

## Abstrak

Pada masa sekarang, untuk mempelajari sebuah ilmu sangatlah mudah. Akan tetapi, ilmu yang ada di dunia sangatlah banyak, dan untuk membaca keseluruhan makalah ilmu tersebut membutuhkan cukup banyak waktu. Maka dari itu diperlukan sebuah sistem yang dapat memberikan informasi – informasi yang dibutuhkan. Hal ini bisa dilakukan dengan melakukan klasifikasi teks. Proses yang terjadi dalam klasifikasi teks adalah pra-proses data, ekstraksi fitur, pemilihan fitur, dan klasifikasi. Pra-proses merupakan satu hal yang mempengaruhi proses klasifikasi, dimana pada proses ini terjadi proses perubahan data mentah menjadi data siap olah. Hal ini dilakukan dikarenakan data mentah biasanya tidak dapat digunakan langsung ke dalam proses klasifikasi teks. Pra-proses ini sendiri biasa dilakukan secara manual. Pada tugas akhir ini, penulis melakukan perbandingan pra-proses data manual dengan pra-proses otomatis. Data yang dipakai merupakan data yang sama yang berjumlah 30 makalah ilmiah berformat PDF yang didapatkan dari *internet*. Pra-proses dengan cara otomatis dilakukan dengan melakukan konversi PDF. Kemudian pemecahan kalimat, dan pendeteksian kalimat sitasi. Penelitian ini dilakukan dengan 2 kelompok fitur yang dimana kelompok fitur pertama hanya menggunakan fitur yang didapat dengan cara mengaplikasikan RegEx dan mendapatkan hasil lebih dari 97%, sedangkan kelompok fitur kedua mengaplikasikan ReEx ditambah dengan N-gram dan menghasilkan hasil presentase sebesar 93%. Hasil dari penelitian ini menghasilkan presentase rata-rata akurasi sistem lebih dari 95%

**Kata kunci:** klasifikasi text, kalimat sitasi, pra-proses.