CHAPTER 1
INTRODUCTION

1.1 Background

The development of data processing and data storage techniques in this modern era has been growing rapidly. The size of data collected nowadays are very large and from these large scaled data we can get useful information. This information can be used for several things especially can be used for determining aspects of a person's personality from a recruitment interview. From the results, we can store those data and study it and we can get the patterns then those patterns will help classifying person's personality. The method used for this is data mining.

Data mining, according to Wikipedia is "the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems" [1]. Data mining is used in many fields such as in the educational field, the social field, the economic field, and etc. The main goal of data mining is to find patterns from a large scale data. So, it will involve statistics, artificial intelligence, and machine learning for getting information. After we get the information, that information can be used for useful things such as in predicting something new from past data or experience, grouping data with same characteristics, finding relationships between data, and most importantly classifying data with standards that have been determined.

In recruitment interview, to decide whether a person is suitable or not for the job position is we need to analyze the answers from the interview. The answers to the questions given can represent several aspects of a person's personality so that the company can recruit employees who have good personality. Of course the decision process will be biased, because every person sees everything differently. Someone's judgment is not always the same.

The Latent Dirichlet Allocation (LDA) algorithm according to Wikipedia is “a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar” [2]. This algorithm is commonly used to group data based on the similarity of topics discussed. LDA algorithm pictures a document or text is built from a collection of topics and a topic is formed from a collection of words [3]. So in a text it contains several topics and in the topic it is formed from words that represent that topic. For the case of grouping text data, in making an LDA model it usually does not require guidance in grouping data (unsupervised). But LDA can be improved to be a supervised, the first step taken is to give a label to all the texts that indicate which text data belongs to which group. The algorithm is often called L-LDA (Labelled Latent Dirichlet Allocation) which was introduced by Daniel Ramage et al [4].

Based on the research, in creating classification model with L-LDA there are several approaches or variants. The first approach, in a text contains words that do not refer to any topic. Usually the word is referred to as a common or neutral word. So in a text you can imagine that the text contains certain topics and common topics. In the text, there are words that refer to certain topics and there are words that do not refer to any topic so that they are included in common topics which indicate that the word is a neutral word. Or maybe it can also be said that a common word is a word that is general in nature which means the word also can sometimes appear frequently in each document. From the results of the grouping later, the distribution of words will be seen in certain topics and will later be used as a tool for classification.

The second approach, to do the grouping of text data directly and do not pay attention that in the text contains the word common or neutral, so it is directly said that a text only contains 1 topic only. So later after grouping the way to classify it will be like the previous grouping approach which views that the text contains words that are neutral or contain common topics.

For this research, we will see which approach is better in the case of creating classification model data which later from the results of the model will be used to classify aspects of one's personality.

1.2 Problem Formulation

1. Can LDA model produced be used as a classification tool for this case?
2. How do we use LDA algorithm to classify answers to specify person's personality aspect?
3. Which approach is more effective in classifying?

1.3 Scope

In this research, there will be some limitations, such as:

1. The data used are data sourced from a company X and the amount has been determined, total 350 data per aspect.
2. This study will only focus on the LDA algorithm used in classifying documents. But the version used in this case is L-LDA
3. In creating the model, this research will only focus on 2 approaches. The first one, a text contains common or neutral word and the second one a text does not contain common or neutral word.
4. This study will only focus on classifying 3 personality aspects: motivation, work enthusiasm, and self-awareness.
5. The results will be analyzed by comparing the manual classification method with the classification carried out by the program.
6. The program is created using the Python programming language.

1.4 Objective

1. Creating a program that could classify personality aspect from recruitent interview.

2. Explaining how the LDA algorithm works in classifying answers from the interview to determine personality aspects.
3. Showing the effectiveness of two approaches at classifying personality aspects to tell which approach is better.