Abstract

Semi-supervised Learning Method for Clickbait Detection in Bahasa Indonesia

Denniskiu Fortino Kurniawan

17/408286/PA/17639

Due to the growth of online media, online journalists are triggered to create something that can attract the attention of readers. One of the techniques used is clickbait. By using clickbait, content creators can attract more readers to read their content. As time progresses, the art of clickbait is developed, and it had become hard for humans to know which are clickbait and non-clickbait. As machine learning is then used to solve this problem, there is a need for huge amount of labeled data to give enough data for the machine to learn. To gather that much data, there are a few possible steps that can be followed. One of the steps to gather data manually and labeling the data manually as well and another way is to buy a labeled dataset. By gathering the data and labeling the data manually, it will take a lot of time. On the other hand, by buying labeled data, it is usually expensive to buy a huge, labeled dataset. Therefore, to gather huge amount of labeled dataset, it is usually time consuming and expensive.

This research aims to utilize a small size of labeled data to create a big enough dataset to detect clickbait from news articles in Bahasa Indonesia. The method used is semi-supervised learning, with self-training and co-training as its algorithm. Logistic Regression and SVM are used as its classifier.

The best model for this research thus is achieved by self-training using Support Vector Machine. Using 150 seed data and producing 10885 labeled data. Achieving average accuracy of 78%, average precision of 78%, average recall of 78% and average F1-Score of 78%.

Keywords: Semi-supervised learning, Self-training, Co-training, Clickbait Detection, Logistic Regression, Support Vector Machine