Tweet-Based Election Prediction

Master's Thesis, 17 December 2014

Nugroho Dwi Prasetyo

<<Page left blank intentionally>>

Tweet-Based Election Prediction

THESIS

Submitted in the partial fulfillment of
The requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE
TRACK INFORMATION ARCHITECTURE

by

Nugroho Dwi Prasetyo
Born in Jakarta, Indonesia

Web Information Systems
Department of Software Technology
Faculty, EEMCS, Delft University of Technology
Delft, the Netherlands
http://wis.ewi.tudelft.nl

Tweet-Based Election Prediction

Author: Nugroho Dwi Prasetyo
Student ID: 4256786
Email: nugrohodwiprasetyo@student.tudelft.nl

Abstract

Twitter is a microblogging service that has more than 500 million messages on a daily basis. Scholars has been utilizing Twitter to monitor people reactions in political activities, such as debates and campaigns. By doing so, some of them claim that a forecast or prediction to an election can be made. Using the data from 2014 Indonesia Presidential Election, we calculate predictions with many different parameters. Our analysis of the prediction results shows the importance of a proper data collection method, removing spam, incorporating sentiment detection to the tweets, and performing data normalization using demographic information. Although looks very promising, our results show that result prediction is not applicable to any election. Dividing the data into 33 provinces, the data suggests that applying the methodology to provinces with a small dataset leads into inaccurate predictions.

Keywords: Twitter, electoral prediction, sentiment analysis, demographic.

Graduation Committee:
Prof. dr. ir. G.J. Houben, Faculty EEMCS[1], TU Delft
Dr. ir. C. Hauff, Faculty EEMCS, TU Delft
Dr. ir. A. Zaidman, Faculty EEMCS, TU Delft

[1] Electrical Engineering, Mathematics and Computer Science

Preface

This thesis report describes my final work at the Information Architecture track of the Computer Science Master Program, TU Delft. I started my study in August 2012 with full scholarship support from the Indonesian Ministry of Communication and Information. The courses in the Information Architecture track are very broad and challenging, from low level studies such as distributed computing to the high level concept such as multi actor system design. This creates a wide range of options for the final project and opportunity to incorporate many knowledge domains to provide a solutions to a real world problem.

Following courses from the web information system group made me realize the growth and direction of the web and how to utilize its potential. Because of that, for my final work, I chose to conduct a research about one important domain of the World Wide Web, the social media. Receiving the scholarship from the government, I tried to use the power of social media to solve one of the problems happening in my country, the inaccuracy of many election polling results. Although similar studies had been conducted since few years ago, many aspects are not yet clear. My study summarized those previous works, implemented in the 2014 presidential election, and hopefully give guidance to future work in this domain. I hope you enjoy reading my thesis.

First of all, I would like to thank the Indonesian Ministry of Communication and Information, especially the Human resource development team who gave me the opportunity to study at TU Delft University. Then, I really want to take this chance to give a sincere thank my supervisor dr. Claudia Hauff whom I think is very helpful, patience, and detail when giving a guidance and suggestions to my work. I would like to thank prof. Geert-Jan Houben for the discussion and the direction of the topic of my thesis, to dr. Andy Zaidman as one of the committee of my thesis, and to dr. Bozzon and Omicron group for their input at the meetings. I would also like to thank all of my friends and family. I could not have finish my study without your love and support.

All Praise and Thanks to the Almighty and All Caring for His Guidance.

Contents

List of Figures

List of Tables

1 Introduction

Election is a very important part in the democracy. It is the main instrument of democracy where the citizens communicate with the representatives. One important element in an election is the election polls/survey. (Lewis-Beck, 2005) stated that the main purpose of an election survey is to provide information to the curious citizens and also to interested parties so that they can make adjustments they feel necessary. For example, the campaign strategy can be changed based on the result of the polls.

Opinion polls has existed since the early 19th century, based on (Hillygus, 2011). And currently, there are many scientifically proven statistical models to forecast an election, as shown in (Lewis-Beck, 2005). But sometimes, even in the developed countries, the polls failed to accurately predict the election outcomes. (Fumagalli, 2011) listed several failed polls result such as in the 1992 British General Elections, the 1998 Quebec Elections, the 2002 and 2007 French presidential elections, the 2004 European elections in Portugal, the 2006 Italian General Elections, and the 2008 Primary Elections in the States. In developing countries such as Indonesia, this phenomenon is still happening, our records show that most of the polls in the 2012 Jakarta (Indonesia's capital city/province) governor election, the 2013 Bandung (West Java capital city) major election, and 2014 General elections, failed to predict the winner or have a large gap between the forecasts and the election results. [2]

Researchers suggested several explanations to this issue. For example, (Arzheimer, 2014) explain that in every democratic countries, there are some pollsters that have a reputation for 'leaning' towards one party or even one political camp, resulting in lower overall accuracy. They called this as the 'house effect'. Based on their study, it can be caused by non-random sampling, sampling over a very short period of time, or post-stratification strategies (adjustment raw polling data). Especially in Indonesia, the 'house effect' becomes more apparent as explained by (Trihartono A. , 2014) that political actors are more interested in exploiting the polls as a political weapon for the sake of political victory rather than hearing the public's sentiments. They exploit polls as a device for obtaining a political vehicle from political parties and to invite bandwagon effect (polling outcome positively influences voters/mass media/businessman tendencies to support the candidate who, according to the polling, has the greatest chance of winning). A solution was offered by (Fumagalli, 2011) who proposed to use statistical matching and weighting procedures to cope with non-random sampling or to accommodate the less representative objects in the sample.

Another problem in developing countries such as Indonesia is corruption. In (Fauzi, 2014), the ex-Minister of Internal Affairs explained that his study proves that direct local elections

[2] We discuss this phenomenon in more detail in Section 2.1.3

have significant effects on corruption committed by the heads of district government. The high cost to win the election needs to be 'returned' even by means of corruption. One part of that high cost is for electoral polls. Each campaign requires many polls to be conducted in order to devise or modify the campaign strategy. According to AROPI (Indonesian Association of Public Opinion Research)[3], in Indonesia, a city level election polling costs between 10,000 - 15,000 US dollars and a province level election polling costs about 50,000 US dollars. To overcome the cost issue, several methods such as web-based polls (forum/survey), SMS polls, and telephone polls were developed. For example, (Down, 2003) developed a method for SMS polls and found that the main challenges for it are the representativeness of the sample, the sample selection, and the response rates. Especially in Indonesia, (Trihartono A. , 2013) studied that, though SMS and web-based polls have the advantages such as, fast, timely, low cost per response, and interesting news value, but they have not been reliable instruments to reflect the voice of the people. He argued that it was caused by the lack of proper methodology and the problem of representativeness.

Trying to resolve the accuracy and high cost issues, we study the possibility of using data from social media as the data source to predict the outcome of an election. Social media has become the most popular communication tool on the internet. Hundreds of millions of messages are being posted every day in the popular social media sites such as Twitter[4] and Facebook[5]. (Pak, 2010) stated in their paper that social media websites become valuable sources for opinion mining because people post everything, from the details of their daily life, such as the products and services they use, to opinions about current issues such as their political and religious views. The social media providers enable the users to express their feelings or opinions as much as possible to increase the interaction between the users and their sites. This means that the trend on the internet is shifting from the quality and lengthy blog posts to much more numerous short posts that are posted by a lot of people. This trait is very valuable as now we can collect different kind of people's opinions or sentiments from the social web.

One of the social media that allows researchers to use their data is Twitter. Twitter is a microblogging web service that was launched in 2006. Now, it has more than 200 million visitors on a monthly basis and 500 million messages daily[6]. The user of twitter can post a message (tweet) up to 140 characters. The message is then displayed at his/her personal page (timeline). Originally, tweets were intended to post status updates of the user, but these days, tweets can be about every imaginable topic. Based on the research in (Dann, 2010), rather than posting about the user's current status, conversation and endorsement of a content are more popular. The advantages of using tweets as a data source are as follows; first, the number of tweets is very huge and they are available to the public. Second, tweets contain the opinion of people including their political view. In Section 2.2.2, we

[3] http://aropi.or.id/ & http://m.inilah.com/read/detail/64553/banyak-lembaga-survei-nakal
[4] https://twitter.com/
[5] https://www.facebook.com/
[6] https://about.twitter.com/company

discuss the use of Twitter in political domain such as for monitoring the reactions of people at a political activity and for forecasting an election.

The research about predicting an election using Twitter was started by (O'Connor, 2010) and many has conducted similar experiment since then. In most of the experiments, the winners were predicted correctly with low error compared to the election result. But there are still many issues regarding this topic. First, the researchers were using different methods in their experiments, from the way they selected the data until the calculation of the prediction. Second, applying the same prediction method, the accuracy of the prediction can vary widely when applied in different elections. Besides those issues, (Gayo Avello, 2011) pointed out several other issues such as, large data does not make such collections statistically representative samples of the overall population. Second, not all tweets are trustworthy, there are many spam tweets and campaign tweets that do not represent the sentiments or opinions of the users. Third, several research, for example (Metaxas, 2011) and (Gayo-Avello D. M., 2011), showed that simple lexicon-based sentiment analysis is not suitable for the complex political tweets. They also suggested that we should carefully evaluate positive reports before assuming that the methods are applicable to any similar scenario. For example, (Tumasjan, 2010), predicting the election result of 6 main parties in Germany, claimed that the number of mention in the tweets correlate with the share of vote in the election. But (Jungherr, 2012) showed that it was not the case when all parties was included in the calculation. The Pirate Party ranked last in the election, but had the most mentions in their data.

In this thesis, we will try to address those issues by, first, summarizing previous research to understand the current progress on this topic. Then predict the election result using many different methods that were conducted previously by other researchers to understand the effects and differences between each method.

1.1 Research Objectives and Questions

The goal of predicting an election result using Twitter is to create an alternative to current polls with lower cost but similar accuracy, and reliability. Comparing to the result of offline polls, most of the results form Twitter-based prediction are still lacking behind in term of accuracy. Only one research, (Ceron A. C., 2014), claimed that their result outperform the result of offline polls. They explained that their results in the US presidential election are better in 8 states, has same result in 2 states, and worse in 2 states. Other results, though predicted the winner correctly, still have bigger Mean Absolute Error (average difference between prediction and the election result in vote share of each candidate) than the polls. For example, in Germany general election, (Tumasjan, 2010) prediction had an MAE of 1.65% while the pollster had an MAE of 0.84% to 1.28%. (Sanders, 2013) prediction in the Dutch Senate Election got an MAE of 2.4% while the pollster had 1.1% of MAE.

While in several elections (as stated in the previous section) offline polls failed to correctly predict the election result, in most cases polls in developed countries are reliable and

accurate. (Beauchamp, 2013) and (O'Connor, 2010) even used the poll results as the ground truth to be compared with their prediction. This is different with Indonesia who has many pollster and based on our data, the results of those polls can differ greatly.[7] Tweet based prediction has a great opportunity to be an alternative method to do opinion poll. This leads to our first research question as follows:

Research Question 1: How effective is the tweet based election prediction compared to Indonesian offline polls?

From the previous research, we understand that researchers employed different methods to calculate the prediction. Some of these works rely on very simple techniques, for example (Tumasjan, 2010) argued that the relative number of tweets mentioning each party's name is a good predictor for its vote share. In other research, (O'Connor, 2010) only used positive tweet, detected by lexicon-based sentiment analysis, to predict the election result. Another example of the difference in the method was in the keyword used when selecting the data. Most researchers used the candidates' name/popular name as the keywords in the data selection, but (Nooralahzadeh, 2013) used more complete keywords by adding the campaign and election hashtags.

There are also some suggestions on how to improve the prediction, for example, examine the trustworthiness of each tweet to detect spam and removing tweet from non-personal (company or institution) users as explained in (Waugh, 2013) and (Makazhanov, 2014). To handle the demographic bias of Twitter users, the methods used by researchers also different between researchers, (Sang, 2012) tried to reduce this bias by using the number of users rather than the number of the tweets while (Choy M. C., 2011) used census correction in their model.

With these differences, it is very important to understand the effect choosing each variable in the prediction model. This leads up to the following research questions:

Research Question 2: What are the most important factors that influence the result of predicting an election based on Twitter data?

Research Question 3: What is the difference in utilizing different parameters when collecting Twitter data?

Research Question 4: Can the accuracy of Twitter-based election prediction be increased by incorporating the users' demographic information and the tweets' sentiment information?

1.2 Research Approach

This research is divided into three phases as seen in Figure 1. The first phase of this research is building a conceptual model to predict an election result using Twitter data. As

[7] Discussed in more detail in Section 2.1.3

the basis for building the model and implementing it, we will conduct a thorough literature survey related to this domain, explained in Chapter 2. The model itself is explained in detail in Section 3.1 and Section 3.2. Based on the model, empirical hypotheses about the assumptions made when building the model will be stated in Chapter 3.

As a use case, the prediction model is implemented on the 2014 Indonesia's presidential election. That election is selected because of several reasons; (1) the needs of lower cost and better accuracy of election prediction; (2) Large number of Twitter user in Indonesia; and (3) Interesting demographic distribution and social media penetration. The application in the 2014 Indonesian presidential election is described in Chapter 4.

To test the hypotheses, we compared the prediction result with the real election result, traditional polls, and previous experiment result described in Section 5.1 and Section 5.2. The prediction results then will be analyzed to show whether the hypotheses hold or not. The analyses of all hypotheses are explained in Section 5.3 and Section 5.4.

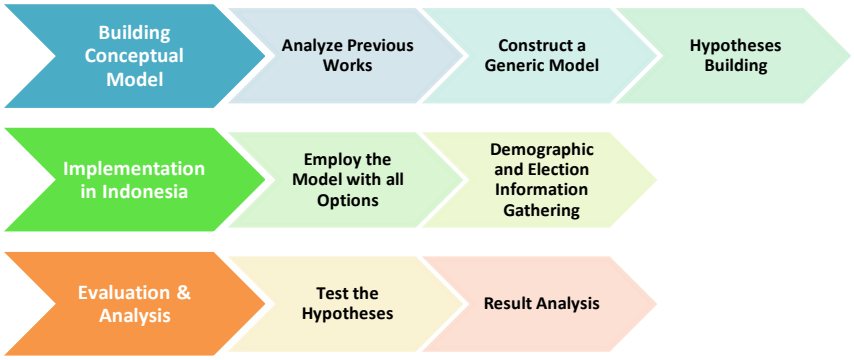

Figure 1 Research Approach Diagram

1.3 Scientific Relevance

This research fits in with the current on-going discussions about the "predictive power of Twitter" literature. In a recent paper, (Gayo-Avello D. , 2013) argued that "while simple approaches are assumed to be good enough, core problems are not addressed." This is precisely the research gap we are hoping to fill. The added value of our work will be incorporating data filtering, sentiment analysis, users' demography (age, gender, and location) in our methodology. We also try to investigate a correlation between the number of data used and the result of the prediction.

1.4 Main Contributions

We understand that even though this topic has received attention few years ago, it is still in the development stage. There are still many things to be improved. In this thesis, rather

than add one more prediction model or one more experiment, we want to analyze the differences between methods that were developed by previous researchers. We will carefully compare and observe the application of different prediction model. In summary, the contribution from our research for this topic will be:

- An analysis of different methods for each step in the prediction model. For example, the difference between using only names and names plus popular hashtags for the keyword selection.
- Implement a prediction using many steps such as data filtering, un-biasing the sample, and sentiment detection.

1.5 Thesis Outline

In the next chapter, a review of the current literature will be provided. The chapter will start with an illustration of conventional/offline election prediction, then shift to Twitter and research about politics in Twitter. After that, it will focus on Twitter-based election prediction and ended by recognizing a research gap.

In the third chapter, the methodology to predict an election will be developed by combining previous methodology in the literature and comparing it to the offline predictions. The methodology that will be used to carry out the research and also the empirical hypotheses will be presented. The fourth chapter will explain about the implementation of this model in an Indonesian election and the issues involved in it will be presented.

This leads to the fifth chapter where we will have finally reached the stage of result interpretation and analysis. The thesis concludes with chapter six, then reviews the contributions and limitations of the work, and provides suggestions for further research.

2 A Review of Literature

This chapter contains background information that are helpful for this research. First, we introduce the basic concept of opinion polls. We will discuss a little bit about the history and related literatures about election in Indonesia in sub-Section 2.1.3 because we use the election and polls result from Indonesia. Then, the next section describes a basic information about Twitter. Related to our topic, the demography and politic in Twitter will have more focus in this section. Next, the related works about Twitter-based election prediction are discussed. In the last section, the literatures about extracting information from Twitter, either the information of the users or the sentiment from the tweets, are described.

2.1 Opinion Polling in the Elections

In this section, we will discuss about opinion polling in an election before researchers start using social media as the data source. Opinion polling itself, based on (Hillygus, 2011), was first used to predict the result of the US presidential election on 1824, by taking informal trial heat tallies in scattered offices, and public meetings. But the scientific method was started in the beginning of 1940s when Gallup Poll first released a presidential pre-election survey polling. As summarized by (Lewis-Beck, 2005), the study of statistical model for pre-election prediction started to appear around 1980 in the US, UK and France.

In general, a polling is conducted by asking questions/questionnaire to random respondents and the general steps to do it are: (1) developing questions to be used in measuring opinion; (2) selecting probability samples that can accurately represent a population; (3) Collecting data by interviewing respondents; (4) Performing statistical analyses using standard principle and procedure; (5) Interpreting and reporting the results. Relating to the topic of this thesis, next we will discuss about the types, methods, calculation and evaluation of election polls.

2.1.1 Election Poll's Types & Methods

One way to classify the election polls is based on the timing of the poll. A poll conducted at the beginning of a campaign is called benchmark poll. It is also quite often that a poll is conducted before the candidate announce their candidacy. This polls can give the candidate the information about their standing/situation before campaigning, so that they can spend most of their resource in the most effective manner. After the candidates are announced, repeated polls in a fixed interval, for example: weekly polls. These polls are called tracking polls. In a competitive race, there are also polls taken between regardless of the time interval to find out the result of a particular message or technique in a campaign. On the election day, there are other polls such as entrance poll and exit poll (taken before and after the voters cast their vote). These polls and quick count are important in a manual count

voting but less significant in an e-voting because the official result of an e-voting can be published immediately on the same day as the election while the official result of a manual voting is published 1-4 weeks after the election, based on the condition of the country.

Classifying based on how the polls are conducted, (Trihartono A. , 2013) divided polls based on the pollster into: public opinion institution and media polling. The first category use face to face interview, though there are differences in the sample selection and question building. The second category conduct the polling through the telephone, SMS and website. Main advantages of the second category is the lower cost and shorter time needed to conduct the polling. Face to face interview uses higher cost and longer time because it needs to send trained interviewers across the country (in the case of Indonesia, there are geographical difficulties), pay elements, such as honoraria, transports expenses, accommodation cost and insurance, and give gifts to the interviewee. But the second category has problem of less representativeness. For example, only a portion of people has a phone, let alone the internet and most of those who use the technology, live in the urban area. The comparison between these polls method (in Indonesia) can be seen in Table 1.

	Means	Price	Length
Mass Media	SMS	< USD 1,000	1-3 days
	Telephone	< USD 10,000	3-5 days
	Online/web	< USD 1,000	1-5 days
Face to Face Interview	National	USD 30,000 – 50,000	10-21 days
	Province	USD 20,000 – 30,000	3-14 days
	City	USD 5,000 – 15,000	3-14 days

Table 1 Price and time needed for offline polls. * (Trihartono A. , 2013)

2.1.2 Calculation and Evaluation of election polls

In (Lewis-Beck, 2005), the author proposed four criteria to evaluate a forecasting model: accuracy, lead, parsimony and reproducibility. While in his evaluation model, accuracy has more weight (3 times) than other factors, he also thought that the farther in advance a model produces accurate forecasts, the better (lead). In term of variables in the model, a few well-specified variables will work better than many questionable ones (parsimony). Lastly, he argued that parsimonious models are easier to understand, and therefore easier to reproduce.

As the most important factor in an opinion poll, we will focus on the accuracy of a poll. Polls, which based on samples of the whole population, are subject to an error that reflect the uncertainty in the sampling process. The error can be caused by coverage bias, response bias, or non-response bias. Coverage bias is happened when the samples are not representative to the population. It could be caused by the methodology used or the sample selected is not entirely random. Non-response bias is also a representativeness problem, but

it is caused by the people that do not want to answer in the interview. Response bias is when the interviewer or the poll's institution fail to understand the answer given by the respondents. This is a result of word selection or question's order in an interview.

Considering those factors, based on the probabilistic theory, the sampling error or usually called the "margin of error" is calculated using Equation 1. Where p is the sample proportion, n is the sample size, and z is the appropriate critical value for the desired level of confidence. For an example, in a poll using a random sample of 1000 people with 95% confidence interval, the margin of error will be about 3%. It means, if this procedure is conducted many times, 95% of the time, the result will be estimated plus or minus 3%. To reduce the margin of error into 1%, we need to increase the number of sample into 10,000 of people.

$$Margin\ of\ Error = z\sqrt{\frac{p(1-p)}{n}}$$

Equation 1 Margin of error in an opinion poll

In several countries, there are several factors that can improve the result from a poll. For example, in the US, based on (Lewis-Beck, 2005), combining the poll result with economic growth can increase the predictor accuracy. He argued that including economic in the prediction model improve the prediction performance. For example, the predicted vote share of a candidate is created from a sum between the candidate popularity and the percentage GDP growth. In different countries, the variable can be different. In the UK, the author argued that the inflation six months prior and party vote share are the most significant variables. While in France, he argued that the important factor to improve the prediction result is the unemployment rate.

2.1.3 Election polling in Indonesia

In Indonesia, opinion polls and survey institutions begin to emerge in 1997, after the downfall of the New Order (the regime of a president who ruled for 32 years). According to the Indonesian Association for Public Opinion Research (AROPI) and the Indonesian Association of Public Opinion Survey (PERSEPI), the number of pollsters increases from 6 in 1998 to more than 60 in 2008. Based on (Trihartono A. , 2013), the pre-election polling activities in those years were tested by the public, openly examined by the mass media and are critically discussed in academic circles. This made those institutions became accepted and trusted as an important player in the elections. In the 2014 presidential election, there are 48 institutions that are legally allowed to do a survey/polls.

Based on our data, low accuracy became a huge problem in Indonesia's main elections. For example, we can see that in Table 2, polls results for the 2012 Jakarta (Indonesia capital city/province) governor election that were published to public, most of the pollster failed to correctly predict the winner. In table 3, the predictions for the 2014 Indonesia general

election also did not have good accuracy. From 7 pollsters that publish their results publicly, 6 of them were correctly predict the party that has the biggest vote share. But only 3 who were able to predict the second or third winner correctly. In the case of 2014 presidential election between 2 candidates, we can see another anomaly. Several pollster results were always in favor of a candidate while there were also other pollster who favor the other candidate.[8] Even the quick count results (after the election) differ, 6 institutions claimed that candidate no 1 wins, while 7 institutions claimed otherwise.

Candidate	Poll Results			Election Result
	LSI	Indo Barometer	Puskaptis	
Fauzi	56.5%	44.6%	59.2%	34.1%
Hendarji	1.6%	1.7%	2.9%	2.0%
Jokowi	18.6%	21.8%	18.9%	42.6%
Hidayat	10.7%	22.5%	12.9%	11.7%
Faisal	7.5%	5.0%	4.0%	5.0%
Alex	5.0%	4.4%	2.0%	4.7%
Predict the Winner Correctly	No	No	No	
MAE	8.5%	7.1%	9.1%	

Table 2 Pre-election polls result for the 1st round of 2012 Jakarta governor election

Political Party	LSI	LSIN	PolTracking	Kompas	Charta Politica	INES	CSIS	Election Result
Nasdem	2.5%	4.1%	3.0%	8.1%	3.0%	6.9%	3.9%	6.7%
PKB	5.7%	5.1%	6.6%	6.0%	8.4%	2.6%	8.1%	9.0%
PKS	5.2%	6.9%	4.1%	2.7%	3.7%	2.1%	4.1%	6.8%
PDIP	23.3%	19.6%	26.4%	25.7%	24.8%	26.7%	24.2%	19.0%
Golkar	25.4%	18.4%	24.1%	19.5%	19.2%	14.8%	19.0%	14.8%
Gerindra	8.2%	11.4%	9.4%	13.6%	14.0%	26.6%	13.6%	11.8%
Demokrat	12.2%	14.9%	12.6%	8.5%	9.4%	4.3%	7.0%	10.2%
PAN	6.5%	5.7%	2.9%	3.8%	5.3%	2.6%	5.8%	7.6%
PPP	5.7%	4.7%	4.9%	2.8%	6.0%	3.6%	4.2%	6.5%
Hanura	4.2%	4.9%	5.0%	7.8%	5.6%	7.5%	8.1%	5.3%
PBB	0.7%	2.8%	1.0%	1.3%	0.5%	1.2%	1.6%	1.5%
PKPI	0.4%	1.6%	0.1%	0.1%	0.1%	1.1%	0.6%	0.9%
Predict Top 3 Winner Correctly	No	No	No	Yes	Yes	No	Yes	
MAE	2.82%	1.87%	3.19%	2.87%	2.14%	4.20%	2.37%	

Table 3 Pre-election polls result for 2014 Indonesia general election

Above phenomenon had been studied in (Trihartono A. , 2014) that concluded that polling has been used beyond capturing the voice of the people. Political actors in Indonesia's elections have been more interested in the short-term exploitation of polling solely to win

[8] Described in more detail in Section 5.1

elections. First, political actors have exploited polling as a tool for gaining a political vehicle. Showing high popularity from polls, they asked for supports from political parties to be selected as candidates for elections. Second, they used poll result as a map for soliciting bribes, as a map for guiding the mobilization of supporters. Third, polls are used for inviting a bandwagon effect. A bandwagon effect occurs when the poll prompts voters to back the candidate shown to be winning in the poll and the underdog occurs when people vote, out of sympathy, for the candidate that has a very low vote share. The first evidence about the effect was from (Simon, 1954) that showed that there are changes on the vote share if the voter saw the poll results before the election.

2.2 Twitter

Twitter is an online social media micro-blogging service where its user can post a short messages (maximum of 140 characters) called tweets. It has more than 250 millions of users monthly[9] and is still growing. Its popularity attracts not only the attention of people as the user but also the researchers. (Zimmer, 2014) summarized the research related to Twitter and showed that there are many research topics such as politic, consumer behavior, marketing, tracking live events, finance, et cetera. In this section, we will discuss several aspects of Twitter related to our topic such as the demography and politics in Twitter.

2.2.1 Demography of Twitter

Twitter has a very great potential as a data source with the huge number of tweet/user and the content inside each tweet. Based on (Mislove, 2011), more than 91% of its users makes their profile and tweets available to public, thus allowing researchers to access them. To be used for an election prediction, first we need to understand the representativeness of Twitter user towards the overall population, because, based on (Bakker, 2011), socioeconomic traits of social media users do not exactly match the actual demographics of the whole population. People on social media are generally younger, more highly educated, concentrated in urban areas, as well as more politically active.

An example of study about the demography of Twitter was shown in (Mislove, 2011). The authors compared the US Twitter users and the US population based on three information: geographic, gender, and ethnic. In geographic, they found that in some counties, Twitter users are less than 0.01% of the population, while in others are more than 10% of the population. They also found a strong bias towards male users where about 72% of the users are male. For ethnicity, they divided the user into four: Caucasian, Hispanic, African-American, and Asian. In their data, African-American and Caucasian have an oversampling problem while Hispanic and Asian have an under sampling problem. These results are in line with the Pew Research Center results[10]. Furthermore, Pew Research Center results

[9] https://about.twitter.com/company
[10] http://www.pewinternet.org/2013/12/30/demographics-of-key-social-networking-platforms/

showed that the demographic bias from education and income is low, while looking from age and urbanity information, Twitter does have representativeness issues.

Indonesia has a different demography than the US. First, the internet penetration in Indonesia is still very low, only about 24% of the total population, based on the research from On Device[11]. But most of the internet users are engaged in the social media. On the same research, it was known that about 80% of internet users joined the social media. Unfortunately, Indonesia also have issue with sampling because most of the users live in big cities such as Jakarta, Bandung, and Jogjakarta.

2.2.2 Politics in Twitter

In the political domain, there are several research areas regarding the use of social media. First is to understand the role of social media in an election. With the success of the campaign of Obama in the US presidential election that utilized social media, election candidates in many countries tried to make use of social media for their campaigns. This encourages researchers to study the implications of social media in politics. An example is shown in (Morozov, 2009) where the authors studied an event when the USA tried to start a revolution in Iran via twitter. (Burgess., 2011) studied API data from twitter under the #ausvotes hashtags in the 2010 Australian election to described the key patterns of activity in social media. These studies showed that public discussion is able to be constructed when a handful of users coming to a consensus on a shared hashtag to include in their tweets.

The second area involves monitoring reactionary content in social media during a political event such as a speech or a debate. For example, (Goggins, 2012) examined Twitter messages during a "Republican Primary Debate" in November 2011 and (Shamma, 2010) characterized the 2008 US presidential debate in terms of Twitter sentiment. These studies show that Twitter and its features such as retweet, reply and hashtags are an effective source of data for identifying important topics.

The third area is result forecasting. An example of this is predicting election results. There have been many researchers who have predicted election results in several countries. For example, (Tumasjan, 2010) found that during the German federal election in 2009, the share of volume on Twitter accurately reflected the distribution of the real votes in the election between the six main parties. Another prediction for the Dutch senate election in (Sang, 2012) produced a good result via volume counting and normalization. (Bermingham, 2011) tried to improve the prediction model in the Irish general election using sentiment analysis but the result showed that the prediction is still not competitive with the traditional polling methods.

[11] https://ondeviceresearch.com/blog/indonesia-social-media-capital-world

2.3 Election Prediction using Twitter

In this section, we go into more detail discussing related works about predicting the result of an election using Twitter. We noticed that researchers use a different approach regarding this problem. There are researchers who try to discover the political or ideology preference of a user, then relate it to the election and there are others who use selected tweet related to the upcoming election and figure out vote preference of the user using that data.

Different strategies such as profile information, user behavior, user graph, Twitter specific feature (reply/re-tweet), and sentiment from tweet content can be used for inferring political leaning. For example, In (Wong F. M., 2013), the authors used tweet containing parties' name in several political events to assign a political/ideological leaning of the user who posted the tweets. Similar to the previous method, (Boutet, 2012) used the tweets and retweets of a user regarding a political party to infer the political leaning. (Golbeck, 2011) assigned a score to every congress member which a Twitter user is following, then a political preference is assigned based on that score. In (Pennacchiotti, 2011), the authors compared several features such as user's bio and avatar, posting behavior, linguistic content, follower, reply and retweet. They found out that the combination between user profile and linguistic outperform other feature. They then applied to classify the ethnicity of the user and whether the user is a Starbucks fan, but their result showed that information from user bio is more accurate for classifying Starbucks fan, and user's avatar for classifying user's ethnic.

The second approach is by using selected data just days or weeks prior to the election. The prediction could be derived by comparing the number of tweets mentioning each candidate or by comparing the number of tweets that has positive sentiments towards each candidate. The earliest research stated that the number of tweets mentioning a party reflects the election result was shown in (Tumasjan, 2010) where they found out that the prediction result from Twitter were only slightly worse than offline election polls. While (O'Connor, 2010) is the first research in which argued that sentiment detection approach from Twitter can replace the expensive and time intensive polling.

Researches have tried to compare these two methods, for example, (Gayo-Avello D. M., 2011) that tried to predict congress and senate election in several states of the US. They showed that though the method is the same, the prediction error can vary greatly. The research also showed that lexicon based sentiment analysis improve the prediction result, but the improvement also vary in different states. Same result was shown in (Bermingham, 2011) where they predict the result of Irish general election using both methods and (Ceron A. C., 2014) which predict the Italian primary election. All of the research showed that sentiment detection do reduce the error of the prediction result. Because of that, several researchers focused on improving the sentiment analysis, such as (Ceron A. C., 2014) and (Ceron A. C.,

2013) who used more sophisticated sentiment analysis than lexicon based in the US presidential election, France legislative election, and Italy primary election.

Other than using sentiment analysis, the prediction result from Twitter can be improved by using user normalization. This is based on the fact that in an election, one person only have one vote. (Sang, 2012) implemented this method and showed that the prediction result of 2011 Dutch senate election was improved. (Choy M. C., 2012) takes further step by adding census correction on the user normalization. (Gaurav, 2013) also implemented this method in several south American countries. He collected more than 400 million of tweets, and got a very good result (low difference with the election result) predicting Venezuela presidential election. But when applying in Ecuador and Paraguay presidential election that has much less dataset, the error of the prediction increase significantly.

Other methods proposed by researchers are by (1) utilizing interaction information between potential voter and the candidates and (2) creating trend line from the changes in follower of the candidates. (Makazhanov, 2014) used interaction information such as the number of interaction, the frequency of interaction, the number of positive and negative terms in the interactions in the Canadian legislative election. The candidates were grouped into four parties, and based on their result, they argued that that the generated content and the behavior of users during the campaign contain useful knowledge that can be used for predicting the user's preference. (Cameron, 2013) tried to utilize the size of candidates' network (follower in Twitter and friend in Facebook), but the result showed that it was not a good predictor of election results. One interesting result from their research is that despite the huge size of social media, it has small effect on the election results. Therefore, it only make a difference in a closely contested elections.

However, there are several researchers arguing that research in this area is still premature and requires a lot of development before it can give satisfying prediction result. (Jungherr, 2012) argued that prediction model using Twitter only able to predict the result from the top candidates/parties and slight variable changes in the model did impact the prediction result. In (Gayo-Avello D. , 2012), the authors listed several drawback of the research in this topic such as, most predictions are actually a post-hoc analysis, no commonly accepted way exists for "counting votes", the sentiment analysis methods are not reliable, no data cleansing step, demography and self-selection bias has not been addressed. In (Gayo-Avello D. , 2013), in addition to previously stated drawbacks, gave several suggestions such as the importance of geographical and demographical bias, the noise in the social media, the reproducibility of proposed methods, and MAE should be use rather than only winner prediction.

2.4 Extracting Information from Twitter

From previous sections, it was explained that implementing sentiment analysis to the content of the tweet can reduce the error of the prediction and one of the major issue of the current prediction model is the representativeness of the sample. Because of that, in this

section, related works of sentiment analysis for Twitter and also related works about detecting demographic information of the Twitter users will be discussed.

2.4.1 Detecting Age and Gender of a Twitter User

Before the era of social media, researchers has developed a method for detecting demographic information from a text. For example, (Boulis, 2005) was the first to propose a lexical model for identifying gender from telephone conversations. On Twitter, (Rao, 2012) was the first to study about classifying the users based on their attributes. The authors develop several methods to detect the users' age, gender, ethnic/origin, and political orientation. They use manual annotation as the ground truth of those attributes. Their study showed that social network element in Twitter such as the ratio of followers and followees, the number of follower, and the number of followees have no relation with the gender and age of the user. There are also no significant difference from tweet, retweet, and reply frequencies of the users between male and female users. For the content of the tweet itself, they applied two models, the first is socio-linguistic model that classify the user based on their lexical choices, and the second is n-gram model where they use bigram from the text as the feature for the classification. The accuracy of those two model do not differ very much for age, gender, and origin classifications but for political orientation, the n-gram outperform the other model. The main challenges in the model are the informal nature of Twitter and the limited size of the tweet.

A different approach can be seen in (Burger, 2011) where the authors compared the classification results from machine learning and crowdsourcing. The ground truth for this experiment was obtained from the users who put their gender information on their blog bio. Their results showed that using trigrams from combination of screen name, full name, description, and tweets as the feature in the machine learning perform much better than crowd sourcing. Only 5% of the humans performed classifications with higher accuracy than the machine. Another approach was shown in (Nguyen, 2013) where the authors classified the users based on their age and life stages. From their research, they concluded that a simple system using only unigram features can already achieve high performance. The result also showed that detecting age regression rather than age group was difficult.

2.4.2 Sentiment Analysis

Sentiment analysis, also called opinion mining, is a type of language processing that examine opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, and topics. Sentiment analysis is still a famous topic because it is technically challenging and useful in many ways. For example, companies always want to understand the people's opinions about their products and services, and which particular features are popular with certain demographics. Customers also want to know opinions of others before purchasing a product or using a service.

Based on the object to be classified, (Liu, 2012) categorized sentiment analysis into document, sentence, and aspect-based sentiment classification. Document-level sentiment classification classifies an opinion document, usually a review of a product, as expressing a positive or negative opinion. One important assumption is that the document expresses opinions on a single object and the opinions are from a single opinion holder. In sentence sentiment classification, the first task is to determine whether the sentence is subjective or objective. Then we can determine the sentiment on the subjective sentence. In some applications, opinion classification in document and sentence level do not prove the necessary detail. A document or sentence with positive opinion does not necessarily mean that the author likes all aspect of the object. Although the general sentiment of a document is positive, in general product review document, the author also writes negative opinions on several aspects of the object. In this case, aspect based sentiment classification is used. This classification needs to employ aspect extraction before the sentiment classification.

There are several methods to conduct the sentiment analysis. The most common method used by researchers are the supervised learning methods. Any existing supervised learning methods can be applied to sentiment classification. In (Vinodhini, 2012), the authors found out that Naive Bayes algorithm and SVM (Support Vector Machine) are widely used algorithm for document classification and using unigrams (a bag of individual words) as features in classification performed well with either naive Bayesian or SVM. Other features that can be used or can be combined together are terms and their frequency, part of speech, and negations words. Other popular methods is by utilizing lexical resources available such as SentiStrength[12], or by using a small set of seed opinion words and an online dictionary. (Chaovalit, 2005), based on their research, argued that the machine learning approach is more accurate but requires a significant amount of time to train the model, while lexical approach is slightly less accurate but is more efficient to use in real-time applications. Especially for micro-blogging such as Twitter, (Kouloumpis, 2011) developed specific features such as emoticons, abbreviations, intensifiers and also various internet slang dictionaries.

For detecting the user's voting intention from their tweets, the most popular method used in previous studies is lexicon based polarization because of the simplicity.[13] Only several experiments such as (Bermingham, 2011), (Fink, 2013) and (Ceron A. C., 2014) employed supervised classification sentiment analysis. Though applying sentiment analysis improve the accuracy of the prediction in most studies, (Gayo-Avello D. , 2013) argued that those sentiment analysis have poor performance and miss the subtleties of political language.

[12] sentistrength.wlv.ac.uk
[13] Detailed discussion can be found in Section 4.5

3 The Prediction Model

In this chapter, we combine and compare tweet-based prediction methods that were previously done by other researchers. As mentioned in the previous chapters, many researchers has tried to create election prediction models using Twitter as the data source. In Table 4, there is a summary from more than 20 published articles that were performed an election prediction. Besides those research, there are several articles that pointed out several challenges and gave suggestions into this topic. All of those information are used to develop the prediction model in this chapter.

The process of predicting the election result is divided into four steps, data collection, data cleaning/filtering, data normalization, and prediction calculation. Those steps will be discussed in Section 3.1. Next, the evaluation of the prediction is discussed in Section 3.2.

No	Authors and Published Year	Country	Election Type	Number of Candidates	Method(s)	Mean Absolute Error
1	(O'Connor, 2010)	US	Presidential	2	Sentiment Analysis	#N/A
2	(Tumasjan, 2010)	Germany	Federal	6	Count Tweets/Hashtags	1.7%
3	(Choy, 2011)	Singapore	Presidential	4	Count Tweets & Sentiment Analysis	6.1%
4	(Chung, 2011)	US	Senate	2	Sentiment Analysis	#N/A
5	(Gayo-Avello D. M., 2011)	US	Senate	2	Count Tweets & Sentiment Analysis	0.9% - 39.6%
6	(Bermingham, 2011)	Ireland	General	5	Count Tweets & Sentiment Analysis	3.7% - 5.9%
7	(Sang, 2012)	Dutch	Senate	12	Count Tweets	1.3%
8	(Choy M. C., 2012)	US	Presidential	2	Count Tweets	1.7%
9	(Jungherr, 2012)	Germany	Federal	6	Count Hashtags & Sentiment Analysis	#N/A
10	(Cameron, 2013)	New Zealand	General	453	Number of Friends/Followers	#N/A
11	(Nooralahzadeh, 2013)	US & French	Presidential	2	Sentiment Analysis	#N/A
12	(Ceron A. C., 2013)	Italy & France	Presidential	7 (Italy) & 2 (France)	Count Tweets & Sentiment Analysis	2.4% - 5.7%
13	(Mejova, 2013)	US	Republican nomination	7	Count Tweets & Sentiment Analysis	#N/A
14	(Gaurav, 2013)	Venezuela, Paraguay & Ecuador	Presidential	2 (Venezuela), 3 (Paraguay), & 2 (Ecuador)	Count Tweets & User	0.1% - 19%
15	(Wong F. T., 2013)	US	Presidential	2	Count Tweets/Retweets & Sentiment Analysis	#N/A
16	(Beauchamp, 2013)	US	Presidential	2	Sentiment Analysis	#N/A
17	(Sanders, 2013)	Netherlands	Parliament	11	Count Tweets	2.4%
18	(Jensen, 2013)	USA	Republican Nomination	4	Count Tweets	3.1%
19	(Fink, 2013)	Nigeria	Presidential	4	Count Tweets & Sentiment Analysis	11.0%

20	(Makazhanov, 2014)	Canada	General	4	Count of Interactions, Followers/Followees	#N/A
21	(Ceron, 2014)	US & Italy	Presidential	2 (US) & 5 (Italy)	Count Tweets & Sentiment Analysis	0.4% - 9.7%

Table 4 List of research used for building the conceptual model

3.1 The Process of Predicting an Election

The research about predicting election results has started since few years ago. The earliest experiments, such as (Tumasjan, 2010) only consist of collecting tweets containing the parties' names then using the number of those names mentioned in the tweets for calculating the prediction. In (O'Connor, 2010), the authors used a lexicon-based approach to classify positive and negative tweets, then use the number positive tweets as the vote share of the prediction. Many researchers have modified or added new techniques since then. For example, (Choy M. C., 2012) adjust the prediction using the census information, (Wong F. M., 2013) extended the keywords rather than only using the candidates' names, (Makazhanov, 2014) removed tweets that are considered as spam, and (Ceron A. C., 2014) employed more sophisticated sentiment detection rather than only comparing them to lexicon databases.

Although employing different methods, those research can be broken down and each process can be categorized into four steps: data collection, data filtering, de-biasing of the data, and the prediction calculation. Data collection contains information such as selected API type, the number of tweets/user, the duration for collecting the data, and the keywords/hashtags. Data filtering focus on cleaning the data such as deleting spam, non-political tweets, and removing non-potential voters, etc. Normalization intended to reduce the bias on Twitter data sample. The last step is to calculate the prediction result. Several ways to do the calculation are count of mention in tweets and count of sentiment-extracted tweets. The model can be seen in Figure 2 and each step is described in more detail in the next sections.

Figure 2 Twitter Based Election Prediction Model

3.1.1 Data Collection

The data collection is the process to get the tweets from Twitter that have a relation to the election. Based on the studies listed in Table 4, there are several major differences on how each study conduct the experiment: the election type/number of candidates, data collection methods, data collection duration, and keyword selection. The categorization result is listed in Table 5 while the detailed discussion of this categorization can be seen below.

No	Category	Parameter	Article Number *refer to Table 4
1	Election Type	Presidential	[1][3][8][11][12][13][14][15][18][19][21]
		General/Party	[2][6][10][20]
		Parliament/Senate	[4][5][9][12][16][17]
2	Method	Search API	[3][11][12][13][17][18][20]
		Stream API	[1][5][7][9][14][15][16][19][21]
3	Duration	< 1 Month	[3][4][5][6][17][18][21]
		1 – 2 Months	[2][7][10][11][12][20]
		3 – 6 Months	[8][9][16]
		> 6 Months	[1][13][14][15][19]
4	Keywords	Candidate/Party names	[2][3][4][5][7][11][12][13][14][15][16][17][18][9]
		Campaign Hashtags	[6][9][11][15][19]
		Election Hashtags	[9][11]

Table 5 Data Collection Parameter

Election type

There are several types of election that were chosen by previous researchers to be studied/predicted based on the data in Table 5. The most common one is the presidential election which select one winner from several candidates, usually from two to three candidates. The other types of election is general election that select political parties (5-10 parties) and parliamentary election that select for the people representatives in the parliament (more than 20 candidates). The most noticeable difference is the number of candidates. In (Metaxas, 2011) and (Gayo-Avello D. , 2013), the authors argued that it is harder to get a better prediction in an election that have more candidates. In case of random guessing, it is true that predicting the winner of an election is harder when there are more candidates, but the Mean Absolute Error (MAE, calculated using Equation 6) will be lower because the total error is divided with more candidates. Besides the number of candidates, the election types also affect other categories such as the keyword selection and evaluation method. The accuracy calculation between predicting which candidates can get a seat in the parliament is very different and predicting how many vote share will a candidate get in a presidential election.

Data Collection Method

There are two methods on how to connect and collect tweets from Twitter. The first method is by searching tweets matching to the keywords. The second method is by collecting all the tweets provided by Twitter through streaming API, or all the tweets in a specific language, or all the tweets in a specific location then put all of them into the database. Both methods have their own advantages and disadvantages. For example, the first method requires only small storage as the data are relatively small. The downside is that researcher cannot get data from other keywords (if he needs to) from an earlier time. Twitter allows the search API only for 7 days backwards. This data collection method is suitable if the focus of the research is on the feature extraction or the prediction method. With the second method, researcher can apply many set of keywords to get the best result. The drawback is that there are so many unused data being stored. For example, in (Fink, 2013), only 0.2% of the total tweets are related to the political election. The biggest data set was shown in (Nooralahzadeh, 2013), where the authors collect 13 billion tweets in JSON format. How to efficiently store and search the dataset was quite a challenge on itself. This method is suitable when the author wants to focus more on how to correctly identify the political tweets. Figure 3 shows the diagram for both methods.

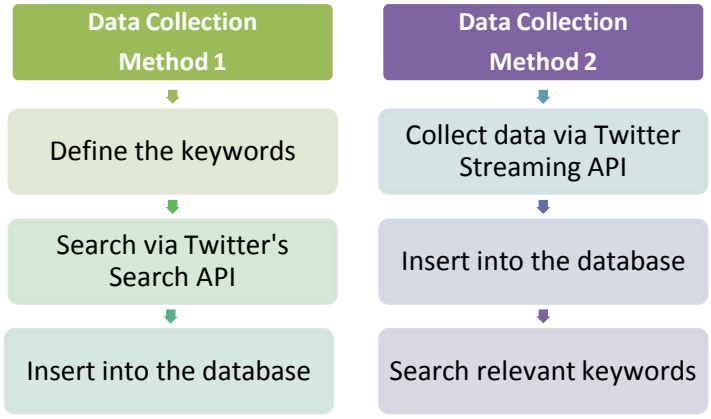

Figure 3 Methods for Data Collection

Data Collection Duration

Other variables to be noticed are the duration of the data collection and what data is used to create the final prediction. Several research such as (Wong F. M., 2013), (Gaurav, 2013), and (Mejova, 2013) collect the tweets more than 6 months prior to the election, while other such as (Ceron A. C., 2014) and (Sanders, 2013) collect the tweets for less than one month. Moreover, (Jensen, 2013) only collected data 2 days prior in their research. Long duration in the data collection does not necessarily mean that there are more data used in the election. Besides (Sang, 2012) that used 1 month of tweets as the data source for the prediction calculation, others use maximum 7 days of data in their calculation. (Bermingham, 2011) compared the result from using different days of data in the calculation, and showed that using 1 days of data produced better result than 5 days of data

and 7 days of data. They tried to push it by only using recent 1000, 2000, and 5000 tweets but the results didn't get better when using only a fraction of tweets in a day.

Keywords for Data Collection

One more important variable is the keywords selection as the tweets collected are really depend on the keywords used by the researchers. Most research use candidate/party names and abbreviations as the keywords. Several research such as (Nooralahzadeh, 2013) went further to include candidate names, candidate addresses, election related hashtags, and campaign hashtags. The assumption is that using more keywords related to the election can collect more tweets and in turn, increase the prediction accuracy.

From above discussion, there are several assumptions that can be taken as follows, first, it is harder to predict the winner in an election that have more candidates and different types of election require different approaches. Second, both search API and stream API have limitations; Twitter search API is limited to Twitter search engine; and Twitter streaming API does not return all public tweets. Third, using data closer to the election produce better prediction result. Current best practice is using as few as 1 day of data, but we should not reduce it to, for example, last 1000 or 2000 of tweets. Last, the prediction result can also be improved if we use more keywords rather than only using the candidates/parties names.

3.1.2 Data Filtering

The goal in this data filtering step is to reduce the noise in the dataset. Based on our review from the list on Table 4, only 5 article stated that they filtered their data before started the calculation of the prediction. (Makazhanov, 2014) trained the data set to recognize spam by using a training set. The authors also manually check the identified spam user and several non-spam user. About 0.3% of total users were labeled as spam. In (Gaurav, 2013), the authors filtered irrelevant tweets by modifying their approach to only consider tweets that contain the name as well as specific keywords like 'eleccion' or 'election'. (Sanders, 2013) cleaned the data by removing stop words, numbers, html references, punctuation symbols and candidate names and addresses. They delete all the re-tweets from the dataset (in order to remove duplicate tweets) as well. (Bermingham, 2011) removed tweets that report poll results.

This research area of detecting spam or bot in Twitter is also interest several researchers. (Chu, 2010) defined four tests that can be used to distinguished humans from bots as follows, entropy test (measure retweet intervals), Spam and Miscreant Test (check for benign or malicious content), Account Properties (Does the Account have subscriber details), and the combination of Entropy, Spam, and Account Properties. (Cook, 2014) developed previous method of detecting bots into 9 tests to distinguish fake tweets, listed in Table 6. This method was implemented in 2013 Australia Federal Election. One of their results showed that in 2 days the followers of on candidate (Tony Abbot) increased from 165 thousands to 234 thousands. The tests revealed that more than 28000 followers are bots.

Nine Way Test for Twitter entities in an Election.
1. Entropy Test - Measure Retweet intervals
2. Spam and Miscreant Test - Check for Benign or Malicious content
3. Account Properties – Does the Account have subscriber details or does it look hollow
4. Accounts Created on or about August the 4th 2013. (Announcement of Election)
5. Inactivity before the election
6. Inactivity after the election
7. Follower alignments – Bots don't follow other Bots
8. Mass retweets on policy-specific days and times
9. Discrimination Analysis – combining Entropy, Spam, and Account, Inactivity, Alignments, and Mass retweets.

Table 6 Nine tests to distinguish fake tweets (Cook, 2014)

To make sure that tweets used in the data set represent the real voter, the keyword based dataset should utilize geo-location information provided by Twitter API. This geo location can also be achieved automatically when the dataset language is not universal, for example tweets written in French or German language. Other approach is to remove non-personal user. Non-personal users frequently use location names, abbreviations and business related terms, and personal users frequently use person names. (Makazhanov, 2014) identified that about 1.8% of users in their dataset were identified as non-personal user. In summary, we have discussed the importance of applying data filtering to remove spam, to make sure only tweets from users that have the right to vote is selected in the dataset.

3.1.3 Reducing the Bias of Twitter Users

Addressing data bias is an important aspect in this process. In the previous chapter, it was described that users of the social media do not represent the global population. Because of that, several research has tried to determine the demographic strata where the users belong to and weighting their tweets accordingly before the calculation process.

In Section 2.2.1, we have explained that Twitter users are not a perfect representatives for the real population. But, there are only few researcher who had include this issue into their prediction model. For example, (Gayo Avello, 2011) attempted to un-bias data according to user age by crossing their full names and county of residence with online public records and by using the age groups to adjust the prediction, their prediction accuracy/MAE was reduced from 13.10% to 11.61%. Their result of identifying about 2,500 users showed that younger people were overrepresented in Twitter.

Another approach was conducted in (Choy M. C., 2011), rather than detecting demographic information from the sample, to predict the outcome of a presidential election, they use external information such as percentage of the population in an age group, percentage of social media user in an age group, percentage of computer literacy of an age group, and vote share of the party at the previous general election who support the candidate. They use those information to adjust the prediction by considering the people who do not use Twitter

or do not connect to the internet. Their calculation for the prediction can be seen in Equation 2 below.

Let Tx = percentage of people for candidate X.
Let TSix = percentage of people for candidate X using social media in age group i.
Let NTSix = percentage of people for candidate X using computer but not social media in age group i.
Let OSix = percentage of people for candidate X who does not use computer in age group i.

$$T_x = \sum_{1}^{i} TS_{ix} + NTS_{ix} + OS_{ix}$$

Equation 2 Vote share calculation (Choy M. C., 2011)

Besides age, Twitter users also have representativeness issues in term of gender, ethnic, and urban information. So it is also important to normalize the dataset based on those parameters. For that we have developed a weighting calculation to adjust the weight for each tweet according to the demographic traits of the user posting it. The calculation can be seen in Equation 3. Each user and tweet will be adjusted accordingly. For example, all users will have a value of 1 without a weighting, then after weighting, people living in a dense area will have less weight (less than 1) and female user will have more weight (more than 1) with the assumption that female users are under presented in Twitter.

Let Wxy = weight of users living in province X and having demographic trait Y.
Let Ux = the number of users in province X.
Let Uxy = the number of users in province X having demographic trait Y.
Let Px = the population in province X.
Let Pxy = the population in province X having demographic trait Y.

$$W_{xy} = \frac{U_x}{U_{xy}} \times \frac{P_{xy}}{P_x}$$

Equation 3 Weight calculation based on demographic information

3.1.4 Calculating the Prediction Result

Based on the features used by the authors, we can divide the calculation methods into two main categorizes, parameter count and sentiment analysis. In parameter count, counting tweets is the most common feature used by researchers followed by counting re-tweet, user, and interaction between candidate and potential voter. One unique method is found in (Cameron, 2013) where the author use the number of candidate follower and its changes over time as the data source to predict the outcome of the election. Detail of this categorization and the articles that implement them can be seen in Figure 4.

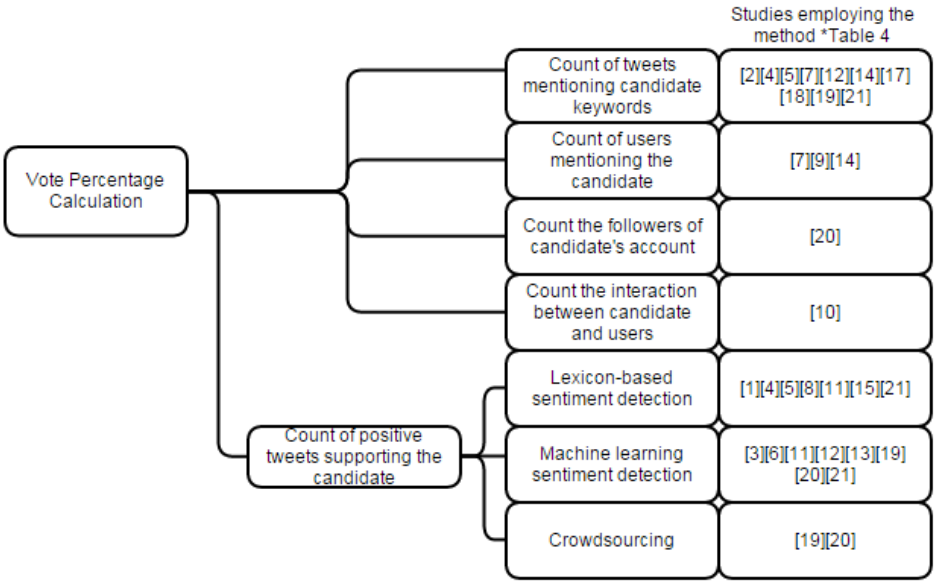

Figure 4 vote calculation method categorization

Count of Tweets and Users

The most common method to calculate the prediction is by using the assumption stated in (Tumasjan, 2010) that candidates" vote proportions in an election correspond with the number of tweets mentioning candidate/party names. Most recent research that using other features such as sentiment analysis (Ceron A. C., 2014) or census correction (Choy M. C., 2012) used the result from tweet counting to check whether their feature improved the result or not. Several modifications of this method are found in (Sang, 2012) who argued that counting the number of user is better than tweet counting because each user is only has one vote. (Makazhanov, 2014) did not count all of the tweets, but only tweet that had an interaction (retweet, mention, reply) with the candidates/parties account.

The number of Followers

Few researchers employ different methods in their prediction. For example, (Cameron, 2013) used social media parameters such as followers in Twitter and friends in Facebook. They had an assumption that the vote proportion in the election correspond with the number of followers or the number of friends of the candidates/parties' social media account. They also use the changes in the number of followers/friends prior to the election. Their stated that there is a relationship between the sizes of online social networks and election results. However, it is not linear and social media presence only make a difference in closely contested elections.

Sentiment Detection

The second category is applying sentiment detection in each tweet to classify positive, negative and/or neutral tweets. This could be performed by using several approaches such as lexicon-based, supervised machine learning, and crowdsourcing. In lexicon based

approach, each word in the tweet is compared with a lexicon database to know the sentiment value of each word. If the total value is positive, then that tweet will be categorized as positive tweet. Using a training data, supervised machine learning uses a classifier to divide between positive and negative tweet. Some researchers compare the result from machine learning with crowdsourcing, such as (Burger, 2011), and their results showed that the performance of machine learning is better than crowdsourcing. We will then discuss briefly about how lexicon-based detection and machine learning techniques.

Lexicon Based Sentiment Detection

Each words in a lexicon-based approach contain a polarity (positive, negative, neutral) and strength. For example, in SentiStrength the strength of a word could be between 5 (very positive) and -5 (very negative) while in SentiWordNet[14], the strength is between 1 and -1. The value can be used to detect the polarity in the whole text or tweet. When there are more positive words in a text or the total value of all words in a text is positive then the text is classified as positive and vice versa. The lexicon database does not contain all of the available words, so words in the tweet that are not in the database are regarded as neutral or have a strength of 0. (Taboada, 2011) explained that the challenges in this method are negation and intensification. Negation words can reduce the strength of a word or even can reverse the polarity of the word, for example "*not* spectacular" and "*none* of them are good". Intensification words such as *slightly* or *very* can either increase or reduce the strength of the subsequent word.

Machine Learning Sentiment Detection - Naïve Bayes

Naïve Bayes is a supervised classifier that applies Bayes' theorem with a (naïve) assumptions that the features are independent with one another. In the Bayes' theorem, the probability of a document (d) belong to a class (c) can be calculated using **Error! Reference source not found.** Then using the conditional independent assumption of the features (f), the probability in the Naïve Bayes can be calculated using **Error! Reference source not found.** Then the classifier assign the class that has the highest probability ($c^* = \arg\max P(c|d)$) (Pang, 2002).

$$P(c|d) = \frac{P(c)\ x\ P\ (d|c)}{P(d)}$$

Equation 4 Bayes Theorem

$$P_{NB}\ (c|d) = \frac{P(c)\ x\ (\prod_{i=1}^{m} P(f_i|c)^{n_i(d)})}{P(d)}$$

Equation 5 Probability of a document in a class using Naive Bayes

[14] http://sentiwordnet.isti.cnr.it/

Naïve Bayes calculate the probability of a class belong to a certain group based on the feature of the text/tweet. In the sentiment analysis of the tweet, each word in the text is the feature. Naïve Bayes use the presence of a word rather than the meaning or position of the word to predict whether a text is likely to be in a certain class or have a certain polarity (positive or negative).

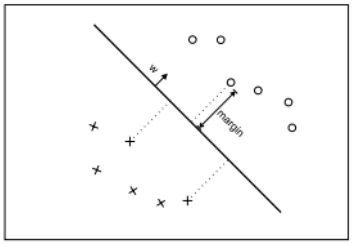

Figure 5 A simple linear Support Vector Machine (Tong, 2002)

Machine Learning Sentiment Detection - Support vector machine (SVM)

SVM is another machine learning model that are able to classify an object (text/word) based on the pattern recognized from a training set. (Tong, 2002) defined SVM as hyper planes that separate the training data by a maximal margin. Given a training set consists of objects, each one belong to one of two categories, an SVM training algorithm builds a model that assigns new examples into one of those categories. An example of a linear SVM can be seen in Figure 5, where all objects lying on one side of the hyper plane are labeled as −1 (negative), and all objects lying on the other side are labeled as 1 (positive).

Summarizing the results from above experiments, we understand that counting tweets is used as the baseline method to be compared with other methods. Both counting the candidates' follower and counting the interaction between users is not a good predictor for an electoral prediction. All researchers who performed user normalization, showed that user counting perform better than tweet counting. As for sentiment analysis, the trend is moving from lexicon based sentiment detection to the more complex machine learning sentiment classification. Although the results were already better than the baseline, (Gayo-Avello D. , 2013) argued that the methods need improvements such as how to detect sponsored/campaign tweets and sarcasm.

3.2 Evaluation of the Prediction

Most authors provide an evaluation of their research in the articles. We classify them based on what the prediction are compared with and how they measure the difference from that comparison. In Table 7, we can see that most research compare their prediction result to the election real result and also to poll result before the election. The common goal is to find out whether prediction using tweets can have the same or better result than the offline poll survey. There are also research such as (Beauchamp, 2013) and (Sanders, 2013) that gave

a series of predictions than compared only to the poll results with the same published date. (Wong F. T., 2013) tried to compare the social media with other media such as news or television network.

No	Category	Parameter	Article Numbers *refer to Table 4
1	Comparison	actual result	[2][3][5][6][7][8][9][11][14][17][18][19][21]
		survey polls	[1][2][3][7][12][13][16][17][19][21]
		sentiment media	[15]
2	Evaluation Value	Mean Absolut Error	[2][5][6][8][9][12][16][17][18][19][21]
		Root Mean Square Error	[14]
		correlation coefficient	[1][14]
		number of winner	[3][7][10]

Table 7 Evaluation Method

For the evaluating the result, most of the prediction are compared to the real election result, then evaluated on how large the distance is between the prediction and the real result. The distance is called Mean Absolute Error and calculated using Equation 6 where n is the number of candidates, P is the prediction result percentage, and R is the real result percentage. (Gaurav, 2013) used its different calculation, Root Mean Square Error, that has similar underlying principle. Research that tried to predict a general election, such as (Cameron, 2013) and (Sang, 2012), used the number of winner or acquired seats as their evaluation.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | P_i - R_i |$$

Equation 6 MAE calculation

Other parameter is suggested in (Metaxas, 2011), where the authors argued that incumbent candidate gets re-elected about 9 out of 10 times, so incumbency should be taken into account as well in the prediction. As explained before in Section 2.1.2, it is applicable in the US but might not work in other countries. One other factor that should be taken into consideration is the granularity level. For example, a prediction could be correctly predicts the winner of a national level election while in state level, the prediction could have a huge error, as shown in (Choy M. C., 2012). As the proven scientific tweet-based prediction method is yet to be developed, we argue that it is important to compare the prediction result to both the more established offline polls and the actual election result. And when it is possible, evaluation with calculating the error should be applied rather than only predicting the winner.

4 Predicting Election in Indonesia

This chapter describe about the implementation of a tweet-based electoral prediction model that was described in the previous chapter. For the use case, we will implement it in the 2014 Indonesian presidential election. Based on the study by Semiocast[15] in 2012, Indonesia ranked as the 5th in the world in term of the number of Twitter accounts. This is important to make sure that there are plentiful tweets to be processed for the prediction. One other important reason is because Indonesia has an interesting demographic information related to the internet users. In the same study, we know that several cities in Indonesia such as Jakarta and Bandung are very active on Twitter (in the top 10 list of cities in the world in term of posted tweets). But the internet penetration in the country is very low, around 24% of the total population. It means that there are also many cities where the citizens rarely use the internet, let alone social media such as Twitter. This fact shows that there are a possibility of high bias in Indonesia's Twitter users and makes the de-biasing step more crucial.

Related to the research question that want to understand which factor that can improve the prediction result, all possible parameter in the model are performed and will be compared to each other in the next chapter. This chapter consists of all steps in the model plus a preliminary and secondary manual annotation to understand the demographic information of the users in the dataset.

4.1 Data Collection

4.1.1 Collecting the Tweets

Tweets Examples	Translated Tweets
tolong di RT ya teman, semoga pak Jokowi jadi presiden #PilihNo2_utkNKRICerdas	Please RT friends, hopefully Jokowi will be president #ChooseNo2_forSmartIndonesia
Jokowi: Siapapun Pemimpinnya Kita Tetap Bersaudara #Salam2Jari #akhirnyamilihjokowi <~♥♥ ;)…	'Jokowi: Whoever the winner, we are still family' #2fingers #IChooseJokowi @2014president
Jokowi Bicara Berantas Premanisme, Kader PDIP Segel Kantor TV One #akhirnya2anarkis #akhirnyaduaserangSATU	Jokowi said eradicate thugs, but the people of his party attack TV ONE office #No2anarchy #No2attacksNo1
Prabowo dan Jokowi bukan Superman atau Suparman.Mereka kelak tidak akan berhasil, kalau tidak didukung rakyat.Ya... kita-kita ini, rakyat.	Prabowo and Jokowi are not superman. They won't succeed without the support of the people. We are the people.
Yeahhhhhhh @baharzhakuv: #PRABOWO-HATTA1	Yeahhhhh @baharzhakuv: #Prabowo-Hatta1
Saya, Istri, Ibu, Ibu Mertua, dan Bapak Mertua semua pilih Prabowo. Jadi dikeluargaku Prabowo menang telak...:)	Me, my wife, mother, mother in law, and father in law choose Prabowo. He already won in our family.
Mas Prabowo itu orang besar,dan punya standar etika. @prabowo @ratu_adil	Prabowo is a great person and have high standards of ethics. @prabowo @ratu_adil
Download Need4Speed Most Wanted 2014 New Android App #CoblosNomor1_PrabowoHatta @NFSworld	Download Need4Speed Most Wanted 2014 New Android App #ChooseNomor1_PrabowoHatta @NFSworld

Table 8 Examples of the tweets in the dataset

[15] http://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US

The election takes place on July 9th, 2014. There are two candidates in this election, the first is Joko Widodo and Jusuf Kalla as his running mate, and the second is Prabowo Subianto and Hatta Rajasa as his running mate. Both announce their candidacy on May 20th, 2014. We collected the tweets starting from April using both candidate popular name, "Jokowi" and "Prabowo". We store the tweet, username, and screenname of each tweet. We called this dataset as "POLDATA". To select tweet only from potential voters, we collect tweets originating from Indonesia. The location information is one of the parameter in the Twitter API[16]. Twitter estimates the location from 'geotagging' information and from the location stated in the users' profile. There are very few users that activate geo-location in Twitter, but most of the user fill the location information in their profile. Even though several users use abstract location such as 'the world', 'home sweet home', etc. (Gayo Avello, 2011) used this location information and they claimed that they were able to identify the location of more than 30% us users using this information. Other researchers used language approaches to handle this problem, for example, (Sang, 2012) only use tweet that has the Dutch word "het" and (Gaurav, 2013) use only tweet containing the word "elecciones". The overview of our dataset can be seen in Table 9 and Figure 6. Several examples of our collection can be seen in Table 8.

Parameter	Value
The number of electoral tweets	7020228
Max tweets in one day	375064
The number of users	490270
Max users in one day	148135
Average electoral tweets per user per day	3.04
Average length of the tweets	107.27
Percentage of retweet	29.24%

Table 9 several parameters of the dataset

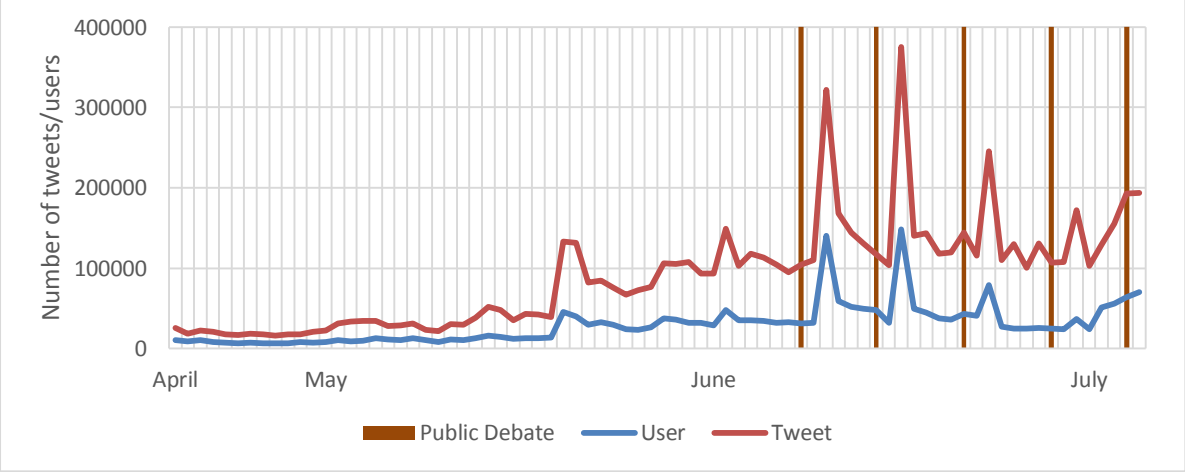

Figure 6 the Number of Tweets and Users. *sudden increase of tweets on the next day after every debate

[16] https://dev.twitter.com/docs/api/1.1/get/search/tweets

One parameter that we try to compare is the keyword selection. Most of the researchers use the candidates' name, but several researchers such as (Fink, 2013) and (Nooralahzadeh, 2013) also use campaign and election hashtags. For this research, we mainly collect tweets mentioning the candidates or vice-candidates name, and we also collect tweet containing campaign hashtags to add more tweets to our dataset. The keywords that are selected come from Twitter trending topics, we checked the Twitter trending topics manually and added to the keyword list if there is any hashtags related to the election. The keyword list can be seen in Table 10.

Category	Keywords	
Popular Names	Jokowi	Prabowo
	Kalla	PrabowoHatta
	jokowi-jk	hatta
Twitter Username	@Pak_JK	@Prabowo08
	@jokowi_do2	@hattarajasa
Campaign Hashtags	IndonesiaHebat	DukungPrabowoHatta
	revolusimental	SelamatkanIndonesia
	salamDUAjari	indonesiabangkit
	JKW4P	SalamSATUjari
	JKWJK	Salam1jari
		prabowoforpresident

Table 10 Keywords used for data collection

From studies such as (Burger, 2011) and (Rao, 2012) that were previously discussed in Section 2.4, we understand that using the tweets content produce better result than using social media feature such as number of followers, post frequencies, etc. For that reason, 100 last tweets of each user is collected as well. But not all of the users published their tweets to public. As stated in Section 2.2.1, there are about 91% of users who make their profile and tweets available to public, or do not protect their timeline. In our experiment, there are 73 thousand of 490 thousand users which tweets cannot be obtained whether because of protected timeline or the account has been suspended. In total, about 42 millions of tweets are collected in the dataset that we called "TLDATA".

4.1.2 Demographics Information of Indonesia

For comparison and calculation in the next chapter. We have gathered demographic information from Indonesian Statistics Institution (2010 BPS[17]). On Table 11 below, we list the number of population, the gender percentage, and age group percentage of each province in Indonesia.

[17] www.bps.go.id

Province	Population	Gender (%)		Age (%)			Province	Population	Gender (%)		Age (%)		
		Male	Female	0-19	20-49	50++			Male	Female	0-19	20-49	50++
Aceh	4 494 410	50.05%	49.95%	41.84%	45.45%	12.71%	Nusa Tenggara Barat	4 500 212	48.53%	51.47%	40.62%	44.53%	14.85%
Sumatera Utara	12 982 204	49.95%	50.05%	42.98%	43.31%	13.70%	Nusa Tenggara Timur	4 683 827	49.67%	50.33%	46.45%	38.85%	14.70%
Sumatera Barat	4 846 909	49.60%	50.40%	41.02%	41.90%	17.07%	Kalimantan Barat	4 395 983	51.12%	48.88%	41.00%	45.75%	13.25%
Riau	5 538 367	51.53%	48.47%	42.07%	47.76%	10.17%	Kalimantan Tengah	2 212 089	52.15%	47.85%	39.89%	48.95%	11.16%
Jambi	3 092 265	51.34%	48.66%	39.32%	47.86%	12.82%	Kalimantan Selatan	3 626 616	50.64%	49.36%	38.16%	48.56%	13.28%
Sumatera Selatan	7 450 394	50.91%	49.09%	39.43%	46.77%	13.81%	Kalimantan Timur	3 553 143	52.67%	47.33%	38.87%	50.34%	10.79%
Bengkulu	1 715 518	51.12%	48.88%	39.87%	47.02%	13.11%	Sulawesi Utara	2 270 596	51.08%	48.92%	36.30%	45.38%	18.32%
Lampung	7 608 405	51.48%	48.52%	38.43%	46.45%	15.12%	Sulawesi Tengah	2 635 009	51.27%	48.73%	41.86%	45.18%	12.96%
Kepulauan Bangka Belitung	1 223 296	51.92%	48.08%	37.81%	48.48%	13.71%	Sulawesi Selatan	8 034 776	48.85%	51.15%	40.08%	43.81%	16.11%
Kepulauan Riau	1 679 163	51.34%	48.66%	36.22%	55.24%	8.54%	Sulawesi Tenggara	2 232 586	50.25%	49.75%	44.62%	43.14%	12.23%
DKI Jakarta	9 607 787	50.69%	49.31%	32.41%	54.59%	13.00%	Gorontalo	1 040 164	50.17%	49.83%	41.88%	44.56%	13.56%
Jawa Barat	43 053 732	50.88%	49.12%	38.21%	46.66%	15.14%	Sulawesi Barat	1 158 651	50.20%	49.80%	45.27%	42.09%	12.64%
Jawa Tengah	32 382 657	49.70%	50.30%	34.67%	44.79%	20.54%	Maluku	1 533 506	50.57%	49.43%	45.46%	41.34%	13.20%
DI Yogyakarta	3 457 491	49.42%	50.58%	30.23%	46.20%	23.57%	Maluku Utara	1 038 087	51.20%	48.80%	44.74%	43.88%	11.37%
Jawa Timur	37 476 757	49.37%	50.63%	32.62%	46.52%	20.86%	Papua Barat	760 422	52.92%	47.08%	43.36%	47.77%	8.87%
Banten	10 632 166	51.15%	48.85%	39.51%	49.36%	11.13%	Papua	2 833 381	53.14%	46.86%	44.82%	48.72%	6.47%
Bali	3 890 757	50.42%	49.58%	33.21%	47.94%	18.85%							

Table 11 Population per Province in Indonesia

4.2 Manual Annotation

4.2.1 Data Filtering

In data filtering step, we want to make sure that only tweet related to the election and contain opinion of users are used in the prediction process. We already use keywords and location based tweets to limit our data so that only electoral tweets from potential voters located in Indonesia are selected. But this still do not remove the spam in our dataset, POLDATA. Other studies, such as (Makazhanov, 2014), used machine learning to detect the spam tweets based on manual annotation as their gold standard. (Sanders, 2013) deleted all retweets in their dataset because they did not want duplicate tweets. In Australia, there is a law that prohibit their citizens to mislead or deceive other voters, and it is also applicable in tweets. (Cook, 2014) called the phenomena as 'slactivism' in Twitter and

developed several methods to detect fake tweets such as measuring the tweets interval, checking the tweets' content, activity before and after election, and combination of those methods.

The spam tweets in Indonesian election could appear in different form. Because of that, manual annotation was conducted to get the idea of how spammer behave and what is the ratio of spam in the dataset. We manually annotate tweets from randomly selected 600 users from all of the users, we use the content of their latest tweets in their timeline and the information from their profile to distinguish spammer. The result can be seen below:

- 7.4% of the users from the annotation are spammer both humans and bots. The most evident type of spammer is the one whose tweets only consist of retweet and sell a number of follower for an amount of money. The second type uses hashtags from famous topics but the rest of their tweets do not relate to the hash tags at all.
- 3.8% of the users from the annotation are a non-personal user. This is known from their explanation from their bio and pattern in their tweets where they always put a URL link or only consist of promotion.
- 2.1% of the users from the annotation are the one that called 'slacktivist' by (Waugh, 2013) and (Cook, 2014). We can detect them either by their activities prior to the election, newly created account and all of the tweets are electoral tweets, and unmatched between the profile and the content of their tweets. An example of the last type is when an 8th grader retweet/posts complicated electoral tweets in all of their latest tweets.

If we combine all of the results above, the users that need to be filtered based on our annotation is more than 10%. These shows the importance of data filtering and every applications that use tweets as the data source do need to consider this step. We implement data filtering for all of our data set, and the result can be seen in Section 4.3.

4.2.2 Demographic Information

In the previous chapter, in Section 3.1.3, we have described the importance of handling the sample bias in the dataset. Several possible bias in Twitter users, based on Pew research center[18], are gender, ethnic, age, education, income/economic, and residence location type (urban/rural). Based on their data, age and gender are more influential compared to others. To confirm that we manually annotate the same 600 randomly selected users to understand the proportion of age and gender in the dataset. We use user's photos, profile information and tweets to distinguish them. While gender can be easily determined from photos and names, we need to go deeper to find special attributes to detect their age group. We find that middle/high-school students like to post picture wearing their uniform. Many of them post their birthday/age and school name in their bio. University students also post their university name in their bio, but they do not put their age information. They also put their

[18] http://www.pewinternet.org/2013/12/30/demographics-of-key-social-networking-platforms/

hobbies or occupation in their bio and pictures. Elder people mostly use their family as their profile picture and put their position in an organization. The result of this annotation, as seen in Figure 7, shows that there are data bias in our dataset in term of age and gender. In term of age, elder people have very low representatives and in term of gender, female users are less representative.

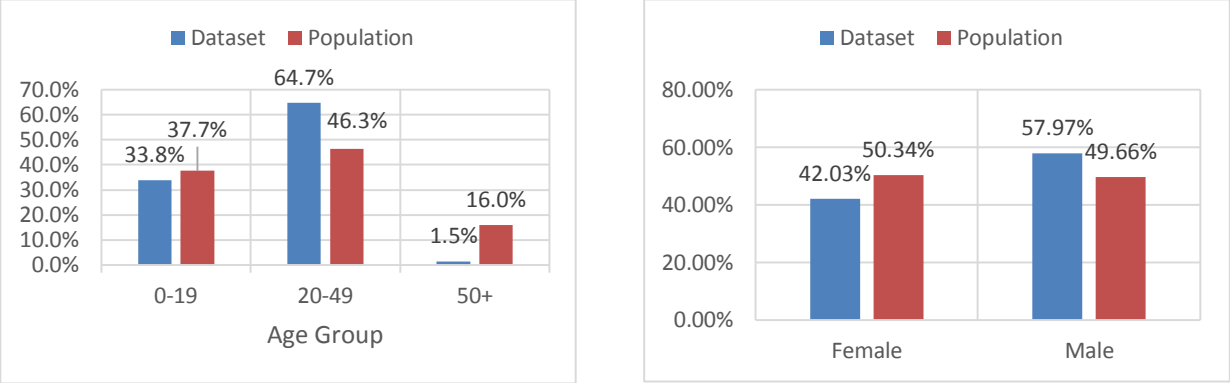

Figure 7 Users' Gender & Age Group of the dataset

4.3 Data Filtering

For the application of data filtering we follow the method from (Chu, 2010) and the result of our annotation explained in Section 4.2.1. Chu et al (2010) proposed a four way test to differentiate Twitter phonies from humans. In the first instance, measuring the time intervals between retweets can reliably detect automated messaging. Scanning for signs of spam was also consistent since humans seldom deliberately send spam, as was examining the account properties of each subscriber, since those subscribers with no real account details, pictures, or descriptors, rarely indicated discrete personages. Moreover, bots are far more likely to post URLs than humans. An examination of these variables in concert therefore, assisted to establish the authenticity of retweeting followers.

Based on the manual annotation result in Section 4.2.1, there are several types of spammers in our POLDATA users. Using our TLDATA that contain the tweets of those users, we develop several criteria to detect spammers as follows:

- We collect users who only post retweets or always use hashtags in their last 100 tweets. Then we manually read the content of their tweets, and find that there are 771 of users in this category.
- We check the content of each tweets and consider them who has a URL link in every tweets as spammer. The number of users in this category is 3551 users. In this step, we do not consider tweet that has URL link from other social media such as YouTube, Path, Instagram, etc. as spammer. We randomly check about 100 of users to verify the result of this method and find that this method is reliable.

- Each tweet that has the words 'twitter', 'follower', and '10rb' or '1k' (equals to 1 USD) is marked. Users that have marked tweets more than 10% of their tweets are considered as spammer. The number of users in this category is 729 users.
- Total aggregate users that needs to be filtered is 4323 users.

4.4 Users' Demographics

From the previous annotation, we know that the users in the dataset do not represent the real population. Besides age and gender, we add one more possible source for data bias, the location of users. The number of Twitter users might not correspond with the number of population in a location. This is important because of the fact that internet penetration in Indonesia is very low. Only people from big cities are connected to the internet and social media, and not every province have big cities. So in this section we will describe about detecting location, age and gender of the users.

4.4.1 Location

In the data collection steps, we get 490,270 users from POLDATA. From those users, we try to identify the location (province) for each user. The desirable way to do this is by using the geo-location of the user provided by Twitter. The geo-location information (latitude and longitude), can be converted into a nearest city/province using a service such as Google Reverse Geocoding[19]. The second source is from location information in the users' profile. For this case, we use a list of cities, village, and district to map the location written in the profile into provinces. The complete result can be seen in Table 12.

Province	Users	Province	Users	Province	Users
Aceh	1864	Jawa Barat	47190	Kalimantan Timur	2139
Sumatera Utara	10012	Jawa Tengah	24449	Sulawesi Utara	2700
Sumatera Barat	3203	DI Yogyakarta	12148	Sulawesi Tengah	353
Riau	4873	Jawa Timur	9100	Sulawesi Selatan	1973
Jambi	1698	Banten	16279	Sulawesi Tenggara	353
Sumatera Selatan	4306	Bali	1959	Gorontalo	588
Bengkulu	708	Nusa Tenggara Barat	2359	Sulawesi Barat	56
Lampung	4001	Nusa Tenggara Timur	420	Maluku	201
Kepulauan Bangka Belitung	304	Kalimantan Barat	2353	Maluku Utara	191
Kepulauan Riau	1727	Kalimantan Tengah	133	Papua Barat	161
DKI Jakarta	58973	Kalimantan Selatan	601	Papua	325
Indonesia (No Province detail)	166585	Unknown Location	105769	Total	490270

Table 12 Dataset location demography

[19] https://developers.google.com/maps/documentation/geocoding/#ReverseGeocoding

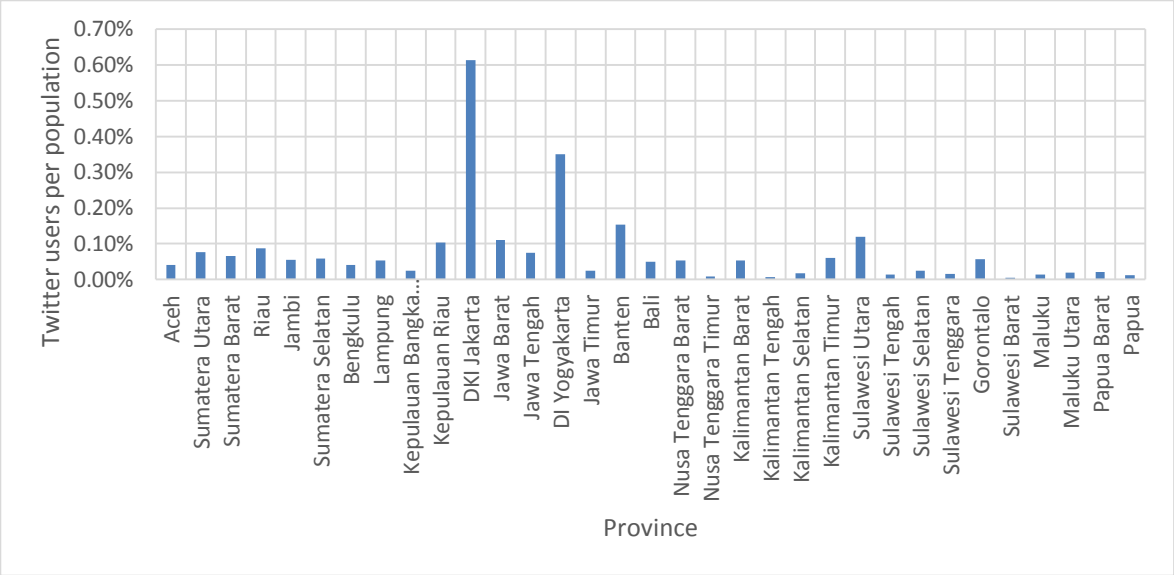

Figure 8 Comparison between twitter user in the dataset and population in the provinces

In Figure 8, we can see that there are big difference between provinces on the ratio of Twitter users in our dataset. Only 5 provinces have more than 0.1% of its people as the samples. DKI Jakarta and Jogjakarta are two provinces with high representatives in this dataset. This is not a representation of all active Twitter users in Indonesia, because first, only users who post tweet related to the election are selected and second, the dataset only contain the users who activate geo-location or put location information in their profile.

4.4.2 Gender

Based on the manual annotation in Section 4.2.2, there are about 60% of male users, while based on statistics data, male population is about 50.3% of the total population. Because of this difference, we want to determine the gender of all users in the dataset so that we can weight each user and tweet accordingly.

Name list

We first identify the gender using a comparison between Twitter screen name (full name, not the username) and a name list. We collect about 6500 popular Indonesian names and use it as the golden truth to identify the gender. Most Indonesian do not have first name-family name format, because of that all part of the names are checked against the list. If more male string found, then we classify the user as male, vice versa. The result can be seen in Table 13. Using this method, we are able to identify the gender of about 140 thousand of 490 thousand users. Unknown gender means that the number of female and male substring are equals or the user's name is not on the list. To verify the precision of this method, we manually check about 2500 identified users selected randomly and find out that the precision of this method is quite high.

	Result (user)	Manual Annotation (Num of user & %)		
Male	83729	True	1538	96.01%
		False	64	
Female	56807	True	1008	98.53%
		False	15	
Unknown	349734			

Table 13 Gender identification result based on a name list

Although using a name list has a high precision, but the identified gender is still less than 30% of the total users. In Section 2.4.1, the results from several research such as (Rao, 2012) and (Nguyen, 2013) shown that social media attributes such as the number of follower/followees or the post/tweet/re-tweet frequencies did not produce good result for extracting demographic information and machine learning classification based on the content of their tweet perform better than them. Because of that we will apply machine learning to identify the gender of the other users.

Machine Learning Classification

As for the classifiers, we follow the study of (Ting, 2011) who compared many text classification approaches, such as k-nearest-neighbor, Naïve Bayes, support vector machines, decision tree, and neural network. Their result showed that the top two text classifier were Naïve Bayes and SVM, but Naïve Bayes text classifier were more widely used because its simplicity, less time consuming and proved effective enough to classify the text in many domains. In the Naïve Bayes classifier, each text is viewed as a collection of words and the order of words is considered irrelevant. Naïve Bayes models allow each word to contribute towards the final decision equally and independently from other word, in which it is more computational efficient when compared with other text classifiers. In the NB classifier, each text is given a weight/probability that the text belongs to a category/class.

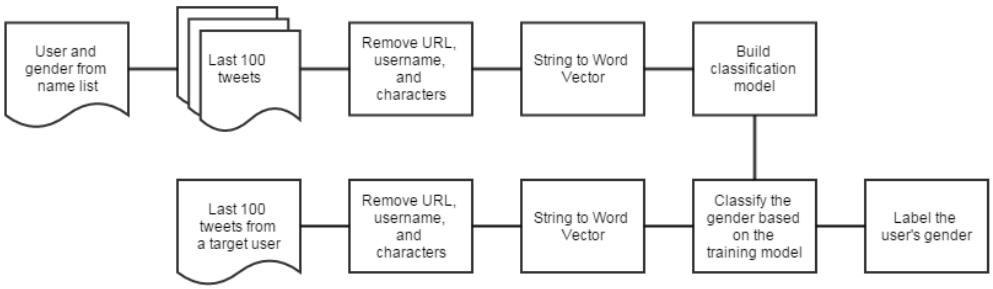

Figure 9 Gender classification process

Training & test data set

To classify the users into two nominal classes, male and female, 100 tweets of each 140 thousand gender identified users from TLDATA will be used as the training set for the machine learning. As for the test set, 100 tweets from all unidentified the users are used.

All the tweets from a user is cleaned by removing the url and username (mention), then combined together as a document. In the training set, each document is labelled either male or female, while the documents in the test set are not labelled. The classification process can be seen in Figure 9.

Classification

To perform the Naïve Bayes classification, WEKA[20] library is used. Each word in the tweets are used as the features in the classification process. Because of that, word tokenization is conducted before the classification. Using WEKA, a classification model is built from the training set. It contains the information about the presence (mean and standard deviation) of each word in both classes, male and female. That information will be used to calculate the probability of the user in the test set as a male and female. When the probability to be in the male class is higher than female class, that user is categorized as a male. Likewise, the user is categorized as a female when the probability to be in the female class is higher.

Evaluation

The performance of the classification is considered good when most of the objects are classified to the correct class/category. It is measured by performing cross validation on the training set. For example, in n-fold cross validation, the training set is divided to n data sets. The classification is performed n times and in each process, the n^{th} data set become the test set and the other data set become the training set. In a two class classification (positive and negative), the number of correctly predicted object in the positive class is called true positive (TP) and in negative class is called true negative (TN). The number of incorrectly predicted object of positive class is called false negative (FN) and incorrect object of negative class is false positive (FP). The accuracy (a), calculated using Equation 7, is a proportion of correct classifications from all classifications result. Accurate classification is represented by a high accuracy. Other measurement is called precision (p), that calculate the proportion of true positives from all messages that are predicted as positive.

$$a = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 7 Accuracy of a classification

$$p = \frac{TP}{TP + FP}$$

Equation 8 Precision of the calculation

From 10-fold cross validation of the training set, we get 0.679 as the accuracy, 0.686 as the precision of male class and 0.671 as the precision of female class. We decide to not use the result because using this low accuracy/precision result does not necessarily mean that the

[20] WEKA (Waikato Environment for Knowledge Analysis) - http://www.cs.waikato.ac.nz/ml/weka/

demographic bias is reduced. Instead, in the calculation process, we will use the data derived from the name list that has high precision.

4.4.3 Age

Similar method as the machine learning gender classification is implemented to identify the age group of the users. The difference is in this classification, the data will be predicted into three classes of age group, "age 0-19", "age 20-49", and "age 50+".

Training & test data set

In Section 4.2.2, we have manually annotated users into three age groups, 0-19, 20-49, and 50+. Using TLDATA, tweets from those users are selected to be the training set and tweets from other users as the test set. All tweets follow the same data cleansing method as the gender classification, removing the url and username (mention).

Classification

Same as gender classification, Naïve Bayes library and String to Word library from WEKA are used for classifying the age group. In this three nominal class classification, a user is categorized to a class/age group when the probability to that class is higher than the probability of being to the other two groups.

Evaluation

The evaluation to a three class classification is a little bit different compared to binary classification. The classification is called true (T) when an object is classified to the correct class, and called error (E) when it is classified to other classes. The confusion matrix for three class problem can be seen in Table 14. While the accuracy and precision are calculated using Equation 9.

		Predicted Class		
		A	B	C
Known class (from training data)	A	T_A	E_{AB}	E_{AC}
	B	E_{BA}	T_B	E_{BC}
	C	E_{CA}	E_{CB}	T_C

Table 14 Confusion Matrix for three-class classification

$$a = \frac{T}{T+E} \; ; \; p_A = \frac{T_A}{T_A + E_{AB} + E_{AC}}$$

Equation 9 Accuracy and precision of three-class classification

Our precision and accuracy for this classification is low, 0.61. We argued that it was because the training set that we used only consist of 600 users. We have the same decision that the classifier result is not enough to be implemented in reducing the bias of Twitter user.

4.5 Sentiment Analysis

Most studies showed that understanding sentiment in the tweet and include it on the calculation can lower the error of the prediction result. As explained in Section 2.4.2, the trend in sentiment analysis is moving from lexicon-based into machine learning sentiment analysis. For detecting sentiment of the tweets in the dataset, lexicon-based is not performed because Indonesian lexicon database is not available yet and it is too time consuming if English lexicon database is used. Naïve Bayes is chosen again because the amount of tweets needed to be classified is very huge so that it is less feasible to use SVM. The process of detecting sentiment is explained below and visualized in Figure 10.

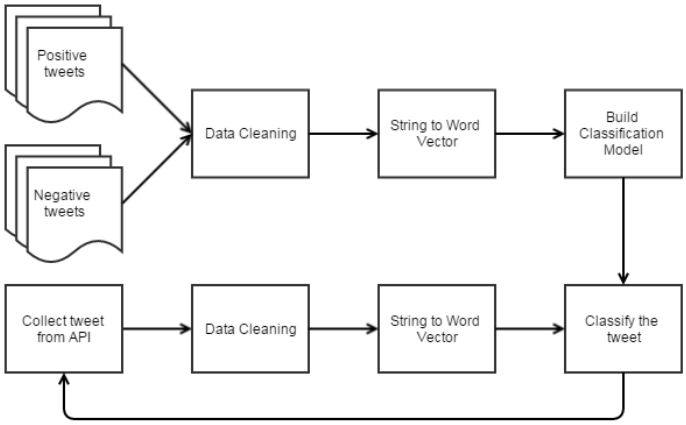

Figure 10 Sentiment analysis method

Training & test data set

As the training set, a larger amount of tweets is used, about 20 thousand (positive: 12287, negative: 7290). Tweets that contain positive emoticon, such as :) or :D, or contain negative emoticon, such as :(, =(or :'(are chosen to be the training set. Same as previous classification, the tweets is cleaned by removing mention/username, URL, electoral keywords, smileys, and characters. We also remove the candidates' name and aliases so that they do not used as the features in the classification. The test set is the new tweets that are received from Twitter API. When new tweet arrived, the sentiment is predicted then the tweet and the sentiment are stored in the database.

Classification

WEKA library for JAVA is used for this classification. Word tokenization is performed on the content of the training set, then the mean and variance of each word's presence are calculated. Those values become the basis for calculating the probability of a tweet having a positive or negative sentiment using Bayes theorem, as explained in Section 2.4.2. Following the experiment of (He, 2011), we use a class prediction probability threshold of

0.8 to filter out low confidence prediction result. It means that a tweet is not considered to be have a sentiment when the probability of having a positive or negative sentiment is lower than 0.8.

Result and Evaluation

The classification model from the training set is evaluated using the same confusion matrix and precision calculation as the gender classification because both are two-class/binary classification. Using 10-fold cross validation (the training model is tested 10 times), the average precision of the training model is 0.799. We decide that the value is quite high and then performed the classification of the whole test set (about 7 million tweets). Especially In the last day data from POLDATA (about 193 thousand of tweets), we are able to identify 54 thousand positive tweets, 21 thousand negative tweets and 117 thousand unidentified tweets.

5 Result & Analysis

In the previous chapters, we built a model to predict an election using Twitter data based on an extensive literature study. We implemented the model on the presidential election in Indonesia and also divided the data into provinces based on each tweet/user location. Although there are many limitations of the data, we believe that this data, if analyzed thoroughly, can potentially tell us a lot, and importantly can empirically test some of the assumptions made when building the model. To do this, we have developed a set of hypotheses, listed in Table 15. We will try to answer these hypotheses through analyzing our data and also other data from previous research. The next section in this chapter will provide the information about the official result of the election and the election polls that published before the election. After that, we display the result from previous research so that we can compare our result with them. Then, the next section will consist of analysis of the prediction result from the model using the hypotheses. The chapter will end with a summary of the tested hypotheses.

Research Question	No	Hypothesis
RQ #1	H1	Tweet-based election prediction using baseline method is comparable to offline polling in term of error/distance with the real election result.
RQ #3	H2	Using several days of data is better than only use the data from the last day before the election.
RQ #3	H3	Using more keywords describes the situation better in Twitter and in turn improves the accuracy of the prediction.
RQ #3	H4	Filtering the data increase the prediction accuracy.
RQ #4	H5	Counting the user instead of counting the tweet improves the prediction result.
RQ #4	H6	Weighting each tweet based on the location improves the prediction result.
RQ #4	H7	Incorporating the users' demographic information reduce the data bias and improve the prediction result.
RQ #4	H8	Implementing Sentiment Analysis in the dataset improves the prediction result.
-	H9	The error/distance between the prediction and the real election result has a relation with the number of user in the dataset.

Table 15 List of hypotheses to be tested

5.1 Election and Polling Result

The evaluation of the prediction model mainly can be divided into validation of method used in the model and to check whether the result is comparable to the offline election polls. In order to do that, both prediction result from this model and from offline polls will be compared to the real election result. There are two main points that used in the evaluation. First is to check if the prediction correctly predicts the election's winner and second, how much is the error or the distance between the prediction and the real result. In this section,

both the official election results (in country level and in province level) and the offline polls results are shown.

The Election Result published by the Election Committee

Presidential election in Indonesia was conducted manually by selecting a candidate's name/picture in a paper. Though the quick count results are available on the same day of the election, the official result was published by the election committee at their website[21] on July 22nd, 2014. The result can be seen in Figure 11, and the result per province in Table 16. We will treat the data in each province as an election and apply the prediction model at each province.

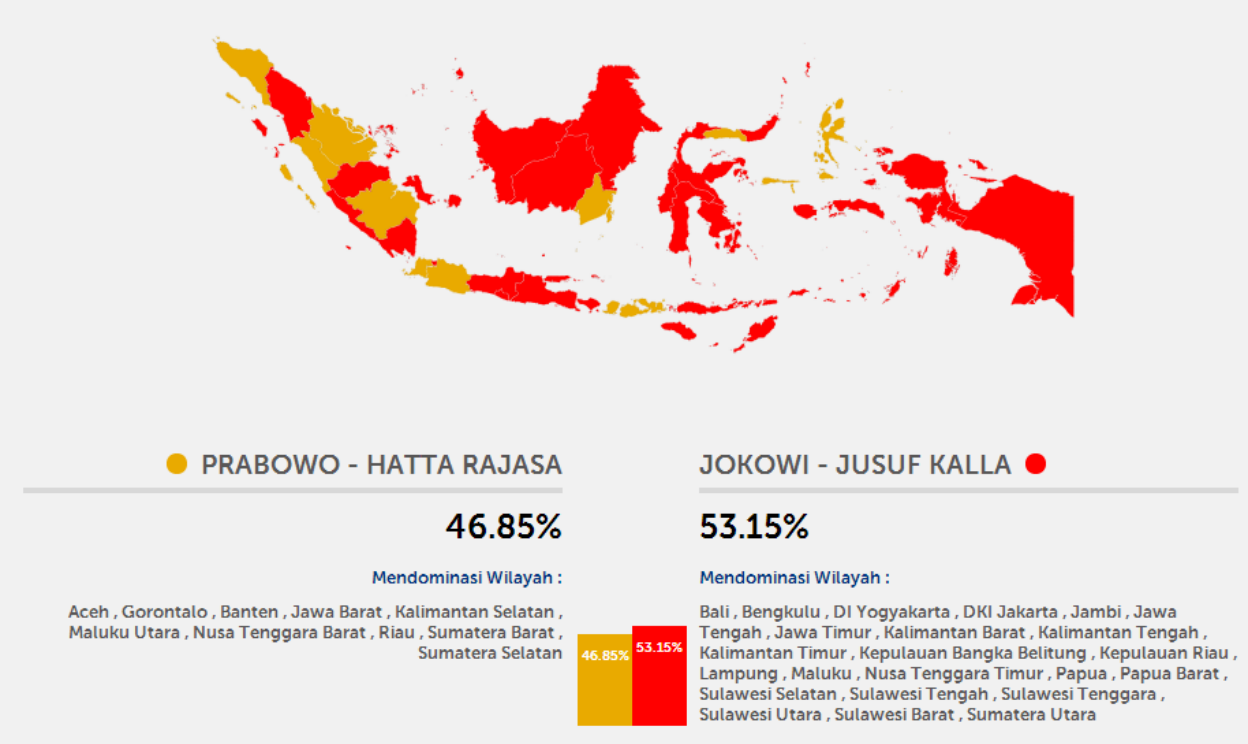

Figure 11 Election Result per Province[22]

As seen in Table 16, Indonesia is divided into 33 provinces. The population/voters in the provinces are differ greatly, for example, in Jawa Barat, there are 23 million voters while in Papua Barat have only 500 thousand voters. The first candidate, the losing one, wins in 10 provinces and the second candidate, wins in 23 provinces. In some provinces, such as Sumatera Barat and Nusa Tenggara Barat, the first candidate wins with more than 70%, while in the other hand, the second candidate also wins more than 70% in some provinces such as Bali and Sulawesi Selatan. Those facts will be interesting to be analyzed further in

[21] http://kpu.go.id/koleksigambar/PPWP_-_Nasional_Rekapitulasi_2014_-_New_-_Final_2014_07_22.pdf

[22] http://news.detik.com/pemilu2014/realcountpilpres

the next section. Another interesting fact is that the winner is supported by a coalition of parties that in total have 35.62% vote in legislative election early this year, while the other candidate is supported by a coalition of parties that have 52.73% vote. There are also several political parties that decided not to join any coalition.

Province	Prabowo-Hatta		Jokowi-JK	
	%	Vote	%	Vote
Aceh	54,39%	1089290	45,61%	913309
Sumatera Utara	44,76%	2831514	55,24%	3494835
Sumatera Barat	76,92%	1797505	23,08%	539308
Riau	50,12%	1349338	49,88%	1342817
Jambi	49,25%	871316	50,75%	897787
Sumatera Selatan	51,26%	2132163	48,74%	2027049
Bengkulu	45,27%	433173	54,73%	523669
Lampung	46,93%	2033924	53,07%	2299889
Kepulauan Bangka Belitung	32,74%	200706	67,26%	412359
Kepulauan Riau	40,37%	332908	59,63%	491819
DKI Jakarta	46,92%	2528064	53,08%	2859894
Jawa Barat	59,78%	14167381	40,22%	9530315
Jawa Tengah	33,35%	6485720	66,65%	12959540
DI Yogyakarta	44,19%	977342	55,81%	1234249
Jawa Timur	46,83%	10277088	53,17%	11669313
Banten	57,10%	3192671	42,90%	2398631
Bali	28,58%	614241	71,42%	1535110
Nusa Tenggara Barat	72,45%	1844178	27,55%	701238
Nusa Tenggara Timur	34,08%	769391	65,92%	1488076
Kalimantan Barat	39,62%	1032354	60,38%	1573046
Kalimantan Tengah	40,21%	468277	59,79%	696199
Kalimantan Selatan	50,05%	941809	49,95%	939748
Kalimantan Timur	36,62%	687734	63,38%	1190156
Sulawesi Utara	46,12%	620095	53,88%	724553
Sulawesi Tengah	45,17%	632009	54,83%	767151
Sulawesi Selatan	28,57%	1214857	71,43%	3037026
Sulawesi Tenggara	45,10%	511134	54,90%	622217
Gorontalo	63,10%	378735	36,90%	221497
Sulawesi Barat	26,63%	165494	73,37%	456021
Maluku	49,48%	433981	50,52%	443040
Maluku Utara	54,45%	306792	45,55%	256601
Papua	27,51%	769132	72,49%	2026735
Papua Barat	32,37%	172528	67,63%	360379
Internasional	46,26%	313600	53,74%	364257

Table 16 Election Result per Province

Offline survey before the election

We need to compare the prediction using Twitter with the current common prediction, offline polls. There are a lot of institutions who published the result of their polls. On Table 17, we compiled the offline poll results that collect sample in the country/national level. All of the pollsters stated that they randomly select samples in several random provinces. This method is called Multistage Random Sampling. The number of sample used in the polls affect the margin of error in each poll.

Institution	Candidate Jokowi	Candidate Prabowo	Swing Voter	DataCollection Start	DataCollection End	Num. of Sample	Margin of Error
LSI Network	35.42	22.75	41.83	20140501	20140509	2400	2.00%
Alvara Research Center	38.8	29	32.2	20140518	20140528	1440	2.64%
Populi Center	47.5	36.9	15.6	20140524	20140529	1500	2.53%
Cyrus Network	53.6	41.1	5.3	20140525	20140531	1500	2.60%
PDB	32.2	26.5	41.3	20140521	20140601	2688	5.00%
Pol Tracking	48.5	41.1	10.4	20140526	20140603	2010	2.19%
SSSG	42.65	28.35	29	20140526	20140604	1250	2.78%
Indobarometer	49.1	36.5	14.4	20140528	20140604	1200	3.00%
LSN	38.8	46.3	14.9	20140601	20140608	1070	3.00%
LSI Network	45	38.7	16.3	20140601	20140609	2400	2.00%
PDB	29.9	31.8	38.3	20140606	20140611	1200	2.80%
Kompas	42.3	35.3	22.4	20140601	20140615	1950	2.20%
Roy Morgan Research	52	48	0	20140615	20140615	3117	1.80%
Vox Populi	37.7	52.8	9.5	20140603	20140615	4898	1.80%
Median	44.3	46.2	9.5	20140615	20140620	2200	2.10%
IRC	43	47.5	9.5	20140614	20140620	1200	2.80%
ISI	45.75	54.25	0	20140615	20140621	999	3.00%
Puskaptis	42.79	44.69	12.52	20140616	20140621	2400	2.00%
LIPI	43	34	23	20140605	20140624	790	3.51%
LSI Network	43.5	43	13.5	20140620	20140625	2400	2.00%
LSN	39.9	46.6	13.5	20140623	20140626	1070	3.00%
PolcoMM	45.3	46.8	7.9	20140623	20140627	1200	3.10%
IDM	34.4	48.7	16.9	20140622	20140630	3324	1.80%
PDB	32.3	40.6	27.1	20140623	20140701	1090	3.00%
INES	37.6	54.3	8.1	20140625	20140702	7000	1.31%
LSI Network	47.8	44.2	8	20140702	20140705	2400	2.00%
SSSG	51	43.4	5.6	20140621	20140705	1250	2.78%
ISI	46.9	53.1	0	20140702	20140706	999	3.00%

Table 17 Offline polling results

On Table 17, we can see that the number of samples used by most of the polling institution corresponds with the margin of error. Based on Equation 1, using more sample means the margin of error if the poll use the same confidence interval. Several polling institution such

as PDB might use different method to calculate their margin of error because their margin of error is quite high.

As seen on the table, though the polling are conducted relatively at the same time, the result are vary greatly between each pollster. The differences are greater than the margin of error of the polling. In Figure 12 and Figure 13, poll results are divided based on which candidates is leading. Based on those figures, we know that several poll institutions always favouring candidate Prabowo, while other institutions always favouring the other candidate. This condition is in line with the explanation in (Hitchens, 2009), where it argued that polls are actually a device for influencing public opinion. As explained in 2, the result from election poll in Indonesia had high difference, more than the margin of error, started in presidential election on 2009 and became worse in governor election on 2012 and legislative election in 2014.

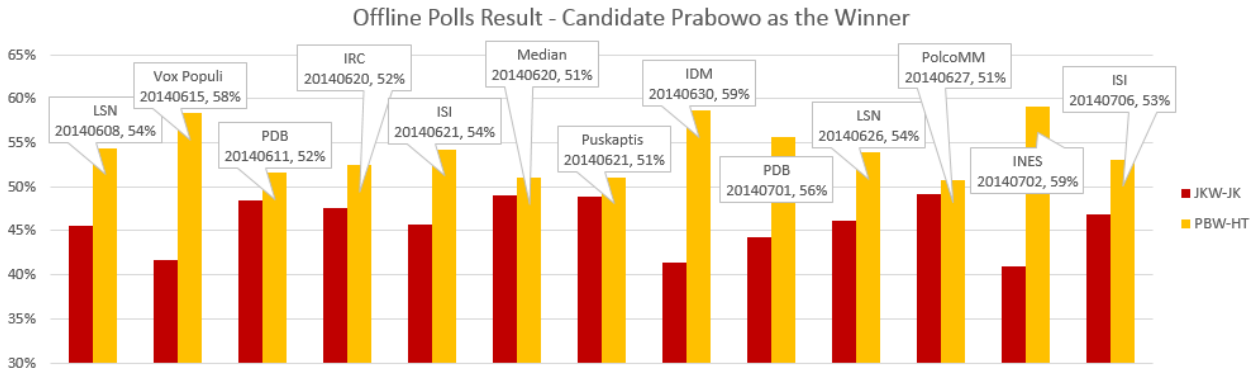

Figure 12 Election polls in favour of candidate Prabowo

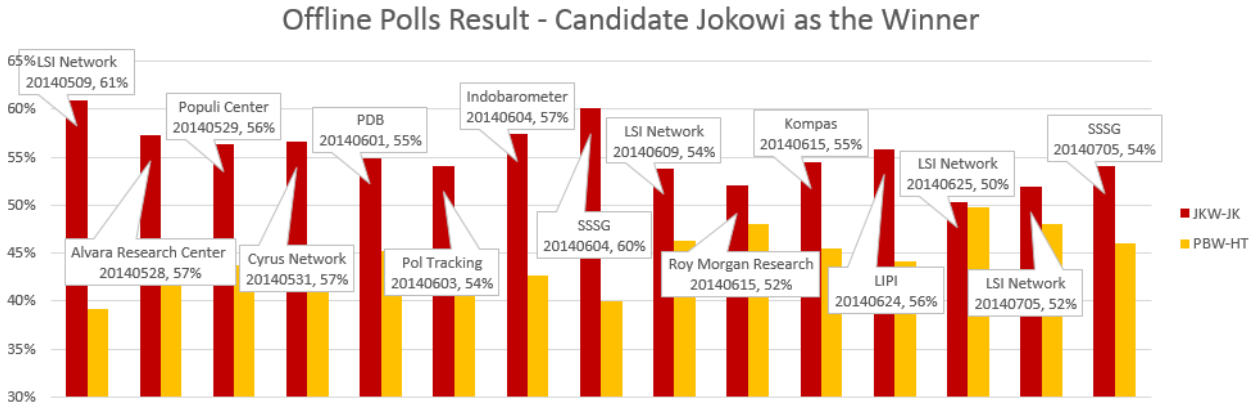

Figure 13 Election polls in favour of candidate Jokowi

5.2 Results from Previous Research

Researchers had conducted experiments in this research domain previously and their results were published in journals. Those results are compiled and used as a comparison for our prediction result. The selected experiments are the experiments that use similar methods such as counting tweets, counting user, sentiment analysis, and population normalization. Their prediction results were compared to the real election results using Mean Absolute Error. Their data collection periods ranged from 1 year to 4 days. One experiment used as few as 7 thousand tweets as their data while another use 50 million tweets as their data. The number of twitter users detected also ranged from 6 thousand to 195 thousand. Several experiments produced predictions that have less than 1% error, but there are a prediction that has 39% of error too. Our compilation has been listed before in Chapter 3 at Table 4 and more detailed information can be seen at Table 34 in Appendix A.

5.3 Testing the Hypotheses

In this section, the result of the Twitter-based prediction that has been explained in Chapter 4 is discussed. Several hypotheses listed in Table 15 are used as the basis to analyze the prediction result. All hypotheses will be tried to be answered in this section. Most of the prediction will be compared to the real election result, then evaluated on how large the gap is between the prediction and the real result or called the Mean Absolute Error.

5.3.1 Hypothesis #1: Tweet Count

"Tweet-based election prediction using baseline method is comparable to offline polling in term of error/distance with the real election result."

This hypothesis aims to answer whether the baseline method of counting tweets is comparable to offline polls or not. We select the last poll conducted by each polling institution to be compared with the result from tweet based prediction. The data used in this hypothesis is only 1 day of data (the last day in the POLDATA). In that data, we count the tweets that mention each candidate name or keywords. Then calculate the MAE of this prediction. Besides calculating the prediction in country level, we also divide the prediction based on its location and calculate their MAE.

Result on National Level:

	No of Tweet	% of Tweet	% Vote	MAE
Candidate Jokowi	109172	56.45%	53.15%	3.30%
Candidate Prabowo	84228	43.55%	46.85%	

Table 18 Tweet Count Result

In national level, Table 18 show that the prediction correctly predict the winner and the MAE of the counting tweets mentioning candidates are at 3.30%. From the offline polls list at Table 17, the MAE of each polling institution are also calculated. Then by comparing them, we can see that tweet based prediction has lower MAE than 13 survey institutions but greater MAE than 7 institutions. The complete list can be seen at Table 19.

Institution	Mean Absolute Error
SSSG	0.9%
Pol Tracking	1.0%
Roy Morgan Research	1.2%
LSI Network	1.2%
Kompas	1.4%
LIPI	2.7%
Populi Center	3.1%
Count of Tweet	**3.3%**
Cyrus Network	3.4%
PolcoMM	4.0%
Alvara Research Center	4.1%
Median	4.2%
Indobarometer	4.2%
Puskaptis	4.2%
IRC	5.6%
ISI	6.3%
LSN	7.0%
PDB	8.8%
Vox Populi	11.5%
IDM	11.8%
INES	12.2%

Table 19 Mean Absolute Error of Counting Tweets and Offline Polling

Result on Province Level:

For each province, comparison was done between the prediction result per province and the real result per province. The prediction correctly predicts the winner at 23 provinces from total 33 provinces, with mean absolute error varies between 0.2 % until 26%.

Though counting tweets does not correctly predict the winner in all provinces, it can be seen that in provinces with great margin, the winning candidates have more than 70% votes (Sumatera Barat, Nusa Tenggara Barat, Bali, Sulawesi Selatan, Papua, and Sulawesi Selatan), counting tweets correctly predict the winner. In provinces where the margin is relatively low (the winner have less than 55% votes), the tweet counting correctly predict the winner in 8 provinces and has incorrect prediction in 7 provinces. We can also see that, in general, candidate Jokowi has more support in Twitter because in 10 provinces where

candidate Prabowo actually wins, the prediction from tweet counting only correctly predict the winner in 3 provinces.

Province	Prabowo-Hatta		Jokowi-JK		Winner Predicted	Mean Absolute Error
	Election Result (%)	Tweet Count	Election Result (%)	Tweet Count		
Aceh	54.39%	49.11%	45.61%	50.89%	Incorrect	5.28%
Sumatera Utara	44.76%	43.08%	55.24%	56.92%	Correct	1.68%
Sumatera Barat	76.92%	50.83%	23.08%	49.17%	Correct	26.09%
Riau	50.12%	49.85%	49.88%	50.15%	Incorrect	0.27%
Jambi	49.25%	47.62%	50.75%	52.38%	Correct	1.63%
Sumatera Selatan	51.26%	43.36%	48.74%	56.64%	Incorrect	7.90%
Bengkulu	45.27%	49.05%	54.73%	50.95%	Correct	3.78%
Lampung	46.93%	51.51%	53.07%	48.49%	Incorrect	4.58%
Kepulauan Bangka Belitung	32.74%	30.89%	67.26%	69.11%	Correct	1.85%
Kepulauan Riau	40.37%	48.79%	59.63%	51.21%	Correct	8.42%
DKI Jakarta	46.92%	39.96%	53.08%	60.04%	Correct	6.96%
Jawa Barat	59.78%	46.91%	40.22%	53.09%	Incorrect	12.87%
Jawa Tengah	33.35%	43.38%	66.65%	56.62%	Correct	10.03%
DI Yogyakarta	44.19%	43.56%	55.81%	56.44%	Correct	0.63%
Jawa Timur	46.83%	46.08%	53.17%	53.92%	Correct	0.75%
Banten	57.10%	42.50%	42.90%	57.50%	Incorrect	14.60%
Bali	28.58%	42.86%	71.42%	57.14%	Correct	14.28%
Nusa Tenggara Barat	72.45%	52.22%	27.55%	47.78%	Correct	20.23%
Nusa Tenggara Timur	34.08%	11.06%	65.92%	88.94%	Correct	23.02%
Kalimantan Barat	39.62%	47.97%	60.38%	52.03%	Correct	8.35%
Kalimantan Tengah	40.21%	30.77%	59.79%	69.23%	Correct	9.44%
Kalimantan Selatan	50.05%	45.16%	49.95%	54.84%	Incorrect	4.89%
Kalimantan Timur	36.62%	44.16%	63.38%	55.84%	Correct	7.54%
Sulawesi Utara	46.12%	44.94%	53.88%	55.06%	Correct	1.18%
Sulawesi Tengah	45.17%	45.99%	54.83%	54.01%	Correct	0.82%
Sulawesi Selatan	28.57%	41.05%	71.43%	58.95%	Correct	12.48%
Sulawesi Tenggara	45.10%	65.77%	54.90%	34.23%	Incorrect	20.67%
Gorontalo	63.10%	41.77%	36.90%	58.23%	Incorrect	21.33%
Sulawesi Barat	26.63%	46.15%	73.37%	53.85%	Correct	19.52%
Maluku	49.48%	65.57%	50.52%	34.43%	Incorrect	16.09%
Maluku Utara	54.45%	62.50%	45.55%	37.50%	Correct	8.05%
Papua	27.51%	30.77%	72.49%	69.23%	Correct	3.26%
Papua Barat	32.37%	41.03%	67.63%	58.97%	Correct	8.66%

Table 20 Tweet counting result

Findings:

From above result, we can conclude that in national/country level, tweet counting is comparable to offline polls where the margin of error usually ranged from 1% to 5% and especially in Indonesia where the error can be as high as 10%. This also in line with previous research, (Ceron A. C., 2014), (Jensen, 2013), (Sanders, 2013), (Gaurav, 2013), and (Bermingham, 2011), where the MAE from only counting tweets are between 2% and 19%.

In the province level, counting tweets cannot correctly predict the winner in all provinces. And in 5 provinces, the MAE are bigger than 20%. This result also in line with previous research such as in (Gayo-Avello D. M., 2011) and (Choy M. C., 2012). In their experiment, the MAE can be as high as 39% and tweet counting also cannot have correct prediction in all provinces. For this result, we argue that there are several items that could be the reason, first is that it is normal to incorrectly predict the winner in an election with a close result between the candidates. Second, the second candidate, Joko Widodo, was more popular in Twitter. Third, different Twitter demographic and the number of Twitter used affect the result in each province. This will be answered in Hypothesis #9: The Number of User. Other than that, the methods that can increase the prediction accuracy will be analyzed in the next hypotheses.

5.3.2 Hypothesis #2: Keyword Selection

"Using more keywords describes the situation better in Twitter and in turn improves the accuracy of the prediction."

The keywords' selection plays an important part in the model because it decides what data that will be used in the prediction. In early research in this area such as (Tumasjan, 2010) and (Choy M. C., 2011), the authors used only the parties or candidates name. Later research such as (Wong F. T., 2013) and (Gaurav, 2013) incorporated more keywords related to the election.

To answer this hypothesis, the keywords are reduced then the prediction accuracy is compared with the baseline method. First, we reduce the keywords into 1 keyword per candidate using the candidates' popular name, 'prabowo' for candidate Prabowo Subianto and 'jokowi' for candidate Joko Widodo. We also select 5 random keywords per candidate and re-calculate the prediction.

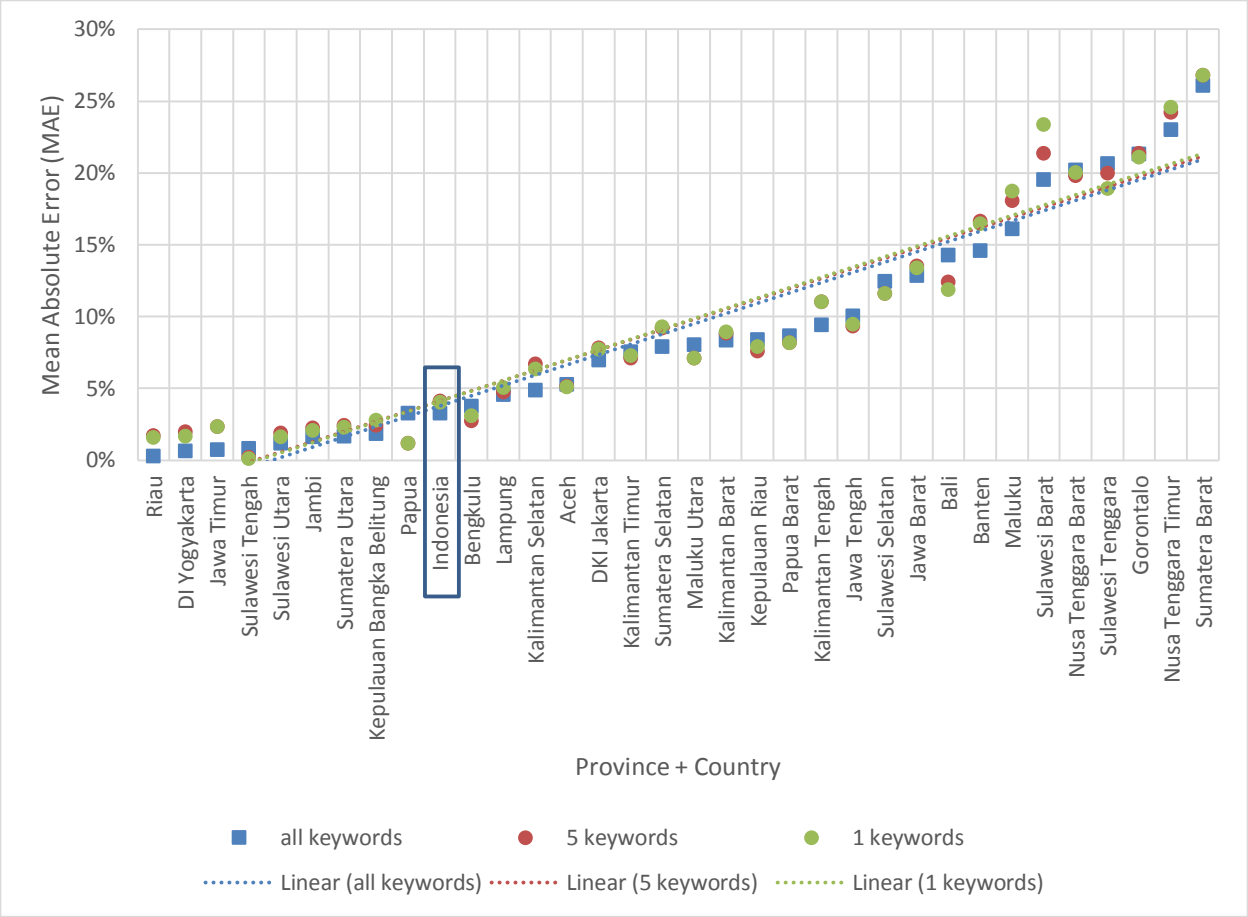

Figure 14 Changes of MAE when using different keywords

Province	MAE 2 Keywords	MAE 10 Keywords	MAE All Keywords
Indonesia	4.05%	4.12%	3.30%
DKI Jakarta	7.75%	7.81%	6.96%
Jawa Barat	13.38%	13.52%	12.87%
Jawa Tengah	9.46%	9.32%	10.03%
Banten	16.48%	16.64%	14.60%
Sumatera Utara	2.30%	2.42%	1.68%
DI Yogyakarta	1.69%	1.97%	0.63%
Jawa Timur	2.32%	2.36%	0.75%
Riau	1.56%	1.73%	0.27%
Sumatera Selatan	9.31%	9.19%	7.90%
Lampung	5.06%	4.76%	4.58%

Table 21 MAE of prediction using different keywords in 10 provinces with most users

Result and Findings:

Using only 1 keyword per candidate (2 keywords in total), the data is reduced by 5 thousands of tweet and about 4 thousands of user. That numbers equals to about 2.6% of total tweets

and 5.7% of total users. If we compare it with the result from counting the tweets using all of the keywords, this prediction accuracy in national level decrease from 3.30% to 4.12%. In Table 21, we can see same results in 9 of 10 provinces with most Twitter users.

The comparison of the prediction accuracy using different keywords in all provinces can be seen at Figure 14. From that figure, we can see that using 10 keywords (5 words per candidate) are slightly better than 2 (1 word per candidate) keywords and using all of the keywords are better than 10 keywords. In provinces such as Sulawesi and Sumatera the names/popular names of the vice president candidates are more influential because that's where they come from. But in provinces such as Jawa Tengah and Bali, hashtags such as 'save indonesia', 'mental revolution', or '2 finger greeting' (a theme song created by many famous artist) have more influence in this hypothesis. The detailed result can be seen in Appendix B at Table 37.

5.3.3 Hypothesis #3: Duration of the data

"Using several days of data is better than only use the data from the last day before the election."

From previous hypothesis, we know that using more data can improves the accuracy of the prediction. In this hypothesis, we want to know whether adding more data from previous days can improves the prediction or not. This issue was also one of the main concerns in (Gayo-Avello D. , 2013). Besides the data duration, we also need to consider how to combine data from several days. First, we can use moving average, where we can calculate the average percentage of the prediction in each day, second is by aggregating all data from several days, and then calculate the percentage for each candidate.

		1 day	7 days	14 days	21 days	30 days
Tweet Count	MAE Moving Average 7-days	3.30%	5.37%	8.50%	7.28%	6.11%
	MAE Aggregate	5.34%	10.22%	4.92%	2.09%	6.20%
User Count	MAE Moving Average 7-days	0.60%	5.37%	2.97%	1.52%	0.08%
	MAE Distinct User	1.30%	1.42%	3.62%	0.57%	0.02%

Table 22 Prediction results using more than 1 day of data

Result and Findings:

As shown in Table 22, we can see that using data further back from the time of election doesn't necessarily reduce the error/deviation from the real election result. The error increase when we use 7 days and 14 days data, then decrease when more data is used. This results are in line with the twitter trend shown in Figure 15. In those figures, we can see that the gap between the candidates are widening starting about 20 days before the election. Similar result happened in 2009 German election where (Jungherr, 2012) got 2.7% of MAE

using the last day of data and (Tumasjan, 2010) got 1.6% of MAE when using data 1 week prior to the election. In our case, the prediction will be better if we use 20 days of data before the election. But this is just a mere chance, it doesn't mean that using prior data can improve the prediction result.

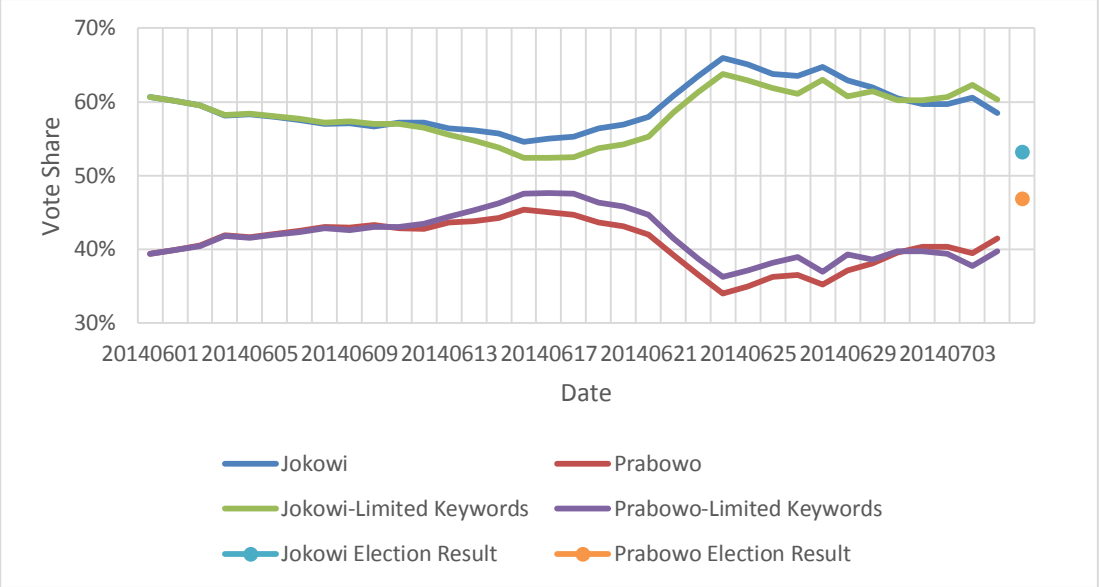

Figure 15 Daily prediction result based on tweet counting

5.3.4 Hypothesis #4: Spam Filtering

"Filtering the data increase the prediction accuracy."

In (Waugh, 2013), the author conclude that during election time, there are a large number of automated and non-trustworthy users who posted or retweeted messages to support the candidates. In previous chapter, we have explained how we check for spam in the Twitter data. With this hypothesis, we try to create a prediction using that filtered data.

	All Data	Filtered Data	%
Num of Tweet	193394	30920	15.99%
Num of Users	70184	2889	4.12%

Table 23 the Number of Filtered Users

Candidate	Negative	Positive	Neutral
Prabowo	1504	6859	12874
Jokowi	1709	7987	

Table 24 Sentiment Analysis of the Filtered Users

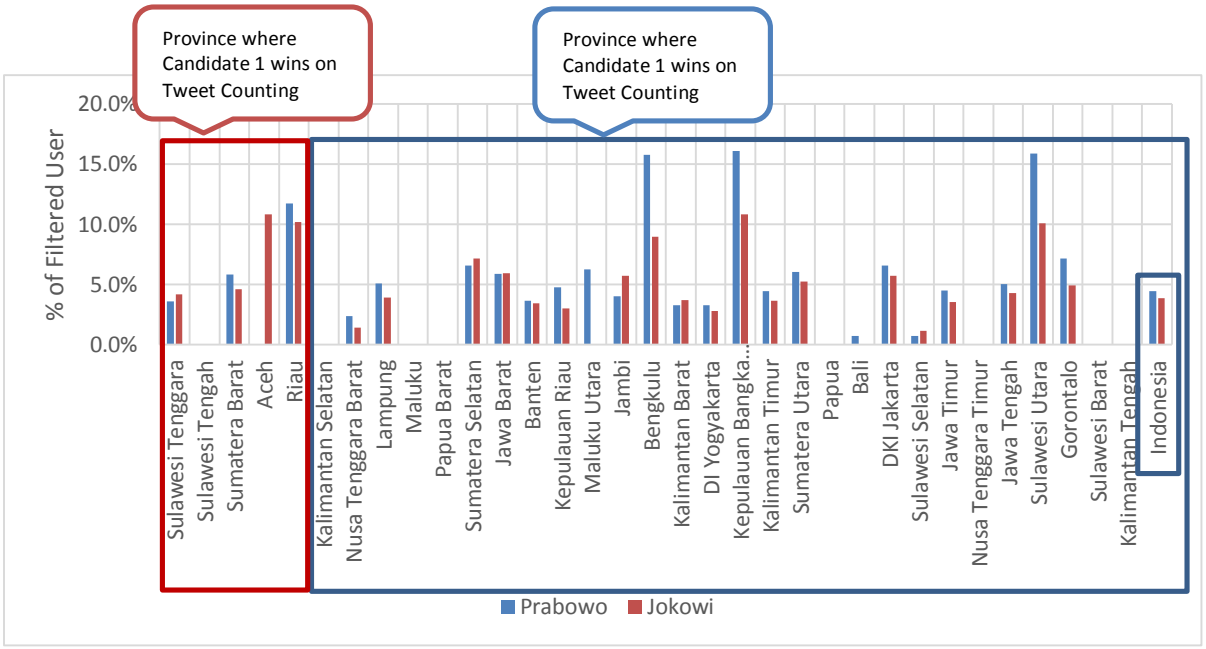

Figure 16 Percentage of Filtered Users per Province

Province	Filtered Users		Winner Predicted	MAE Spammer	Mean Absolute Error	Differences
	Prabowo-Hatta	Jokowi-JK				
Indonesia	1416	1473	Correct	1.44%	1.30%	-0.15%
DKI Jakarta	381	433	Correct	3.62%	3.41%	-0.21%
Jawa Barat	233	257	Incorrect	12.03%	12.05%	0.01%
Jawa Tengah	93	106	Correct	9.38%	9.56%	0.18%
Banten	51	54	Incorrect	10.02%	9.97%	-0.05%
Sumatera Utara	50	53	Correct	0.06%	0.27%	0.21%
DI Yogyakarta	30	30	Correct	1.51%	1.64%	0.13%
Jawa Timur	30	31	Correct	3.85%	3.60%	-0.25%
Riau	54	46	Correct	0.00%	0.43%	0.43%
Sumatera Selatan	22	26	Incorrect	3.03%	3.20%	0.16%
Lampung	15	12	Correct	1.85%	2.15%	0.31%

Table 25 MAE of prediction after data filtering in top 10 provinces with most users

Result and Findings:

In the whole data, we are able to identify about 6 thousands of 'spam' users out of 500 thousands of users or about 1.2% of all users. Many of those users were active in the last day of our data. From about 70 thousands of users, the number of filtered users are almost 3 thousands or are about 4.1%. But these users post about 16% of the tweets, as seen in Table 23.

We can see from Figure 16 that in most of the provinces, spammer who support candidate Prabowo are higher than those who support the other candidate. But there are also some

provinces where spammer supporting candidate Jokowi are higher, such as in Aceh and Kalimantan Barat. The number of spammer doesn't make the candidate that they support have greater tweet in their province. For example, in Bengkulu and Kep. Bangka Belitung, there are many spammer who support candidate Prabowo, but candidate Jokowi still had more tweet and more user in those provinces. From Table 24, we know that the filtered user post positive tweets much more than negative tweets.

The detailed prediction for all provinces created using this filtered data can be seen in Appendix B at Table 41 while the prediction for top 10 provinces can be seen in Table 25. The predictions improves in 7 out of 10 provinces and in all provinces, 20 out of 33 provinces. This results are interesting as we have not seen other research that compare prediction results before and after filtering their data.

5.3.5 Hypothesis #5: User Count

"Counting the user instead of counting the tweet improves the prediction result."

Based on the assumption that one user represent one vote in the election, we should count the users rather than the tweets, because no matter how much tweets a user post, he/she only has one vote. A user is considered to favor one candidate if he/she mentions that candidate more often than other candidates. Same as previous hypothesis, we use 1 day of data for the prediction, so when a user does not post/mention a candidate on that day, his/her vote will not be included in the calculation.

Result on National Level:

It predicts the winner correctly with Mean Absolute Error of 1.295%. This result MAE improve/reduce about 2% compared to the result from counting the tweets. This prediction has lower MAE than 16 survey institutions and greater MAE than 4 institutions.

Institution	Mean Absolute Error
SSSG	0.9%
Pol Tracking	1.0%
Roy Morgan Research	1.2%
LSI Network	1.2%
Count of User	**1.3%**
Kompas	1.4%
LIPI	2.7%
Populi Center	3.1%
Cyrus Network	3.4%
PolcoMM	4.0%
Alvara Research Center	4.1%
Median	4.2%
Indobarometer	4.2%

Puskaptis	4.2%
IRC	5.6%
ISI	6.3%
LSN	7.0%
PDB	8.8%
Vox Populi	11.5%
IDM	11.8%
INES	12.2%

Table 26 Mean Absolute Error of Counting Users and Offline Polling

Result on Province Level:

The prediction correctly predicts the winner at 24 provinces from total of 33 provinces. Increased from tweet counting where it predicts 23 provinces correctly. The MAE in all provinces ranged between 0.2% and 25%. The detail result can be seen in B. The predictions' MAE decrease in 18 provinces and increases in 15 provinces.

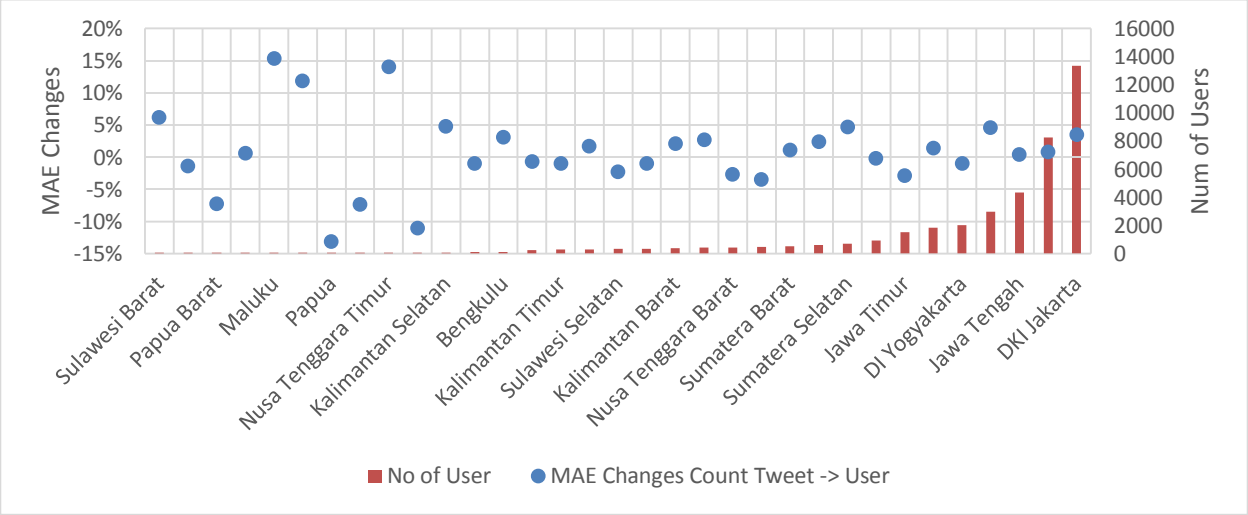

Figure 17 Relation between Changes in MAE and the Number of Users

Findings:

In the national level, counting user increase the accuracy of the prediction from MAE 3.298% to 1.295%. This result is similar to the experiment conducted in (Gaurav, 2013) and (Sang, 2012). In those research, the author were able to decrease the MAE in Venezuela presidential election from 2.0% to 0.5%, and in Ecuador presidential election from 19% to 3%. In the province level, counting user does not improve the prediction accuracy in all province but when aggregating to national level, the accuracy does improve. This was because in province DKI Jakarta and Jawa Barat (where more than 50% of Twitter users lived), counting users do decrease the MAE. We also try to find a correlation between the detected user in our data and the changes between MAE from counting tweets and MAE

from counting users. In Figure 17, we can see that in provinces that has many users, MAE from counting users are lower than MAE from counting tweets. But in provinces with less users, the changes in MAE randomly increase or decrease.

5.3.6 Hypothesis #6: Population Weight

"Weighting each tweet based on the location improves the prediction result."

Users' location is one of the causes of data bias/sampling error in the dataset. As seen in previous chapter, the number of Twitter users does not correspond with the number of population in a province. We will give a weight for each chapter based on the user's location. With the same number of user/tweet, user/tweet located in a populated area have higher weight than located in a less populated area and the higher the number of user/tweet in an area, the lower the weight.

Equation to calculate the weight of each tweet:

$$Tweet\ Weight\ in\ Province\ (x) = \frac{All\ Tweet}{Tweet\ in\ Province\ (x)} \times \frac{Population\ in\ Province\ (x)}{All\ Population}$$

Equation 10 Weight Calculation for Tweet in a Province

$$User\ Weight\ in\ Province\ (x) = \frac{All\ User}{User\ in\ Province\ (x)} \times \frac{Population\ in\ Province\ (x)}{All\ Population}$$

Equation 11 Weight Calculation for User in a Province

	Election Result	Count Tweets	Weighted Tweets	Count Users	Weighted Users
Candidate Prabowo	46.85%	43.55%	44.86%	45.55%	45.74%
Candidate Jokowi	53.15%	56.45%	55.14%	54.45%	54.26%
Prediction MAE		3.30%	1.99%	1.30%	1.11%

Table 27 Prediction Result with Population Weight

Result and Findings:

The detail result of population weight can be seen at Table 38 for tweet count with population weight and at Table 39 for user count with population weight, and the summary of the result for national level is shown in Table 27. Population weight improve the accuracy in both prediction; Tweet count MAE decreased from 3.30% to 1.99% and user count MAE decreased from 1.30% to 1.11%. This result correspond with previous research, (Choy M. C., 2011), that show census correction did improve their prediction in Singaporean election. The same author then tried to predict election in the USA, (Choy M. C., 2012), but in this

research, census correction did not improve their non-debiased prediction. They argued that, one of the reason was because "*The political tweets were not collected on a geographical basis*".

5.3.7 Hypothesis #7: Demographic of Twitter User

"Incorporating the users' demographic information reduce the data bias and improve the prediction result."

Many researchers, such as (Fink, 2013) and (Gayo-Avello D. , 2013), argue that data bias can affect the predictive *skills* of social media. In fact, they argue that sample bias is one of the fundamental challenges of using social media for political analysis. If we compare to offline polling, there are several item that we should consider:

- *Sample selection*. In offline polls, one sample is counted only once. In Tweet counting, a user can post many messages and will be counted several times. For this bias, we have normalized this by counting a user only once. The result is discussed in Hypothesis #5: User Count.
- *Location of the sample.* As explained in Chapter 2, offline polls sample use Multi Stage Random Sampling where they randomly select several districts in a province, then continue by selecting random houses from those districts. But in Twitter, we cannot limit the user who post a message based on its location. So we give different weights to each tweet based on its location so that the tweets can represent the number of population in the province. This bias has been discussed before in Hypothesis #6: Population Weight.
- *Age of the sample*. In each houses of their sample, offline polls randomly select one sample; it could be the grandfather/parent/the teenage children. In our Twitter data, younger sample are more than the elders. In Chapter 4, we plan to classify the user based on their age, then give each tweet a corresponding weight. We do not continue as we only have low precision from the classifier.
- *Gender of the sample*. The gender distribution in each province in Indonesia is different than in Twitter. Twitter has much more male user than female. Using name list, we have classified one third of all our data, explained in Chapter 4. Then using only the classified data, a new prediction is created. The weight calculation are based on Equation 12.

$$Weight\ of\ User\ with\ gender\ (y)\ in\ Province\ (x) = \frac{User\ in\ (x)}{User\ (x,y)} \times \frac{Population\ (x,y)}{Population\ (x)}$$

Equation 12 Weight Based on Gender

	Male	Female	M vs F (%)
All dataset	103952	69848	59.81% vs 40.18%
1 day of data	20520	11716	63.35% vs 36.65%
Weighted data	16231	16005	50.35% vs 49.65%
Candidate Prabowo	9357	5198	64.29% vs 35.71%
Candidate Prabowo Weighted	7401	7101	51.04% vs 48.96%
Candidate Jokowi	11163	6518	63.14% vs 36.86%
Candidate Jokowi Weighted	8830	8904	49.79% vs 50.21%

Table 28 Gender Weighted User

Result:

The classification result in Table 28 show that male user in all of our dataset is about 60% while in the last day is about 63% of all the users sample. After weighting the user, we can see that more male users support candidate Prabowo and more female users support candidate Jokowi. In Indonesia, male population are slightly higher than female population (50.35% vs 49.65%). In both candidate, we can see that the deviation after applying the weight is not high (less than 1%).

In Figure 18 we can see the distribution of the users' gender in each province. We see one anomaly in the Kep. Bangka Belitung Province where the female user are more than the male user. In Sulawesi Barat, we cannot detect any female user and able to detect 4 male users in the last day data.

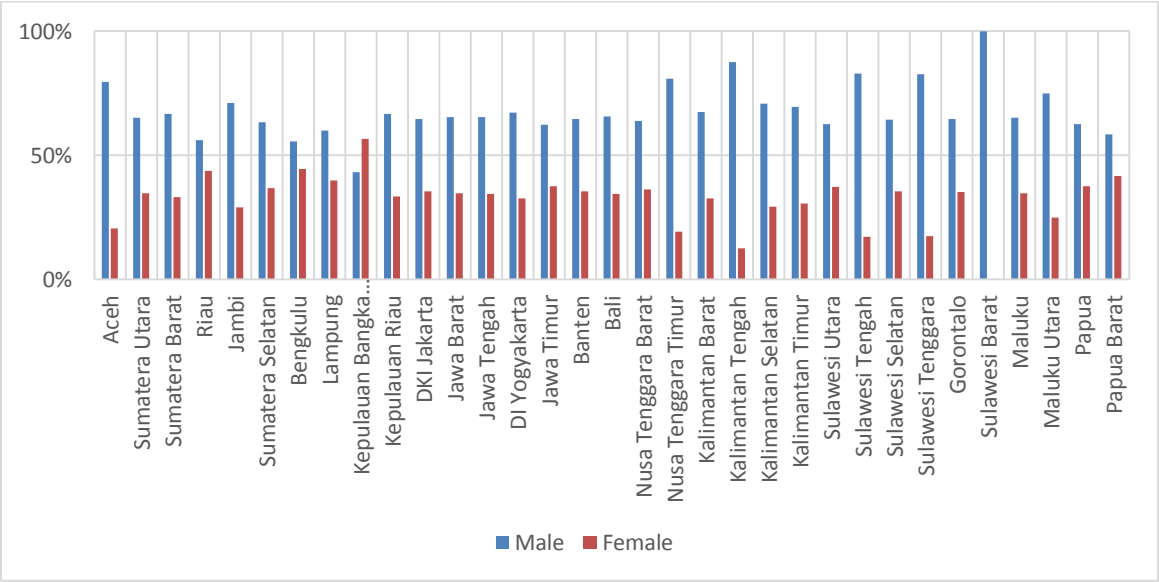

Figure 18 Twitter User in Provinces Classified by Gender

In Appendix B, Table 42, the complete prediction is displayed. In national level, the prediction accuracy decrease from 1.3% to 1.8%. In province level, the prediction accuracy increase in 20 provinces and decrease in 13 provinces.

Findings:

In our experiment results, we cannot conclude that gender normalization improve the accuracy in our prediction because in national level and in 13 provinces the accuracy decreases. Different from user normalization and population weighting, there are not many researchers that incorporated age and gender in their prediction model. In general, (Mislove, 2011) conclude that post-hoc correction could be applied to improve Twitter-based predictions. (Gayo Avello, 2011) used age data in his prediction, and it reduced the prediction MAE from 13.1% to 11.6%.

5.3.8 Hypothesis #8: Sentiment Analysis

"Implementing Sentiment Analysis in the dataset improves the prediction result."

Rather than only counting the tweets, several researchers argued that it is important to understand the sentiment of each tweet whether it has positive sentiment towards the candidate he/she mention in his/her tweet. The methods to understand the tweet's sentiment can be divided into; (1) an affective word list as shown in (Tumasjan, 2010) and (Gayo-Avello D. M., 2011), (2) automated/machine learning sentiment analysis such as (Ceron A. C., 2014), (3) annotators/crowd source like in (Fink, 2013), or (4) by the combination of previous methods (Bermingham, 2011).

This experiment uses automated sentiment analysis to classify the tweets. Because the candidates in this election are only two, besides using only positive tweets, we also use the assumption, that negative sentiment toward a candidates means that the tweet are supporting the other candidates. This assumption was also used in (Gayo-Avello D. M., 2011).

For each tweet we divided the tweets into 5 groups: neutral, positive to candidate 1, negative to candidate 1, positive to candidate 2, and negative to candidate 2. For each user, we take all of his/her tweets in a day, then decide their sentiment based by comparing the number of tweet favoring each candidates. When a user has the same number of tweets with sentiment analysis between candidate 1 and candidate 2, we consider him/her as a neutral user.

Result on National Level:

As seen in Table 29 and Table 30, semantic analysis improves the prediction result from both tweet counting and user counting. We also compare the result when we use only the positive tweets and both positive and negative tweets. While all of them still produce less MAE than the baseline method (count mention in tweets or count the user), we see that

there are much more negative tweet towards candidate Jokowi than towards candidate Prabowo, but the number of user who post negative tweets are more or less the same. This makes the prediction by counting the tweets in this election using only positive tweets have better MAE.

Result on Province Level:

In the top 10 provinces shown in Table 31, the prediction MAEs improve in 6 provinces, but in all of provinces, MAE only improves in 10 province, and reduces in 23 province. While it seems that semantic analysis doesn't produce better predictions in most provinces, note that in province where most of the twitter user lived, DKI Jakarta, the prediction's accuracy increases. The detailed result can be seen in Appendix B at Table 40.

	Cand Jokowi	Cand Prabowo	Cand Jokowi (%)	Cand Prabowo (%)	MAE	Non SA MAE
positive tweet	29287	24971	53.98%	46.02%	0.83%	3.29%
post tweet + other cand's neg tweet	38289	37269	50.67%	49.33%	2.48%	3.29%

Table 29 Prediction from Tweet-Based Semantic Analysis *Neutral Tweets: 117842

	Cand Jokowi	Cand Prabowo	Cand Jokowi (%)	Cand Prabowo (%)	MAE	Non SA MAE
user from positive tweet	28130	23945	54.02%	45.98%	0.87%	1.30%
user from positive tweet + other cand's neg tweet	29120	25033	53.77%	46.23%	0.62%	1.30%

Table 30 Prediction from User-Based Semantic Analysis *Neutral User: 16035

Findings:

In the national level sentiment analysis improve the prediction accuracy. It is in line with most previous research that employed sentiment analysis in the prediction, for example in (Ceron A. C., 2014) the prediction accuracy of an election in USA improve from 17.9% (Counting tweets MAE) to 1.29% (Counting tweets with sentiment analysis) and in Italy from 9.72% to 8.65%.

Province	Detected tweets	MAE Sentiment Analysis	Mean Absolute Error	Changes
DKI Jakarta	30.51%	3.29%	3.41%	0.12%
Jawa Barat	25.40%	12.06%	12.05%	-0.01%
Jawa Tengah	24.52%	9.27%	9.56%	0.30%
Banten	23.78%	11.16%	9.97%	-1.19%

DI Yogyakarta	22.49%	0.40%	1.64%	1.24%
Sumatera Utara	24.12%	-0.51%	0.27%	0.78%
Jawa Timur	23.96%	3.09%	3.60%	0.51%
Riau	27.52%	0.60%	0.43%	-0.17%
Sumatera Selatan	25.11%	4.18%	3.20%	-0.98%
Lampung	20.87%	1.23%	2.15%	0.92%

Table 31 Review of per province Sentiment Analysis

In province level, we cannot conclude that sentiment analysis can reduce the MAE because, the MAE in most provinces increase or in other words the prediction become more inaccurate. This is probably because the classifier cannot distinguish most of the tweets in several We argued that automated sentiment analysis must be conducted differently per province or per city, because the national language, Indonesian, is only used in daily basis at the capital city and other big cities. (Indonesia has more than 700 languages[23]). The improvement of MAE in national level is because the MAEs at the capital city/province, where most of the Twitter users' lived, is reduced.

To confirm this argumentation, we check the sentiment analysis result and find that in several province the sentiment analysis does not perform well. For example, it failed to detect the sentiment in Sulawesi Barat and only detect less than 10% of the tweets in Sulawesi Tenggara and Nusa Tenggara Timur. We then perform language detection using Google language-detection in our dataset. The result shows that 88% of the tweets detected as Indonesian language, and there are many tweets containing unofficial or accented Indonesian detected as other language (polish, Tagalog, Swahili, Somalia, etc.).

Language	Number of tweets
Indonesian	170349
English	9564
Polish	7164
Tagalog	1876
Swahili	1270
Somali	373
Unknown	353
Slovene	330
Estonian	329
Other	1792

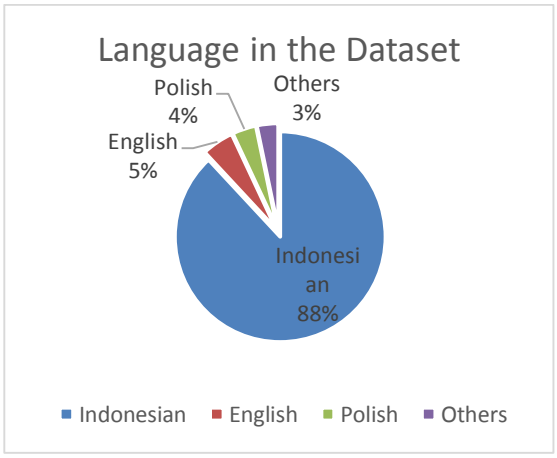

Table 32 Detected language in the dataset

[23] http://en.wikipedia.org/wiki/Languages_of_Indonesia

5.3.9 Hypothesis #9: The Number of User

"The prediction's accuracy has a relation with the number of user in the dataset."

In the statistics of a survey/polling, the estimated margin of error is related to the confidence interval and the number of sample used in the polling. The equation to calculate the margin of error based on simple random samples can be seen in the equation below:

$$Margin\ of\ error = \frac{Z\ (\propto/2)}{2 \times \sqrt{n}}$$

Equation 13 Margin of error in a polling

Where Z is the critical value of a confidence interval using standard normal distribution, and n is the number of sample. In case of a fix confidence interval, we can calculate the margin of error only based on the number of sample. This approach is conducted in (Gayo Avello, 2011), he calculated the margin of error in several states in the US, and showed that the MAE in those provinces (0.42% - 20.34%) are mostly higher than the calculated margin of error (1.46% to 3.87%).

In this hypothesis, we want to know whether there is a correlation between the number of users and the margin of error. We combine our user count per-provinces prediction results and the results from (Fink, 2013), (Gaurav, 2013), and plot it in Figure 19 below.

Result and Findings:

As shown in Figure 19, the errors of the prediction are reduced with the increase of the users. Similar with the result in (Gayo-Avello D. M., 2011), we cannot use offline polls' margin of error calculation (red line at the Figure) to calculate the margin of error of Twitter-based prediction. But we can see that if the prediction only use samples less than 10,000 users, its error can vary from 0% to 25%. These results are in line with (Ceron A. C., 2013) conclusion where they stated that any growth of the information available online improved the predictive skills. In their case, an increase of 1000 in the number of tweets analyzed lowered their error by approximately a quarter point.

From the figure, we see that for the same number of user, twitter based prediction has a lot higher margin of error than offline polling. But tweet based prediction still has the advantages, because, for example, it takes about 1 week to get data from 2000 random sample in an offline polling. While using Twitter, we can get about 200,000 tweets or 70,000 users in a few hours.

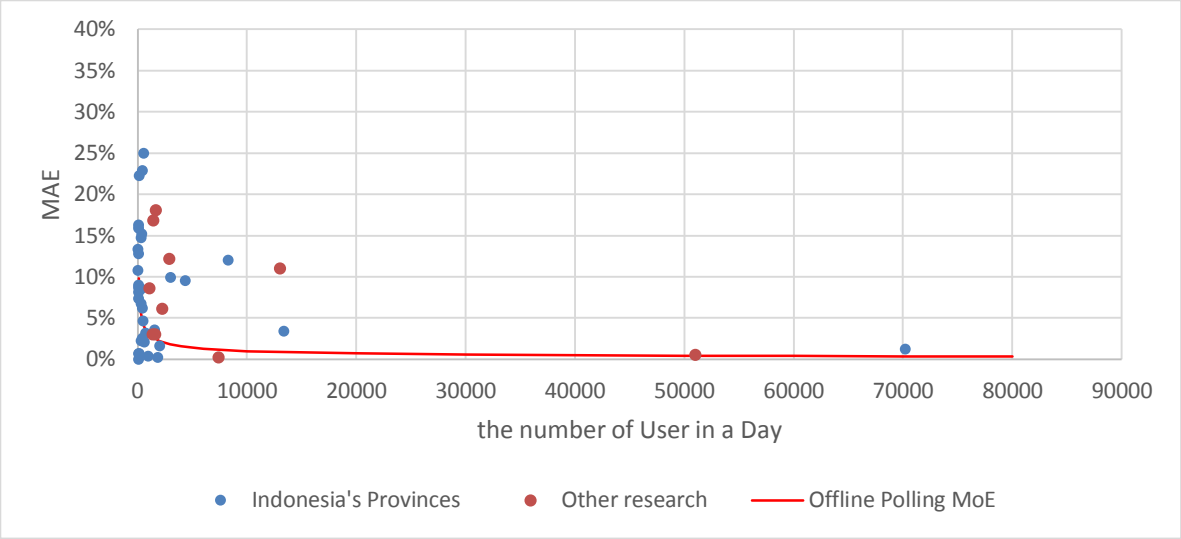

Figure 19 Number of User and Mean Absolute Error

5.4 Summary

The previous calculation results show several points which can be used for future research in this topic. We have summarized the results in Table 33. Using only tweet count, in the national level we get an MAE of 3.3%. For a country like Indonesia, this is a good result when compared to offline polls. But in developed countries such as Germany in (Tumasjan, 2010) or the Netherlands in (Sang, 2012), that number is not enough as the offline polling in those countries are credible and do produce a low MAE.

As for data collection, we show that using more keywords increase the number of data and opinion of the users and in turn improve the prediction result. Although we believe that our method on selecting the keywords is very simple. More sophisticated method is required to completely collect all tweets related to the election. On the duration needed for the calculation for the prediction, we show that the use 1-day of data is still the best practice.

Several methods to handle the user bias are implemented. In data filtering, we explain on how to detect non-personal users and spammers. We detect about 16% of the tweets in the last day were spam. We also try to reduce the user bias by detecting the location and the gender of the users, then weight each user accordingly. The result of location normalization is positive while with gender, the results accuracy in some provinces are improved but it decreases in others.

Implementing user count to accommodate the fact that one person only have one vote reduces the error greatly from 3.3% to 1.3% in national level. In the province level, the MAE randomly increase or decrease at the provinces with small number of samples. Same thing happen when the sentiment analysis is implemented to understand the polarity of the tweet (positive

or negative). In national level, the MAE reduce from 1.3% to 0.6%, while the results vary in provinces level. We argue that the number of samples correlate with this issue. We examine all previous research results and find that the research with a very high number of samples produce a good prediction result with low MAE and research with a low number of samples have varying error.

Hypothesis	Mean Absolute Error	Summary
#H1: Tweet Count	MAE 3.30%	tweet-based prediction accuracy is better than 13 polling institutions
#H2: Duration of Data	-	use 1 day of data better than 7/14/21/30 days
#H3: Keyword Selection		use all keywords better than 1 or 5 keywords Same results in 9 of 10 top provinces
#H4: User Count	MAE 1.30%	user normalization improve the prediction accuracy the accuracy is better than 16 polling institutions increase the accuracy in 7 of 10 top provinces
#H5: Data Filtering	MAE 1.44%	16% of tweets from 4% users are filtered increase the accuracy in 7 of 10 top provinces
#H6: Population Weight	MAE 1.11%	79% of users' location is detected Weighting tweet based on the ratio of user and population increase the accuracy
#H7: Gender Information	MAE 1.7%	29% of users' gender are predicted Weighting tweet based on the gender only improve the accuracy in 5 of 10 top provinces
#H8: Sentiment Analysis	MAE 0.62%	Only use sentiment detected tweets increase the prediction accuracy Same result found in 6 of 10 top provinces
#H9: The Number of User	-	Small dataset produce unpredictable error

Table 33 Summary of the prediction result

6 Conclusions

In this concluding chapter, we will revisit our research questions, and see how they can be answered based on the empirical evidence (i.e., results of the data analysis). The first is a descriptive research question, while the last three can be answered based on the empirical findings. In this section, we will discuss the research questions one by one, including the corresponding hypotheses. The contribution of this research to the domain will be described in the next section. Meanwhile the ideas that were not performed in this experiment become an interesting source for the future research direction.

Research Question 1: How effective is the tweet based election prediction compared to Indonesian offline polls?

Based on the offline polls published by the pollster, we conclude that in the country level tweet-based election can outperform most of the offline polls. With about 200 thousand of tweets and 70 thousand of users, even the simple tweet counting perform better than 13 pollsters (from total 20 pollsters) and user counting is better than 16 pollsters' results. Dividing the data into provinces, we conclude that tweet based prediction is not applicable in all provinces. The prediction incorrectly predict the winner in 9 provinces. One of the factors that we studied is the number of sample. Employing the same prediction model, the MAE can vary from 0.05% to 25.01% in the provinces where the number of users are less than 5 thousand.

Research Question 2: What are the most important factors that influence the result of predicting an election based on Twitter data?

We try to answer the question by comparing it with the established and proven methods of offline polling. An offline polling is conducted by developing clear question to avoid misunderstanding, selecting representative samples, interviewing and data analysis. With Twitter-based prediction, we can map that into extracting user preference by sentiment analysis, normalizing/un-biasing the users, collecting the tweets, and calculating the prediction. In data collection, we emphasize the importance of keyword selection, collection duration, and removing spam from the data. Representative samples can be obtained by normalizing the users by their location, urbanity, age, gender, economic condition, education, and ethnicity. One more issue is to incorporate the opinion of people who do not use Twitter. In Twitter, rather than developing a clear questions for the interview, sentiment analysis of the tweets is performed to understand the vote preference of the users. In Chapter 3, how the twitter-based prediction is performed have been described based on the literatures study.

Research Question 3: What are the differences in utilizing different parameters when collecting Twitter data?

In Chapter 3, we have discussed different methods used by previous researchers for collecting the tweets from functional aspect. By collecting all the tweets in a duration, researchers can conduct many experiments with different parameters to get more tweets related to the election. But, collecting tweets from search API requires much less processing and storing resources. For example, in their work, (Gaurav, 2013) collected 13 billion of tweets in total and found about 500 million of electoral tweets. But only 58 thousand of tweets from 51 thousand users are used. Researchers use different types of keywords to collect electoral tweets, such as names, aliases, and hashtags. In Chapter 5, our results show that utilizing more keywords leads to better accuracy. In term of the number of tweets/users, we conclude that a high number of data tend to give an accurate prediction and only tweets that posted closest to the election are significant. Our results show that using tweets posted several days/weeks prior to the election do not improve the prediction accuracy. We also perform data cleaning in our data set by removing tweets from spammers and from non-personal users. Though it does not directly related to the prediction accuracy, it makes sure that only the real potential voters' opinion are calculated in the prediction.

Research Question 4: Can the accuracy of Twitter-based election prediction be increased by incorporating the users' demographic information and the tweets' sentiment information?

Twitter users are not representative to the real population. Based on our data, Twitter users who post electoral tweets mostly lived in Java Island, especially in the big cities such as Jakarta and Bandung. Other important demographic biases, based on Pew research, are age and gender, while other information such as education and economic situation do not have much influence. We have able to identify the location (cities/district, then mapped to province) from about 78% of 490 thousand of users and the gender of 29% of the users. Our result in Section 5.3 shows that incorporating location improve the accuracy of the prediction, but incorporating gender information slightly reduce the accuracy. This might be caused by our inability to identify most of the users' gender.

Understanding sentiment information in the analysis and calculation is important because not every tweet mentioning a candidate/party name means the user support that candidate. A user could also post negative opinion about that candidate. Almost all researchers argued that incorporating sentiment information improve the prediction accuracy and our result supports this by showing that sentiment analysis reduce the prediction error from 3.3% to 2.5%. The combination of user normalization, demographic information, and sentiment analysis greatly reduce the prediction error from 3.3% to 0.6%.

6.1 Contributions

We believe that our work is beneficial for Twitter based prediction research especially in the electoral prediction because of several aspects. First, we perform predictions of an election

with Twitter in a country with an interesting demographics where the internet penetration is still low, about 24%. When looking deeper into province-level, we can see more gap between provinces. For example, in one province, there are almost 1% of its population post their political preference on Twitter while in other province, it is only about 0.01%. This makes the data has a high probability of bias and makes the data normalization process more important. We perform the prediction model in national level and province level, so in total we have 34 prediction result. Our results strengthen the assumption made in (Metaxas, 2011) and (Gayo-Avello D. M., 2011). We show that Twitter based prediction cannot produce high accuracy in every election but we can expect a good accuracy when the sample is very large.

Second, we perform as many as possible methods for each step in the prediction model. Previous researchers employed many different methods in their prediction model and claimed that their method is better. We compare several methods to give an evidence that one method do perform better than the other. For example, we compare the result of different keyword selection; only candidates' names, names and aliases, campaign hashtags, and all possible keywords. We show that by using more keywords, we can gather more user opinions and in turn, increase the prediction accuracy. In the duration of data collection, we reinforce (Bermingham, 2011) argument, stating that using 1 day of data is more accurate than using several days or weeks of data. We also agree with (Sang, 2012) who stated that user normalization or counting the user is better than counting the tweets because in reality, one person only represent one vote. In our result, user counting improve the prediction accuracy greatly. Last, we perform many steps in our prediction model from tweets collection, keywords and duration selection, data filtering, reducing the bias (gender and location), and sentiment polarity of the tweets. We show that using all of those methods, the prediction error in our experiment can be reduced from 3.3% to 0.6%.

6.2 Suggestion on Future Works

We believe that this research domain in this topic is far from complete. There are many aspects that are still unclear and need to be discussed. In this research, although we have covered many items that are doubtful in the previous research, we only answered on how to improve the prediction accuracy. There are still many studies needed before the prediction using Twitter as the data source can be accepted scientifically.

Focusing on this research, there are several things that limit the methods chosen in our experiment. First is the language problem. Many powerful language processing technique is English-based and it is too time consuming to be implemented for our data. Indonesian originally have hundreds of local languages, and until now the people in different cities have different 'style' or 'accent' in their Indonesian language. This makes the process of language model building become more difficult, and lowering the classification accuracy.

This research topic can be extended in several ways. The first is to improve the method explained in this research or implement another feature that can improve the prediction accuracy. For example, collect tweets from Twitter 'paid' firehose. (Gaurav, 2013) explained

that Twitters public API provides only 1% or less of its entire, without control over the sampling procedure, which is likely, insufficient for accurate analysis of public sentiment. In term of sentiment analysis, we are able to differentiate between positive and negative tweets but we still have not able to differentiate between campaign/propaganda and the opinion of voters. To un-bias the sample, there are many information that we have yet to explore, such as economic condition, education, ethnicity, and urban/rural information. When all of those information are incorporated in the prediction, hopefully the sample bias can be reduced.

This topic also has other areas to be studied such as the detection of swing voters. In many occasions, election results were determined by the choice of swing voters. While it is very easy to detect their share in offline/interview-based polling, it is hard to detect them based only on tweets. Related to this issue, there is no proper method in quantify the effect of self-selection bias yet. Not everyone use Twitter, and their opinion need to be taken into account.

Based on our prediction results, the accuracy of the prediction greatly vary in different provinces. Even though all of the prediction use the same process. Conclusion that can be derived from it is that the tweet-based prediction is not applicable in every election unless we also know the estimated error of the prediction. In a polling that use random sampling, the estimated margin of error can be calculated only using the number of sample and a confidence value, while in the tweet-based prediction, more elements need to be considered. This is an important direction to be pursued because people trust can be easily gained when most of tweet-based prediction results have an error that still within the margin of error. One more interesting research direction is on how to reduce the variables needed in the prediction model without reducing its accuracy. So that people can easily reproduce the model and create a reliable real-time system.

References

Arzheimer, K. &. (2014). A new multinomial accuracy measure for polling bias. *Political Analysis, 22(1)*, 31-44.

Bakker, T. P. (2011). Good news for the future? Young people, Internet use, and political participation. *Communication Research,* .

Beauchamp, N. (2013). Predicting and interpolating state-level polling using twitter textual data. *Meeting on automated text analysis, London School of Economics*.

Bermingham, A. &. (2011). On using Twitter to monitor political sentiment and predict election results.

Boulis, C. O. (2005). A quantitative analysis of lexical differences between genders in telephone conversations. *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics,*, 435-442.

Boutet, A. K. (2012). What's in your Tweets? I know who you supported in the UK 2010 general election. *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Burger, J. D. (2011). Discriminating Gender on Twitter. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1301-1309.

Burgess., A. B. (2011). ausvotes: How twitter covered the 2010 aus-. *Communication, Politics and Culture, 44(2)*, 37-56.

Cameron, M. P. (2013). Can Social Media Predict Election Results? Evidence from New Zealand. *No. 13/08*.

Ceron, A. C. (2013). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society, 16(2)*, 340-358.

Ceron, A. C. (2014). Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters—Evidence From the United States and Italy. *Social Science Computer Review*.

Chaovalit, P. &. (2005). Movie review mining: A comparison between supervised and unsupervised classification approaches. *System Sciences, 2005. HICSS'05.*, 112c-112c.

Choy, M. C. (2011). A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. *arXiv preprint arXiv:1108.5520*.

Choy, M. C. (2012). US Presidential Election 2012 Prediction using Census Corrected Twitter Model. *arXiv preprint arXiv:1211.0938.*

Chu, Z. G. (2010). Who is tweeting on Twitter: human, bot, or cyborg? *In Proceedings of the 26th annual computer security applications conference*, 21-30.

Chung, J. E. (2011). Can collective sentiment expressed on twitter predict political elections? *AAAI*.

Cook, D. M. (2014). Twitter Deception and Influence: Issues of Identity, Slacktivism, and Puppetry. . *Journal of Information Warfare, 13(1)*, 58-71.

Dann, S. (2010). Twitter content classification. *First Monday, 15(12).*

Down, J. &. (2003). SMS polling. A methodological review. *ASC*, 277-286.

Fauzi, G. (2014). THE CONTRIBUTION OF DIRECT ELECTIONS FOR LOCAL LEADERS TO THE CORRUPTION IN INDONESIA: LESSONS LEARNED FROM INDONESIAN DEMOCRATIC SYSTEM. *International Journal of Education, 7(2)*, 103.

Fink, C. B. (2013). Twitter, Public Opinion, and the 2011 Nigerian Presidential Election. *2013 International Conference on Social Computing (SocialCom). IEEE.*, 311-320.

Fumagalli, L. &. (2011). The total survey error paradigm and pre-election polls: The case of the 2006 Italian general elections. *ISER Working Paper Series. 2011-29*.

Gaurav, M. S. (2013). Leveraging candidate popularity on Twitter to predict election outcome. *Proceedings of the 7th Workshop on Social Network Mining and Analysis. ACM.*, 7.

Gayo Avello, D. (2011). Don't turn social media into another'Literary Digest'poll. *Communications of the ACM, 54(10)*, 121-128.

Gayo-Avello, D. (2012). No, you cannot predict elections with twitter. *Internet Computing, IEEE, 16(6)*, 91-94.

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review*.

Gayo-Avello, D. M. (2011). Limits of electoral predictions using twitter. *ICWSM.*

Goggins, C. M. (2012). Twitter as virtual town square: Citizen engagement during a nationally televised republican primary debate. *APSA 2012 Annual Meeting Paper*.

Golbeck, J. &. (2011). Computing political preference among twitter followers. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

He, Y. &. (2011). Self-training from labeled features for sentiment analysis. *Information Processing & Management, 47(4),*, 606-616.

Hillygus, D. S. (2011). The evolution of election polling in the United States. *Public opinion quarterly, 75(5),*, 962-981.

Hitchens, P. (2009). *The Broken Compass.* Bloomsbury Publishing.

Jensen, M. J. (2013). Psephological investigations: Tweets, votes, and unknown unknowns in the republican nomination process. *Policy & Internet, 5(2)*, 161-182.

Jungherr, A. J. (2012). Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment". *Social Science Computer Review, 30(2),*, 229-234.

Kouloumpis, E. W. (2011). Twitter sentiment analysis: The good the bad and the omg! *ICWSM, 11*, 538-541.

Lewis-Beck, M. S. (2005). Election forecasting: principles and practice. *The British Journal of Politics & International Relations, 7(2),*, 145-164.

Liu, B. &. (2012). A survey of opinion mining and sentiment analysis. *Mining Text Data*, 415-463.

Makazhanov, A. R. (2014). Predicting political preference of Twitter users. *Social Network Analysis and Mining*, 1-15.

Mejova, Y. S. (2013). GOP primary season on twitter: popular political sentiment in social media. *Proceedings of the sixth ACM international conference on Web search and data mining*, 517-526.

Metaxas, P. T.-A. (2011). How (not) to predict elections. *Privacy, security, risk and trust (PASSAT), IEEE third international conference on social computing (SocialCom)* (pp. 165-171). IEEE.

Mislove, A. L. (2011). Understanding the Demographics of Twitter Users. *ICWSM, 11, 5th*.

Morozov, E. (2009). Iran: Downside to the" twitter revolution". *Dissent, 56(4),* , 10-14.

Nguyen, D. G. (2013). " How Old Do You Think I Am?" A Study of Language and Age in Twitter. *ICWSM*.

Nooralahzadeh, F. A. (2013). 2012 Presidential Elections on Twitter--An Analysis of How the US and French Election were Reflected in Tweets. *Control Systems and Computer Science (CSCS), 2013 19th International Conference on*, 240-246.

O'Connor, B. B. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM, 11*, 122-129.

Pak, A. &. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREC*.

Pang, B. L. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 79-86.

Pennacchiotti, M. &. (2011). Democrats, republicans and starbucks afficionados: user classification in twitter. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* , 430-438.

Rao, D. Y. (2012). Classifying Latent User Attributes in Twitter. *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 37-44.

Sanders, E. &. (2013). Relating Political Party Mentions on Twitter with Polls and Election Results. *DIR*, 68-71.

Sang, E. T. (2012). Predicting the 2011 dutch senate election results with twitter. *the Workshop on Semantic Analysis in Social Media* (pp. 53-60). Association for Computational Linguistics.

Shamma, N. A. (2010). Characterizing debate performance via aggregated twitter sentiment. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1195-1198.

Simon, H. A. (1954). Bandwagon and underdog effects and the possibility of election predictions. *Public Opinion Quarterly, 18(3)*, 245-253.

Taboada, M. B. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics, 37(2)*, 267-307.

Ting, S. L. (2011). s Naive Bayes a good classifier for document classification? *International Journal of Software Engineering and Its Applications, 5(3)*, 37.

Tong, S. &. (2002). Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research, 2.*, 45-66.

Trihartono, A. (2013). A Vox Populi Reflector or Public Entertainer? Mass Media Polling in Contemporary Indonesia. *Procedia Environmental Sciences 17*, 928-937.

Trihartono, A. (2014). Beyond Measuring the Voice of the People: The Evolving Role of Political Polling in Indonesia's Local Leader Elections. *Southeast Asian Studies, 3(1)*, 151-182.

Tsfati, Y. (2001). Why do People Trust Media Pre-Election Polls? Evidence from the Israeli 1996 Elections. *International Journal of Public Opinion Research, 13(4)*, 433-441.

Tumasjan, A. S. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM, 10*, 178-185.

Vinodhini, G. &. (2012). Sentiment analysis and opinion mining: a survey. . *International Journal, 2(6).*

Waugh, B. A. (2013). The Influence and Deception of Twitter: The Authenticity of the Narrative and Slacktivism in the Australian Electoral Process. *14th Australian Information Warfare Conference.* Perth: SRI Security Research Institute.

Wong, F. M. (2013). Quantifying Political Leaning from Tweets and Retweets. *ICWSM*.

Wong, F. T. (2013). Media, pundits and the us presidential election: Quantifying political leanings from tweets. *In Proceedings of the International Conference on Weblogs and Social Media.*

Zimmer, M. &. (2014). A Topology of Twitter Research: Disciplines, Methods, and Ethics. *Aslib Proceedings (Vol. 66, No. 3)*, 2-2.

Appendix A: The Complete Experiment Result

A. Result from previous research

Author	Country	Election Type	No of Tweets	No of User	Data Duration	Error
(Ceron A. C., 2014)	US	Presidential	50 million	Unknown	6 weeks	0.4%
(Ceron A. C., 2014)	Italy	primary	500 thousand	Unknown	6 weeks	9.7%
(Ceron A. C., 2014)	Italy	primary	500 thousand	Unknown	6 weeks	8.7%
(Fink, 2013)	Nigerian	Presidential	26 million	160 thousand	1 year	11.0%
(Jensen, 2013)	US/Iowa	Presidential	697 thousand	195 thousand	4 days	3.1%
(Sanders, 2013)	Netherlands	Parliamentary	170 thousand	Unknown	10 days	2.4%
(Gaurav, 2013)	Venezuela	Presidential	400 million	Unknown	1 week	2.0%
(Gaurav, 2013)	Ecuador	Presidential	397 million	Unknown	1 week	19.0%
(Gaurav, 2013)	Paraguay	Presidential	395 million	Unknown	1 week	3.0%
(Gaurav, 2013)	Venezuela	Presidential	400 million	Unknown	1 week	0.1%
(Gaurav, 2013)	Ecuador	Presidential	397 million	Unknown	1 week	3.0%
(Gaurav, 2013)	Paraguay	Presidential	395 million	Unknown	1 week	3.0%
(Ceron A. C., 2013)	Italy	Presidential	107 thousand	Unknown	1 Month	5.7%
(Ceron A. C., 2013)	France	Legislative	244 thousand	Unknown	1 week	2.4%
(Tumasjan, 2010)	Germany	federal	104 thousand	Unknown	1 Month	1.7%
(Choy M. C., 2011)	Singapore	Presidential	16 thousand	Unknown	8 days	6.1%
(Gayo-Avello D. M., 2011)	US/Massachusetts	senate	234 thousand	56 thousand	1 week	6.3%
(Gayo-Avello D. M., 2011)	US/Massachusetts	senate	234 thousand	56 thousand	1 week	1.6%
(Gayo-Avello D. M., 2011)	US/Colorado	congress			1 week	21.8%
(Gayo-Avello D. M., 2011)	US/Colorado	congress			1 week	9.7%
(Gayo-Avello D. M., 2011)	US/Nevada	congress			1 week	0.9%
(Gayo-Avello D. M., 2011)	US/Nevada	congress			1 week	1.9%
(Gayo-Avello D. M., 2011)	US/California	congress			1 week	5.7%
(Gayo-Avello D. M., 2011)	US/California	congress			1 week	4.4%
(Gayo-Avello D. M., 2011)	US/Kentucky	congress			1 week	39.6%
(Gayo-Avello D. M., 2011)	US/Kentucky	congress			1 week	1.2%
(Gayo-Avello D. M., 2011)	US/Delaware	congress			1 week	24.5%
(Gayo-Avello D. M., 2011)	US/Delaware	congress	13 thousand	6 thousand	1 week	17.8%
(Bermingham, 2011)	Ireland	general	Unknown	Unknown	2 weeks	5.9%
(Bermingham, 2011)	Ireland	general	Unknown	Unknown	2 weeks	3.7%
(Sang, 2012)	Dutch	senate	7 thousand	Unknown	N/A	1.3%
(Choy M. C., 2012)	US	congress	7 million	Unknown	6 weeks	1.65%

Table 34 Data from Previous Experiments

B. Complete Tweet Based Prediction Result

Province	Prabowo-Hatta		Jokowi-JK		Winner Predicted	Mean Absolute Error
	Election Result (%)	Tweet Count	Election Result (%)	Tweet Count		
Aceh	54.39%	49.11%	45.61%	50.89%	Incorrect	5.28%
Sumatera Utara	44.76%	43.08%	55.24%	56.92%	Correct	1.68%
Sumatera Barat	76.92%	50.83%	23.08%	49.17%	Correct	26.09%
Riau	50.12%	49.85%	49.88%	50.15%	Incorrect	0.27%
Jambi	49.25%	47.62%	50.75%	52.38%	Correct	1.63%
Sumatera Selatan	51.26%	43.36%	48.74%	56.64%	Incorrect	7.90%
Bengkulu	45.27%	49.05%	54.73%	50.95%	Correct	3.78%
Lampung	46.93%	51.51%	53.07%	48.49%	Incorrect	4.58%
Kepulauan Bangka Belitung	32.74%	30.89%	67.26%	69.11%	Correct	1.85%
Kepulauan Riau	40.37%	48.79%	59.63%	51.21%	Correct	8.42%
DKI Jakarta	46.92%	39.96%	53.08%	60.04%	Correct	6.96%
Jawa Barat	59.78%	46.91%	40.22%	53.09%	Incorrect	12.87%
Jawa Tengah	33.35%	43.38%	66.65%	56.62%	Correct	10.03%
DI Yogyakarta	44.19%	43.56%	55.81%	56.44%	Correct	0.63%
Jawa Timur	46.83%	46.08%	53.17%	53.92%	Correct	0.75%
Banten	57.10%	42.50%	42.90%	57.50%	Incorrect	14.60%
Bali	28.58%	42.86%	71.42%	57.14%	Correct	14.28%
Nusa Tenggara Barat	72.45%	52.22%	27.55%	47.78%	Correct	20.23%
Nusa Tenggara Timur	34.08%	11.06%	65.92%	88.94%	Correct	23.02%
Kalimantan Barat	39.62%	47.97%	60.38%	52.03%	Correct	8.35%
Kalimantan Tengah	40.21%	30.77%	59.79%	69.23%	Correct	9.44%
Kalimantan Selatan	50.05%	45.16%	49.95%	54.84%	Incorrect	4.89%
Kalimantan Timur	36.62%	44.16%	63.38%	55.84%	Correct	7.54%
Sulawesi Utara	46.12%	44.94%	53.88%	55.06%	Correct	1.18%
Sulawesi Tengah	45.17%	45.99%	54.83%	54.01%	Correct	0.82%
Sulawesi Selatan	28.57%	41.05%	71.43%	58.95%	Correct	12.48%
Sulawesi Tenggara	45.10%	65.77%	54.90%	34.23%	Incorrect	20.67%
Gorontalo	63.10%	41.77%	36.90%	58.23%	Incorrect	21.33%
Sulawesi Barat	26.63%	46.15%	73.37%	53.85%	Correct	19.52%
Maluku	49.48%	65.57%	50.52%	34.43%	Incorrect	16.09%
Maluku Utara	54.45%	62.50%	45.55%	37.50%	Correct	8.05%
Papua	27.51%	30.77%	72.49%	69.23%	Correct	3.26%
Papua Barat	32.37%	41.03%	67.63%	58.97%	Correct	8.66%

Table 35 Tweet counting result

Province	Prabowo-Hatta		Jokowi-JK		Winner Predicted	Mean Absolute Error
	Election Result (%)	User Count	Election Result (%)	User Count		
Aceh	54.39%	51.78%	45.61%	48.22%	Correct	2.61%
Sumatera Utara	44.76%	45.03%	55.24%	54.97%	Correct	0.27%
Sumatera Barat	76.92%	51.91%	23.08%	48.09%	Correct	25.01%
Riau	50.12%	50.55%	49.88%	49.45%	Correct	0.43%
Jambi	49.25%	46.97%	50.75%	53.03%	Correct	2.28%
Sumatera Selatan	51.26%	48.06%	48.74%	51.94%	Incorrect	3.20%
Bengkulu	45.27%	45.97%	54.73%	54.03%	Correct	0.70%
Lampung	46.93%	49.08%	53.07%	50.92%	Correct	2.15%
Kepulauan Bangka Belitung	32.74%	45.59%	67.26%	54.41%	Correct	12.85%
Kepulauan Riau	40.37%	47.12%	59.63%	52.88%	Correct	6.75%
DKI Jakarta	46.92%	43.51%	53.08%	56.49%	Correct	3.41%
Jawa Barat	59.78%	47.73%	40.22%	52.27%	Incorrect	12.05%
Jawa Tengah	33.35%	42.91%	66.65%	57.09%	Correct	9.56%
DI Yogyakarta	44.19%	45.83%	55.81%	54.17%	Correct	1.64%
Jawa Timur	46.83%	43.23%	53.17%	56.77%	Correct	3.60%
Banten	57.10%	47.13%	42.90%	52.87%	Incorrect	9.97%
Bali	28.58%	43.81%	71.42%	56.19%	Correct	15.23%
Nusa Tenggara Barat	72.45%	49.53%	27.55%	50.47%	Incorrect	22.92%
Nusa Tenggara Timur	34.08%	43.08%	65.92%	56.92%	Correct	9.00%
Kalimantan Barat	39.62%	45.84%	60.38%	54.16%	Correct	6.22%
Kalimantan Tengah	40.21%	29.41%	59.79%	70.59%	Correct	10.80%
Kalimantan Selatan	50.05%	50.00%	49.95%	50.00%	Incorrect	0.05%
Kalimantan Timur	36.62%	45.15%	63.38%	54.85%	Correct	8.53%
Sulawesi Utara	46.12%	41.45%	53.88%	58.55%	Correct	4.67%
Sulawesi Tengah	45.17%	53.33%	54.83%	46.67%	Incorrect	8.16%
Sulawesi Selatan	28.57%	43.31%	71.43%	56.69%	Correct	14.74%
Sulawesi Tenggara	45.10%	53.85%	54.90%	46.15%	Incorrect	8.75%
Gorontalo	63.10%	40.78%	36.90%	59.22%	Incorrect	22.32%
Sulawesi Barat	26.63%	40.00%	73.37%	60.00%	Correct	13.37%
Maluku	49.48%	48.72%	50.52%	51.28%	Correct	0.76%
Maluku Utara	54.45%	47.06%	45.55%	52.94%	Incorrect	7.39%
Papua	27.51%	43.86%	72.49%	56.14%	Correct	16.35%
Papua Barat	32.37%	48.28%	67.63%	51.72%	Correct	15.91%

Table 36 User Counting result

Province	Prabowo-Hatta			Jokowi-JK			MAE 2 Keywords	MAE 10 Keywords	MAE All Keywords
	1 Keyword	5 Keywords	All Keywords	1 Keyword	5 Keywords	All Keywords			
Aceh	679	683	716	699	704	742	5.12%	5.15%	5.28%
Sumatera Utara	2194	2203	2333	2973	3000	3082	2.30%	2.42%	1.68%
Sumatera Barat	479	484	521	477	482	504	26.82%	26.82%	26.09%
Riau	1213	1218	1327	1285	1299	1335	1.56%	1.73%	0.27%
Jambi	337	338	360	377	381	396	2.05%	2.24%	1.63%
Sumatera Selatan	873	880	950	1208	1212	1241	9.31%	9.19%	7.90%
Bengkulu	120	120	129	128	130	134	3.12%	2.73%	3.78%
Lampung	732	732	767	676	684	722	5.06%	4.76%	4.58%
Kepulauan Bangka Belitung	56	57	59	131	131	132	2.79%	2.42%	1.85%
Kepulauan Riau	224	224	241	240	243	253	7.91%	7.60%	8.42%
DKI Jakarta	19804	19915	21420	30754	31007	32186	7.75%	7.81%	6.96%
Jawa Barat	11126	11199	11877	12853	13012	13442	13.38%	13.52%	12.87%
Jawa Tengah	4507	4541	4852	6020	6102	6333	9.46%	9.32%	10.03%
DI Yogyakarta	1841	1849	2034	2491	2530	2635	1.69%	1.97%	0.63%
Jawa Timur	1374	1387	1524	1713	1732	1783	2.32%	2.36%	0.75%
Banten	2798	2828	3245	4090	4161	4390	16.48%	16.64%	14.60%
Bali	223	228	249	328	328	332	11.89%	12.43%	14.28%
Nusa Tenggara Barat	424	429	447	385	386	409	20.04%	19.81%	20.23%
Nusa Tenggara Timur	39	41	47	372	375	378	24.59%	24.22%	23.02%
Kalimantan Barat	418	422	438	443	449	475	8.93%	8.83%	8.35%
Kalimantan Tengah	7	7	8	17	17	18	11.04%	11.04%	9.44%
Kalimantan Selatan	52	52	56	67	68	68	6.35%	6.72%	4.89%
Kalimantan Timur	343	344	359	438	443	454	7.30%	7.09%	7.54%
Sulawesi Utara	507	508	542	632	641	664	1.61%	1.91%	1.18%
Sulawesi Tengah	82	83	86	100	100	101	0.12%	0.19%	0.82%
Sulawesi Selatan	194	194	204	289	289	293	11.60%	11.60%	12.48%
Sulawesi Tenggara	64	69	73	36	37	38	18.90%	19.99%	20.67%
Gorontalo	134	134	137	185	187	191	21.09%	21.36%	21.33%
Sulawesi Barat	12	12	12	12	13	14	23.37%	21.37%	19.52%
Maluku	73	77	80	34	37	42	18.74%	18.06%	16.09%
Maluku Utara	48	48	50	30	30	30	7.09%	7.09%	8.05%
Papua	29	29	32	72	72	72	1.20%	1.20%	3.26%
Papua Barat	15	15	16	22	22	23	8.17%	8.17%	8.66%
Indonesia	78011	78581	84226	104275	105305	109168	4.05%	4.12%	3.30%

Table 37 MAE with different Keywords

Province	MAE	Population	Prabowo-Hatta		Jokowi-JK		MAE Weigthed
			Original Tweet	Tweet Weighted	Original Tweet	Tweet Weighted	
Aceh	5.28%	4494410	716	1190	742	1233	N/A
Sumatera Utara	1.68%	12982204	2333	3015	3082	3983	N/A
Sumatera Barat	26.09%	4846909	521	1328	504	1285	N/A
Riau	0.27%	5538367	1327	1488	1335	1497	N/A
Jambi	1.63%	3092265	360	794	396	873	N/A
Sumatera Selatan	7.90%	7450394	950	1741	1241	2275	N/A
Bengkulu	3.78%	1715518	129	454	134	471	N/A
Lampung	4.58%	7608405	767	2113	722	1989	N/A
Kepulauan Bangka Belitung	1.85%	1223296	59	204	132	456	N/A
Kepulauan Riau	8.42%	1679163	241	442	253	464	N/A
DKI Jakarta	6.96%	9607787	21420	2070	32186	3110	N/A
Jawa Barat	12.87%	43053732	11877	10887	13442	12322	N/A
Jawa Tengah	10.03%	32382657	4852	7572	6333	9884	N/A
DI Yogyakarta	0.63%	3457491	2034	812	2635	1052	N/A
Jawa Timur	0.75%	37476757	1524	9310	1783	10892	N/A
Banten	14.60%	10632166	3245	2436	4390	3295	N/A
Bali	14.28%	3890757	249	899	332	1198	N/A
Nusa Tenggara Barat	20.23%	4500212	447	1267	409	1159	N/A
Nusa Tenggara Timur	23.02%	4683827	47	279	378	2246	N/A
Kalimantan Barat	8.35%	4395983	438	1137	475	1233	N/A
Kalimantan Tengah	9.44%	2212089	8	367	18	826	N/A
Kalimantan Selatan	4.89%	3626616	56	883	68	1072	N/A
Kalimantan Timur	7.54%	3553143	359	846	454	1070	N/A
Sulawesi Utara	1.18%	2270596	542	550	664	674	N/A
Sulawesi Tengah	0.82%	2635009	86	653	101	767	N/A
Sulawesi Selatan	12.48%	8034776	204	1778	293	2553	N/A
Sulawesi Tenggara	20.67%	2232586	73	791	38	412	N/A
Gorontalo	21.33%	1040164	137	234	191	327	N/A
Sulawesi Barat	19.52%	1158651	12	288	14	336	N/A
Maluku	16.09%	1533506	80	542	42	285	N/A
Maluku Utara	8.05%	1038087	50	350	30	210	N/A
Papua	3.26%	760422	32	126	72	284	N/A
Papua Barat	8.66%	2833381	16	627	23	901	N/A
Indonesia	3.30%	237641326	84226*	57472	109168*	70631	1.99%

Table 38 Counting Tweets with Population Weight *Including unkown province location

Province	Mean Absolute Error	Population	Prabowo-Hatta		Jokowi-JK		Mean Absolute Error
			Original User	User Weighted	Original User	User Weighted	
Aceh	2.61%	4494410	218	401	203	374	N/A
Sumatera Utara	0.27%	12982204	829	1008	1012	1230	N/A
Sumatera Barat	25.01%	4846909	258	434	239	402	N/A
Riau	0.43%	5538367	461	483	451	472	N/A
Jambi	2.28%	3092265	124	250	140	283	N/A
Sumatera Selatan	3.20%	7450394	335	617	362	667	N/A
Bengkulu	0.70%	1715518	57	136	67	160	N/A
Lampung	2.15%	7608405	294	644	305	668	N/A
Kepulauan Bangka Belitung	12.85%	1223296	31	96	37	115	N/A
Kepulauan Riau	6.75%	1679163	147	136	165	153	N/A
DKI Jakarta	3.41%	9607787	5807	721	7538	935	N/A
Jawa Barat	12.05%	43053732	3949	3542	4324	3878	N/A
Jawa Tengah	9.56%	32382657	1859	2395	2473	3186	N/A
DI Yogyakarta	1.64%	3457491	917	273	1084	323	N/A
Jawa Timur	3.60%	37476757	664	2792	872	3667	N/A
Banten	9.97%	10632166	1395	864	1565	969	N/A
Bali	15.23%	3890757	145	294	186	377	N/A
Nusa Tenggara Barat	22.92%	4500212	210	384	214	391	N/A
Nusa Tenggara Timur	9.00%	4683827	28	348	37	460	N/A
Kalimantan Barat	6.22%	4395983	182	347	215	410	N/A
Kalimantan Tengah	10.80%	2212089	5	112	12	269	N/A
Kalimantan Selatan	0.05%	3626616	42	313	42	313	N/A
Kalimantan Timur	8.53%	3553143	135	276	164	336	N/A
Sulawesi Utara	4.67%	2270596	189	162	267	229	N/A
Sulawesi Tengah	8.16%	2635009	32	242	28	212	N/A
Sulawesi Selatan	14.74%	8034776	136	600	178	785	N/A
Sulawesi Tenggara	8.75%	2232586	28	207	24	178	N/A
Gorontalo	22.32%	1040164	42	73	61	106	N/A
Sulawesi Barat	13.37%	1158651	6	80	9	120	N/A
Maluku	0.76%	1533506	19	129	20	136	N/A
Maluku Utara	7.39%	1038087	16	84	18	95	N/A
Papua	16.35%	760422	25	57	32	74	N/A
Papua Barat	15.91%	2833381	14	236	15	253	N/A
Indonesia	1.30%	237641326	31972*	18736	38212*	22222	1.11%

Table 39 Counting Users with Population Weight *Including unkown province location

Province	Prabowo-Hatta		Jokowi-JK		Winner Predicted	Tweet + SA MAE	Non SA MAE	MAE Difference
	Election Result (%)	Tweet + SA Count	Election Result (%)	Tweet + SA Count				
Aceh	54.39%	53.20%	45.61%	46.80%	Correct	1.19%	5.28%	4.10%
Sumatera Utara	44.76%	51.57%	55.24%	48.43%	Incorrect	6.81%	1.68%	-5.13%
Sumatera Barat	76.92%	53.88%	23.08%	46.12%	Correct	23.04%	26.09%	3.05%
Riau	50.12%	57.65%	49.88%	42.35%	Correct	7.53%	0.27%	-7.26%
Jambi	49.25%	55.18%	50.75%	44.82%	Incorrect	5.93%	1.63%	-4.30%
Sumatera Selatan	51.26%	49.95%	48.74%	50.05%	Incorrect	1.31%	7.90%	6.59%
Bengkulu	45.27%	49.54%	54.73%	50.46%	Correct	4.27%	3.78%	-0.49%
Lampung	46.93%	29.76%	53.07%	70.24%	Correct	17.17%	4.58%	-12.59%
Kepulauan Bangka Belitung	32.74%	35.38%	67.26%	64.62%	Correct	2.64%	1.85%	-0.79%
Kepulauan Riau	40.37%	48.94%	59.63%	51.06%	Correct	8.57%	8.42%	-0.15%
DKI Jakarta	46.92%	43.17%	53.08%	56.83%	Correct	3.75%	6.96%	3.22%
Jawa Barat	59.78%	49.73%	40.22%	50.27%	Incorrect	10.05%	12.87%	2.82%
Jawa Tengah	33.35%	52.76%	66.65%	47.24%	Incorrect	19.41%	10.03%	-9.38%
DI Yogyakarta	44.19%	47.98%	55.81%	52.02%	Correct	3.79%	0.63%	-3.17%
Jawa Timur	46.83%	57.84%	53.17%	42.16%	Incorrect	11.01%	0.75%	-10.26%
Banten	57.10%	48.41%	42.90%	51.59%	Incorrect	8.69%	14.60%	5.91%
Bali	28.58%	77.46%	71.42%	22.54%	Incorrect	48.88%	14.28%	-34.61%
Nusa Tenggara Barat	72.45%	61.58%	27.55%	38.42%	Correct	10.87%	20.23%	9.36%
Nusa Tenggara Timur	34.08%	71.43%	65.92%	28.57%	Incorrect	37.35%	23.02%	-14.33%
Kalimantan Barat	39.62%	54.23%	60.38%	45.77%	Incorrect	14.61%	8.35%	-6.26%
Kalimantan Tengah	40.21%	90.00%	59.79%	10.00%	Incorrect	49.79%	9.44%	-40.35%
Kalimantan Selatan	50.05%	34.38%	49.95%	65.63%	Incorrect	15.68%	4.89%	-10.79%
Kalimantan Timur	36.62%	56.47%	63.38%	43.53%	Incorrect	19.85%	7.54%	-12.31%
Sulawesi Utara	46.12%	50.19%	53.88%	49.81%	Incorrect	4.07%	1.18%	-2.90%
Sulawesi Tengah	45.17%	65.38%	54.83%	34.62%	Incorrect	20.21%	0.82%	-19.40%
Sulawesi Selatan	28.57%	65.15%	71.43%	34.85%	Incorrect	36.58%	12.48%	-24.11%
Sulawesi Tenggara	45.10%	33.33%	54.90%	66.67%	Correct	11.77%	20.67%	8.90%
Gorontalo	63.10%	62.09%	36.90%	37.91%	Correct	1.01%	21.33%	20.32%
Sulawesi Barat	26.63%	50.00%	73.37%	50.00%	Incorrect	23.37%	19.52%	-3.85%
Maluku	49.48%	46.67%	50.52%	53.33%	Correct	2.81%	16.09%	13.28%
Maluku Utara	54.45%	66.67%	45.55%	33.33%	Correct	12.22%	8.05%	-4.17%
Papua	27.51%	61.90%	72.49%	38.10%	Incorrect	34.39%	3.26%	-31.14%
Papua Barat	32.37%	62.50%	67.63%	37.50%	Incorrect	30.13%	8.66%	-21.47%

Table 40 Counting Tweets with Sentiment Analysis

Province	Filtered Users		Prabowo-Hatta		Jokowi-JK		Winner Predicted	MAE Spammer	Mean Absolute Error	Differences
	Prabowo-Hatta	Jokowi-JK	Election Result (%)	User Count	Election Result (%)	User Count				
Aceh	0	22	54.39%	50.27%	45.61%	49.73%	Correct	4.12%	2.61%	-1.51%
Sumatera Utara	50	53	44.76%	44.82%	55.24%	55.18%	Correct	0.06%	0.27%	0.21%
Sumatera Barat	15	11	76.92%	51.59%	23.08%	48.41%	Correct	25.33%	25.01%	-0.32%
Riau	54	46	50.12%	50.12%	49.88%	49.88%	Correct	0.00%	0.43%	0.43%
Jambi	5	8	49.25%	47.41%	50.75%	52.59%	Correct	1.84%	2.28%	0.44%
Sumatera Selatan	22	26	51.26%	48.23%	48.74%	51.77%	Incorrect	3.03%	3.20%	0.16%
Bengkulu	9	6	45.27%	44.04%	54.73%	55.96%	Correct	1.23%	0.70%	-0.54%
Lampung	15	12	46.93%	48.78%	53.07%	51.22%	Correct	1.85%	2.15%	0.31%
Kepulauan Bangka Belitung	5	4	32.74%	44.07%	67.26%	55.93%	Correct	11.33%	12.85%	1.52%
Kepulauan Riau	7	5	40.37%	46.67%	59.63%	53.33%	Correct	6.30%	6.75%	0.45%
DKI Jakarta	381	433	46.92%	43.30%	53.08%	56.70%	Correct	3.62%	3.41%	-0.21%
Jawa Barat	233	257	59.78%	47.75%	40.22%	52.25%	Incorrect	12.03%	12.05%	0.01%
Jawa Tengah	93	106	33.35%	42.73%	66.65%	57.27%	Correct	9.38%	9.56%	0.18%
DI Yogyakarta	30	30	44.19%	45.70%	55.81%	54.30%	Correct	1.51%	1.64%	0.13%
Jawa Timur	30	31	46.83%	42.98%	53.17%	57.02%	Correct	3.85%	3.60%	-0.25%
Banten	51	54	57.10%	47.08%	42.90%	52.92%	Incorrect	10.02%	9.97%	-0.05%
Bali	1	0	28.58%	43.64%	71.42%	56.36%	Correct	15.06%	15.23%	0.17%
Nusa Tenggara Barat	5	3	72.45%	49.28%	27.55%	50.72%	Incorrect	23.17%	22.92%	-0.25%
Nusa Tenggara Timur	0	0	34.08%	43.08%	65.92%	56.92%	Correct	9.00%	9.00%	0.00%
Kalimantan Barat	6	8	39.62%	45.95%	60.38%	54.05%	Correct	6.33%	6.22%	-0.11%
Kalimantan Tengah	0	0	40.21%	29.41%	59.79%	70.59%	Correct	10.80%	10.80%	0.00%
Kalimantan Selatan	0	0	50.05%	50.00%	49.95%	50.00%	Incorrect	0.05%	0.05%	0.00%
Kalimantan Timur	6	6	36.62%	44.95%	63.38%	55.05%	Correct	8.33%	8.53%	0.20%
Sulawesi Utara	30	27	46.12%	39.85%	53.88%	60.15%	Correct	6.27%	4.67%	-1.60%
Sulawesi Tengah	0	0	45.17%	53.33%	54.83%	46.67%	Incorrect	8.16%	8.16%	0.00%
Sulawesi Selatan	1	2	28.57%	43.41%	71.43%	56.59%	Correct	14.84%	14.74%	-0.10%
Sulawesi Tenggara	1	1	45.10%	54.00%	54.90%	46.00%	Incorrect	8.90%	8.75%	-0.15%
Gorontalo	3	3	63.10%	40.21%	36.90%	59.79%	Incorrect	22.89%	22.32%	-0.57%
Sulawesi Barat	0	0	26.63%	40.00%	73.37%	60.00%	Correct	13.37%	13.37%	0.00%
Maluku	0	0	49.48%	48.72%	50.52%	51.28%	Correct	0.76%	0.76%	0.00%
Maluku Utara	1	0	54.45%	45.45%	45.55%	54.55%	Incorrect	9.00%	7.39%	-1.60%
Papua	0	0	27.51%	43.86%	72.49%	56.14%	Correct	16.35%	16.35%	0.00%
Papua Barat	0	0	32.37%	48.28%	67.63%	51.72%	Correct	15.91%	15.91%	0.00%
Indonesia	1416	1473	46.85%	45.41%	53.15%	54.59%	Correct	1.44%	1.30%	-0.15%

Table 41 Prediction Result with Filtered Users

Province	Prabowo-Hatta		Jokowi-JK		Winner Predicted	MAE Limited Keywords	Mean Absolute Error	Difference
	Election Result (%)	User Count	Election Result (%)	User Count				
Aceh	54.39%	47.33%	45.61%	52.67%	Incorrect	7.06%	2.61%	-4.45%
Sumatera Utara	44.76%	45.23%	55.24%	54.77%	Correct	0.47%	0.27%	-0.20%
Sumatera Barat	76.92%	50.20%	23.08%	49.80%	Correct	26.72%	25.01%	-1.72%
Riau	50.12%	52.84%	49.88%	47.16%	Correct	2.72%	0.43%	-2.29%
Jambi	49.25%	49.35%	50.75%	50.65%	Correct	0.10%	2.28%	2.18%
Sumatera Selatan	51.26%	48.64%	48.74%	51.36%	Incorrect	2.62%	3.20%	0.58%
Bengkulu	45.27%	46.13%	54.73%	53.87%	Correct	0.86%	0.70%	-0.17%
Lampung	46.93%	50.53%	53.07%	49.47%	Incorrect	3.60%	2.15%	-1.44%
Kepulauan Bangka Belitung	32.74%	42.36%	67.26%	57.64%	Correct	9.62%	12.85%	3.22%
Kepulauan Riau	40.37%	45.30%	59.63%	54.70%	Correct	4.93%	6.75%	1.81%
DKI Jakarta	46.92%	42.97%	53.08%	57.03%	Correct	3.95%	3.41%	-0.55%
Jawa Barat	59.78%	47.76%	40.22%	52.24%	Incorrect	12.02%	12.05%	0.02%
Jawa Tengah	33.35%	42.81%	66.65%	57.19%	Correct	9.46%	9.56%	0.10%
DI Yogyakarta	44.19%	44.88%	55.81%	55.12%	Correct	0.69%	1.64%	0.95%
Jawa Timur	46.83%	40.94%	53.17%	59.06%	Correct	5.89%	3.60%	-2.29%
Banten	57.10%	48.09%	42.90%	51.91%	Incorrect	9.01%	9.97%	0.96%
Bali	28.58%	38.31%	71.42%	61.69%	Correct	9.73%	15.23%	5.50%
Nusa Tenggara Barat	72.45%	50.96%	27.55%	49.04%	Correct	21.49%	22.92%	1.43%
Nusa Tenggara Timur	34.08%	36.08%	65.92%	63.92%	Correct	2.00%	9.00%	6.99%
Kalimantan Barat	39.62%	44.37%	60.38%	55.63%	Correct	4.75%	6.22%	1.48%
Kalimantan Tengah	40.21%	55.30%	59.79%	44.70%	Incorrect	15.09%	10.80%	-4.29%
Kalimantan Selatan	50.05%	48.51%	49.95%	51.49%	Incorrect	1.54%	0.05%	-1.49%
Kalimantan Timur	36.62%	44.61%	63.38%	55.39%	Correct	7.99%	8.53%	0.54%
Sulawesi Utara	46.12%	41.58%	53.88%	58.42%	Correct	4.54%	4.67%	0.14%
Sulawesi Tengah	45.17%	49.83%	54.83%	50.17%	Correct	4.66%	8.16%	3.50%
Sulawesi Selatan	28.57%	37.49%	71.43%	62.51%	Correct	8.92%	14.74%	5.82%
Sulawesi Tenggara	45.10%	51.32%	54.90%	48.68%	Incorrect	6.22%	8.75%	2.52%
Gorontalo	63.10%	51.74%	36.90%	48.26%	Correct	11.36%	22.32%	10.96%
Sulawesi Barat	26.63%	42.86%	73.37%	57.14%	Correct	16.23%	13.37%	-2.86%
Maluku	49.48%	51.69%	50.52%	48.31%	Incorrect	2.21%	0.76%	-1.44%
Maluku Utara	54.45%	49.40%	45.55%	50.60%	Incorrect	5.05%	7.39%	2.34%
Papua	27.51%	36.86%	72.49%	63.14%	Correct	9.35%	16.35%	7.00%
Papua Barat	32.37%	49.11%	67.63%	50.89%	Correct	16.74%	15.91%	-0.83%
Indonesia	46.85%	44.99%	53.15%	55.01%	Correct	1.86%	1.30%	-0.57%

Table 42 Gender Weighted Prediction Result