CHAPTER 3
RESEARCH METHODOLOGY

There are several steps taken in working on this project.

1. Identification Problems and Literature Study

The first stage of a series of stages of research carried out. The identification stage of this problem is the stage where researchers conduct further data deepening of existing case studies. The literature used is mostly in English because to find literature in Indonesian is very minimal. To find literature through the internet by typing the url address https://scholar.google.com. Researchers collected a minimum of 8 pieces of literature as a basis for a topic. From each of these journals there is a publication year from 2014 to 2019. The steps in gathering this literature to be studied as a reference and source of information in solving the problems faced.

2. Collecting data

In this research, data is needed as an analysis. The data used is false data and fact data. Turnbackhoax.id website is a site that provides these data, also provides hoax news and facts that have been classified based on the results of discussions and fact-finding carried out by a group of members who are only to clarify a data.

In addition to turnbackhoax.id, there are also website resources such as liputan6, tribunnews and BBC. All three websites will later be taken only in the title and text sections. To retrieve data from a website a web scrapping technique is needed. Web scrapping is a technique to get data from website sources. To be able to retrieve data on the website, you can use the API or by using a library of python, beautifulsoap. The researcher uses an API to retrieve the turnbackhoax.id site. API is the Application Programming Interface, this API as connecting 2

different systems so that they can be connected to each other. To be able to enter the website system, a code name is required in order to be able to access data in the site.

From the scrapping results obtained from the website turnbackhoax.id obtained 1395 hoax news and 454 real news, while the tribunnews.com, liputan6.com and BBC.com websites used the beutifulsoap library, the fact the data was obtained 866 data. The amount of data that can be determined according to the source on the website.

This research has been labeled between fake news and fact news, data are classified into 2 parts. Before it is divided into 2 parts, each document will be checked with other documents to avoid duplicate data. Document for research is 1100 data, the study was carried out as many as 4 times the tests that have different portions.

3. Perform data pre-processing

After getting data from several websites, it then enters the Pre-processing stage. This pre-processing stage is the initial stage in processing a data. The pre-processing stage consists of several steps, namely folding, tokenizing, filtering, stemming. The researcher added the stages of changing each document to the ASCII code form as well as removing '[]' in each document. Here are some steps when doing pre-processing.

a. Change to ASCII type

This stage changes the sentence to the ASCII form so that it can be processed into the next process. ASCII is an international standard in letters and symbols.

b. Remove sign '[]'

This stage removes the '[]' mark and the contents of the words contained in the character. At this stage, using the regular expression technique.

Full sentence:

[SALAH] "Yel-yel Kecebong Yang Merencanakan Untuk Menyusup di CFD".

The results of the sentence after going through the process :

"Yel-yel Kecebong Yang Merencanakan Untuk Menyusup di CFD".

c. Case Folding

This stage turns all letters into letters in the document into lowercase letters. Letters received ‘a’ through ‘z’.

Full sentence:

"Yel-yel Kecebong Yang Merencanakan Untuk Menyusup di CFD".

The results of the sentence after going through the Folding process :

"yel-yel kecebong yang merencanakan untuk menyusup di cfd".

d. Case Stemming

This stage is a process of changing the words in a document into basic form words.

Full sentence:

"yel-yel kecebong yang merencanakan untuk menyusup di cfd".

The results of the sentence after going through the Stemming process :

yel kecebong yang rencana untuk susup di cfd.

e. Case Filtering

This filtering stage is to remove words that contain conjunctions such as ‘di’, ‘ke’, ‘dari’, ‘dan’, ‘atau’, etc.

Full sentences:

yel kecebong yang rencana untuk susup di cfd.

The results of the sentence after going through the filtering process :

yel kecebong rencana susup cfd.

4. Calculate TF-IDF on each document

After going through the process of pre-processing techniques, researchers conducted TF-IDF calculations. This process will produce a weight of each word that has been cut from the document. To produce TF-IDF values, there are several stages of calculation, namely the calculation of Term Frequency, Document Frequency, Inverse Document Frequency, and finally calculation of Frequency-Inverse Document Frequency.

a. Term Frequency

This stage counts for the number of times the term or keyword appears in each document.

b. Document Frequency

Document Frequency is the number of documents based on the terms that appear.

c. Inverse Document Frequency

$$idf_i = \log(N / df_i)$$

N = Number of documents in the collection.

df = number of documents containing the term (document frequency).

d. Term Frequency - Inverse Document Frequency

$$W_{ij} = tf_{ij} * idf_j$$

TF-IDF = Term frequency results are multiplied by the inverse document frequency results.

5. Making a data classification system

At this stage the researcher made a classification of each document. In this system built using the Python programming language, for the method using the Support Vector Machine algorithm and Random Forest. Both of these algorithms have the same function which can do the classification. But the way these two algorithms work is very different.

Linear support vector machine is a linear kernel in the support vector machine algorithm where hyperplane lines are usually straight in the shape of dividing two classes (Maulina Dina, Sagara Rofie, 2018). The training data will later be studied by the machine, to compare with each value of the testing data, then the data will be grouped by the line separator that is the hyperplane line. Hyperplane lines are lines that divide both two vector groups and are usually valued $w * x - b = 0$ and margin worth $w * x - b = 1$ or $w * x - b = -1$.

SVM linear equation : (Maulina Dina, Sagara Rofie,2018)

$$w * x - b = 0$$

w = Hyperplane parameters are sought.

x = SVM input data.

b = Hyperplane parameters are sought (bias).

Whereas Random Forest is an operation of the decision tree method that produces the class of each tree (Cuşmaliuc et al., 2019). This algorithm also requires training data that has been labeled to train the machine in processing test data. Random Forest formula calculation document (Putri et al., 2019):

$$Entropy(Y) = -\sum_i P(C|Y) \log_2 P(C|Y)$$

$P(C|Y)$ = Total proportion of samples for class C.

6. System Testing and Analysis

After all systems are formed, this process is tested on the system. This test requires training data and test data. Training data is data that is used to be studied by machines that have been labeled in the data. While the training data has not been labeled, so the data wants to be analyzed and predicted according to the machine learned from the training data. Both of these data have previously produced a weight value of TF-IDF, so that in this process, the machine can classify fake or factual news.

When the testing process is complete, then enter the analysis phase. In this analysis phase the system will calculate the performance of each algorithm to be compared. The researcher uses the calculation of accuracy, precision, recall, and F1-Score as the analysis process. The formula used to do the calculation.

In calculating accuracy, precision, recall (Putri et al., 2019) :

a. $Akurasi = \frac{tp+tn}{tp+tn+fp+fn}$

b $Precision = \frac{tp}{tp+fp}$

c. $Recall = \frac{tp}{tp+fn}$

d. $F1 = \frac{2 \times recall \times precision}{recall + precision}$

Explanation :

TP: True Positive (the actual document is True but the prediction results are True too).

FP: False Positive (documents that are actually True but predictions are False).

TN: True Negative (documents that are actually False but predictions are True).

FN: False Negative (the document is actually False but predictions are False).

7. Report writing

At this last stage, the researcher writes a report of the process of the research work steps from the beginning to the end of the system testing and analysis phase. Also write a performance analysis of the two algorithms, namely Support Vector Machine and Random Forest to solve false news problems. After writing the analysis, the researcher also added suggestions for further research.