Sentiment Analysis Of The Holding Of 2019 General Election In Indonesia on Social Media Twitter Using The Classification Method Of Support Vector Machine and Lexicon Based Features

Dr. Miftah Andriansyah Ssi., MMSi, Roni Budianto

Faculty of Computer Science and Information Technology, Gunadarma University

Jakarta, Indonesia

Andriansyah.miftah@gmail.com

**Abstract**

Indonesia has finished holding elections in 2019. Elections are a process in all citizens in a democratic country that uses its voting rights to elect a president and some leadership in a new government held simultaneously. in this digital era where everyone is free to talk on social media about a variety of things, about personal affairs, work, politics, event how the general election will also lead to many opinions that arise. Based on the above considerations, the researcher wants to find out how the public response about the course of the general election in Indonesia uses sentiment analysis. Sentiment analysis will be carried out on Twitter social media using 1000 data taken in the post-election period (28 April - 28 May) 2019. Sentiment analysis is performed using the support vector machine classification method and lexicon based features using 80% composition of training data and 20 % testing data. The sentiment produced was 589 (58.9%) negative sentiment and 411 positive (41.1%) with an accuracy rate for positive sentiment of 77.5%, negative sentiment accuracy of 82.5%, overall accuracy of 80.55% and precision of 74%.

Keywords : support vector machine, svm, twitter, indonesia general election, sentiment analysis

# Introduction

Indonesia has completed elections in 2019. Elections are a process in all citizens in a democratic country that uses its voting rights to elect a president and some leadership in a new government held simultaneously. in this digital era where everyone is free to talk on social media about a variety of things, about personal affairs, work, politics, event how the general election will also lead to many opinions that arise. Twitter users in Indonesia are quite numerous. According to the site we are social, Twitter users as of January 2019 totaled 6.43 million people. Twitter is used to express thoughts which are then known as tweets. Based on the above considerations, the author wants to find out about people's responses about this using sentiment analysis on twitter social media. Sentiment analysis is often used to monitor social media because it allows us to get a broad picture of public opinion on a particular topic. The purpose of sentiment analysis is to determine the behavior or opinion of the author (twitter account who made tweets) by paying attention to the topic being discussed. therefore, sentiment analysis will be conducted on the indonesian general election using #pemilu, #pemilu2019, #kpu , and #kpuid Hastag. This research takes and uses data through Twitter API with a total of 1000 data and using bahasa indonesia. Data was taken on vulnerable period from 28 April 2019 to 28 May 2019 (post-election) and processed using the support vector machine classification method and lexicon based features.

# The Review of Related Literature

## Machine Learning

Machine learning is a technique for inferring data with a mathematical approach that reflects data patterns [12]. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

1. Supervised learning

Describes a class of problem that involves using a model to learn a mapping between input examples and the target variable. Models are fit on training data comprised of inputs and outputs and used to make predictions on test sets where only the inputs are provided and the outputs from the model are compared to the withheld target variables and used to estimate the skill of the model.

1. Unsupervised learning

Describes a class of problems that involves using a model to describe or extract relationships in data.Compared to supervised learning, unsupervised learning operates upon only the input data without outputs or target variables. As such, unsupervised learning does not have a teacher correcting the model, as in the case of supervised learning.

1. Reinforcement learning

Describes a class of problems where an agent operates in an environment and must learn to operate using feedback. This model is used for situations where machines are drilled to make specific decisions regarding business needs with the primary goal of maximizing efficiency.

## Sentiment Analysis

Sentiment analysis is the automated process of analyzing text data and classifying opinions as negative, positive or neutral. Sentiment analysis is done to see an opinion or tendency of opinion on a problem or object by someone, whether it tends to be negative or positive.

In general, Sentiment analysis is divided into 2 general categories:

1. Coarse-grained sentiment analysis

this category carries out the analysis process at the document level. In short, classification is done with the orientation of a document as a whole.

1. Fined-grained sentiment analysis

this category is currently popular. The object to be classified is not at the document level but in a sentence in a document.

The basic task of sentiment analysis is to group the text in a document or sentence and then determine the opinion found in the sentence or document whether it is positive or negative.

## Support Vector Machine

SVMs are supervised learning models that are used to build classifiers and regressors. An SVM finds the best separating boundary between the two sets of points by solving a system of mathematical equations[3]. Support Vector Machine (SVM) Is one of the algorithms on machine learning that is most effective in terms of both practical and theoretical. SVM tries to find hyperplane by maximizing the distance between classes.

The learning process of SVM is to determine support vectors, only enough to know the kernel functions used and do not need to know the form of non-linear functions. SVM can manage data linearly and nonlinearly. This research using hard margin svm because the machine just classified the data into two classes.

SVM can separate data linearly and nonlinearly:

1. Hard Margin

Hard margin or linearly sparable data is data used for svm calculations that can be linearly separated. Hard margin SVM can work only when data is completely linearly separable without any errors (noise or outliers). In case of errors either the margin is smaller or hard margin SVM fails.

1. Soft margin

If data cannot be separated linearly, it can be transformed data into feature space dimensions so that it can be linearly separated in the feature space. Transformation is done by *xk 🡪* (*xk*) mapping functions from input space into feature space according to the class.

## TF-IDF Term Weighting

Tf-idf stands for Term frequency-inverse document frequency. The tf-idf weight is a weight often used in information retrieval and text mining. Variations of the tf-idf weighting scheme are often used by search engines in scoring and ranking a document's relevance given a query.

The TF-IDF weighting model is a method that integrates the term frequency (tf) and reverse document frequency (idf) models. Term frequency (tf) is a process for counting the number of occurrences of terms in one document and the frequency of inverted documents (idf) used to count terms that appear in various documents (comments) which are considered as general terms, which are considered not important [17].

The formula used is:

TFIDF = TF x IDF = TF x loge

Where,

TF = Term Frequency, the number of words

DF = Document Frequency, the number of documents at least 1 document.

| D | = Total of all documents

This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus (data-set).

## Lexicon Based Features

The lexicon-based method works by creating a dictionary of opinion words (lexicon). The words contained in the dictionary are used to identify whether a sentence contains an opinion or not. If the total number of positive values is greater than the number of negative values, the sentence will be positively oriented. But if the total number of positive values is less than the number of negative values, the sentence will be negatively oriented [14].

The formula used is:

Sentencesentiment

As explained before if the total number of positive values is greater than the number of negative values, the sentence will be positively oriented and vice versa.

## Confusion Matrix

Confusion matrix is a table of contingencies. in the context of statistical modeling, they typically describe the label prediction versus actual labels. It is common to output a confusion matrix (particularly for multiclass problems with more classes) for a trained model as it can yield valuable information about classification failures by failure type and class.

Confusion matrix is a table that use to understand the performance of a classification model. This helps to understand how testing data can be classify into different classes[3]. Confusion matrix also help to find out accuracy and precision of a prediction result.

The formula used is :

Accuracy = \* 100%

Precision = \* 100%

The detail shown completely in fig 2.1

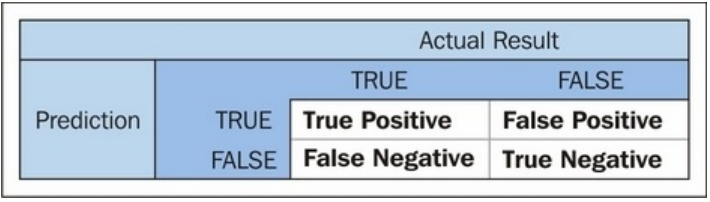


Figure 1 Confusion Matrix Table

# Method

## Tweet Extraction

Extraction is done by using a web crawler to retrieve data automatically. Web crawlers or also known as web spiders or web robots are programs that work with certain methods and automatically collect all the information contained in a website [18]. Before data retrieval, several codes are required to get access rights to Twitter. To get the access rights, there are several steps needed to get the access rights code. Below is the steps:

1. Regist to <https://developer.twitter.com/en/apps> to get acces to developer option
2. Get the API keys and acces token in Keys and Tokens Tab
3. Use the api with python language to crawling tweets

If the data search process is successful, then the search results will produce a json file that contains all search data from Twitter hashtags that use the python language. The search process takes time depending on the amount of data you want to retrieve.

## Preprocessing Data

Preprocessing is done to change the shape of the data to be more structured and in accordance with the needs for further processing. There are several steps that must be done to make data clean and structured, including:

1. Cleansing Tweet

Cleansing data is one step in the preprocessing process to clean up 'dirty' data so that it can be used for further processing. what is done at this stage is Deleting URLs in tweets, Deleting hashtags, Deleting RTs and Deleting @

1. Case Folding

Case folding is the step of changing all data to lowercase or lowercase. This is done to facilitate the program to carry out its functions and avoid case sensitive errors.

1. Remove Double Tweet

This step is done to eliminate the twin tweets that are taken while crawling data

1. Tokenizing

Tokenizing is the process of breaking a sentence into words, the words forming a sentence will be broken down into a single word. This step is done to do labeling or scoring that requires a word for its parameters.

1. Stopword Removal

Stopword Removal is the process of deleting word lists that have no meaning. This step is done to reduce the number of words stored by the system. Stopwords used are owned by masdevid which can be downloaded at the link <https://github.com/masdevid/ID-Stopwords>.

1. Stemming

Stemming is the process of searching for a basic word, which is to change the word that has an influence into the basic word. In this stemming process a python library called Sastrawi is used. Sastrawi is the latest indonesian language stemmer.

## Lexicon Based Features Classification

This classification matches each word with a positive dictionary or a negative dictionary previously created first. Classification is done by finding the number of positive and negative words in a tweet, if in one sentence the number of positive words is greater than negative sentiment words, will be classified as positive sentiment and given a score of 1. However, if otherwise the sentence will be classified as negative sentiment and given score of -1.

The database for the Indonesian lexicon used in this study is the masdevid lexicon which can be downloaded at the link <https://github.com/masdevid/ID-OpinionWords>.

## TF-IDF Term Weighting

In this study using the TF-IDF (Term Frequency - Inverse Document Frequency) method to do word weighting. 3 sentences will be used as a calculation example.

Table 1 TF-IDF Tweet Example

|  |  |
| --- | --- |
| **Dokumen** | **Terms** |
| D1 | Kpu berkerja sangat baik |
| D2 | Dukung kawal perhitungan kpu semangat |
| D3 | Situng kpu dinilai tambah kacau brutal |

After calculating using the function described in the previous chapter, the following results is:

Table 2 TF Calculation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **TF** | | | **DF** |
| **D1** | **D2** | **D3** |
| Kpu | 1 | 1 | 1 | Log(3/3)= 0.159 |
| Berkerja | 1 | 0 | 0 | Log(3/1)= 0.477 |
| Sangat | 1 | 0 | 0 | 0.477 |
| Baik | 1 | 0 | 0 | 0.477 |
| Dukung | 0 | 1 | 0 | 0.477 |
| Kawal | 0 | 1 | 0 | 0.477 |
| Perhitungan | 0 | 1 | 0 | 0.477 |
| Semangat | 0 | 1 | 0 | 0.477 |
| Situng | 0 | 0 | 1 | 0.477 |
| Dinilai | 0 | 0 | 1 | 0.477 |
| Tambah | 0 | 0 | 1 | 0.477 |
| Kacau | 0 | 0 | 1 | 0.477 |
| Brutal | 0 | 0 | 1 | 0.477 |

Table 3 TF-IDF Calculation

|  |  |  |  |
| --- | --- | --- | --- |
| **Term** | **W = TF x IDF** | | |
| **D1** | **D2** | **D3** |
| Kpu | 0.159 | 0.159 | 0.159 |
| Berkerja | 0.477 | 0 | 0 |
| Sangat | 0.477 | 0 | 0 |
| Baik | 0.477 | 0 | 0 |
| Dukung | 0 | 0.477 | 0 |
| Kawal | 0 | 0.477 | 0 |
| Perhitungan | 0 | 0.477 | 0 |
| Semangat | 0 | 0.477 | 0 |
| Situng | 0 | 0 | 0.477 |
| Dinilai | 0 | 0 | 0.477 |
| Tambah | 0 | 0 | 0.477 |
| Kacau | 0 | 0 | 0.477 |
| Brutal | 0 | 0 | 0.477 |

the results of this calculation will be used for future calculations.

## Support Vector Machine Classification

Before the calculate with SVM algorithm, the training data from the TF-IDF weighting phase is converted into vector form, sentences that enter the positive class are expressed in number 1 and the negative class is expressed in number -1. The next step is to calculate kernel functions. The kernel used in this process is a linear kernel. The linear kernel function is shown below:

*K(xi , xj ) = xj*

This kernel is a simple kernel that is used to classify data that is divided into 2 classes in a linear manner separated by hyperplane.

The kernel will iterate and multiply each x with another x. After the iteration process is finished, the results of the iteration multiplication of each x are as shown in table 4.

Table 4 x Iteration Score

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **x1** | **x2** | **x3** | **x1** | **x2** |
| 0.707 | 0.025 | 0.025 | 0.025 | 0.935 |
| **x3** | **x1** | **x2** | **x3** |
| 0.025 | 0.025 | 0.025 | 1.162 |

the result of multiplying all the values of x, then the result of multiplication is made a matrix of  **x1.** The same way is done for the value of y so that it gets a matrix  **yj.**

The next step is to find the x and y values for each document by adding up the matrix rows of  **xj** and  **yj.** So that the results of x and y are shown in tables 5 and 6

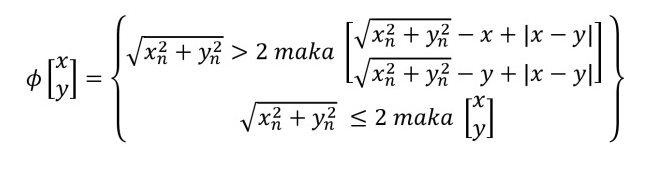
Table 5 XD values

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| X | 0.758 | 0.985 | 1.21 |

Table 6 YD Values

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| Y | 1 | 1 | -1 |

After the values of x and y are known, the next step is to find out the value of the transformation vector by entering the values of x and y into the following equation:



by performing calculations with functions, the transformation values for each document are as follows:

= , = , =

The next step, each vector transformation value is given a bias value of 1 to help get the hyperplane value.After getting the transformation vector value given bias, the next step is to find the value of ai by multiplying each document by the value of a. Which means it will be multiplied iteration for each ai with each D Process and the results of the calculation of the equation above are as follows:

D1 Calculation :

a1  \* = 2.575

a1  \* = 2.747

a1  \* = 0.920

D2 Calculation:

a2  \* = 2.747

a2  \* = 2.972

a2  \* = 1.196

D3 Calculation :

a3 \* = 0.920

a3 \* = 1.196

a3 \*= 3.472

The results of the above equation are substituted so that they get a value for each a. The results of substitution are as follows:

*a*1= -0.460, *a*2= 0.961, *a*3= -0.490,

the value of ai used to the next stage is zero or positive, because a1 and a3 are negative then the value used is only a2 and can be determined if a2 is a support vector

Furthermore, the support vector value is entered into the formula to get the value w (weight) and b hyperplane. The equation used is:

Di

after calculating the value obtained is:

*w* = 0.961 =

Then, it is found that the value of *w* = and *b* = 0.961. value b is a hyperplane for classifying positive and negative values. the same way is done to test the other tweet data by adding and recalculating tf-idf. if the results of the vector transformation are smaller than the hyperplane, the tweet will be categorized negative and if it is larger will be categorized positive.

## Test The Prediction Using Confusion Matrix

Tweet data that has been classified, the prediction results are then searched for accuracy and precision using the confusion matrix with the formula in the previous chapter.

# Result

* Crawled Text Tweet

Table 7 Crawled Data Tweet

|  |  |
| --- | --- |
| **No** | **Text Tweet** |
| 1. | RT @Yswn\_Kawilarang: TPS 10 Gerung utara kec Gerung, Lombok barat NTB. Input di web KPU tidak sesuai dengan C1 asli |
| 2. | Ini \ud83d\udc4e malah lebih dahsyat salah inputnya:\n\nhttps://t.co/7lrEZTrcbV" |
| 3. | Kesalahan KPU Input Data 5 Formulir C1 Itu\u00a0Manusiawi https://t.co/l52MGashcr https://t.co/xDZArQaYwB |
| 4. | RT @imayadewi: Saya tidak salahkan @KPU\_ID KPU. Saya pertanyakan Komisi II DPR yang memilih manusia goblok kayak gini jadi komisioner KPU !\u2026 |
| ... | ... |

Data taken from Twitter uses the Python language and the Twitter API is taken from 1500 data. The data is saved into a file with a .csv extension and then used for preprocessing. Data shown on table 7.

* Cleaned Tweet Data

Table 8 Cleaned tweet data

|  |  |
| --- | --- |
| **No** | **Text Tweet** |
| 1. | TPS 10 Gerung utara kec Gerung Lombok barat NTB Input di web KPU tidak sesuai dengan C1 asli |
| 2. | lebih dahsyat salah inputnya |
| 3. | Kesalahan KPU Input Data Formulir C1 ItuManusiawi |
| 4. | Saya tidak salahkan KPU Saya pertanyakan Komisi II DPR yang memilih manusia goblok gini jadi komisioner KPU |
| ... | ... |

The data that has been taken is then carried out by the preprocessing process to clear the tweet so that it can be used for further processing. The results of preprocessing can be seen in Table 8 above.

* Labeled Tweet

Table 9 Labelled Tweet

|  |  |  |
| --- | --- | --- |
| **No** | **Text Tweet** | **Label** |
| 1 | kpuid bawasluri mantapwa**slur ap** | **1.0** |
| 2 | legowo sama hasil situng emng kalah legowo pilgub dki aman | 1.0 |
| 3 | liat timeline kpu bakal tabrak hehe | -1.0 |
| 4 | komplek mega mai mana suaranya salah input kpu tps lurah padang sarai komplek mega mai kota | -1.0 |
| 5 | tps gerung utara gerung lombok barat ntb input web kpu suai c asli janggal c | -1.0 |
| **6** | **malah lebih dahsyat salah inputnya** | **-1.0** |
| ... | ... | ... |

Tweet data is then labeled using the lexicon base classification feature, because this study uses two classes, 0 required data is needed, resulting in 1000 data that is labeled positive or negative. Results can be seen in Table 9 above.

* Predicted Tweet

Table 10 Predicted Tweet

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Text Tweet** | **Label** | **Predict** |
| 1 | jokowi bakal menang kalau curang maya curang manana sama kpuid c | -1 | -1 |
| 2 | ketua kpuid harusnya mundur | -1 | -1 |
| 3 | fadlizon kpuid harusnya lgsg ditembak tempat tpstps salah input c c valid | -1 | -1 |
| 4 | kpuid nyata sekelas kpuid seleranya akun bokep penghina rosululloh | -1 | -1 |
| 5 | manfaatnya dilakukn pihak jawab mungkin pihak | 1 | 1 |
| **6** | kpuid nyata mana kau dusta kerja kotor kau bayar mahal goblok | -1 | 1 |
| ... | ... | ... |  |

By using an 80: 20 composition, a test was carried out using 20% of the total data, and resulted in a prediction of 200 data as shown in table 10.

* Confusion Matrix

|  |  |  |
| --- | --- | --- |
| Prediction | True Values | |
| True | False |
| True | TP  62 | FP  18 |
| False | FN  21 | TN  99 |

With all the processes that have been carried out, a confusion matrix calculation is made with the configuration previously done for the calculation and accuracy of the svm model. the resulting accuracy and precision are:

1. Positive Sentiments Accuracy

Accuracy = \* 100 = \*100 = 77.5%

The results of positive sentiment accuracy reached 77.5% and the error rate was 22.5%.

1. Negative Sentiments Accuracy

Accuracy = \* 100 = \*100 = 82.5%

The results of negative sentiment accuracy reached 82.5% and the error rate was 17.5%.

1. Overall

Research resulted in an overall level of accuracy and precision with an accuracy of 80.5% and 74% precision

Accuracy = \* 100 = 80.5%

Precision = \* 100 = 74%

* Wordcloud



Wordcloud visualization is a combination of several terms (words) based on the number of frequencies. The more the number of frequencies, the bigger the word size. The bigest word appears is ‘kpu’, meaning the word ‘kpu’ appears the most out of the many words available.

# Discussion

1. How to do sentiment analysis using python programming on twitter?

Sentiment analysis is done by collecting data or documents from social media in their respective ways. for twitter, Twitter API and access token are used to retrieve data, then the data is processed and sentiment analysis is done using the desired method

1. What is the response of the public to the running of the 2019 indonesia general elections?

Based on research conducted, negative sentiment results were 589 (58.9%) and positive sentiments were 411 (41.1%), which means that the indonesia general election 2019 got more negative responses on the Twitter social media.

1. How to apply the Support Vector Machine algorithm in classifying responses and visualizing data from the number of tweets?

Support vector machine method is done by cleaning the data (preprocessing) obtained and then labeling the data, so that the data is ready to proceed to the svm process, to display the data, the python language can be used for further processing.

# Conclusion

Based on the research that has been done, it can be concluded that the sentiment analysis research on the Indonesian general election using the support vector machine (SVM) and lexicon based features classification method results in negative sentiment of 589 (58.9%) and positive sentiment of 411 (41.1%), which means that indonesian general election get more negative responses on twitter social media. The word ‘kpu’ appears the most out of the many words available in wordcloud, which means that kpu (indonesia general election institution) get the most opinion tweeted. This research also resulted in an accuracy rate for positive sentiment of 77.5%, negative sentiment accuracy of 82.5%, overall accuracy of 80.55% and precision of 74%. with a composition of 80% training data and 20% test data.

# References

[1] Anton Subagia. 2016*. Membuat Web dengan Php 7 dan Database PDO MysqLi*. Elex Media Komputindo, Jakarta.

[2] F. A. Pozzi et al*.* 2017. *Challenges of Sentiment Analysis In Social Networks: An Overview, Sentiment Analysis in Social Networks.* Elsevier, Camridge.

[3] Joshii Prateek. 2016. *Python: Real World Machine Learning*. Packt Publishing Ltd, Birmingham.

[4] Purbo, Onno W. 2019. *Text Mining – Analisis Medsos, Kekuatan Brand, & Intelijen di Internet*. ANDI, Yogyakarta

[5] Rosa A.S. 2014. *Rekayasa Perangkat Lunak Terstruktur dan Berorientasi Objek*. Informatika, Bandung.

[6] Schneider, K.-M. 2005. *Techniques for Improving the Performance*. *In Proceedings of CICLing, 5-7.*

[7] Suyanto. 2019. *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Informatika, Bandung.

[8] Teguh Wahyono. 2018. *Fundamental of Python for Machine Learning*. Gava Media, Yogyakarta.

[9] Yuniar Supardi dan Ading Hermawan. 2018. *Semua Bisa Menjadi Programmer CodeIgniter Basic*. Elex Media Komputindo, Jakarta.

[10] Abu Ahmad. 2017. *Mengenal Artificial Intelligence, Machine Learning, Neural Network, dan Deep Learning*.

[11] Imam Fahru., dkk. 2012. *Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi, Jurnal EECCIS Vol. 6, No. 1.*

[12] Jan Wira Gotama Putra. 2019. *Pengenalan Konsep Pembelajaran Mesin dan Deep Learning.*

[13] Moch Ali Ramdhani. 2015. *Permodelan Proses Bisnis Sistem Akademik Menggunakan Pendekatan Business Process Modelling Notation (BPMN) (Studi Kasus Institusi Perguruan Tinggi XYZ)*.

[14] Pamungkas, E. W & Putri, D. P. 2016*. An Experimental Study of Lexicon-Based Sentiment Analysis on Bahasa Indonesia. Proceeding of The* 6th*International Annual Engineering Seminar (INAES).*

[15] Ramadhan W.P., dkk. 2017. *Analisis Sentimen Menggunakan Support Vector Machine dan Maximum Entropy*. Telkom University.Bandung, Indonesia

[16] Sumartini Saraswati. 2011. *Text Mining dengan Metode Naive Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis*. Universitas Udayana.

[17] Wibawanto, Wanda. 2017. *Desain dan Pemrograman Multimedia Pembelajaran Interaktif*. Cerdas Ulet Kreatif, Jember.

[18] Wijaya, Vijam. 2016. “*Pentingnya Web Crawling sebagai Cara Pengumpulan Data di Era Big Data”*. URL: http://www.teknologi-bigdata.com/2016/07/web-crawling-di-era-big-data.html. Accessed September 2nd, 2019.