ABSTRAK

Kemajuan IT (*Information Technology*) yang semakin pesat menyebabkan jumlah pengolahan data terkomputerisasi semakin meningkat, begitupun dengan jumlah data digital yang diolah. Untuk mendapatkan suatu informasi yang diinginkan tentu bukanlah hal yang mudah. *Clustering* merupakan salah satu teknik dalam *Text mining* yang dapat menjadi solusi untuk mengatasi masalah ini. *Clustering* bertujuan untuk mengelompokkan data/dokumen ke dalam suatu kelompok/klaster yang memiliki informasi yang sejenis/hampir sama. Efektifitas dan hasil dokumen klaster dari proses *clustering* ditentukan oleh algoritma yang digunakan, salah satu algoritma yang cukup populer dalam *clustering* dokumen teks adalah *Suffix Tree Clustering (STC).* Namun, sebenarnya algoritma STC sendiri masih memiliki beberapa kekurangan, salah satunya yaitu algoritma STC tidak memiliki pengukuran kesamaan yang efisien untuk menghitung kesamaan antara *inter-cluster* maupun *intra-cluster*. Maka dari itu pada penelitian ini akan digunakan algoritma modifikasi dari STC yaitu STHAC (*Suffix Tree Hierarchical Agglomerative Clustering*). Pada algoritma ini akan menambahkan 2 proses tambahan pada algoritma STC biasa, yaitu adanya proses *clusters ranking and filtering*, serta proses *cluster cleaning*. Hasil dari *clustering* dianalisis dengan uji validasi *Precision, Recall,* dan *F-measure* untuk mengetahui kualitas performansinya. Setelah dilakukan beberapa pengujian diperoleh bahwa secara umum algoritma STHAC memiliki hasil klasterisasi dan performansi yang lebih baik dibandingkan dengan algoritma STC. Nilai *Recall* dan *F-measure* pada hasil pengujian STHAC lebih tinggi dari STC. Sedangkan untuk nilai *Precision,* algoritma STHAC masih lebih rendah daripada STC.

Kata Kunci: *Clustering, Text mining, Suffix Tree Hierarchical Agglomerative Clustering, Precision, Recall, F-measure.*