Twitter as Source of Auxiliary Information in Small Area Estimation (A Case Study about Estimation Electability of The Candidate Pairs Of President and Vice-President Of The 2019 President Election)

Fathi Abdul Muhyi[1], Anang Kurnia[2], Bagus Sartono[3]
{muhyifathi@gmail.com[1], anangk@ipb.ac.id[2], bagusco@gmail.com[3]}

Department of Statistics, IPB University, Bogor, 16680, Indonesia, Phone/Fax (0251) 8624535[123]

Abstract. When it comes to topic about electability people are become excited. People always wonder who has higher electability. During campaign of president election 2019, politics becomes one of the very interesting topic especially in social media. People talk about each pair candidate's president of Indonesia in social media. Topics that people talk can contain whether it is positive topic or negative topic. And now, what about linking electability with topics that people talk about in social media. It would be very interesting if there are some information in social that related to electability. Information extracted from social media can come in handy as auxiliary information to help estimating electability in each province using small area estimation.

Keywords: electability, small area estimation, social media, twitter.

1 Introduction

A survey about electability was design to estimate electability of The Candidate Pairs of President and Vice-President of The 2019 Presidential Election at National Level. The survey was never planned to estimate electability at province level but however electability at provincial level can be very interesting. Electability at provincial level was very necessary so strategies in elections will be managed more efficiently. The amount of data from some provinces are insufficient because sample size at some provinces were too small. Sample size at some provinces are only twenty. Increasing sample size is not a proper solution because it is very inefficient in terms of time, cost, and energy. Small area estimation method can be used to overcome sample size problem. Small area estimation can provide estimation with adequate precision without increasing sample size [1].

In context of survey sampling, a subpopulation or area estimate is usually referred to as a "direct estimate" if it is based only in the specific sample data coming from that area[2]. Electability can be estimated indirectly by small area estimation which combining data from another source, usually called auxiliary information. Source used in this study is from Twitter. Twitter is micro-blogging service letting people stream information in individual feeds known as tweets [3]. Towards 2019 Presidential Election, Twitter's user in Indonesia are become very interested in the topic about electability. It would be very interesting to generate auxiliary information from Twitter then used it in small area estimation. Tweets can be obtained from certain location In order to extract auxiliary information from each region [4].

ICSA 2019, August 02-03, Bogor, Indonesia
Copyright © 2020 EAI
DOI 10.4108/eai.2-8-2019.2290497

2 Data and Methodology

2.1 Data

Data used in this study are Electability Of The Candidate Pairs of President and Vice President from The 2019 Presidential Election survey data, tweets from Twitter with certain keywords, and The 2019 Presidential Election's Result. Survey was held during February 2019. Survey was done in each province. Respondents were given a question about which candidate pairs would be chosen if presidential election held that time. The options of this question are Joko Widodo – Maruf Amin (Jokowi-Maruf), Prabowo Subianto - Sandiaga Uno (Prabowo-Sandi), and refused/undecided. Respondents who answered refused/undecided will not be included in analysis. Sampling method used in the survey is stratified random sampling with province as stratum. Sample taken proportionally from each province. Auxiliary information will be extracted from tweets from each province in Indonesia. Tweets were collected from March 1st 2019 to March 13th 2019. Tweets can be collected with limited time range, only seven days before. In analysis Riau and Kepulauan Riau will be merged. The 2019 Presidential Election's Result will be used as validation for the electability estimation result.

2.2Small Area Estimation with Big Data

Small area estimation and big data are rapid growing topic and surly will get more attention in the future. Big data has big potential as source of auxiliary infomration for small area estimation [5].In Indonesia, small area estimation is usually aplied by using auxiliary information from SUSENA (*Survei Ekonomi Nasional*), BPS (*Badan Pusat Statistik*), and another administrative data [6]. There are two types of small area models based on the availability of the auxiliary information. Unit level small area model used when auxiliary information available for each unit of the populatiom while Area level small area model used when auxiliary information available in form of aggregation in each area [1]. Due to technical problems and legal restrictions, it is unfeasible at this stage to use unit level small area model when using auxiliary information from big data [5]. Auxilari information used in this research is extracted form twitter. Some research previously done using twitter besides as source of auxiliary information is using tweets to detect bots accounts [7].

2.3Tweets Extraction

Tweets will be extracted to generate auxiliary information for small area estimation. Extraction process divided into three major steps, those steps are tweets collection from each province, text preprocessing and sentiment scoring and aggregation at province level.

Tweets Collection. There are three things needed in tweets collection, keywords, time collection, and geocode. Tweets will be collected based on two topics. Topics about each candidate pairs (Jokowi-Maruf and Prabowo-Sandi). Keywords used to collect tweets with topics about Jokowi-Maruf are "@KHMarufAmin", "@jokowi", "jokowi", "cebong" and keywords used to collect tweets with topics about Prabowo-Sandi are "@prabowo", "@sandiuno", "prabowo", "kampret". There are tweets that contain both topics. Those tweets will not be included in analysis.

Tweets collected from each province were arranged based on radius and certain spot in the corresponded province. Radius used in some province often sliced with surrounded province, so there will be tweets redundancy.

Text Preprocessing and Sentiment Scoring. In this process, each topics will be identified the sentiment at province level. Here is the process to transforms tweets to sentiment :

1. Lowercased tweets
2. Remove some characters :
   a. Digits
   b. Punctuation
   c. URLs
   d. HTML characters
   e. Stopwords, stopwords are collection of words which do not have any tendency to be negative or positive (Tala 2003).
   f. Double whitespace
3. Stemming, stemming is a process to change affixed words into the basic word [8].
4. Compute sentiment score for each tweet by parsing each word with the collection of positive words and the collection of the negative words by using formula :

$$Score_{ijk} = \frac{pos_{ijk} - neg_{ijk}}{t_{ijk}},\ i=1,2,\ ...33,\ \ j = 1,2,\dots,L_{ik},\ \ k = 1,2 \qquad (1)$$

with $Score_{ijk}$, sentiment score j-th tweets at i-th province and topic about k-th candidate pair, pos_{ijk} count of positive words, neg_{ijk} count of negative words, t_{ijk} count of words after text preprocessing point one to three, and L_{ik} count of tweets at i-th province and topic about k-th candidate pair.

Aggregation at Province Level. Sentiment score will be aggregated for each province with formula :

$$X_{ki} = \frac{\sum_{j=1}^{j=L_{ik}} Skor_{ijk} w_{ijk}}{\sum_{j=1}^{j=L_{ik}} w_{ijk}},\ i=1,2,\ ...33,\ \ j = 1,2,\dots,L_{ik},\ \ k = 1,2 \qquad (2)$$

with $w_{ijk} = \frac{1}{f_{ijk}}$, X_{ki} provincial level sentiment at i-th province and topic about k-th candidate pair, f_{ijk} the count of the similar tweets which similar to the j-th tweet at i-th province and topic about k-th candidate pair and w_{ijk} weight of the tweet. Tweets that unintentionally collected several times will be weighted smaller in the computation of the sentiment score mean for each province.

2.4 Small Area Model Development

Small area model used in this research is area level small area model. Parameters from small are models will be estimated using integral approximation Gauss-Hermit quadrature [9].

Estimated value of the parameter will be corrected using formula presented by Stefanski and Carrol [10]. Steps needed to develop small area model are :

1. Estimate electability at the national level with direct estimation:

$$\hat{p}_1 = \sum_{i=1}^{33} \frac{N_i}{N} \hat{p}_{1i} \; i=1,2,\ldots, 33 \tag{3}$$

and

$$\hat{p}_2 = 1 - \hat{p}_1 \text{m} \tag{4}$$

with $\hat{p}_1$ direct estimated electability for Jokowi-Maruf and $\hat{p}_2$ for Prabowo-Sandi, $\hat{p}_{1i}$ direct estimated electability at provincial level for Jokowi-Maruf, N_i population in i-th province, and N total population at the national level.

2. Develop small area model with measurement error:

$$log\left(\frac{\hat{p}_{1i}}{1-\hat{p}_{1i}}\right) = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i, \; i=1,2,\ldots, 33 \tag{4}$$

$$\boldsymbol{X_i} = \boldsymbol{x_i} + \boldsymbol{\eta_i} \tag{5}$$

with $\boldsymbol{X}_i' = (1 \;\; X_{1i} X_{2i})$ as observed auxiliary information vector at i-th province, $\boldsymbol{x_i}$latent auxiliary information vector at i-th province, $\boldsymbol{\eta_i}$measurement error vector at i-th province with $\boldsymbol{\eta}_i \sim iid \; \boldsymbol{N}(\boldsymbol{0}, \sigma_{\eta_i}^2 \boldsymbol{I})$, $\boldsymbol{\beta} = (\beta_0 \beta_1 \; \beta_2)'$ **as** regression parameter vector, v_i random effect at i-th province with $v_i \sim iid \; N(0, \sigma_v^2)$, and σ_v^2 provincial random effect variance.

3. estimate electability at the provincial level with indirect estimation
4. aggregate indirect estimated electability to estimate electability at the national level with formula :

$$p_1^H = \sum_{i=1}^{M} \frac{N_i}{N} p_{1i}^H \; i=1,2,\ldots, 33 \tag{6}$$

and

$$p_2^H = 1 - p_1^H \tag{7}$$

with p_1^H direct estimated electability for Jokowi-Maruf and p_2^H for Prabowo-Sandi.

5. compare indirect estimated electability at provincial level with The 2019 Presidential Election's result.
6. compare both direct estimated electability and indirect estimated electability at national level with The 2019 Presidential Election's result.

2 Results and Discussion

Results. Tweets gathered in early March 2019. The number of tweets collected is 634115. Tweets will be separated based on topics. Tweets contained topics about Jokowi-Maruf is 336449 and Tweets contained topics about Prabowo-Sandi is 247150. Tweets contained both topics will not be include in the analysis further. Each group of the topics will be derived to obtain sentiment for each candidate pair both at the provincial level or the national level. Keywords, collection time, and geocode are needed to collect tweets. Some problem may occur in this process. Keyword selection is very important to collect tweets representing about the topics. Tweets also needed to be collected at the right time. In this research, tweets were collected adjusted to the survey time. Time gap between survey time and tweets collection time can affect the result of the analysis. For example, if tweets collected months after survey there is a chance that tweets become irrelevant with survey data.

Sentiment score is basically computed by parsing words in tweets with collection of positive words and collection of negative words. Problem may occur in this process when

there are any tweets either typed incorrectly or typed with non-standard language. Same problem also may occur in removing stopwords. The problem described above lead to error in computing sentiment score. Another problem may occur when there is a sentence with negation, for example "kamu tidak hebat". In this sentence there is a word detected in the collection of positive words so the sentiment score will be positive but this sentence doesn't actually have positive sentiment.

Tweets are identified into three character, positive sentiment, negative sentiment, and neutral. Tweets with topics about Jokowi-Maruf have 27.35 % positive tweets, 16.68% negative tweets, and 55.97% neutral tweets. Tweets with topics about Prabowo-Sandi have 23.04 % positive tweets, 20.95% negative tweets, and 56.01% neutral tweets.

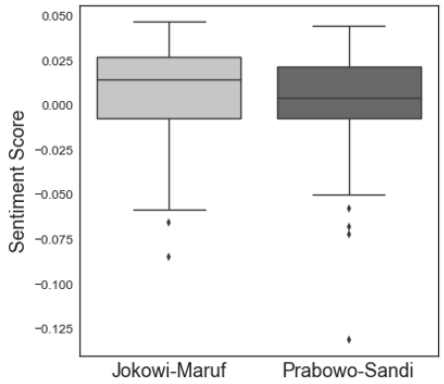

Fig. 1. Comparisson of the provincial sentiment score from each candidate pairs.

Sentiment score from each tweets will be aggregated for each province. Sentiment score from each tweets will be aggregated at the provincial level for each topics. From Figure 1 we can see that Jokowi-Maruf have higher sentiment than Probowo-Sandi. Some province have extreme sentiment for certain candidate pair. If we look at survey data in figure 2, this is in line with the superiority of Jokowi-Maruf in the survey.

There are some province that have very different value of the sentiment score among another provinces. Bali has the lowest sentiment score for Prabowo-Sandi and Lampung has the highest sentiment score for Jokowi-Maruf. These two province is known to be fanatical about Jokowi-Maruf. Kalimantan Selatan, Aceh, Maluku, and Nusa Tenggara are the four lowest sentiment for Jokowi-Maruf.

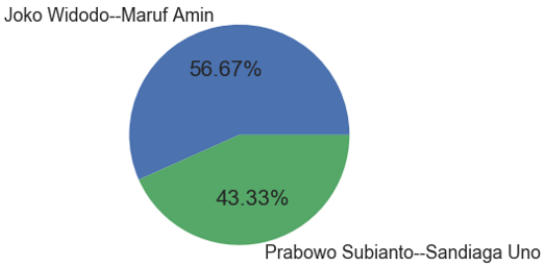

Fig. 2. Percentage of the the repondent who chose each candidate pairs.

Sentiment score will be compared with The 2019 Presidential Election's Result. Comparison will be done by look at who is better based on sentiment score at province level and who win in The 2019 Presidential Election's Result at provincial level.

There are 14 province which have higher sentiment for Jokowi-Maruf and is won by Jokowi-Maruf based on election's result. While, There are 9 province which have higher sentiment for Prabowo-Sandi and is won by Prabowo-Sandi based on election's result. Overall, there 23 of 33 province that can be predicted correctly who will be the winner if we used sentiment score.

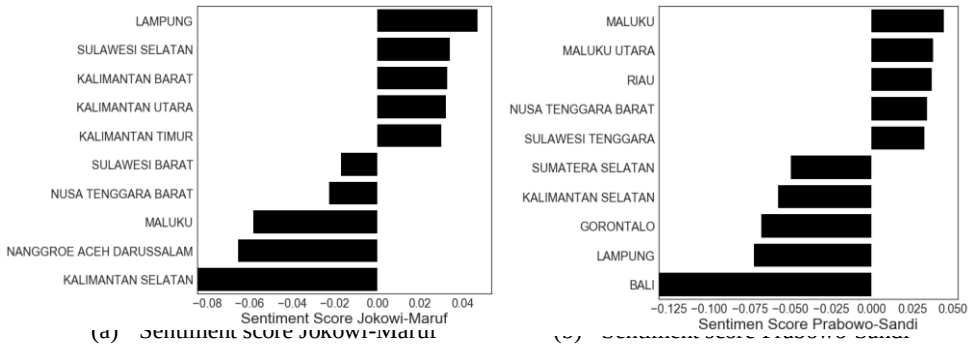

Fig. 3. Barchart of the sentiment score for each candidate pair, (a) top five and bottom five of Jokowi-Maruf's sentiment, (b) top five and bottom five of Prabowo-Sandi's sentiment.

Small area model area level will be developed to estimate electability indirectly. Subject in this modeling is 33 province in Indonesia. Auxiliary information used are Jokowi-Maruf sentiment and Prabowo-Sandi sentiment. Small area model used is binomial Generalized Linear Mixed Model with province as random effect. The candidate pair that become the indicator at this modelling is Jokowi-Maruf.

Table 1. Comparisson of sentimen score and The 2019 Presidential Election'result at province level.

Higher sentiment at the provincial level	Winner at The 2019 Presidential Election at the provincial level	
	Jokowi-Maruf	Prabowo-Sandi
Jokowi-Maruf	14	4
Prabowo-Sandi	6	9

Jokowi-Maruf sentiment have positive significant effect to The Jokowi-Maruf's electability at the 5% level. Prabowo-Sandi sentiment have negative significant effect to The Jokowi-Maruf's electability at the 5% level. Sentiment can describe movement of the electability at provincial level. Based on the modeling, indirect estimate used to predict electability for each candidate pair can be seen below:

$$p_{1i}^{H} = \frac{exp(-0.1818+17.9443x_{1i}-11.0036x_{2i}+\hat{v}_i)}{1+exp(-0.1818+17.9443x_{1i}-11.0036x_{2i}+\hat{v}_i)}, i = 1,2,\dots,33 \quad (8)$$

$$p_{2i}^{H} = 1-p_{1i}^{H} \quad (9)$$

Table 2. Estimated model parameters.

Parameters	Estimated	stdev	z-score	Pr(>|z|)
Intercept	-0.1818	0.1739	-1.049	0.29572
Jokowi-Maruf Sentiment	17.9443	6.0928	2.945	0.00323
Prabowo-Sandi Sentiment	-11.0036	4.7046	-2.339	0.01934
Random effect variance	0.7431			

The 2019 Presidential Election's result will be used to validate estimated electability both at the provincial level and the national level. The Pearson correlation coefficient between the estimated indirect electability in each province and The 2019 Presidential Election's result in each province is 0.7212. Direct estimate method has smaller different in result than Indirect estimate but indirect estimate is able to estimate electability at provincial level.

Table 3. The national level estimate.

Method	Jokowi-Maruf (%)	Prabowo-Sandi(%)
Direct	56.430	43.570
Indirect (small area estimation)	57.048	42.951
The 2019 Presidential Election's Result	55.310	44.690

3 Conclusion

The 2019 Presidential Election's result will be used to validate estimated electability both at the provincial level and the national level. The Pearson correlation coefficient between the estimated indirect electability in each province and The 2019 Presidential Election's result in each province is 0.7212. Direct estimate method has smaller different in result than Indirect estimate but indirect estimate is able to estimate electability at provincial level.

Based on survey's result, Candidate pair number 01 is more superior than Candidate pair number 01 in the national scale. This was accompanied by a better sentiment from the Candidate pair number 01 at the national level. Sentiment in each province obtained from Twitter is able to describe the electability of candidate pairs in each province. Twitter is quite capable to become a source of auxiliary information in small area estimation to estimate

electability of the candidate pairs. Sentiment score from each province area capable to predict correctly who will win in Presidential election 2019 in 23 of 33 province in Indonesia. Sentiment score from national capable to predict correctly who will win in Presidential election 2019. The Pearson correlation coefficient between the estimated indirect electability in each province and the vote acquisition result of the 2019 presidential election in each province is 0.7212.The value indicates that the estimation of the electability used small area estimation method by utilizing Twitter as proved of auxiliary variable capable of describing the 2019 Presidential Election's result.

Based on summary above, twitter can quite desribe electability of the candidate pairs of president Republik Indonesia. Twitter can be used as source of auxliary information to estimate the electability of the candidate pairs of president Republik Indonesia.

Table 4. The provincial level estimate.

No	Province	ID	Jokowi-Maruf (%)	Prabowo-Sandi (%)	$\widehat{rmse}$ (%)
1	Aceh	11	32.47	67.53	9.11
2	Sumatera Utara	12	23.64	76.36	4.52
3	Sumatera Barat	13	19.42	80.58	8.43
4	Riau danKep. Riau	14	42.97	57.03	7.38
5	Jambi	15	46.68	53.32	8.11
6	Sumatera Selatan	16	43.10	56.90	6.38
7	Bengkulu	17	57.56	42.44	8.61
8	Lampung	18	51.08	48.92	9.43
9	Kep. Banka Belitung	19	41.03	58.97	7.78
10	Dki Jakarta	31	57.87	42.13	3.87
11	Jawa Barat	32	62.71	37.29	2.95
12	Jawa Tengah	33	71.05	28.95	2.99
13	DI Yogyakarta	34	72.74	27.26	10.30
14	JawaTimur	35	76.94	23.06	2.71
15	Banten	36	37.49	62.51	5.68
16	Bali	51	82.16	17.84	7.16
17	NTB	52	13.04	86.96	6.52
18	NTT	53	84.21	15.79	9.19
19	Kalimantan Barat	61	32.90	67.10	8.15
20	Kalimantan Tengah	62	36.41	63.59	8.16
21	Kalimantan Selatan	63	21.47	78.53	7.42
22	Kalimantan Timur	64	50.75	49.25	8.72
23	Kalimantan Utara	65	55.44	44.56	8.52

24	Sulawesi Utara	71	78.40	21.60	8.92
25	Sulawesi Tengah	72	53.53	46.47	10.21
26	Sulawesi Selatan	73	52.23	47.77	6.74
27	Sulawesi Tenggara	74	23.31	76.69	9.74
28	Gorontalo	75	41.13	58.87	10.43
29	Sulawesi Barat	76	39.32	60.68	8.25
30	Maluku	81	9.54	90.46	5.35
31	Maluku Utara	82	40.31	59.69	7.57
32	Papua	91	73.96	26.04	9.53
33	Papua Barat	92	82.46	17.54	9.04

References

[1] Rao, J.N.K., Molina, I.:Small Area Estimation. New Jersey (US) : John Wiley & Sons, Inc (2015)

[2] Notodiputro, K.A., Kurnia, A.:Development Of Small Area Estimation Research In Indonesia. *IcoMs* (2007)

[3] Kwartler, T.: Text Mining in Practice with R. Chennai (India): John Wiley & Sons, Inc. (2017)

[4] Porter, A.T., Holan, S.H., Wikle, C.K., Cressie, N.: Spatial Fay-Herriot Models for Small Area Estimation with Functional Covariates. NIASRA. Vol. 28, pp. 8-13 (2013)

[5] Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Padreschi, D., Rinzivilio, S., Pappalardo, L., Gabrielli, L.: Small Area Model-Based Estimators Using Big Data Sources. *JOS*. Vol. 32, pp. 263-281 (2015)

[6] Sadik, K., Notodiputro, K.A.: Small Area Estimation with Time and Area Effect Using Dynamic Linear Model. *IcoMs*(2008)

[7] Dickerson, J.P., Kagan, V., Subrahmanian V.S., Using Sentiment to Detect Bots on Twitter: Are Humans more Optionated than Bots?. ASONAM (2014)

[8] Tala, F.Z.: A Study Of Stemming Effects On Information Retrieval In Bahasa Indonesia. Amsterdam (NL):Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands (2003)

[9] Stroup, W.W.: Generalized Linear Mixed Models Modern Concepts, Methods and Aplications. Boca Raton: CRC Press (2013)

[10] Stefanski, L.A., Carrol, R.J.: Covariate Measurement Error in Logistic Regression. The Annals of Statistics. Vol 13, pp. 1335-1351 (1985)