INTISARI

KLASIFIKASI BERITA BERKATEGORI OLAHRAGA DENGAN ALGORITMA *MULTIVARIATE BERNOULLI NAÏVE BAYES* DAN *MULTINOMIAL NAÏVE BAYES*

Oleh

Adhika Wisaksono
13/347489/PA/15261

Bertambahnya pengguna internet di Indonesia menyebabkan muncul dan berkembangnya situs-situs penyedia informasi yang banyak ditemukan dalam bentuk berita. Jenis berita yang dicari oleh masyarakat bermacam-macam. Oleh karena itu banyak kategori disajikan oleh situs penyedia berita. Antara satu situs berita dengan situs berita lainnya tidak sama dalam menentukan kategorisasi berita. Tetapi terdapat kategori yang cukup populer dan ada hampir di setiap situs berita, yaitu kategori olahraga menjadi dasar penentuan kelas pada penelitian ini.

Pada penelitian ini dilakukan klasifikasi teks berita menggunakan algoritma *Multinomial Naïve Bayes* dan *Multivariate Bernoulli Naïve Bayes*. Kategori yang dipakai dalam penelitian ini adalah 9 kelas cabang olahraga utama, 1 kelas cabang olahraga lain dan 1 kelas non-olahraga. Proses dimulai dari pengumpulan data dengan metode *scraping*, *preprocessing*, lalu seleksi fitur. Tahap selanjutnya dibentuk model klasifikasi pada tahap *training*. Prediksi dan evaluasi dilakukan pada tahap *testing*. Untuk memvalidasi uji performa atas data *training* yang terkumpul, digunakan metode 10-*fold* cross validation.

Didapatkan bahwa nilai performa maksimum algoritma *Multinomial Naïve Bayes* didapatkan saat digunakan metode seleksi fitur TF-IDF dengan *top-n* = 2000 yang menghasilkan rata-rata nilai akurasi 97,1%, presisi 97,3%, *recall* 97,1% dan *f-measure* 97%. Sementara metode klasifikasi multivariate *bernoulli* mendapatkan nilai performa maksimal saat menggunakan metode seleksi fitur *most common* dengan *top-n* = 1500 yang menghasilkan rata-rata nilai akurasi 94%, presisi 94,8%, *recall* 94% dan *f-measure* 93,9%. Metode klasifikasi *Multinomial Naive Bayes* memiliki nilai performa yang lebih tinggi daripada *Multivariate Bernoulli Naive Bayes* pada kasus ini. Metode seleksi fitur TFIDF meraih performa maksimal jika jumlah kata fitur lebih banyak pada model *multinomial* dan lebih sedikit pada model *bernoulli*. Sementara *most common* memiliki performa maksimal saat jumlah kata fitur banyak pada kedua model.

Kata Kunci : klasifikasi teks, text mining, *Multinomial Naive Bayes*, *Multivariate Bernoulli Naive Bayes*

ABSTRACT

SPORT NEWS CLASSIFICATION USING MULTIVARIATE BERNOULLI NAÏVE BAYES AND MULTINOMIAL NAÏVE BAYES ALGORITHM

by

Adhika Wisaksono
13/347489/PA/15261

As the number of internet users in Indonesia grows, there has been noticeable increase in the emergence and development of information providing sites that are mostly found in the form of news sites. People search for various types of news. Therefore, news sites classify their posts into different categories. News categorization itself varies from one site to another. Nonetheless, sport is a particular category that can be found in almost every news site and therefore served as the base for class determination in this research.

This research has done news text classification using Multinomial Naïve Bayes algorithm and Multivariate Bernoulli Naïve Bayes algorithm. The categories used in this research are 9 primary sport branches, 1 class of other sport and 1 class of non-sport. The process starts with collecting the data using scraping method, preprocessing, and then selecting feature. The next phase is training phase in which classification model is formed. Prediction and evaluation done in testing phase. To validate performance test value from training data collected, 10-fold cross validation method is used.

It is found that the maximum performance value of Multinomial Naive Bayes algorithm is obtained when it used TF-IDF feature selection method using top-n value = 2000 which has average performance value of 97,1% accuracy, 97,3% precision, 97,1% recall and 97% f-measure. On the other hand, Multivariate Bernoulli Naive Bayes method reached its maximum performance value when it used most common feature selection method with top-n value = 1500 which has average performance value of 94% accuracy, 94,8% precision, 94% recall and 93,9% f-measure. The Multinomial Naive Bayes method has higher performance value than Multivariate Bernoulli Naive Bayes method in this case. TFIDF method reached its maximum performance when it used more features in multinomial model but fewer features in bernoulli model. While most common method reached its maximum performance when it used more features on both models.

Keyword : text classification, text mining, *Multinomial Naive Bayes*, *Multivariate Bernoulli Naive Bayes*