CHAPTER 1
INTRODUCTION

1.1. Background

Nowadays, the amount of news has increased every day. They also grow rapidly. This situation makes it difficult for the editor to categorize manually and with this manual categorization makes the categories incorrect for the news. They also wasted the time choosing the category for the news which were published. Even if there is automatic news categorization, the prediction is less accurate. It happens because many categorizations use supervised learning that the dataset is labeled manually with people that spend a lot of time and money.

For that problem, people just only classify small partial data and the rest should be done by the computer. So, it can reduce the time and the money for labeling the dataset. This method calls semi-supervised learning with the method pseudo labeling or in the theory of the book called self-training.

Self-training does with the only partial dataset labeled by the human and then the classifier learns from that partial dataset and does the prediction which it appends on the training set for the next prediction (real prediction). Self-training also does with the model or classifier that can be taken from supervised learning.

In this research, we proposed news categories classification using machine learning with semi-supervised learning with pseudo labeling method. We extract the feature from the news and then list the term as a keyword for each category. This can be used for labeling the dataset, which can be used for the training data until testing the data. We also used News from the Indonesian language. Of course, these datasets are from different categories. We used an Adaptive Boosting (Adaboost) classifier that can boost the data from misclassified. This classifier is usually used for the binary classification problem, but We used it for a multi-classification problem.

1.2. Problem Formulation

1. How can semi-supervised learning increase accuracy?
2. How can AdaBoost work in topic classification?

1.3. Scope

This project starts from NLP (Natural Language Processing) from preprocessing datasets like tokenizing, stemming, and lemmatization until processing the dataset with TF-IDF and vectorization. To make a variation training set, I used ratio dataset proportion and compared between supervised and semi-supervised learning with Adaboost. For modeling, I use the Adaboost classifier from scratch which is based on mathematical formulation.

1.4. Objective

The objective of this research is the system can classify news topics with algorithms using semi-supervised learning combined with AdaBoost classifier. So, that for the accuracy of classification can be performed well.