ABSTRAK

Wira Kusuma Jaya, I Gede (2020), Klasifikasi Teks Berita Berbahasa Indonesia dengan Algoritma Naive Bayes melalui Kombinasi Lavenhstein Distance dan Terjemahan Kata Bahasa Inggris-Indonesia. Tesis, Ilmu Komputer, Program Pascasarjana, Universitas Pendidikan Ganesha.

Tesis ini sudah disetujui dan diperiksa oleh Pembimbing I : Dr. Gede Rasben Dantes, S.T., M.T.I. dan Pembimbing II: Dr. Gede Indrawan S.T., M.TI.

Kata-kata kunci: algoritma naive bayes, modifikasi algoritma, klasifikasi teks

Penelitian ini bertujuan membandingkan tingkat akurasi dan kecepata proses klasifikasi antara algoritma Naive Bayes dan algoritma Naive Bayes dengan beberapa modifikasi. Modifikasi yang dimaksud adalah dengan penambahan proses perubahan kata menggunakan algoritma Levenhstein Distance dan terjemahan kata dalam bahasa Inggris ke dalam kata dasar berbahasa Indonesia yang dilakukan pada tahap preprosesnya. Data latih dan data uji yang digunakan diambil dari sebuah situs berita online yang terdiri dari 7 kategori berita yang telah ditentukan, kemudian dilakukan preproses teks sebelum dilakukannya tahap klasifikasi teks. Untuk mengukur tingkat akurasi, *precision* dan *recall*, digunakan sebuah metode yang bernama *confusion matrix.* Secara berturut-turut dalam lima kali percobaan yang telah dilakukan, rata-rata algoritma Naive Bayes setiap kategori adalah Fenomena (76,46%), Film (71,04%), Internet (69,92%), Keuangan (65,67%), Musik (58,29%), Olahraga (97,54%), Otomotif (88,88%), dan sedangkan algoritma Naive Bayes Modifikasi adalah sebesar Fenomena (98,50%), Film (95,04%), Internet (98,25%), Keuangan (94,79%), Musik (87,08%), Olahraga (95,29%), Otomotif (95,29%). Hasil rata-rata yang diperoleh untuk akurasi klasifikasi Naive Bayes adalah 75,40% sedangkan untuk algoritma Naive Bayes Modifikasi adalah 94,95%. Kecepatan proses untuk klasifikasi Naive Bayes sebesar 0,16870 detik dan Naive Bayes Modifikasi adalah 0,13512 detik. Peningkatan waktu proses terjadi hanya 19,91%, hal ini dipengaruhi oleh jumlah kata yang digunakan dan pengukuran kecepatan proses dilakukan saat klasifikasi saja dan bukan saat preproses berlangsung. Penelitian selanjutnya dapat dikembangkan dengan mencoba algoritma perubahan kata lainnya dan pembaharuan kamus data kata asing atau kata tertentu yang tidak diubah karena telah memiliki makna tersendiri.

ABSTRACT

Wira Kusuma Jaya, I Gede (2020), Indonesian Language News Text Classification using the Naive Bayes Algorithm through the Combination of Lavenhstein Distance and English-Indonesian Word Translation. Thesis, Computer Science, Graduate Program, Ganesha University of Education.

Keyword : *algoritma naive bayes, modifikasi algoritma, klasifikasi teks*

This study aims to compare the accuracy and speed of the classification process between the Naive Bayes algorithm and the Naive Bayes algorithm with several modifications. The modification referred to is the addition of the word change process using the Levenhstein Distance algorithm and the translation of English words into Indonesian basic words which are carried out at the preprocessing stage. The training data and test data used were taken from an online news site consisting of 7 predefined news categories, then the text preprocessing was carried out before the text classification stage was carried out. To measure the level of accuracy, precision and recall, a method called confusion matrix is used. In five consecutive experiments that have been carried out, the average Naive Bayes algorithm in each category is Phenomenon (76.46%), Film (71.04%), Internet (69.92%), Finance (65.67%). %), Music (58.29%), Sports (97.54%), Automotive (88.88%), and while the Modified Naive Bayes algorithm is Phenomenon (98.50%), Film (95.04%) , Internet (98.25%), Finance (94.79%), Music (87.08%), Sports (95.29%), Automotive (95.29%). The average result obtained for the Naive Bayes classification accuracy is 75.40% while for the Modified Naive Bayes algorithm is 94.95%. The processing speed for the Naive Bayes classification is 0.16870 seconds and the modified Naive Bayes classification is 0.13512 seconds. The increase in processing time was only 19.91%, this was influenced by the number of words used and the measurement of processing speed was carried out only during classification and not during the preprocessing process. Further research can be developed by trying other word change algorithms and updating the data dictionary for foreign words or certain words that are not changed because they already have their own meaning.