Abstrak

Dalam pangkategorisasian teks, biasanya terdapat *outlier* dalam *data training*, seperti *mislabeled* pada data, data yang tekategori pada perbatasan antara dua kategori berita, data yang memang tidak terkategorikan, dan lain- lain. Oleh karena itu, perlu dilakukan outlier detection untuk meningkatkan performansi sebuah dokumen teks.

Salah satu metode dalam outlier detection adalah Distance Based Outlier Detection, yaitu mencari outlier berdasarkan jarak antar data dalam data set. Metode Distance Based Outlier yang sering digunakan adalah metode dengan menggunakan k-Nearest Neighbor, yang memiliki tiga pengertian atau cara, yaitu: *outlier* adalah contoh yang lebih kecil atau lebih besar dari p contoh lainnya dalam jarak d, *outlier* adalah n objek yang memiliki jarak yang terjauh pada ke-*k-nearest neighbor*, dan o*utlier* adalah n objek yang memiliki jarak rata-rata yang terjauh pada *k-nearest neighbor*.

Hasil dari sistem ini adalah perbandingan performansi sistem dalam mendeteksi outlier dari ketiga pengertian di atas. Selain itu, sistem ini juga menghasilkan performansi pengkategorian dokumen sebelum dan sesudah outlier dihilangkan.

Kata kunci :.*data mining, outlier detection, k- nearest neigbhor*