IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.13, No.1, January 2019, pp. 73~82
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.42531

Levels of Political Participation Based on Naive Bayes Classifier

Rumaisah Hidayatillah*[1], Mirwan[2], Mohammad Hakam[3], Aryo Nugroho[4]
[1,2]Department of Informatics Engineering, Universitas Narotama, Surabaya, Indonesia
[3]Department of Computer Systems, Universitas Narotama, Surabaya, Indonesia
[4]Faculty of Computer Science, Universitas Narotama, Surabaya, Indonesia
e-mail: ***[1]rumaisah@fik.narotama.ac.id**, [2]mirwan.14@fasilkom.narotama.ac.id,
[3]hakam@fik.narotama.ac.id, [4]aryo.nugroho@narotama.ac.id

Abstrak

Dewasa ini, media sosial tumbuh secara pesat dan menyeluruh hingga menjadi bagian penting di dalam kehidupan masyarakat. Pada masa kampanye pemilu di Indonesia, kandidat dan partai pendukungnya memanfaatkan media sosial sebagai media kampanye. Media sosial seperti Twitter telah dikenal sebagai media mikroblog politik yang menyediakan data terkait peristiwa-peristiwa politik yang sedang terjadi. Semua informasi dapat diperoleh dari tweet atau kiriman status pengguna. Dengan memanfaatkan Twitter sebagai sumber data, penelitian ini menganalisis partisipasi publik selama periode kampanye pemilihan kepala daerah provinsi Jawa Timur tahun 2018. Tujuan dari penelitian ini adalah mengetahui seberapa besar reaksi publik terhadap kandidat yang maju ke pemilihan. Data diperoleh dengan menggunakan program crawling, semua tweet yang mengandung nama masing-masing kandidat akan diunduh. Setelah melalui serangkaian tahapan preprocessing, data dapat diklasifikasikan menggunakan Naïve Bayes. Fitur prediktor pada dataset klasifikasi adalah jumlah reply, retweet, dan like. Sementara variabel targetnya adalah reaksi yang dibagi menjadi tiga kategori, yaitu high, medium, dan low. Kategori tersebut ditentukan berdasarkan tingkat reaksi yang diberikan pengguna terhadap suatu tweet.

Kata kunci*—media sosial, kampanye pemilu, naïve bayes*

Abstract

Nowadays, social media is growing rapidly and globally until it finally became an important part of society. During campaign period for the regional head election in Indonesia, the candidates and their supporting parties actively use social media as one of the campaign instruments. Social media such as, Twitter has been known as a political microblogging media that can provide data about current political event based on users' tweets. By using Twitter as a data source, this study analyzes public participation during campaign period for 2018 Central Java regional head election. The purpose of the study is to measure reaction given to each candidate who advanced in the election. Tweets containing certain candidate names were downloaded using the crawling program. After going through a series of preprocessing stages, data was classified using Naive Bayes. Predictor features in classification datasets are the number of replies, retweets, and likes. While the target variable is reaction with three levels, i.e., high, medium, or low. These levels were determined based on users' reaction in a tweet.

Keywords*—social media, election campaign, naïve bayes*

Received June 1st,2012; Revised June 25th, 2012; Accepted July 10th, 2012

1. INTRODUCTION

The election of regional heads in Indonesia is a routine event carried out simultaneously every five years. As a form of democratic event, this activity certainly involves public participation [1]. Therefore, before the elections are held there is a campaign period. Where advanced candidates are given time to attract people's attention and get as much support as possible. In the digital age like now campaigns conducted by political candidates involve various social media that are so close to the society. One of the most popular social networking sites right now is Twitter. In the 2016 United States presidential election, many people express their likes and dislikes of certain presidential decisions using this micro-blogging media [2]. In political science, social media is currently the key to understanding the nature of public opinion and political participation [3]. Right now, campaign activities and political news are increasingly moving to online media platforms, therefore many researchers are beginning to observe the political participation of social media users both in terms of demographics and political relations. For politicians and political parties, Twitter is used extensively in organizing campaigns, referendums, debates, and providing information about elections.

Registered social media users can give likes, comments and share certain topics. In a more comprehensive understanding, this pattern allows the use of social media to influence other users, not only for creating a sense of community [4]. Each social media has different characteristics. But one thing in common is that they connect individuals online. All interconnected individuals usually have a common interest in certain things. It's a sure thing that someone will feel more happy if faced with something that is in line with his opinion, while when faced with conflicting opinions one will feel stressed and forced to be asked to receive it [5]. With this friendship network pattern, information and political opinions are easily disseminated to certain groups, making it easier for election candidates to approach people and certain community. The success of using social media in political campaigns can be seen in the election of US president Barack Obama. Obama's victory was influenced by his social media campaigning in 2008 and 2012 [6]. Some analysts even revealed that Obama's victory was largely influenced by his online campaign strategy.

The most common interaction in social media is to influence other users [7]. Although it does not invite other people into groups or communities but the existence of these interactions gives the effect of social interaction and then form some kind of friendships networking. This will open up opportunities for public figures or political candidates to use Social Media as a tool to attract people's attention and support, by interacting in cyberspace. As it is known, almost a lot of public figures even the president has a social media account like Twitter. There will be interaction and reciprocity between users until finally the interaction data can be used as a source of problems to be examined. Twitter's rapid development has successfully attracted the attention of researchers from various scientific disciplines, in fact there is a research that examined the number of scientific publications mentioned on Twitter [8]. Until now there are still many researches that raise the role of Twitter in various aspects. One interesting aspect to be studied is the role of Twitter in political administration. There is research that study about 115 studies related to Twitter's role in politics [9]. In that study Twitter's role in politics was divided into three topic categories, including Twitter usage by politicians during the campaign, the use of Twitter by the public regarding campaign and election issues, and comments on Twitter related to political campaigns such as broadcast debates, party conventions, and election results.

Political campaign is an important phase where candidates try to get votes from the public. Social media users share a lot of important news and information during key moments of political events [10]. One of the first ways to understand political opinion is to classify sentiments in a tweet. Research has been done using Twitter as a corpus for sentiment analysis. One of the topics raised is about hate speech against immigrants. Tweets related to hate speech

against immigrants are used to create a dataset reference for automatic monitoring systems for hate speech [11]. Since digital communication technologies are increasing, online media can be used as a medium of hate speech that can affect users. So analytic observation is needed.

In recent years social networks are increasingly used as a source of data for the study of political opinion, observing the condition of campaigns, and predicting election results. In terms of data collection, the most widely used method is with the Twitter API. As in the case of a constitutional referendum that took place in Italy, Twitter took on the role of a data source to understand the pattern of the topics being discussed [12]. This research collects approximately one million tweets containing hashtags that refer to the referendum so that the analysis is only done on the relevant text. With that much data volume, an analysis of topic modeling was built using the Latent Dirichlet Allocation (LDA) model. LDA model was chosen because this model is very good if it works on large numbers of documents [13]. Analysis of the data was carried out to find out the most frequent words related to voting. There are positive words that support constitutional changes such as future and change. While words related to opposing voting such as fear and risk.

This research will try to examine the political participation of Twitter users in terms of the level of reaction given. The reaction rate will be classified into three levels, namely high medium, and low based on the parameters of the number of replies, retweets, and likes obtained. In this classification process, the Naive Bayes method is used. Naive Bayes is often used as a baseline, and consistently performs classification tasks very well. Therefore, this method is very popular in machine learning especially in the field of text classification. The step of this research is first to collect all tweets containing the names of each candidate who advanced in the 2018 Central Java gubernatorial election. Data collection is done using a crawling program based on the Python programming language, and collected during the campaign period. The next step is to prepare the dataset for further classification using the Naive Bayes classifier. The target of the classification is to find out how big is the reaction from public – especially Twitter users toward each governor candidates.

2. METHODS

2.1 Data Crawling

The process of collecting data is done using python based crawling tool. This tool does not use Twitter's API (Application Programming Interface). Because if we use the API, the data taken will be limited in number based on account, regional, trending topics, or keywords used [14]. By utilizing the crawling tool, the data gathering process can be done maximally and comprehensively. Data withdrawal is done by entering keywords in the form of each candidate's name into the tools, then the application will pull and download all tweets from Twitter's Search that contain the keywords entered.

Data collection with the crawling tool above is an example of application of data scraping methods, or the method used to extract data from a website [15]. Web scraping tools can access the World Wide Web directly using Hypertext Transfer Protocol, or through a web browser. In addition, web scraping can be done manually by using various programming tools available, such as in this study, that uses the Python programming language. This activity includes copying, where specific data is collected and copied from the web, usually to a local database or main spreadsheet, for the purpose of further analysis.

2.2 Data Preprocessing

Before being classified, the data will go through some preparation steps so that it becomes the desired dataset. This process is called data preprocessing. The aim is to get data with good representation so that it meets the data eligibility requirements. Data preprocessing is a step that must be passed in data mining. This is because, often encountered common problems

when extracting large amounts of data. For example, the information contained in the data is heterogeneous, making it difficult to process. Therefore, it is necessary to do the preprocessing stage called data cleansing, which is the process of filtering, modifying, and removing unnecessary data. Activities in data cleansing also involve adding missing value, reducing noise in the data, solving problems in inconsistent data, and eliminating unneeded parts. In this study, the tweet components taken as datasets are reply, retweet, and like. All of them are data containing numbers. This is intentionally done because the Naive Bayes classifier can provide better accuracy if the parameters are in numerical form [3].

Another problem faced in extracting data is when the values on the labels have a far range. This can affect the accuracy of the classifier. To overcome this, data transformation is done, which is the process of converting data from one format or structure into another format or structure. Data transformation can significantly influence parameters and estimation of uncertainty measurements [16]. The types of data transformations commonly used cannot be separated from the mathematical function equation. There are several types that are most commonly used, including square root transformation, logarithmic transformation, ArcSin transformation, wavelet transformation, and BoxCox transformation. In this study, the type of transformation used is logarithmic transformation, with the formula as follows:

$$\log(x + 1) \quad (1)$$

Where *x* is the original value, while the addition of number 1 is done to cover data that is 0 (zero). After making sure all the samples are in accordance with the needs of the dataset, the next step is to sort by class of reaction (high, medium, and low).

2.3 Naïve Bayes

Naive Bayes is one of the classification methods that is very popular in the machine learning world. Naive Bayes classification method depends on two basic assumptions, first the features are independent from one another. Second, each feature has the same prominence [17]. With these assumptions, Naive Bayes algorithm works based on an existing probability to determine future probabilities.

To understand more about Naive Bayes, it is necessary to first understand Bayes theorem. Bayes theorem is named after the inventor Thomas Bayes. The algorithm works based on conditional probability, which is a measure of the probability that something will happen based on events that have happened before. Here is the probability equation that underlies Naive Bayes:

$$P(c|x) = \frac{P(x|c) \times P(c)}{P(x)} \quad (2)$$

where *P(c|x)* is the posterior probability or the probability that the value we are looking for, *P(c)* is the probability class based on the hypothesis (prior probability), *P(x|c)* is the predictor probability based on the given class (likelihood). *P(x)* is a predictor probability. In simple terms the Bayes probability equation can be written as:

$$posterior = \frac{likelihood \times prior}{evidance} \quad (3)$$

the probability equation of Bayes's theorem can be substituted by the following equation:

$$P(class|data) = \frac{P(data|class) \times P(class)}{P(data)} \quad (4)$$

where *class* is the reaction level consisting of three categories (high, medium, and low). While *data* is input to determine *class.* The input or predictor feature consists of number of replies, retweets, and likes.

Classification is a very useful method in machine learning, for example Naïve Bayes is suitable for the classification of political or business sentiments [18]. Where both are important topics that are often used as research material. Before carrying out the classification process, Naive Bayes will be trained with a dataset that has been prepared. The classification results are very dependent on decision rules. The rules are used to determine the most likely hypothesis, or known as Maximumm A Posteriori (MAP). Naive Bayes Classifier is one method in data mining that combines bayes theorem with MAP [19]. The MAP formula can be written as follows:

$$CMAP = argmax\, P(X|c) \times (P(C) \quad (5)$$

where *C* is the class target, and *P(X|c)* probability of class based on hypothesis, meanwhile *P(C))* is the prior probability. The task is to find the correct class for each sample.

3. RESULTS AND DISCUSSION

3.1 Twitter Crawling

Searching and retrieving data from the web is getting easier. Various methods for downloading data are emerging. One of them is the crawling method. In this research, data collection is done by a crawling application that are built with the Python programming language. The supporting module used is Beautiful Soup. That is a Python package library that can be used to pull data from HTML and XML documents [20]. This library works with a parser to provide idiomatic ways to navigate, search, and modify the parse tree. Its use can save programmer's work time.

The application will make connections with Twitter's search engine. This method allows the application to retrieve data in the form of a tweet containing the desired keywords. The way to run Twitter crawling application is to enter command input on the terminal. The command entered is as follows:

```
1. Python crawl.py –s "Ganjar Pranowo" –o file1.csv --csv
2. Python crawl.py –s "Sudirman Said" –o file2.csv --csv
```

Figure 1 Command Input for Twitter Crawling

All tweets that contain the keywords (sentences with quotation marks) will be collected. The following is a table showing the results of data collection during 29th March 2018 to 11th April 2018.

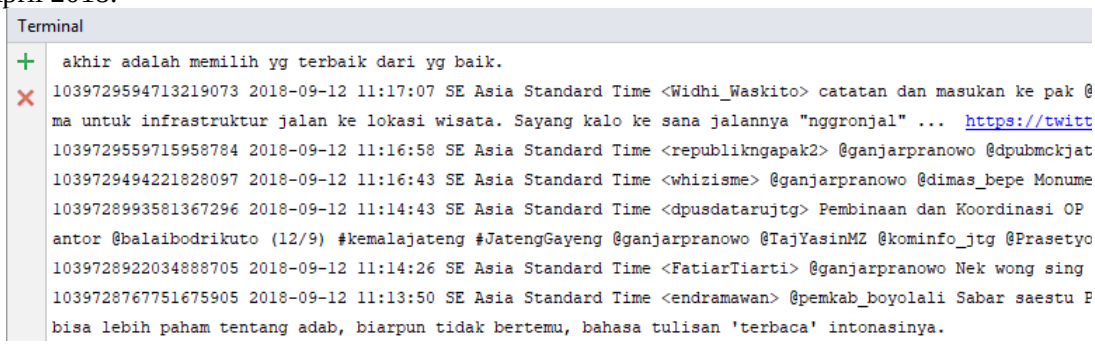

Figure 2 Twitter Crawling Process

Table 1 Data Collection Results

Data	1st Candidate	2nd Candidate
Keyword	Ganjar Pranowo	Sudirman Said
Tweets	9,972	4,821
Replies	6,028	3,537
Retweets	18,776	11,829
Likes	25,646	16,275

3.2 Class and Predictors

Before starting the classification process, the data that has been selected must undergo the preparation stage. At this stage all samples in both data will be labeled based on their class, to find out which class a sample is placed on. The following table will provide an overview of how to determine the label.

Table 2 Determination Of Dataset Labels

Reaction	Reply	Retweet	Like
Low	0	0	1
Low	0	1	0
Low	1	0	0
Medium	0	1	1
Medium	1	0	1
Medium	1	1	0
High	1	1	1

Determination of labels in datasets is divided based on three classes, namely high, medium, and low. All of it represent the categories of reactions given to a tweet. While the predictor feature consists of replies, retweets, and likes, assuming a value of 1 indicates netizen response to status. While 0 indicates no response is given.

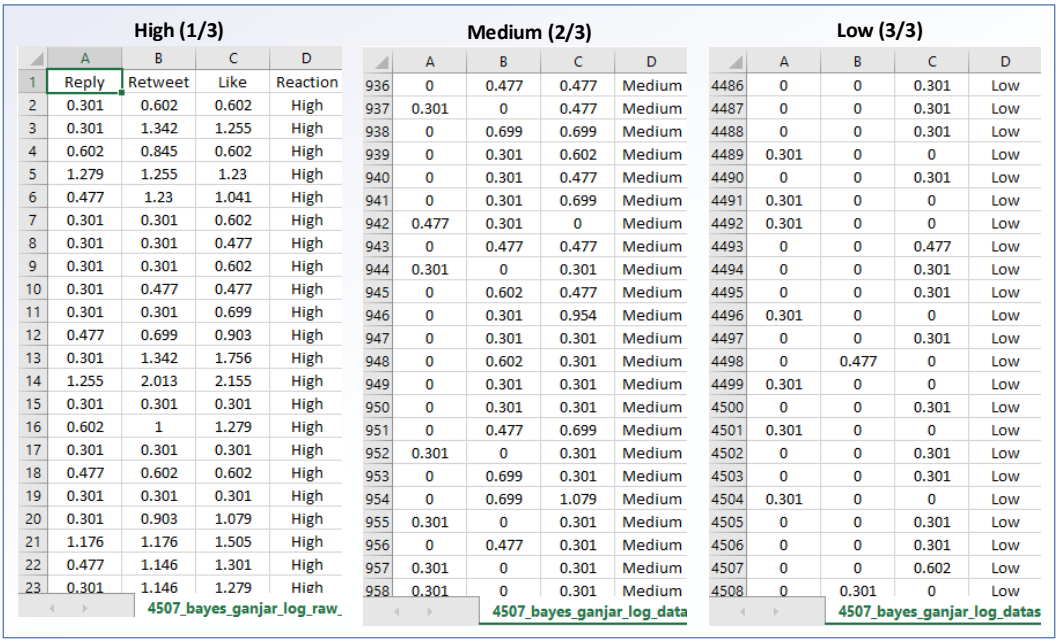

Figure 3 Distribution of Class Label in Dataset

Figure 3 displays the dataset used, it can be seen that there are four interconnected class attributes. All numbers have a value range that is not too far away, this is intentionally done by using logarithmic transformations. The purpose is to improve the performance and accuracy of the Naive Bayes classifier.

3.3 Naïve Bayes Classifier

When the dataset is ready and meets the desired criteria, the classification process can be carried out. The following is a diagram for the classification process from the beginning until the classification results.

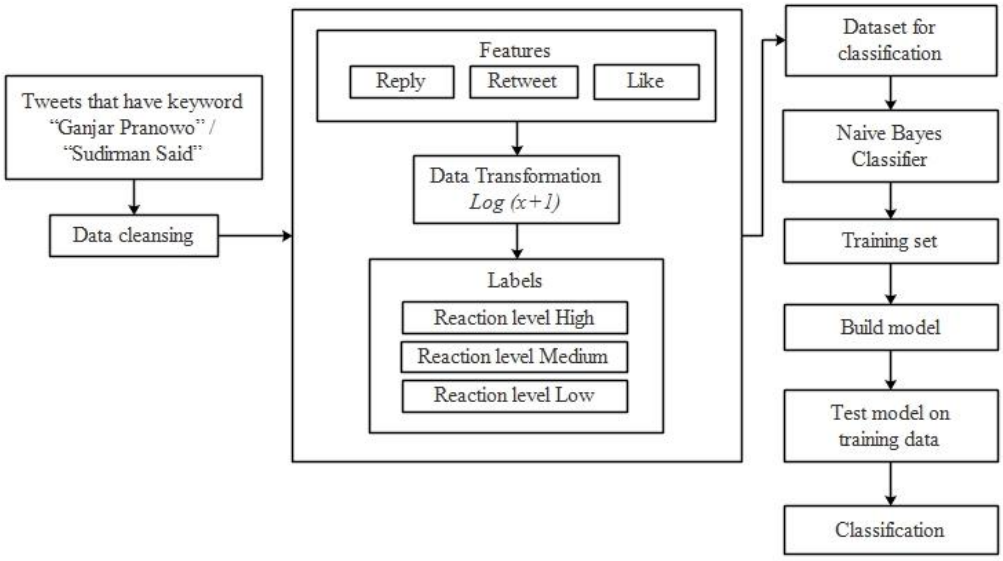

Figure 4 Classification Process Step by Step

The first step in the classification process is to clean the downloaded tweets. Tweets containing keywords Ganjar Pranowo and Sudirman Said are cleaned by removing unnecessary features such as status, username, date, time, etc., leaving only reply, retweet, and like. After that data will go through a transformation process to equalize the value ranges between features, and then the labels are added. After going through these preprocessing stages, the dataset is ready and the classification process can be executed. By using Naive Bayes classifier, the datasets were processed into training set to build a classification model. Then the model was tested using the same training data. After all these processes, datasets can be classified into groups according to the specified label i.e, high, medium, or low.

After the classification results are obtained, the performance of the classifier can be measured using confusion matrix. Confusion matrix is a matrix table containing evaluation of classification results [21]. It describes the distribution of sample values between predicted classes and actual classes. The calculation of classification results using confusion matrix also aims to find the success rate of the classification process. That way the results of calculations that are done manually and by using the application can be known and compared.

Table 3 Confusion Matrix for Ganjar Pranowo Dataset

n = 4,507 (Ganjar Pranowo)		Classification Results		
		High	*Medium*	*Low*
Actual Reactio	*High*	394	289	0
	Medium	70	838	419
	Low	3	267	2,227

Table 4 Confusion Matrix for Sudirman Said Dataset

n = 2,783 (Sudirman Said)		Classification Results		
		High	*Medium*	*Low*
Actual Reactio	*High*	297	436	0
	Medium	86	637	206
	Low	0	140	981

The samples that were successfully classified correctly were 3,459 for Ganjar Pranowo's dataset and 1,915 for Sudirman Said's dataset. For each dataset, Naive Bayes classifier successfully predicts high reaction rates of 394 and 297 samples, mediums of 838 and 637 samples, while low reactions occupy the highest number of samples with 2,227 and 981 samples respectively. Based on the confusion matrix, the accuracy of Naive Bayes classification can be described as follows:

$$(TH + TM + TL) / n = (394 + 838 + 2227) / 4507 = 0.7674 = 76.74\%$$
$$(TH + TM + TL) / n = (297 + 637 + 981) / 2783 = 0.6881 = 68.81\%$$

where TH = True High, TM = True Medium, TL = True Low, n = number of samples.

3.4 Classification Percentage

This research does not want to explicitly predict the winner in the election, but based on the number of reactions and interactions that occur on Twitter, it can give an idea of how people react to certain candidates. And it was found that the number of tweets about certain candidate did not determine level of reaction given. As can be seen in Figure 5 and Figure 6.

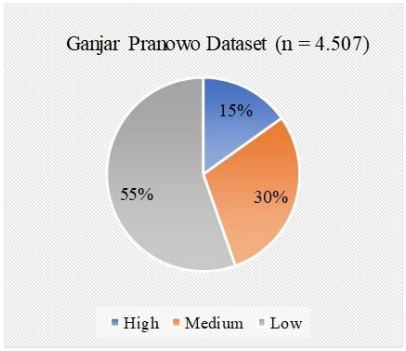

Figure 5 Reaction Levels for Ganjar Pranowo Dataset

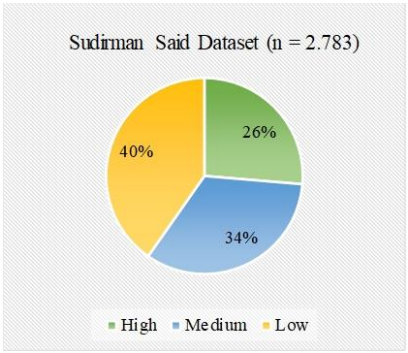

Figure 6 Reaction Levels for Sudirman Said Dataset

It can be seen in the two graphs of percentage, Sudirman Said has a higher reaction rate than Ganjar, even though there are 4,507 Ganjar samples. While Sudirman Said's sample numbered only 2,783. From this, it can be concluded that the large number of tweets does not necessarily determine the number of votes.

4. CONCLUSIONS

From this research it can be concluded that value ranges that are too far affect the accuracy of the Naive Bayes algorithm. This is because the characteristics between classes are difficult to distinguish. For this reason, at the preprocessing stage data transformation is performed to equalize the attribute values. Overall, the preprocessing stage determines the quality of the data, which is an important factor in the success of classification. Because low data quality can result in low accuracy values as well. The accuracy of the Naive Bayes Classifier obtained was 76.74% for the Ganjar Pranowo dataset and 68.81% for the Sudirman Said dataset. A significant difference is largely influenced by the difference in the amount of data between the two candidates, that is 4,507 belonging to Ganjar Pranowo, and only 2,783 for Sudirman Said.

ACKNOWLEDGEMENTS

This research was conducted with the aim of direct and factual information in order to find valid data about public political participation. During this process there are parties who constantly provide support and attention for the success of this research. Therefore, we express our deepest gratitude to all parties involved. Especially to those we respect, Mr. Fuat Zen and Mrs. Eni Citrawati for the financial assistance provided to support this research. Thanks to the support and facilities they provide this research can run well and fulfill the planned schedule.

REFERENCES

[1] A. G. Sooai, A. Nugroho, M. N. A. Azam, S. Sumpeno, and M. H. Purnomo, "Virtual artifact: Enhancing museum exhibit using 3D virtual reality," in *2017 TRON Symposium (TRONSHOW)*, 2017.

[2] 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.

[3] W. Hall, R. Tinati, and W. Jennings, "From Brexit to Trump: Social Media's Role in Democracy," *Computer (Long. Beach. Calif).*, vol. 51, no. 1, pp. 18–27, Jan. 2018.

[4] M. L. Khan, "Computers in Human Behavior Social media engagement : What motivates user participation and consumption on YouTube ?," *Comput. Human Behav.*, vol. 66, pp. 236–247, 2017.

[5] E. Colleoni, A. Rozza, and A. Arvidsson, "Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data," *J. Commun.*, vol. 64, no. 2, pp. 317–332, 2014.

[6] J. Groshek and K. Koc-Michalska, "Helping populism win? Social media use, filter bubbles, and support for populist presidential candidates in the 2016 US election campaign", *Information, Communication & Society*, vol. 20, no. 9, pp. 1389-1407, 2017.

[7] A. Nugroho, S. Sumpeno, M. H. Purnomo, "Visualizing Interaction in Catfiz Indonesian Messenger Using Graph Coloring," *NICOGRAPH International Conference,* pp. 1234–1237, 2015.

[8] B. Álvarez-Bornstein, R. Costas, "Exploring the relationship between research funding

and social media: disciplinary analysis of the distribution of funding," *STI 2018 Conference Proceedings*, pp 1168–1181, 2018.

[9] A. Jungherr, "Twitter in Politics: A Comprehensive Literature Review," *SSRN Electronic Journal*, pp 1–90, 2014.

[10] M. Glowacki, V. Narayanan, S. Maynard, G. Hirsch, B. Kollanyi, L. Neudert, P. Howard, T. Lederer and V. Barash, "News and Political Information Consumption in Mexico: Mapping the 2018 Mexican Presidential Election on Twitter and Facebook", *COMPROP DATA MEMO 2018.2/ JUNE 29*, 2018.

[11] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, and M. Stranisci, "An Italian Twitter Corpus of Hate Speech against Immigrants," in *Proc. of the 11th International Conference on Language Resources and Evaluation (LREC 2018), ELRA (2018), 2018*

[12] S. Fano and D. Slanzi, "Using Twitter Data to Monitor Political Campaigns and Predict Election Results," *The PAAMS Collection - 15th International Conference*, 2017.

[13] K. W. Lim and W. Buntine, "Twitter Opinion Topic Model: Extracting Product Opinions from Tweets by Leveraging Hashtags and Sentiment Lexicon," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 1319–1328, 2014.

[14] W. B. Zulfikar, M. Irfan, C. N. Alam, and M. Indra, "The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter," *2017 5th Int. Conf. Cyber IT Serv. Manag. CITSM 2017*, 2017.

[15] G. Boeing and P. Waddell, "New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings," *J. Plan. Educ. Res.*, vol. 37, no. 4, pp. 457–476, 2016.

[16] M. Sadegh, M. Shakeri Majd, J. Hernandez, and A. T. Haghighi, "The Quest for Hydrological Signatures: Effects of Data Transformation on Bayesian Inference of Watershed Models," *Water Resour. Manag.*, vol. 32, no. 5, pp. 1867–1881, 2018.

[17] Z. Wu, Q. Xu, J. Li, C. Fu, Q. Xuan, and Y. Xiang, "Passive Indoor Localization Based on CSI and Naive Bayes Classification," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 48, no. 9, pp. 1566–1577, 2018.

[18] M. N. M. Ibrahim and M. Z. M. Yusoff, "Twitter sentiment classification using Naive Bayes based on trainer perception," in *2015 IEEE Conference on e-Learning, e-Management and e-Services (IC3e)*, pp. 187–189, 2015.

[19] G. R. Septianto, F. F. Mukti, M. Nasrun, and A. A. Gozali, "Jakarta congestion mapping and classification from twitter data extraction using tokenization and naïve bayes classifier," *Proc. - APMediaCast 2015 Asia Pacific Conf. Multimed. Broadcast.*, pp. 14–19, 2015.

[20] R. P. N. Budiarti, N. Widyatmoko, M. Hariadi, and M. H. Purnomo, "Web scraping for automated water quality monitoring system: A case study of PDAM Surabaya," *Proceeding - 2016 Int. Semin. Intell. Technol. Its Appl. ISITIA 2016 Recent Trends Intell. Comput. Technol. Sustain. Energy*, pp. 641–648, 2017.

[21] O. Aborisade and M. Anwar, "Classification for Authorship of Tweets by Comparing Logistic Regression and Naive Bayes Classifiers," in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 269–276, 2018.