CHAPTER 3
RESEARCH METOLODOGY

3.1. Literature Study

In the study literature stage, a theoretical study of the methods used for problem solving in this study was carried out. The sources taken were obtained from journals, the internet and e-books or previous research. This helps to provide basic knowledge for research. Matters relating to SVM algorithms, sentiment analysis, k-folding cross validation, TF-IDF being the focus of the material.

3.2. Data

The data used in this study amounted to 3000. The form is in the form of a tweet taken using a tweepy connected to the twitter API. A twitter account is required to demonstrate the API to twitter first.

3.3. Coding

The programming language used in this study is python on jupyter notebook with the addition of several libraries to support this research.

3.4. Analysis

Before the data can be used on the algorithm, several processes must be carried out on the data. Here is the detail :

3.4.1. Crawling Data

Tweet data was retrieved by crawling 3000 tweets. Topics on tweets are limited to #IKN, IKN, #ibukotabaru. Data taken randomly in the time span of 4 – 11 June 2022.

3.4.2. Pre Processing

The crawled data still has a lot of noise in it that can disturb when the labeling process is running. For items that are considered noise are mentions, URLs, retweets, characters other than a-z, and Unicode. Case folding is also applied to this process. After everything is removed, separate the sentence by using NLTK. This will transfor sentence to token form. Sastrawi library use in the next process to remove stopword from the data. After that we do stemming. Stemming is a process performed to remove suffixes on tokens that can obscure the true meaning of a particular word. Sastrawi library is used for stemming because the data is in Indonesian. The last is labeling, in this process lexicon Indonesian is used to give polarity to tweet sentences that have been preprocessed.

3.4.3. SVM

In this process, the SVM algorithm will classify the data that has been processed using the Lexicon Based method to test its accuracy in determining the sentiment of a tweet. Here the data will be tested using the k-fold cross validation method with the number of k = 5. The dataset will be divided into two, the meaning is 5 parts with 4/5 parts used for the training process and 1/5 of them used for the testing process. The repetition lasts 10 times with variations in stratified data.

3.4.4. K-Fold Cross Validation

K-Fold Cross Validation is one method that can check overfitting on a model. Data that divided into k parts allow every piece of data stops predicting data is faster than not shared first. In k-Fold Cross Validation, the model that has been created is divided into k equal parts or close to size. Model accuracy will be tested using test data on each fold, and continues to the next fold until it's finished. Accuracy will be totaled and divided by number of k.