Abstrak

Perkembangan informasi digital berkembang sangat cepat, informasi tersebut dapat berupa teks berita, lirik lagu, atau atrikel. Tingginya pengguna internet menyebabkan banyak dokumen yang tidak diketahui siapa atau apa gender dari penulis teks tersebut. Banyak orang menggunakan media sosial untuk menulis dan menyebabkan semakin tingginya teks yang ada di internet. Salah satu masalah yang ada yaitu apakah kita dapat mengetahui gender dari penulis tersebut? Dan apakah penulis tersebut tidak memalsukan gendernya? Analisis *text mining* diperlukan dalam masalah ini. Dengan mengetahui siapa penulis dari sebuah teks dapat mengurangi kasus plagiarisme yang sangat merugikan seorang penulis.

Kegiatan *text mining* yang dapat dilakukan yaitu klasifikasi. Ada beberapa metode klasifikasi yang dapat digunakan seperti Support Vector Machine, Naive Bayes Classsifier, dan Decision Tree. Pada penelitian sebelumnya [1] untuk mengidentifikasi gender pada teks menggunakan tiga metode berbeda yaitu *Support Vector Machine, Bayesian-based logistic regression* dan *AdaBoost decision tree.* Ekstraksi ciri pada penelitian ini menggunakan fitur-fitur sebagai berikut : (1) *character-based; (2) word-based; (3) syntactic; (4) structure-based;* dan *(5) function words.*

Berdasarkan hasil penelitian kombinasi dengan hasil akurasi paling rendah yaitu menggunakan kombinasi fitur *Character-Based Features dan Syntactic-Based Features* dengan akurasi 42%, sedangkan kombinasi dengan hasil akurasi paling tinggi yaitu hasil pengujian menggunakan kombinasi fitur *Syntactic-Based Features, Structurally-Based Features dan Function Word-Based Features* dengan akurasi 76%.

Kata Kunci : klasifikasi dokumen, klasifikasi gender penulis, support vector machine.