Analyzing Netizens' Perceptions Towards Indonesian Presidential Candidates Using Topic Modeling Approach

IOP Conf. Series: Materials Science and Engineering **546** (2019) 052037 doi:10.1088/1757-899X/546/5/052037

Analyzing Netizens' Perceptions Towards Indonesian Presidential Candidates Using Topic Modeling Approach

Devi Karolita[1*], **Ariesta Lestari**[1]

[1]Department of Informatics Engineering, Faculty of Engineering, Palangka Raya University

*Corresponding author: devikarolita@it.upr.ac.id

Abstract. Over the past few years, Twitter has significantly grown as the microblogging platform. Millions of user use this platform to share their attitudes, views, and opinion on a daily basis. This phenomenon has been used to promote people's attention towards some event, such as 2019 Indonesian Presidential Election. In this study, we investigate people's online opinions towards the event through social media. The goal of the study is to discover frequent topics amongst netizens' tweets during the election campaign. We collected tweets containing the names of the candidates, then applied topic modelling approach using Latent Dirichlet Allocation (LDA) method to cluster the topics. Based on the experiment, the tweets are clustered into ten topics with different focuses e.g., a topic discusses the candidate's position towards sensitive issues, a topic about the community supports towards one presidential candidate. Our result shows that topic modelling approach can be used to analyse people's perception in social media towards an important event.

1. Introduction

Indonesia 2019 Presidential Election, which was scheduled on 17 April, will be a competitive battle of two old rivals: Joko Widodo and Prabowo Subianto [1]. This fact raises opinions from the residents of Indonesia towards hot political situation of the upcoming election. Social media had been used both by the candidates and the netizens to express ideas and opinions related to the election. Another advantage of using social media for political campaign is to predict the popularity of candidates among netizens and the demography of the netizens who express their opinions toward the candidates [2].

Twitter as a micro-blogging platform social media which is limited by 280 characters is often seen as an effective channel to express opinions. The rapid use of Twitter as one of publicity tools for political campaign in Indonesia was started in 2012 DKI Jakarta Gubernatorial Election. The usage of Twitter as a platform to express netizens' opinion regarding political situation increased during 2014 Indonesia's Presidential Election. Therefore, Twitter data is considered as a reliable source in observing citizens' sentiment toward current political situation and claimed to have a better prediction than offline polling conducted by several independent survey institutions [3]. Therefore, this phenomenon leads us to focus our study in investigating mostly discussed topics on Twitter in order to gather the opinions toward 2019 Indonesia's Presidential Election.

2. Related Work

Twitter data have characteristics which differentiate it from Internet Web page data, such as very short, sparse, and spreading rapidly. These natures of the data make it challenging to detect current

trending topics on Twitter. Frequent Pattern Mining is one of the approaches that is recently used for topic detection on Twitter. [4]used FP-stream to mine tweets related to swine flu and observe the performance of the algorithm in terms of time and memory consumption. However, FP-stream only considers the frequency of the words and disregard their utility. High Utility Pattern Mining is an enhancement of FP-stream which takes utility of the words into account for detecting topics on Twitter to find groups of words with high frequency and high utility [5].

Topic Modelling is also used for a particular purpose, for instance political use. During the political year, Twitter is also considered as an effective platform both for campaign and expressing opinion. US Presidential Election in 2008 was one of the remarks of Twitter's usage for campaign purpose. The transcript of each candidate's speeches was analyzed to determine overused and underused terms from both a statistical and dynamic perspective[6]. Other research was not only using Twitter data but also messages on Facebook. The objective of the research is to discover each candidate focus on their campaign [7].

3. Method and Results

In this section, we introduce our methods of sampling and collecting tweets. In addition, we also explain Latent Dirichlet allocation (LDA) to discover topics in the tweet dataset.

3.1. Data Sampling and Collection

We used Twitter, an R package which provides access to the Twitter API [8], to acquire data from Twitter. In order to obtain a representative sample of tweets related to the Indonesian Presidential Election, the tweet dataset was gathered during the campaign periods. We used the public streaming Twitter Search API, and able to gather 33,048 tweets, starting on 1st February 2019 until 7 February 2019. Since the purpose of our experiments is to discover the frequent topics discussed by netizens during the Indonesian Presidential Election, keywords contain the candidate names such as "jokowi", "marufamin", "prabowo" and "sandiagauno" were used as the filter. The dataset contains *tweets* with maximal 140 characters, *created date*, *user id*, *retweetcount*, *isRtweet*, *retweeted*.

In order to have a more efficient way to process the data, we administered text processing with the objective to attain cleaner and more compact data. The first step of text preprocessing is tokenisation where the texts are split into smaller pieces or tokens. Secondly, less important words, which is called stopwords, are removed in order to have more focus on important words.

The preprocessing steps were conducted in order to ensure the dataset is clean from error and noise, consistent and ready to be analysed using data mining process (source). First, we cleaned the dataset by removing unnecessary characters, such as mentions @, hashtags #, punctuations, special characters, links and numbers from the tweets. We also removed stopwords from the tweets using the Indonesian stopwords dictionary generated from[9]. We added some common terms in the twitter such as *like*, *rt* and *retweet* in the stopwords dictionary. We converted the tweets into lower cases and removed the whitespaces.

3.2. Dataset Analysis using LDA

We analysed the tweet dataset by performing LDA to produce a topic model. LDA is a model to generate probabilistic of a corpus which is an enhancement of the Bayesian Model. This model characterizes topics by a distribution over words. Moreover, LDA is one of the simplest topic model introduced by [10] and have an illustrative way in presenting the result. Another advantage of LDA is its ability to reduce the dimensionality of the document in which is often considered as a challenge in document classification [11].

We configured LDA to generate the topics distribution and started with five topics then increased the topic number by increments of 5 until it reached 100 topics. We chose LDA model with 10 topics, due to with this number the topic model provided more relevant topics relating to the Indonesian Presidential Elections. InTable 1, we present a result of a 10-topic model trained on the entire tweet dataset. The label of each tweet is manually summarizing based on the top words in each topic.

However, some topics are much harder to label than others. Some topics are simply less clear, partlybecause of the random and unrelated terms inside the topics. For example, there is a lack of clarity in topic 5. Each term in this topic seem to be unrelated, even though the top terms refer to the candidate, other terms have not connected to the candidates.

Table 1.10-Topics disscussed over the tweet dataset

Topic	Focus	Terms
1	Netizenpilihjokowi	caprespilihannetizen, jokowi, pemimpin, media, jujur, masyarakat, pilih
2	Dukunganaminmaruf	amin, maruf, dukung, padang, mui, ketua, cawapres, sumbar, jakarta
3	Dukungan alumni	alumni, media, dukung, pasangan, pangudi, luhur, deklarasi, sekolah, joko, widodo
4	MarufMinang	maruf, jokowi, kyai, ulama, anak, minang, maaf, urang, duo
5	Prabowojokowi	Prabowo, jokowilagi, politik, jokowiamin, lgbt, psi rusia, simbol, gerakan
6	Jokowimaruf	Jokowi, prabowo, jawa, provinsi, kyai, bpn, memilih, bilang, milenial
7	Pilpres Indonesia	Pilpres, indonesia, radikal, perang, moderat, ideologi, ekonomi, keadilan, pertumbuhan
8	Sandiagauno	Sandiaga, uno, paslon, warga, pesantren, pasar, disambut, kedatangan, bukti, emakemak
9	Mendukungaminmaruf	Jokowi, presiden, janji, pengurus, nasional, pengusaha, daerah, partai,
10	Prabowo	Prabowo, rakyat, negara, konsultan, hoax, kubu, pendukung, jusuf, asing, kemenangan

A visualization of the distribution of words for the two meaningful topics is given by the wordclouds in**Figure *1***a and **Figure *1***b. The size of the words corresponds to their relative weights; words having a large weight are more often generated by this topic. These two topics are related to the Indonesian presidential candidate. This visualization helps us understand the terms/words associated with the two candidates. Topic 1 which represents Jokowi has most common words such as "*pemimpin*", "*masyarakat*", "*pilih*", and "*jujur*". While the other candidate, Prabowo, represents in topic 10 which could be associated with terms, "*rakyat*", "*negara*", "*konsultan*", and "*pendukung*.

Figure 1. Wordcloud for topic 1 and 10, represents two President candidate. (a) Wordcloud for Jokowi. (b) Wordcloud for Prabowo.

Next, we determined the number of tweets having the same main topic. We examined the per-document-per-topic probabilities (gamma). Each of these values is an estimated proportion of words from that document that are generated from that topic.[10]mentioned that the basic idea of LDA is that multiple topics in different proportion can be revealed in documents. In this paper, we only interested in discovering the main topic of a document which is defined as the topic with the largest probability. For each tweet in our tweet dataset, we identified the topic index for which the probability is the largest, i.e., the main topic. Grouping by the topic index, counting, and sorting results in the counts of documents per topics plotted inFigure 2. This figure shows that more than 5,000 tweets are clustered in topic 1 and the less frequent discussed topic, which is topic 4, has approximately 2,300 tweets.

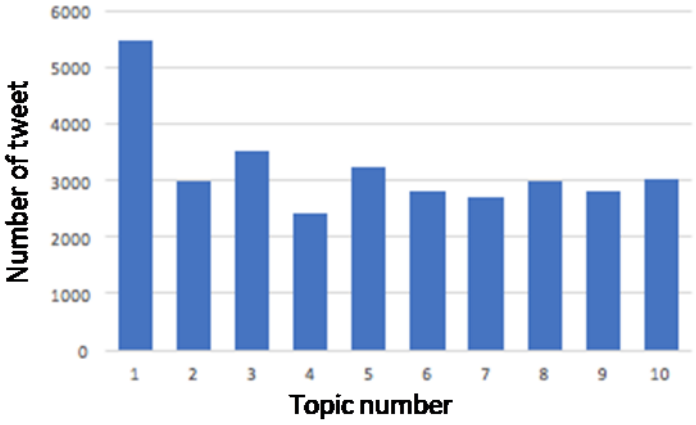

Figure 2. Number of tweet for each topic.

4. Discussion and Conclusion

In this study, we explore the realistic application and effectiveness of LDA to learn more about people's perception in social media towards the Indonesian Presidential Election. As expected, the implementation of LDA over a large dataset of tweets able to provide some topics related to the political issues. As shown in Table 1, there are topics that straightforward mention the candidate names and situation related to their campaign.

Through the topics clustering, we also can see the netizen reaction towards the presidential candidates and their campaign. Our result in Figure 3 shows that topic about Jokowi is the most

frequent discussed topic among the netizens. This result is in line with the polling result from Lembaga Survey Indonesia that stated Jokowi - Maruf electability reached 58%, while the opponent only reached 32% [12].Based on the two topics about the president candidate shown in Figure 1could be used to analyses the perception of people towards the candidates. In topic 1, JokoWidodo is associated with terms "*pemimpin*", "*jujur*", and "*masyarakat*". While topic about Prabowo contains the term "*negara*", "*rakyat*", "*asing*", and "*hoax*". For this study, we gathered 33048 tweets associated with 2019 Indonesian Presidential Election candidates starting from 1st February 2019 until 7 February 2019. These 140 characters-tweets have attributes such as *created date, user id, retweetcount, isRtweet, retweeted*.

The first objective of the study is to cluster the tweets related to the candidates into ten topics that most frequently discussed on Twitter. Then we counted the number of the tweets for each most frequent topics to determine which topic draws most netizens' attention. The result showed that it is reliable to use Twitter data in order to observe netizens' perception regarding the Presidential Election in Indonesia.

For future work, we would like to expand the study by eliminating noise in our data set such as tweets generated by computer bots, also tweets made by paid and fanatic users.

References

[1] G. Fealy, "Commentary: Old rivals face off in Indonesia ahead of elections Read more at https://www.channelnewsasia.com/news/commentary/jokowi-prabowo-indonesia-2019-elections-presidential-parliament-11082762," 2019. [Online]. Available: https://www.channelnewsasia.com/news/commentary/jokowi-prabowo-indonesia-2019-elections-presidential-parliament-11082762. [Accessed: 05-Feb-2019].

[2] L. A. Abdillah, "Indonesian's presidential social media campaigns," ArXiv14098372 Cs, Sep. 2014.

[3] M. Ibrahim, O. Abdillah, A. F. Wicaksono, and M. Adriani, "Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation," (:unav), Nov. 2015.

[4] J. Guo, P. Zhang, JianlongTan, and L. Guo, "Mining Hot Topics from Twitter Streams," Procedia Comput. Sci., vol. 9, pp. 2008–2011, 2012.

[5] H.-J. Choi and C. H. Park, "Emerging topic detection in twitter stream based on high utility pattern mining," Expert Syst. Appl., vol. 115, pp. 27–36, Jan. 2019.

[6] J. Savoy, "Lexical Analysis of US Political Speeches," J. Quant. Linguist., vol. 17, no. 2, pp. 123–141, May 2010.

[7] J. Ryoo and N. Bendle, "Understanding the social media strategies of US primary candidates," J. Polit. Mark., vol. 16, no. 3–4, pp. 244--266, 2017.

[8] J. Gentry, "Package 'twitteR.'" 2014.

[9] F. Z. Tala, "A study of stemming effects on information retrieval in Bahasa Indonesia," M.S. Thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam, 2003.

[10] D. M. Blei, "Probabilistic topic models," Commun. ACM, vol. 55, no. 4, p. 77, Apr. 2012.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993--1022, 2003.

[12] D. Nurita, "Survei LSI Denny JA: Jokowi 58,7 Persen, Prabowo 30,9 Persen," 2019. [Online]. Available: https://pilpres.tempo.co/read/1182053/survei-lsi-denny-ja-jokowi-587-persen-prabowo-309-persen. [Accessed: 05-Mar-2019].