ABSTRACT

Clickbait Detection for News Article in Bahasa Indonesia using Random Forests

Khazita Seiya Sadanti
16/392771/PA/17075

Along with the increasing use and popularity in present as information sources, online media platform triggers the journalist to create contents that are attractive for the readers to read their articles. Therefore, they need to make some catchy headlines for the articles, which are also known as ‘clickbait’. The high amount of clickbait causes some problems to the readers and to the content creators themselves. A necessary prerequisite is machine learning technology which is capable to reliably detect clickbait.

This research aims to classify the news headlines in Bahasa Indonesia to either clickbait or non-clickbait category. In this research, the Random Forest algorithm is going to be used to do the classification task, then its performance will be evaluated.

The result shows that the Random Forest was the best to perform the clickbait detection using unigram features and default value compared to the other cases, such as using Naïve Bayes, Random Forest using bigram and default value, also Random Forest using bigram and the tuned hyperparameter. The evaluation score was 0.97 for precision, recall, f1-score and accuracy score in detecting the clickbait.

Keywords: Random Forest, Clickbait Detection, News Headlines, N-gram, Uni-gram, Bi-gram, TF-IDF

ABSTRAK

Pendeteksi Umpan Klik untuk Artikel Berita dalam Bahasa Indonesia menggunakan Random Forests

Khazita Seiya Sadanti
16/392771/PA/17075

Bersamaan dengan meningkatnya penggunakaan dan popularitas sebagai sumber informasi sekarang ini, media *online* mendorong para jurnalis untuk membuat konten yang menarik pembaca untuk dapat membaca artikel yang mereka buat. Oleh karena itu, mereka perlu untuk membuat judul berita yang menarik perhatian untuk artikel-artikel mereka, yang disebut juga sebagai '*clickbait'* atau (umpan klik). Tingginya jumlah penggunaan umpan klik menyebabkan beberapa masalah untuk para pembaca dan para pembuat konten itu sendiri. Salah satu hal yang dapat dilakukan adalah menggunakan teknologi *machine learning* yang mana dapat digunakan untuk mendeteksi umpan klik.

Penelitian ini bertujuan untuk mengklasifikasikan judul berita dalam Bahasa Indonesia ke dalam kategori umpan klik maupun tidak. Dalam penelitian ini, algoritma Random Forest akan digunakan untuk melakukan pengklasifikasian, kemudian hasil dari performanya akan dievaluasi.

Hasil dari penelitian menunjukkan bahwa Random Forest menunjukkan performa yang paling baik dalam mendeteksi umpan klik menggunakan fitur *unigram* dan nilai standar pada *hyperparameter* dibandingkan pada scenario lainnya, seperti saat menggunakan Naïve Bayes, Random Forest menggunakan bigram dan nilai standar pada *hyperparameter*, serta Random Forest menggunakan bigram dan *hyperparameter* yang telah disesuaikan nilainya sebelumnya. Hasil skor evaluasi performa, yaitu 0,97 untuk nilai presisi, *recall, f1-score* dan akurasi dalam mendeteksi umpan klik.

Kata kunci*: Random Forest, Clickbait Detection, News Headlines, N-gram, Uni-gram, Bi-gram, TF-IDF*