The Framework for Political Communication Text Mining Based on Twitter

Jufri
Department of Software Engineering
STMIK Dipanegara Makassar
Makassar, Indonesia
jufri.ldp@dipanegara.ac.id

Aedah Binti Abd. Rahman
Department of Information and Technology
Asia e University Malaysia
Subang, Malaysia
aedah.abdrahman@aeu.edu.my

Suarga
Department of Information System
STMIK Dipanegara Makassar
Makassar, Indonesia
ssuarga@hotmail.com

2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS) | 978-1-7281-7257-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICORIS50180.2020.9320805

***Abstract*—In recent years, social media as a medium that is widely used in communication in the community. The phenomenon of communication on social media is increasingly being used in political communication. Social networking site services like Twitter and Facebook are believed to have the potential to increase public participation in politics. Twitter is an ideal platform for voters, politicians, political parties, and political institutions to disseminate not only public information but also political opinions to the public through their networks. This research is related to the effectiveness of social media (Twitter) as a means of political communication used by the public, especially in the election of the Mayor of Makassar and other elections in Indonesia. This paper, using two methods for social media analysis on political communication. Support Vector Machine (SVM) to classify predictable words or sentences, and the K-Means method is used to classify words or sentences related to political communication. The results of this study can be used by voters, candidates, political parties, and political institutions, as information and also as a measurement aid in determining the choice of candidates.**

Keywords—Social Media, Twitter, SVM, K-Means, Political Communication

I. INTRODUCTION

Nowadays, the internet dramatically influences human life. Based on data from the Indonesian Internet Service Providers Association or "Asosiasi Penyelenggara Jasa Internet Indonesia" (APJII) in 2018 that internet users in Indonesia amounted to 171,170.000 people out of a total population of 264,160,000. This number shows that half or more than 50 per cent of Indonesia's population has used the internet. From this data, 95% of users access social media. The total of 514 cities in Indonesia spread across 34 provinces, Makassar is the 5th largest user for social media [1].

Social media allows users to communicate with people who are at different places distances. Not only as a communication channel, but social media is also present as a means for people to express themselves in writing known as status updates, or posts posted on social media user profile pages. The types of posts by people vary, for example, by posting food posts at restaurants, when travelling in recreational areas. Besides, the public will also post to urban conditions such as traffic jams, garbage, political issues, and others.

As a communication channel, social media is also used by politicians as a means to carry out campaign activities. They play the issue to attract the sympathy of the city community. These political groups gather opinions to blame those who are not liked and continually explore messages in the scent of competition. Social media in the official network of political groups does not stand alone in exploring information for the benefit of the group, because there are so many social media from supporters and sympathizers who spread false news that is not in line with the official attitude of the institution or political group [2].

Several things make social media a campaign tool [3]:

- The current election competition is heavily influenced by social media in increasing party competition. In some cases, small parties that have limited resources have no influence, especially in the General Elections. But with the internet as a cheap medium, and also more accessible than other communication technologies, they can compete with big parties who have stronger resources. Social media allows small political parties to reach potential supporters, similar to large parties.
- Social media can increase community interaction with political parties and candidates. The community has more access to channel their aspirations to political parties and the candidates they support. At the same time, political parties and candidates can coordinate their supporters more easily and quickly to mobilize them, for example, during the campaign.
- Social media as an institutional adaptation. The meaning of institutional adaptation is a shift in the form of political activity offline to online. Political parties and candidates can utilize social media with the same campaigns as in offline politics. Through social media, political parties and candidates can make communication strategies more effective.

With millions of data generated, social media generates digital data that reveals various information, one of which is political communication obtained from community-driven text data on social media [3].

II. LITERATURE REVIEW

A. Political Communication

The most important part of politics is "competition", so the science of political communication must also develop because in using communication as a political tool it is always adapted to the times and science, or to happens at that time so that communication becomes an important part of politics can go by the competitive system of an era and can influence [2].

In another part, many statements of political experts show the closeness between communication and politics, as experts. Political communication is distinguished as one form of communication among nine other forms, namely: communication between users, communication between groups, rumors, public communication, communication between media, and other communications [2].

It is believed that the science of political communication is the highest phase of the study of the field of communication. Firstly because the science of political communication has touched on the fields of power and the political system, speaking of the highest political system is the study of the highest power system on earth, namely the "State", and secondly, the science of political communication is a discussion that has reached a serious, focused and related part because of macro problems in the life of the nation and state whose purpose and impact is to create a just, civilized (moral) and prosperous society [3].

Based on available data, in the 2013 Makassar City Election, the voter participation rate was 59.94%. Whereas in 2018, voter participation fell by about two points to 57.02% or only around 584,406 who exercised their voting rights from the total permanent voter list (LVP) of 1.011.000 [4]. This great power will be the target to be controlled in the interests of the power of political communicators (elite opinion and attentive public), that great power is called public opinion. From time to time the influence of public opinion has proven to be able to overthrow power in several countries, such as the French revolution, the fall of the New Order in 1998, the overthrow of President Husni Mubarak's regime in Egypt which turned out to be also driven through social media [5].

B. Social Media

Social media networks are human activities in cyberspace, showing human existence without limits that can connect in their activities and collaboration. Social media networks can be used as a media connection (facilitator) that strengthens the relationship between users of social media to facilitate the sharing of information and data [6].

According to the above understanding, it can be interpreted that social media is a tool or media based online, in which individuals or groups can share information and data, as well as facility human relations without being limited in place and time. Besides, users easily participate in it, share, and create messages. In this case, several social media sites are popular today, among others: Twitter, Instagram, Blog, Facebook, WhatsApp, Line, and others.

Social media is a feature of cyber media, because social media is part of cyber media, with the following forms [7]:

1. Network. The connection between one device to another through a media interface, so that communication occurs between users, including the exchange of data and information.
2. Information. A very important part of social media networking is information because the information that results in communication between users, getting to know each other, sharing information, and feedback between users.
3. Archive. Archiving data or content on social media becomes very important so that social media users can access data anywhere with different devices.
4. Interactivity. With social media networks, users on social media can make many friends, expand relationships between people, and allow unlimited information sharing.
5. Society Simulation. Social media networks have different models for their penetration into society. Social media can interact directly with other users, so simulations with social media users are not found on other media.
6. User-Generated Content. Social media users are greatly facilitated in interacting on social media because users are given the freedom of time and place in interacting with other users. Different the mainstream media, users can only see, read, or hear and are not given time users to make feedback.

C. Text Mining

Text mining is defined as the process of text analysis, unstructured data management, or knowledge gained in text form, and stored in digital or electronic format [8].

Many research experts also categorize document categorization as text mining. Although document categorization can provide accurate labels and conclusions on certain documents, this does not produce new facts or relationships [8].

1. Text Pre-processing. In this phase, the process of semantic analysis and text syntax. The purpose of this process is to prepare text into data that will be processed further.
2. Text Transformation. Also called the formation of features to find the form in the desired document. Make changes in the form of text and words into initial forms and reduce the words in the document.
3. Feature Selection. It is a phase in reducing the part in the process of changing text. However, in the previous phase, the deletion of words was done that was not narrative, and also not all words had important meanings.
4. Pattern Discovery. It is important to get patterns or knowledge from the text as a whole. In this process, text mining generally uses technical methods in text mining.

D. Sentiment Analysis and Text Mining

Sentiment analysis is the most widely used text analysis application to date. An important part of sentiment analysis is to include data analysis to gain an understanding and another source of data. Usually, sentiment analysis of the best text that has subjective values does not only have objective values. Because parts of the text that have objective content, generally bring up variations in statements or facts that do not cause emotions or moods. Sentiment analysis is generally used to analyze social media for any information, for users based on the information they use [9].

The purpose of mining text is to obtain unstructured data, issuing sequences of numbers that are interpreted as text. For this reason, the information created in the text can be used in the algorithm. Communication in the data can be issued to get the information contained in the document in the form of a summary—the use of unsupervised to obtain data mining in the predictive form [10].

E. Natural Language Processing (NLP) and Machine Learning (ML)

Natural Language Processing (NLP) is an inseparable part of Computer Science which has advantages in understanding human communication when humans communicate naturally. Communication used by humans is the desired communication, the possibility of using terms, incorrect mention, and everyday language. This model makes it difficult for devices to analyze human language. The use of NLP and the use of Machine Learning (ML) have now made significant improvements in solving this problem [11].

To understand natural language is influenced by several reasons that must be known:

a. Semantics is part of linguistics related to the meaning of the word.
b. The syntax is the arrangement of words that must get the purpose of the word, and more important is the arrangement of words or sentences.
c. Context, being the two main parts, gets the correct meaning, but usually, other parts must be understood so that the intended purpose can be understood. This will be seen in the part where the sentence or word appears.

Can use machine learning to analyze text. But the results are not as expected. The use of ML is quite helpful for detecting entities and sentiments in documents; usually, there is a place for improvement when issuing groups of topics or themes. However, because the NLP and ML techniques have been developed and used for a long time, this problem is handling time.

Text analysis, ML is known as a statistical model that functions to find out patterns, including sentiment, unity, group communication, and other meanings in words. The form of the ML procedure used is, supervised learning and unsupervised learning. Supervised learning is an ML process that can be expressed as a form that can be used in other texts. Various algorithms can be used in large data sets to find what is commonly called unsupervised machine learning. The difference between supervised and unsupervised learning and getting the best part of supervised and unsupervised learning for one system is mainly to get the best results [11].

F. Supervised Learning Method and Support Vector Machine

Supervised learning is a way to train data in the form of labels and then use it to train functions, to produce new data. Experiments in training include information that can show true or false data to create true data. In supervised learning, it is also possible to create a function using the test data labels taken from input data and then produce the output as desired. Then monitor output data according to the desired results to adjust the latest results in accordance with the function. The results of this training can use functions for new information, then issue results, classifications or predictions, which can be received correctly [12].

G. Support Vector Machine (SVM)

SVM is a method that is widely used to carry out classification and regression for various purposes. SVM is a binary classification that can divide data into two classes by hyperplane. Hyperlane functions to separate and maximize between classes., called a margin, and data points that are right at a distance from the hyperplane are called support vectors. In SVM information retrieval is also widely used, especially in the process of data classification. Its ability to process large-dimensional data is an advantage of SVM compared to other classifiers. In the text data retrieval information, the advantages of SVM to process large-dimensional data can be utilized, because of the nature of text data which is usually of a large in dimension [13].

SVM Linear was initially used in formulations for binary classification. Provided training data and related labels (x_n, y_n), $n = 1, \dots, N, x_n \in \mathbb{R}^D, t_n \in \{-1, +1\}$ SVM learning consists of the following constrained optimization shown in (1) [14]:

$$\min_{\omega,\xi_n} \frac{1}{2}\, \omega^T\omega + C \sum_{n=1}^{N} \xi_n \quad (1)$$

$$s.t\ w^T x_n t_n \geq 1 - \xi_n \forall n$$

$$\xi_n \geq 0\ \ \forall n$$

H. Unsupervised Learning Method, K-Means, and Other Methods

Machine Learning (ML) is a group of techniques and algorithms that can give a computer the ability to learn. This method can be used in various fields. Data mining uses the ML method to create insights and predictions from the data [15].

The supervised learning algorithm is considered easier than Unsupervised learning to do the processing. But unsupervised compared to other natural learning methods are considered more predictable. As for unsupervised learning, it is possible without an objective variable [16].

I. K-Means Clustering

K-Means algorithm is a way to get the highest value in every iterative repetition. The first value is the number of groups to be selected. For the way to form groups, it is required to group the data into parts K. Then group K exceeds the sum of the values of the smaller groups, with more parts of the way that are not different. In comparison, the smaller K camp is more part of the granularity, which is considered smaller.

From a group of labels interpret the output of the algorithm. One of the K groups will assign a value to the data. In each group on K-Means, it means the group makes centroid. To add value in a cluster and find the closest position or heart of a computer is called a centroid [16].

Some unsupervised machine learning techniques are [16]:

1. Grouping automatically divides datasets into groups based on their similarity.
2. To examine anomalies in determining the portion of data not contained in a dataset. Like searching transaction data when embezzlement occurs.
3. The mining association identifies the dataset that is normally displayed together in the dataset.
4. In preprocessing the most widely used form of data is to use latent variables. For example, separating each data into several components or separating some features in a dataset.

Fixed model, to produce values from input point groups in N data $x_1, x_2 \dots x_N$ into k groups subset C_i, $i = 1\dots, k$, every includes data points ni, $0 < n_i < N$, following mean square error (MSE) cost functions [14].

$$J_{MSE} = \sum_{i=1}^{k} \sum_{xt \in C_i} \|x_t - c_i\|^2 \quad (2)$$

x_t is a vector describe of *t-th* the point of data in the cluster C_i and c_i is the centroid of geometrical is the cluster C_i. The last, this algorithm to minimize J_{MSE}, where x_t-c_i^2 is a selected measurement distance between the data point x_t and the cluster center c_i. An input data point x_t is assigned to cluster i if it satisfies the following condition:

$$I(x_t,i)=\begin{cases}1 \text{ if } i=\arg\min(\|x_t-c_i\|^2)\ j=1,\ldots,k\\ 0 \text{ otherwise}\end{cases} \quad (3)$$

J. Theoretical Framework

The theoretical framework shown in Fig. 1, explains the process of Extraction and Classification of data from Twitter, resulting in political communication sentiment related to the election of the Makassar mayor which is polarized to positive, negative, and neutral. In the process of selection and Clustering produces words or sentences that are related to political communication in the election of the Mayor of Makassar and words or sentences that have nothing to do with political communication in the election of the Mayor of Makassar.

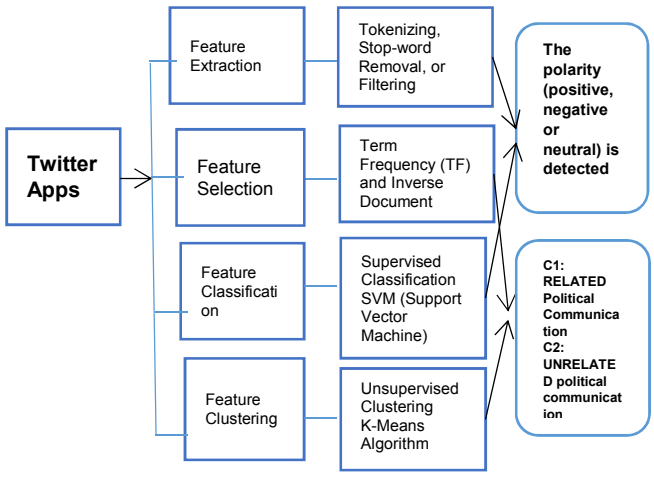

Fig. 1. Theoretical Framework

III. METHOD

This research uses quantitative and qualitative methods. The quantitative method uses a descriptive approach which is a study based on existing data or investigations aimed at solving problems. The qualitative method is also called the naturalistic research method because the research is conducted naturally. Therefore, the qualitative description approach is research that describes qualitative data as it is, and then the data has analyzed the meaning behind the facts that arise [6].

Data collection techniques using instruments; (1) observation, i.e., conduct direct observation, (2) survey, and (3) characteristics of the social life of the community to obtain in-depth data and information. The secondary data obtained from the literature related to research studies. The data analysis will be supported by theories related to the theory of social media usage and the theory of political communication to know, how the Makassar mayor candidates use social media as a medium of political communication and how the public responds to it [6].

This research uses quantitative approaches as a research tool. This research will be conducted in 15 sub-district and 153 villages in the city of Makassar has an area of 175.77 km² and a population of 1.526.677, shown in Table I [17].

TABLE I. SUB-DISTRICT CITY OF MAKASSAR

No.	Sub-District	Village	Population
1.	Mariso	9	60,499
2.	Mamajang	13	61,452
3.	Tamalate	11	205,541
4.	Rappocini	11	170,121
5.	Makassar	14	85,515
6.	Ujung Pandang	10	29,054
7.	Wajo	8	31,453
8.	Bontoala	12	57,197
9.	Ujung Tanah	9	35,534
10.	Sangkarang	3	14,531
11.	Tallo	15	140,330
12.	Panakukkang	11	149,664
13.	Manggala	8	149,487
14.	Biringkanaya	11	220,456
15.	Tamalanrea	8	115,843
	Total Population	**153**	**1,526,677**

A. Research Conceptual

Fig. 2 show the conceptual framework for this research.

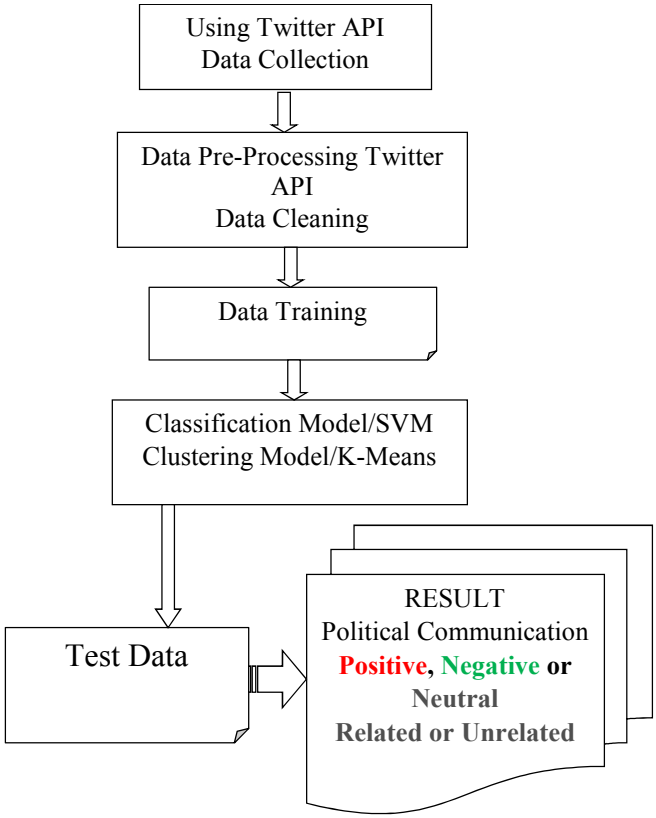

Fig. 2. Conceptual Framework

B. Research Framework

This research proposes a research framework as seen in Fig. 3 for social media in the political domain. At a glance, the framework must function as a kind of guideline for the development of a toolset that aims to collect, store, monitor, analyze, and summarize user-generated content in politically relevant communications from social media for political matters [18].

C. Text Collection Process

Data Collection: Fig. 4 shown the process of using the Twitter API to search with additional filters based on the user account. Next is data processing/cleaning. After all the data tweets have been successfully retrieved, it is necessary to separate the data into two datasets which are training data and

test data. Then in the second part of the data preprocessing the text from the data is done.

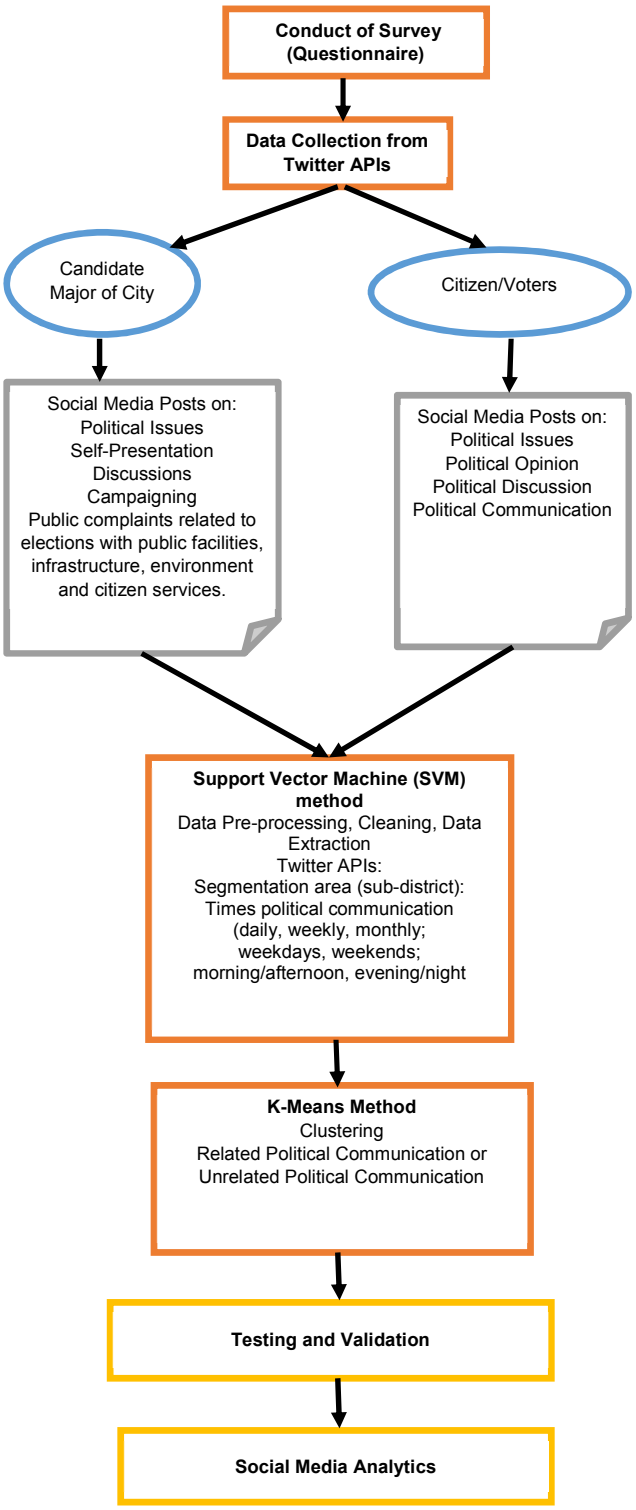

Fig. 3. Research Framework

IV. RESULTS

This study uses data sampling before, and after the election of the mayor of Makassar in November and December 2020, the data used is the big data from social media. This research, collecting social media data about content, usage, and structure through Twitter's application programming interface (API). Sampling data will be used 1,000 users Twitter, 15 sub-district Makassar City.

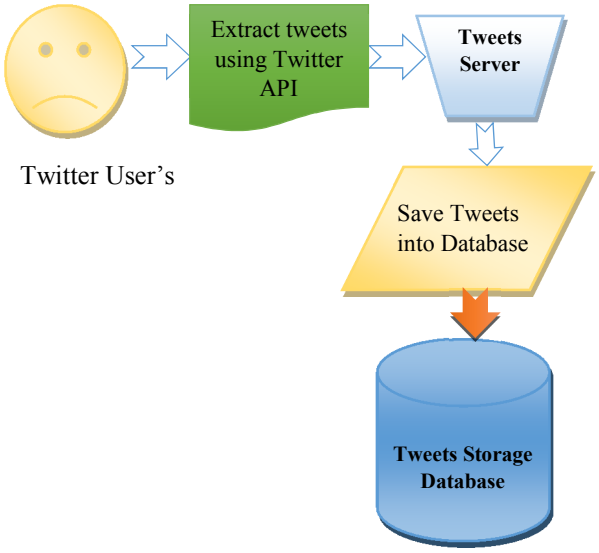

Fig. 4. Text Collection Process

The population is part of the total unit of analysis whose characteristics will be predicted. The population in this study are people in 15 districts in Makassar who use Twitter. The selection of the Makassar city community in this study is based on data requirements related to political communication in the Makassar mayor election. However, the population determined cannot be counted as an amount of accuracy.

The sample taken is a small set of cases selected by researchers from the population who will generalize the population. In this study, sampling was conducted using a probability sampling method Cluster random sampling is a method for determining participants randomly when participants are geographically dispersed. For example, 1,000 participants will be selected from the entire population in the city of Makassar; it is very difficult to get a complete list of all people. Instead, researchers determine the area at random (e.g., city or district) and select at random from within certain limits. Considerations or criteria in taking the sample used in this study are the people of Makassar city aged at least 17 years. The instruments in the study are:

1. The questionnaire aims to find out Twitter user data such as Age, Gender and Religion. This is related because there is difficulty in knowing the data demographic.
2. Analysis to process data using SPSS or Microsoft Excel.
3. Data collection techniques using instruments, observation, survey (questionnaire), and characteristics of the social life of the community to obtain in-depth data and information. Secondary data were obtained from literature related to research studies.

V. CONCLUSION

This research uses descriptive quantitative methods and qualitative methods. The novelty that is expected in this research is the combination of several existing methods, such as Support Vector Machines, K-Means to produce data extraction. Data analysis will be supported by theories related to the theory of the use of social media and political communication theory to find out how candidates for the mayor of Makassar use social media as political media. This research will be conducted in 15 districts and 153 villages in the city of Makassar, which has an area of 175.77 km² and a

population of 1,526,677. Instruments for data collection techniques using, observation, survey (questionnaire), and characteristics of the social life of the community to obtain in-depth data and information. Secondary data were obtained from literature related to research studies.

REFERENCES

[1] APJII, "Asosiasi Penyelenggara Jasa Internet Indonesia," Apjii, 2018. [Online]. Available: https://apjii.or.id/survei. [Accessed: 19-Apr-2020].

[2] E. H. Susanto, "Caleg Pemilik Modal dan Otoritas Parpol (Media Indonesia, 5 September 2013)," Eko Harry Susanto, 17-Sep-2013. [Online]. Available: https://ekoharrysusanto.wordpress.com /2013/09/17/1610/. [Accessed: 19-Apr-2020].

[3] A. Purbolaksono, Media Sosial, Pilkada Serentak dan Pemilu 2019, 27-May-2018. [Online]. Available: https://www.theindonesianinstitute.com/media-sosial-pilkada-serentak-dan-pemilu-2019/. [Accessed: 19-Apr-2020].

[4] L. Herlina, "Partisipasi Pemilih Pilkada Makassar Menurun Di 2018," Media Indonesia, 07-Jul-2018. [Online]. Available: https://mediaindonesia.com/Read/Detail/170778-Partisipasi-Pemilih-Pilkada-Makassar-Menurun-Di-2018. [Accessed: 19-Apr-2020].

[5] M. Shahreza, "Pengertian Komunikasi Politik," Jun-2018. [Online]. Available: https://www.researchgate.net/publication/325686538_ Pengertian_ Komunikasi_Politik. [Accessed: 19-Apr-2020].

[6] I. A. Ratnamulyani and B. I. Maksudi, "Peran Media Sosial Dalam Peningkatan Partisipasi Pemilih Pemula Dikalangan Pelajar Di Kabupaten Bogor," Sosiohumaniora, vol. 20, no. 2, 2018.

[7] R. Nasrullah, Media sosial Perspektif Komunikasi, Budaya, dan Sosioteknologi. Bandung: Simbiosa Rekatama, 2015.

[8] I. Adiwijaya, "Text Mining dan Knowledge Discovery," Kolokium bersama komunitas datamining Indonesia & soft-computing Indonesia, Sep. 2006.

[9] N. Saxena, "Text Mining and Sentiment Analysis - A Primer," Data Science Central, 29-May-2018. [Online]. Available: https://www.datasciencecentral.com/profiles/blogs/text-mining-and-sentiment-analyses-a-primer. [Accessed: 12-Oct-2020].

[10] "What is Text Mining in Data Mining - Process & Applications," DataFlair, 21-Sep-2018. [Online]. Available: https://data-flair.training/blogs/text-mining/. [Accessed: 12-Oct-2020].

[11] Bitext, Natural Language Processing (NLP) vs. Machine Learning, 20-May-2019. [Online]. Available: https://blog.bitext.com/natural-language-processing-vs-machine-learning. [Accessed: 12-Oct-2020].

[12] M. T. Jones, "Supervised learning models," IBM Developer, 26-Feb-2016. [Online]. Available: https://developer.ibm.com/articles/cc-supervised-learning-models/. [Accessed: 12-Apr-2020].

[13] I. K. Purnamawan, "Support Vector Machine Pada Information Retrieval," Jurnal Pendidikan Teknologi dan Kejuruan, vol. 12, no.2, 2015.

[14] K. Korovkinas, P. Danėnas, and G. Garšva, "SVM and k-Means Hybrid Method for Textual Data Sentiment Analysis," Baltic Journal of Modern Computing, vol. 7, no. 1, 2019.

[15] Packt, "Clustering and Other Unsupervised Learning Methods," Packt Hub, 04-Apr-2018. [Online]. Available: https://hub.packtpub.com/clustering-and-other-unsupervised-learning-methods/. [Accessed: 12-Oct-2020].

[16] Guru99, Ed., "Unsupervised Machine Learning: What is, Algorithms, Example," Guru99. [Online]. Available: https://www.guru99.com/unsupervised-machine-learning.html. [Accessed: 12-Oct-2020].

[17] BPS, "Luas Wilayah Menurut Kecamatan di Kota Makassar," Badan Pusat Statistik Kota Makassar. [Online]. Available: https://makassarkota.bps.go.id/dynamictable/2016/10/19/4/luas-wilayah-menurut-kecamatan-di-kota-makassar.html. [Accessed: 12-Oct-2020].

[18] S. Stieglitz and L. Dang-Xuan, "Social media and political communication: a social media analytics framework," Social Network Analysis and Mining, vol. 3, no. 4, pp. 1277–1291, 2012.