Abstrak

Text categorization bertujuan untuk mendefiniskan kategori dari dokumen yang belum diketahui kategorinya. Salah satu algoritma dalam *text categorization* adalah algoritma multinomial naïve bayes yang diketahui memiliki cara kerja yang sangat sederhana, efektif dan mempunyai performansi yang bagus. Pada penelitian-penelitian sebelumnya, *text categorization* hanya sampai pada tahap kategorisasi saja. Pada penelitian ini, file biner dokumen hasil kategorisasi akan disimpan ke dalam basisdata agar pengolahan data dapat dilakukan dengan mudah. Salah satu basisdata jenis NoSQL adalah *document oriented* dan salah satu DBMS dari *document oriented* adalah MongoDB. MongoDB mempunyai fitur GridFS dan *sharding* yang dapat menyimpan file biner secara distribusi ke dalam beberapa mesin. Dengan menempatkan data pada beberapa mesin memungkinkan untuk menyimpan lebih banyak data dan menangani beban lebih besar tanpa diperlukan adanya mesin dengan performansi tinggi. Dari hasil pengujian yang dilakukan, performansi kategorisasi teks berada diatas nilai 88% pada pemakaian 756 data latih dan 84 data uji. Untuk basisdata dokumen, hasil terbaik yang didapatkan dengan nilai *response time* 1,713 detik dan *throughput* 130,869 transaksi per detik.

Kata kunci : *Text categorization,* NoSQL*, document oriented database,* GridFS*, sharding*