249.

[3]. A Montejo-Ráez, "Ranked WordNet graph for Sentimen Polarity Classification in Twitter," *Computer Speech & Language*, vol. 28, pp. 93-107, April 2013.
[4]. M D Conover, "Predicting the Political Alignment of Twitter Users," in *3rd IEEE Conference on Social Computing SocialCom*, Boston, 2011, pp. 192 - 199.
[5]. M Ghiassi, J Skinner, and D Zimbraa, "Twitter brand sentiment analysis: A hybrid system using n -gram analysis and dynamic artificial neural network," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6266–6282, November 2013.
[6]. A Z Arifin, "Topic Identification of Indonesian Language News Article Based on Theme Query," *IPTEK, Journal of Engineering*, vol. 1, no. 1, 2014.
[7]. S Eyheramendy, D D Lewis, and D Madigan, "On the Naive Bayes Model for Text categorization," in *Ninth International Workshop on Artificial Intelligence and Statistics*, Florida, 2003, pp. 332–339.

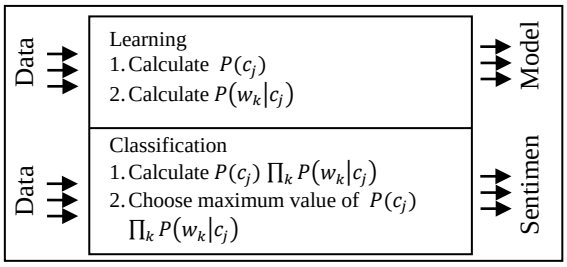

Figure 1. Classification Step using NBC

Figure 2. Prabowo Subianto's lover and hater account

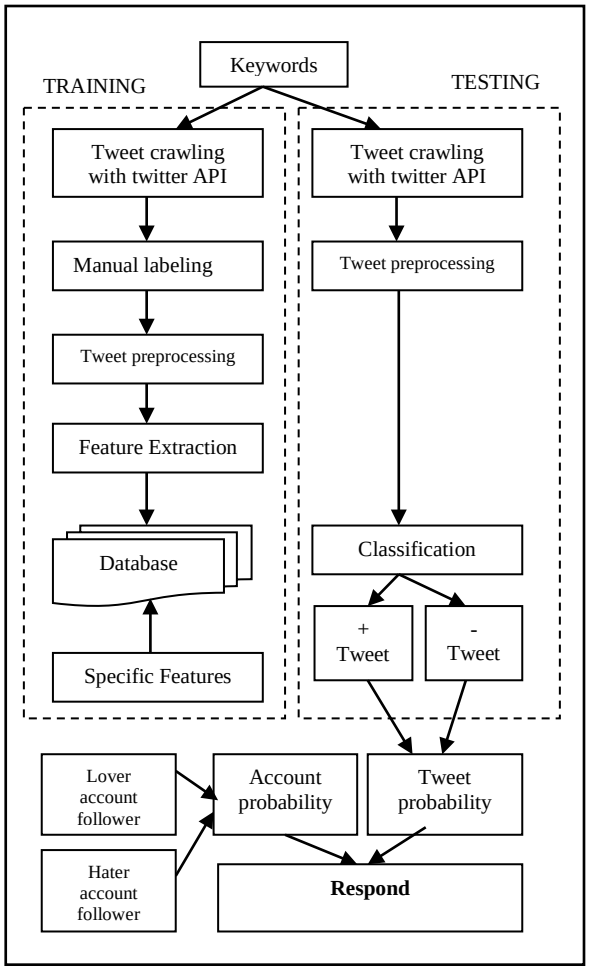

Figure 3. Diagram of public figure sentiment analysis

TABLE 1.
THE DATASET EXAMPLE FROM KEYWORD "JOKOWI" AND "RHOMA IRAMA"

User	Tweet
@Robbiiromeo	**Jokowi** Promosi Blok G di Depan Ibu PKK: Image:... http://t.co/wz0JcDe1GV
@BeritaBaruID	JOKOWI CAPRES : **Jokowi** Top of Mind Survei Pilpres CSIS http://t.co/1EgOY867t2
@GunzVinsLautner	Golkar: Saat Hadapi Buruh Jokowi Liberal: **Jokowi** lebih mirip kader partai beringin yang tumbuh di 'kandang banteng'. http://t.co/oXQKC4SE6j
@aditya_RSH	**Rhoma Irama** for President 2014. Goyang Bang Oma!!! [pic]? https://t.co/NnJI7UkW0E
@alda_eva	Bang haji **rhoma irama** nyapres.. ntr yg mimpin soneta siapa baaang wakakkkkaaaakkkk.saya z ya baang..
@armelDPS	RT @dheniedonat: saya mendukung **rhoma irama** menjadi presiden, biar Indonesia makin ngga jelas

TABLE 2.
EMOTICON DICTIONARY

Sentiment	Emoticon
Positive	:-) :) :o) :] :3 :c) :> =] 8) =) :} :^) :-d :d 8-d 8d x-d xd X-d xd =-d =d =-3 =3 b^d :-)) ^_^ (^_^)/ (^^)/ (^o^) >^_^< (^.^) (^_^.) (^_^) (^^) ;;) ^_^ (#^.^#) (y) ;d :* :") (*^0^*) ;) :') !(^^)! (*^^)v (^^)v (^_^)v
Negative	>:[:-(:(:-c :c :-< -_,- :< :-[:[:{ :-|| :@ qq (>_<) (>_<)> -____- </3.___. :]x \m/ :| (/_;) (t_t) (;_;) (;_:) (;o;) (:_;) (tot) -___- (-.-) (-_-) x_x --" -__- :/ (*_*) (*_*; (+_+) (@_@) :-/ -_- (~o~) (~_~) :-| :'(._.
Neutral	:o :p

Positive Sentiment	Negative Sentiment
pilih	tidak-pilih
dukung	tidak-dukung
coblos	tidak-coblos
vote	jangan-vote
	jangan-dukung
	jangan-coblos

TABLE 3.
CLASSIFICATION RESULT

Tweet	Manual	System
@freddy_ggmu: #Perfect !!!! RT @kompasiana: POLITIK:Prabowo-Jokowi Capres-Cawapres Pemilu 2014? http://t.co/S6GQQHR7W5	Positive	Positive
@NorselMaranden:@PartaiSocmed: Mengenal Sosok Prabowo Subianto Sang Jenderal Terbuang - Bag. 2 @PartaiSocmed http://t.co/EVRdxzYLb8 cc @Vote_Prabowo	Positive	Negative
@Ferdylonge: Pesen gw aja, kalo dukung jangan membabi buta. Open minded. @LoveNiey: Oya? Kata macan? Cape dee @detanamaku: Nah RT Ga tau jokowi korup	Negative	Positive
@TitoHuberto: ni apaan sih mau pilpres 2004 baru deh iklan Gita Wirjawan di segala Web. ga bakalan gua pilih demokrat lagi. butuh rakyat pas pilpres doing	Negative	Negative

TABLE 4.
CONFUSION MATRIX

Sentiment	Classification Result		
	Neutral	Positive	Negative
Neutral	**0** (TNet)	**2** (FP)	**0** (FNeg)
Positive	**2** (FNet)	**408** (TP)	**3** (FNeg)
Negative	**1** (FNet)	**132** (FP)	**52** (TNeg)

TABLE 5.
PROBABILITY OF BOTH ACCOUNT

Tokoh	Number of follower Account		Probability of Account	
	Lover	Hater	Lover	Hater
Aburizal B.	3567	4	0,999	0,001
Dahlan Iskan	31119	13400	0,699	0,301
Gita Wirjawan	3879	0	1	0
Joko Widodo	26667	55	0,998	0,002
Prabowo S.	4129	398	0,912	0,088
Rhoma Irama	1201	29	0,976	0,024

TABLE 6.
RESPOND PROBABILITY

Tokoh	Probability Public Tweet		Probability Account		Mean	
	(+)	(-)	(+)	(-)	(+)	(-)
Aburizal B.	0.878	0.122	0.999	0.001	0.938	0.062
Dahlan Iskan	0.930	0.070	0.699	0.301	0.815	0.185
Gita Wirjawan	0.820	0.180	1.000	0.000	0.910	0.090
Joko Widodo	1.000	0.000	0.998	0.002	0.999	0.001
Prabowo S.	0.960	0.040	0.912	0.088	0.936	0.064
Rhoma Irama	0.859	0.141	0.976	0.024	0.918	0.082

TABLE 7.
ACCURATION RESULTS

Tokoh	Real			Detected		
	(+)	N	(-)	(+)	N	(-)
Aburizal B.	61	0	39	86	2	12
Dahlan Iskan	92	0	8	93	0	7
Gita Wirjawan	30	0	70	82	0	18
Joko Widodo	89	0	11	100	0	0
Prabowo S.	73	2	25	96	0	4
Rhoma Irama	68	0	32	85	1	14