INTISARI

PERBANDINGAN METODE SELEKSI FITUR PADA KLASIFIKASI ARTIKEL BERITA BERBAHASA INDONESIA MENGGUNAKAN NAÏVE BAYES

Gilang Mugiardi
10/305414/PA/13509

Banyaknya fitur yang biasanya muncul pada masalah klasifikasi teks menjadi perhatian utama karena bisa menyebabkan bertambahnya waktu pelatihan. Oleh karena itu, banyak metode seleksi fitur yang dikembangkan untuk memilih fitur-fitur yang ada. Namun demikian, performa dari proses klasifikasi juga akan terpengaruh. Beberapa metode seleksi fitur diajukan baik itu metode seleksi fitur baru atau yang merupakan modifikasi dari metode yang sudah ada dengan tujuan untuk mendapatkan performa terbaik. Beberapa dari metode tersebut diantaranya adalah *Distinguishing Feature Selector*, *Term Significance*, *Improved Information Gain* dan *Improved Mutual Information*.

Pada penelitian ini dilakukan perbandingan performa dari keempat metode seleksi fitur tersebut. Perbandingan dilakukan dengan menggunakan 3 varian dari *Naïve Bayes* yaitu : *Multinomial Naïve Bayes*, *Binarized Multinomial Naïve Bayes* dan *Multivariate Bernoulli Naïve Bayes*. Data yang digunakan adalah artikel berita berbahasa Indonesia. Pada tahap sebelum pemrosesan, dilakukan pembuangan *stop words* dan *stemming*. Pengujian dilakukan untuk mengukur akurasi, *Macro-F1*, *Micro-F1* dan waktu pemrosesan dengan menggunakan metode *k-fold cross validation*.

Hasil pengujian menunjukkan bahwa metode *Improved Mutual Information* merupakan metode seleksi fitur dengan pengaruh performa klasifikasi terbaik pada 15 dari 21 kombinasi pengujian. Algoritma *Multinomial Naïve Bayes* merupakan algoritma dengan performa klasifikasi terbaik ketika jumlah fitur yang digunakan lebih dari atau sama dengan 3000. Algoritma *Multivariate Bernoulli Naïve Bayes* merupakan algoritma dengan performa klasifikasi paling stabil terhadap penggunaan jumlah fitur yang berbeda-beda. Metode *Improved Information Gain* merupakan metode seleksi fitur dengan waktu pemilihan tercepat.

Kata kunci : seleksi fitur, klasifikasi teks, *naïve bayes*, perbandingan

ABSTRACT

COMPARISON OF FEATURE SELECTION METHOD ON CLASSIFICATION OF INDONESIAN LANGUAGE NEWS ARTICLES USING NAÏVE BAYES

Gilang Mugiardi
10/305414/PA/13509

The high number of features that usually appear on text classification problem are one of the main concerns as they can lead to increased training time. Therefore, many feature selection methods being developed to select the available features. However, the performance of the classification process will also be affected. Some feature selection methods are proposed whether it is a completely new feature selection method or that is a modification of an existing one in order to get the best performance. Some of these methods are Distinguishing Feature Selector, Term Significance, Improved Information Gain and Improved Mutual Information.

A performance comparison between the four feature selection methods was done in this research. The comparisons were made by using 3 variants of Naïve Bayes : Multinomial Naïve Bayes, Binarized Multinomial Naïve Bayes and Multivariate Bernoulli Naïve Bayes. The data used is news articles in Indonesian language. In the preprocessing stage, the stop words removal and stemming is done. Testing is done to measure accuracy, Macro-F1, Micro-F1 and processing time by using k-fold cross validation method.

The test results show that the Improved Mutual Information method is a feature selection method with the best classification performance effect on 15 out of 21 test combinations. Multinomial Naïve Bayes algorithm is the best performing classification algorithm when the number of features used is more than or equal to 3000. Multivariate Bernoulli Naïve Bayes algorithm is the most stable classification algorithm against different number of features. The Improved Information Gain method is a feature selection method with the fastest selection time.

Keywords : feature selection, text classification, naïve bayes, comparison