CHAPTER 1
INTRODUCTION

1.1 Background

Before internet era people communicated directly by involving more senses (Meilinda, 2018). Besides that humans can only rely on television, radio and newspapers as information media. But with the changing times of the Information Technology that is currently developing. Media - media are now a lot of convenience in terms of communication, not limited by time and distance to communicate with each other. The media is also a forum for disseminating information that is very influential on society (Juditha, 2018). So that Online Media greatly affects the mindset of the public in consuming information received.

In terms of disseminating information it is very easy to get through social media including facebook, twitter, instagram, whatsapp, and others. So that with the internet, it is very easy for humans to spread information as well as get information received. But with the ease of disseminating this information many individuals or groups are not responsible for spreading fake or often referred to as hoaxes.

A hoax is information or news that contains things that have not been proven true or are uncertain. In Indonesia, Kementerian Komunikasi dan Informasi (Kominfo) in April 2019 identified 486 hoaxes, the highest number of hoaxes since August 2018 (Kementerian Komunikasi dan Informasi, 'Temuan Kominfo: Hoax Paling Banyak Beredar di April 2019', Kementerian Komunikasi dan Informasi 2 May 2019).

During the Coronavirus disease pandemic outbreak (Covid-19), according to Menteri Komunikasi dan Informatika (Menkominfo) Johnny G Plate stated that as many as 554 hoaxes or information about the corona virus (COVID-19) spread across a number of social media platforms (Umah, ‘Kominfo: Ada 554 Hoax Soal COVID-19 dengan 89 Tersangka’, CNBC Indonesia 18 April 2020).

The more problems with hoaxes at the moment, the way to prevent them is also emerging. Regarding the hoax problem, it is needed a technology analysis that is able to classify hoax news and fact news.

With Machine Learning, a branch of artificial intelligence that is able to learn on its own without having to be repeatedly programmed by humans. Machine Learning requires training data (training) before issuing results. This project will also use the Machine Learning process to classify fake and factual news.

In Machine Learning there are 2 parts, namely Supervised Learning, and Unsupervised Learning. Supervised Learning is a method that requires data that has been trained as a process, there are also variables as targets which aim to group data into existing data. While Unsupervised Learning is a method that does not have training data so that the grouping of data can be divided into several parts. In this project the algorithms used are Support Vector Machine and Random Forest, both of these algorithms are included in the Supervised Learning section which requires training data as a Machine Learning process.

In this project also calculates tf-idf to measure the word weight in each document. The result of the tf-idf weight is used for the algorithm stage. The algorithm used is Support Vector Machine and Random Forest, both of these algorithms produce a percentage of accuracy. From the results of the accuracy are compared then formed visualization data. Data visualization is a visual presentation of data so as to communicate information clearly and not monotonously. Not only measure the percentage of accuracy, the two algorithms are able to classify fake news categories and fact news.

1.2 Problem Formulation

Some questions that want to be proved during this project

1. How to get data using web scrapping ?

2. How to do the pre-processing steps after the data is taken ?

3. How do you calculate the weight of TF-IDF data ?

4. How to apply the Support Vector Machine and Random Forest algorithm to classify fake and factual news ?

1.3 Scope

This analysis project the author uses Python version 2.7, training in real news data taken from tribunnews.com, liputan6.com, bbc.com, turnbackhoax.id, while hoax data is taken from turnbackhoax.id. The author uses sastrawi library as a pre-processing process, then the data is divided into 2 parts training data and testing data. After it is divided into 2 parts the next step is to calculate the weight of tf-idf. The weighted results will be used for the Support Vector Machine and Random Forest algorithm stages. This project will focus on analyzing the accuracy of the problem of false news by using both algorithms used, also classifying that the news is false or fact.

1.4 Objective

The first objective of this project is to classify news so that it gets relevant news. The second goal is to get accuracy results and compare the two algorithms that are good for problems in classifying fake news. So that the second goal, can provide a reference for the development of a system that utilizes the Support Vector Machine algorithm and Random Forest.