CHAPTER 3
RESEARCH METHODOLOGY

1. Literature Study

The first step in this research is the literature study. In the implementation carried out by searching scientific journals and articles on certain websites. In this step, the material or reference material sought is related to the LDA algorithm and topic modeling. This step is a very important main step because, through this step, the research process has a clear foundation of implementation.

2. Collecting data

Data or corpus is obtained from labs.semanticscholar.org/corpus, which is the provider of published journal data. However, this study only used half of the data that had been provided. The data used is the title of a scientific journal that on English and has various topics. The data used is in the form of a list of scientific journal titles and stored in CSV files so that all processes that use the corpus lead to calls for CSV. Corpus collection is conducted on September 17, 2018.

3. Pre-Processing

The first step of this research is pre-processing. In pre-processing, the process carried out produces a Bag of Words. According to (Sriurai, Wongkot, 2011) states that a bag of words is the result of gathering certain words or terms and not knowing the true meaning of each word. In Bags of Word itself, repetition of words is not permitted.

Then, according to (Schofield, Alexandra,2017) said that in pre-processing activities, there are several processes such as cleaning, stopwords and stemming. Even in her journal also states that doing stopwords can improve model compatibility and quality.

a) Tokenizing

The first step in pre-processing is tokenizing. Tokenizing is the process of cutting a sentence or text into a word for word which is often called a token. Each word is stored as one called a bag of words.

b) Cleaning

In the cleaning stage, it is used to clear all symbols such as punctuation or numbers. Through this process, the data obtained is only composed of alphabetical. The remaining letters are the result of changing the letters of the existing letters into lowercase letters. One of the ways to do cleaning using library re.

c) Stopwords

This stage is used to get rid of conjunctions or words that are often used but have no meaning. Although there are several words or tokens that are wasted in this stage, it will not affect the meaning of the text or sentence as a whole. In python, one of the ways to do stopwords using library NLTK (Natural Language ToolKit). Import stopwords on NLTK to get a list of useless words. Current research uses the NLTK stopword with English.

d) Stemming

The stage of stemming is the step to change each word into a basic form or become a standard word. In the process of change is a standard word, every prefix or suffix is omitted. There are various ways to do stemming. One algorithm that is quite popular for stemming in English is the Porter Stemmer. In python, the is Porter stemmer is provided on NLTK.

The Porter Stemmer algorithm relies heavily on consonants, vowels, and combinations of vowel-consonants called VC. Deleting a suffix is done if the number of VC is more than 0.

The Bag of Words which is the result of the Pre-Processing process will then be used to apply the Latent Dirichlet Algorithm algorithm. The detailed process of pre-processing will be discussed in Chapter 4.

4. Initialization

The Initialization Phase is a step to assign values to the parameters used. In the Latent Dirichlet Allocation algorithm, it requires the establishment of values on the parameters. Some parameters that need to be set are alpha, betta, number of topics and number of iterations. Alpha parameters are used to distribute topics on each document, while beta parameters are used to distribute topics in each word. The number of topics and iterations can be searched through perplexity.

5. Topic modeling using LDA

At this stage, the Latent Dirichlet Allocation algorithm is used to do topic modeling. The use of the LDA itself is due to the LDA being able to do topic modeling on very sparse data. This algorithm can generate the distribution of words and documents by looking for probabilities. The LDA method used in this study is the Gibbs sampling method. In the Gibbs Sampling method, there are 2 formulas. The first formula is called phi. Phi is the result of the word-topic matrix distribution. Then, the second formula is theta. Theta is the result of the distribution of the topic of the document. The multiplication of these two formulas will produce opportunities. There are various ways to help calculate opportunities. In Python, the library help counting client is Numpy. Numpy is a library for converting arrays into matrices. The use of Numpy can help shorten the calculation time because multiplication and division between matrices are easier than multiplication

and division between arrays. In addition, at this stage, the perplexity calculation process is also carried out for checking. The smaller the value shows the better modeling of the topic.

6. Perplexity

Perplexity is the process of utilizing distribution opportunities to find optimal results without training data. The smaller perplexity value indicates the more precise information from probability calculations. Perplexity can be used to test the suitability of documents for the identity of the topic. Perplexity is a common method used to see the accuracy of the probability of a model in predicting samples and processing during iteration. In this study, perplexity was used to find the number of topics and the optimal number of iterations.