ABSTRACT

CLICKBAIT DETECTION FOR BAHASA INDONESIA USING BI-LSTM

Andika William
16/398497/PA/17458

Clickbait usage has been increasing rapidly over the years along with the rise of online journalism. Since online publishers rely upon clicks in order to generate revenue, there is a growing trend of writing headlines with the sole aim of attracting clicks instead of delivering information. The problem of clickbait has raised concerns in the international community in recent years, however, the majority of studies are still focused on English and are still lacking in other languages.

This research attempts to study the case of clickbait for Indonesian news, as well as develop an automatic detection model using machine learning methods. Following the lack of published datasets on Indonesian News Headlines, this research attempts to contribute by constructing an Indonesian news dataset that was extracted from 12 Indonesian publishers. The "CLICK-ID" dataset is consisted of 46,517 articles data, along with a clickbait corpus of 15,000 annotated headlines. Our labels shows that out of the 12 publishers, *all* publishers were detected of using clickbait, indicating the use of clickbait is already common practice. By using this corpus, a Bi-LSTM model was developed with its best model achieving 0.88 accuracy on testing. Finally, this research made use of the LIME explainer library (Ribeiro *et al.*, 2016) in order to analyze and gain insight on why our machine learning model is able to classify clickbait and non-clickbait.

Keywords: Bi-LSTM, Bahasa Indonesia, News, Clickbait, Text-Classification, LIME Explainer