ABSTRAK

Dengan semakin banyaknya jurnal ilmiah yang dipublikasikan secara daring, kebutuhan akan sistem klasifikasi kategori jurnal yang efisien dan akurat menjadi sangat penting. Dalam mengklasifikasi jurnal dibutuhkan 2 parameter utama yaitu judul dan abstrak. Pada penelitian ini dilakukan 4 tahap. Pertama, pre-processing dataset yang mencakup case folding, tokenizing, stopword removal, dan stemming. Kedua, dilakukan pembobotan kata menggunakan Term Frequency-Inverse Document Frequency (TF-IDF). Setelah pembobotan, dilakukan penerapan algoritma K-Nearest Neighbor (K-NN). Di salah satu proses K-NN terdapat perhitungan jarak metriks. Pada penelitian ini perhitungan jarak cosine similarity dan euclidean distance dibandingkan. Hasil penelitian menunjukkan bahwa terdapat perbedaan dalam kinerja algoritma K-Nearest Neighbor (K-NN) berdasarkan kedua metode perhitungan jarak tersebut. Terakhir, pengujian dengan confusion matrix. Cosine similarity cenderung memberikan hasil yang lebih baik dalam mengukur kemiripan antar dokumen teks, terutama dalam konteks klasifikasi teks yang melibatkan banyak dimensi dan variasi kata. Hasil akurasi terbesar oleh cosine similarity sebesar 0,8249 pada k = 23. Di sisi lain, euclidean distance, meskipun sederhana dan intuitif, kurang efektif dalam menangani data teks yang memiliki dimensi tinggi. Hasil akurasi terbesar oleh euclidean distance sebesar 0,8209 pada k = 23.

Kata kunci: K-Nearest Neighbor (K-NN), Cosine Similarity, Euclidean Distance, Jarak metriks, Text mining, Jurnal

ABSTRACT

With the increasing number of scientific journals published online, the need for an efficient and accurate journal category classification system is very important. In classifying journals, 2 main parameters are needed, namely title and abstract. In this research, 4 stages were carried out. First, dataset pre-processing which includes case folding, tokenizing, stopword removal, and stemming. Second, word weighting is done using Term Frequency-Inverse Document Frequency (TF-IDF). After weighting, the K-Nearest Neighbor (K-NN) algorithm is applied. In one of the K-NN processes, there is a metric distance calculation. In this study, cosine similarity and euclidean distance calculations were compared. The results show that there is a difference in the performance of the K-Nearest Neighbor (K-NN) algorithm based on the two distance calculation methods. Finally, testing with confusion matrix. Cosine similarity tends to give better results in measuring the similarity between text documents, especially in the context of text classification that involves many dimensions and word variations. The largest accuracy result by cosine similarity is 0.8249 at k = 23. On the other hand, euclidean distance, although simple and intuitive, is less effective in handling high-dimensional text data. The largest accuracy result by euclidean distance is 0.8209 at k = 23.

Keywords: K-Nearest Neighbor (K-NN), Cosine Similarity, Euclidean Distance, Metric distance, Text mining, Journal