IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

Summarizing Netizens' Sentiments Towards the 1st Indonesian Presidential Debate using Lexicon Sentiment Analysis

To cite this article: Ariesta Lestari and Devi Karolita 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **546** 052041

Summarizing Netizens' Sentiments Towards the 1st Indonesian Presidential Debate using Lexicon Sentiment Analysis

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

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

Corresponding author: ariesta@it.upr.ac.id

Abstract. Twitter is one of the popular social media platforms in Indonesia. This platform has been used as a media communication and public engagement tool for many purposes, especially in political and governance domains. During the process of 2019 Indonesian Presidential Election, many people use Twitter to express their opinion/sentiment towards the election process. In this paper, we investigate the nature of people's opinion towards the Indonesian Presidential Election after the 1st debate. The goal of this study is to perform exploratory sentiment based analysis of Twitter data, and that was gathered after the 1st debate. We used lexicon sentiment analysis to calculate the sentiment of political tweets collected after the 1st debate. The identification of positive and negative opinion was automatically conducted using the available dictionary. Our result shows that sentiment of the netizen towards the 1st Presidential debate was mostly negative. In addition to this result, a predictive model was generated using CART and logistic regression to predict the netizens' sentiment. This experiment shows that the accuracy of the prediction model reaches 90%. Therefore, our study suggests that Twitter data can be used to analyse citizens' sentiment toward the Indonesian Presidential Debate and can generate a model to predict citizens' future sentiment toward the next debate.

1. Introduction

The use of social media has been increased over the last few years. It is originally used to interact with other people, nowadays social media is also frequently used to express the opinion towards current trending topics. Approximately 88.1 million out of 260 million population in Indonesia had access to the internet in the beginning of 2016 [1]. Around 90 percent of them were active social media users and they spent more time using the social media than watching television. During political year, social media is effective tool for political cyber campaigns and the target is the first voters because the citizens already friendly with the social media and the conventional campaigns became outdated [2]. Twitter is a micro-blogging platform social media which is often used as a media to read and share the news, let alone the opinion sharing. Additionally, proliferation of smartphones has further facilitated the use of this medium, and makes the number of Twitter's users has been increasing rapidly in the last decade. This phenomenon has been recognized by the politicians and political parties for political campaign purpose. In Indonesia, DKI Jakarta gubernatorial election in 2012 has been a turning point on using Twitter as media for campaign. Twitter is considered as an effective media because the opinions expressed go straight to the main point due to its character limitation. In addition, Indonesia's Twitter users, where half of them are adolescents, can easily access latest trending topics through hashtags. For example, in 2012 DKI Jakarta gubernatorial election the trending hashtag #ReplaceTitleSongWithJOKOWI was used to promote one candidate's electability [3]. The awareness of using Twitter for political purpose was growing up during 2014 political year. Three out of ten popular hashtags in 2014 are talking about President-candidate and current Indonesian President Joko 'Jokowi' Widodo, i.e. #AkhirnyaMilihJokowi, #TegasPilih2, and #Salam2Jari which are recorded more than 78 million mentions [4].

Based on the trend of Twitter's usage for expressing opinion regarding current political situation, we attempt to do a research about netizen's sentiment toward Indonesia current political event, in particular 1st debate Presidential Election. In this study we gathered Twitter data after the 1st debate and calculate the sentiment of political tweets using lexicon sentiment analysis.

2. Related works

Twitter is one of the most popular social media, especially because it is a micro-blogging platform where people can express themselves. One of the advantages of using Twitter data is to observe netizens' opinion toward current political situation. Twitter dataset is also proven can be used to observe netizens' sentiment in 2016 US Presidential Election. A method named SentiStrength was used to analyze the sentiment of topics expressed on Twitter related to the election. The result of the study showed that one candidate offered a more optimistic and positive campaign message than the other candidate [5]. [6] used twitter data to analyse the sentiment of Twitter's users towards the candidates of 2016 US Presidential Election. The sentiments are calculated using lexicon and Naive Bayes Machine Learning Algorithm and the result then is compared to the polling data to see how much correlation they share.

Indonesia's netizens also use the platform to express their concerns about political situation that happened currently, especially during election both the gubernatorial and the presidential one. During the Jakarta's gubernatorial election in 2017 the experiment regarding netizens' sentiment was conducted and the result was compared to the result of the election itself [7]. The study used Multinomial Naive Bayes as the predictive approach and Support Vector Machine to classify the dataset. [8] conducted a study on twitter data related to Indonesia Presidential Election using lexicon based approach. In the study, they made an algorithm and method to count important data, top words and train the model and predict the polarity of the sentiment.

3. Methodology

The proposed system consists of two modules: (1) Data pre-processing module (2) Sentiment Analysis module which consists of Lexicon Sentiment Analysis and Classification and Regression Tree (CART).

3.1. Data Pre-processing

3.1.1. Tokenization

The aim of tokenization is to convert each word or term in a document into a distinct attribute. The long strings of the text will be split into smaller pieces or tokens. Tokenization is also referred to as text segmentation or lexical analysis.

3.1.2. Stopword removal

Stopwords are a set of commonly used words in any language. The most common words in text documents are articles, prepositions, and pronouns, etc. that does not give the meaning of the documents. The objective of removing the stopwords is to reduce the dimensionality of term spaces and have more focus on the important words.

3.2. Lexicon Sentiment Analysis

Lexicon Sentiment Analysis (LSA) is an approach that is also called a dictionary approach and relies on a lexicon or dictionary of words with pre-calculated polarity [9]. LSA is considered to be part of the Machine Learning Unsupervised approach as well. The first step of LSA is data pre-processing, then followed by sentiment score calculation. To calculate the score [8], each word from the bag-of-words gets compared against the lexicon. If the word is found in the lexicon, the sentiment score of that word is added to the total sentiment score of the text.

Score = Number of positive words − Number of negative words

If Score > 0, the sentence has an overall positive opinion
If Score < 0, that the sentence has an overall negative opinion
If Score = 0, the sentence is considered to be a neutral opinion

3.3. Classification and Regression Tree (CART)

CART is machine learning method for constructing prediction models from data. The method was first introduced by [10]. According to [11], CART are intuitive methods, often described in graphical or biological terms. A tree is typically shown growing upside down, beginning at its root. The construction of the tree involves choosing the major issues: implementing splitting rules, and pruning process.

4. Experimental Result

4.1. Dataset

In this study, the tweets dataset was gathered using the public streaming Twitter API**.** Since the purpose of our experiments is to analyses the sentiments or twitters' user towards the Indonesian Presidential debate, we used keywords *"#debatpilpres2019"* as the filter. We have collected 3852 tweets, starting on mid-January 2019 until end-January 2019. The dataset contains *tweets* with maximal 140 characters, *created date*, *user id*, *retweetcount*, *isRtweet*, *retweeted*.

The lexicon dataset for analysing the sentiment score were gathered from [12], [13]. This dataset contains positive and negative words as shown in Table 1.

Table 1. Lexicon Dataset

Dataset	Words
Positive	*enthusiastic, aspirations, win, achievement, kind, beloved*
Negative	*apathetic, angry, fanatic, intimidation, disappoint, cheating*

4.2. Data pre-processing

The tweets dataset should be preprocessed 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 [14]. Common terms in the twitter such as *like*, *rt* and *retweet* were also removed, together with some frequent words, such as *www*, *http*, *debat*, *pilpres*, and *debatpilpres*. We converted the tweets into lower cases and removed the whitespaces.

4.3. Sentiment Analysis

The sentiment score was calculated using the equation in subsection 3.2. In this experiment, tweet is labelled as negative if the sentiment score <= -1, labelled as positive if the sentiment score >= 1, otherwise, the tweet is labelled as neutral.

The purpose of sentiment analysis in this paper is to identify the overall sentiment of the Twitter conversations related with Indonesian 1st Presidential Debate 2019. After calculating the sentiment score of all the tweets, we discovered the scores mostly in the range of 0 to -3 with -2 as the median scores (see Figure 1). This scores shows that in general the Indonesian netizens' sentiment towards the 1st Indonesian Presidential Debate is mostly negative.

Table 2 shows some tweets that are labelled as negative, positive and neutral based on the sentiment scores.

The purpose of sentiment analysis in this paper is to identify the overall sentiment of the Twitter conversations related with Indonesian 1st Presidential Debate 2019. After calculating the sentiment score of all the tweets, we discovered the scores mostly in the range of 0 to -3 with -2 as the median scores (see Figure 1). This scores shows that in general the Indonesian netizens' sentiment towards the 1st Indonesian Presidential Debate is mostly negative.

Table 2. Sentiment score of the netizen's tweets

Tweets (translated into English)	Sentiment label
RT @MetroTVNewsRoom: Please watch the battle of vision and mission from both of Presidential Candidate. The theme is energy and food, natural resources and environmental issue	positive
Stupid rezim and hoax, forced to be 2 periods, crazy you bong *#HypocriteCoalition* *#JokowiKHMakrufAminWin2019â€¦ https://t.co/D0fAtJvOYz*	negative
There is new format in #DebatPilpres2019 the second candidate of President/Vice President, Number 3 will be the best. *#redwhite #Pilpres2019 https://t.co/aLr5Wv1hln*	neutral

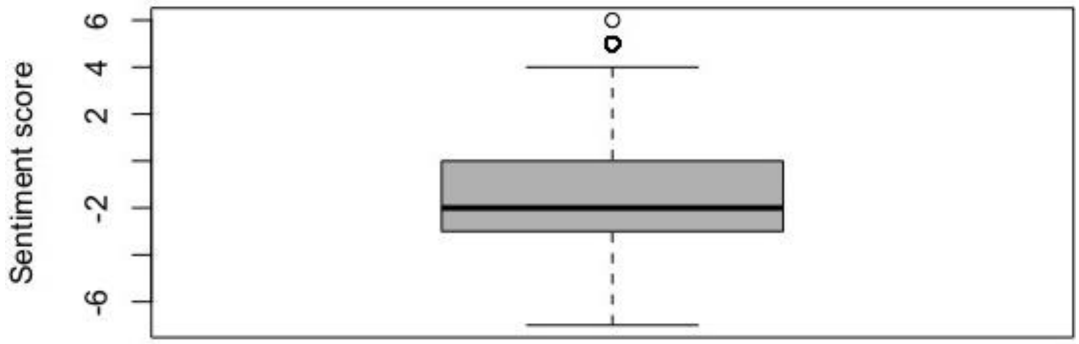

Figure 1. Boxplot of sentiment score

4.4. User behavioral model

In this study, we also aim to analyse the user behavior in creating the content of their tweets. We investigated the netizen's tweet in order to understand whether they are speaking their mind or reusing content or thoughts other users created by simply retweeting them. [15] discovered that number of retweets on Twitter usually ranged from 1.44% to 19.1% of all messages, which means only small number of tweets that retweeted. Our finding shows a contrary result, the number of retweet in the dataset is as high as 80%. This result reveals that users in our dataset are most likely reusing the present information/content instead of creating a new tweet to express their own opinion towards the Presidential Debate.

A study from [16] discovered that tweets with negative sentiment are more viral compared to tweets with positive sentiment. Therefore, in this study we also explored the user behaviour regarding reusing the content to express their opinion. We analyse the sentiment of these retweets, whether positive, negative of neutral sentiment that most likely to be retweeted. Similar with the finding from [16], the analysis shows that more than 70% of the retweets have negative sentiment and only 20% and 18% of the retweets are positive and neutral, respectively (see Table 3). This result then confirms the finding in the previous section, that most of the netizens' sentiment towards the debate is negative.

Table 3. Retweet vs pure tweet

Tweet type	Sentiment		
	Positive	Negative	Neutral
retweet	20%	73%	7%
pure tweet	22%	18%	59%

4.5. Predicts the users' sentiment towards the Indonesian 1st Presidential Debate

In the previous section, we analyzed the users' opinion and their online behaviour towards the Presidential Debate. Here, we propose probabilistic models for classifying the sentiment of a tweet, whether it has positive, negative or neutral sentiment. We used the cleaned dataset from subsection B to score the weight of terms in the tweet dataset. Term Frequency-Inverse Document Frequency (TF-IDF) method is applied to score the weight of terms based on how frequently they appear across multiple tweets.

In order to prepare the dataset for the classification task, each tweets were labelled; positive, negative or neutral; according to the sentiment score calculated in section C. Further pre-processing process were conducted by removing stop words and removing terms that occurred less than five times. After the pre-processing, we split the dataset into training and testing dataset. Training set has 2,696 tweets (898 positive tweets, 1482 negative tweets and 216 neutral tweets), while testing set has 1,156 tweets (393 positive tweets, 635 negative tweets and 128 neutral tweets). In this experiment, we applied supervised machine learning methods using CART technique. We build CART classification regression tree model on the training dataset and test the model using the testing dataset. Table 4 shows the confusion matrix as the result of the experiment based on the test dataset. The accuracy of the model in classifying a tweet whether it has positive, negative or neutral sentiment is quite high, reaching 90%.

Table 4. Confusion matrix of the test dataset

Actual class	Predicted class	
	False	True
False	520	1
True	101	534

Actual class	Predicted class	
	False	True
False	757	6
True	103	290

a. Negative Sentiment

b. Positive Sentiment

Actual class	Predicted class	
	False	True
False	1028	0
True	71	57

c. Neutral Sentiment

5. Conclusion

For this study, we gathered 3,852 Twitter messages associated with Indonesian Presidential Debate of 2019. These tweets were collected during mid-January until the end of January, after 1st debate that was hold on 17 January 2019. The tweets were filtered based on hashtag #debatpilpres2019.

The purpose of the study was to investigate the nature of political discourse that took place on Twitter during the elections in terms of sentiment, content, and users' sentiment prediction. Based on the study we concluded that most netizens' had negative sentiment towards the 1st Indonesian Presidential Debate. A user behavioral model was developed to observe if the users expressed their own opinion or just retweeted the existed tweets. The result showed that most of the content were rather retweeting instead of original content. We also classified the tweets to positive, negative, and neutral ones with 90% of accuracy.

In the future, we would expand our study by comparing the netizens' sentiment during the Presidential Election to the result of the election itself. We also would like to add users attributes such as number of followers and friends, duration of activity on Twitter, number of messages to date and reputation and observe the contribution of these attributes to similar research.

References

[1] A. C. Johansson, "Social Media and Politics in Indonesia," *Stockh. Sch. Econ. Asia Work. Pap.*, vol. 42, 2016.

[2] L. A. Abdillah, "IT based social media impacts on Indonesian general legislative elections 2014," presented at the 9th AASRC international Conference on Innovative Trends in Management, Information, Technologies, Computing and Engineering to tackle A Competitive Global Environment (ITMITCE – 2014), Crown Plaza (Istanbul-Harbiye) Hotel, Istanbul, Turkey, 2014.

[3] Merdeka, "Hashtag #ReplaceTitleSongWithJOKOWI ramai di Twitter," 2012. [Online]. Available: https://www.merdeka.com/teknologi/hashtag-replacetitlesongwithjokowi-ramai-di-twitter.html. [Accessed: 30-Jan-2019].

[4] Tempo, "Twitter Releases 2014 Most Popular Hashtag in Indonesia," 2018. [Online]. Available: https://en.tempo.co/read/628434/twitter-releases-2014-most-popular-hashtag-in-indonesia. [Accessed: 30-Jan-2019].

[5] U. Yaqub, S. Chun, V. Atluri, and J. Vaidya, "Sentiment based analysis of tweets during the us presidential elections," in *Proceedings of the 18th annual international conference on digital government research*, 2017.

[6] B. Joyce and J. Deng, "Sentiment analysis of tweets for the 2016 US presidential election," in *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, 2017, pp. 1-4.

[7] A. S. Nugroho and A. Doewes, "Twitter sentiment analysis of DKI Jakarta's gubernatorial election 2017 with predictive and descriptive approaches," in *2017 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, IEEE, 2017, pp. 89-94.

[8] W. Budiharto and M. Meiliana, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," *J. Big Data*, vol. 5, no. 1, p. 51, 2018.

[9] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," *Comput. Linguist.*, vol. 37, no. 2, pp. 267–307, Jun. 2011.

[10] L. Breiman, J. Friedman, R. Olshen, and C. Tong, "Classification and regression trees," *Wadsworth Int Group*, vol. 37, no. 15, pp. 237-251, 1984.

[11] G. Moisen, "Classification and regression trees," *Encycl. Ecol.*, vol. 1, pp. 582--588, 2008.

[12] D. H. Wahid and S. Azhari, "Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity," *Indones. J. Comput. Cybern. Syst.*, vol. 10, no. 2, pp. 207-218, 2016.

[13] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in *Proceedings of the 14th international conference on World Wide Web*, 2005, pp. 342-351.

[14] 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.

[15] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," in *Fourth international AAAI conference on weblogs and social media*, 2010.

[16] S. Tsugawa and H. Ohsaki, "Negative messages spread rapidly and widely on social media," in *Proceedings of the 2015 ACM on Conference on Online Social Networks*, 2015, pp. 151-160.