ABSTRACT

MODIFIED SELF-TRAINING USING K-MEANS CLUSTERING FOR CLICKBAIT DETECTION IN BAHASA INDONESIA

Rafqi Muhammad Azzaydan

18/429295/PA/18686

Nowadays, the term clickbait is used by individuals or organizations to gain viewer counts advantage on their content by giving viewers or readers misleading and very interesting headlines while the content itself is unrelated to the headlines. To overcome this problem, it can be solved by using Clickbait Detection. Clickbait Detection is a task to classify text or headlines between two classes which are clickbaits and non-clickbaits.

There are thousands of clickbait datasets but most of them are unlabelled. In this research, author proposed a method of using semi supervised-learning method to help label the data and create a corpus of labelled data that can be used for classification purposes.

This research proposed a modified version of semi-supervised learning method using K-means to select high value data before putting them into the classifier to be labelled. A comparison between Fasttext and TF-IDF as the word representation combining with the modified self-training method using K-means clustering method is conducted in this research. The result of clickbait detection on both models gives the best result from the combination of Fasttext and modified version of self-training with 85.20% on accuracy, precision score of 90.80%, recall score of 81.60%, and F1-score of 85.90%

Keywords: Clickbait Detection, Natural Language Processing, Fasttext, TF-IDF, Semi-supervised Learning, K-Means, Support Vector Machine.