ABSTRAK

Banyak penelitian sebelumnya melakukan klasifikasi teks korpus ke dalam topik, sentimen, *genre*, atau penulis. Di dalam penelitian ini, diselidiki bentuk teks korpus yang berbeda. Teks korpus yang digunakan adalah tafsir Al-Quran yang dibuat oleh Al-Jalalyn. Teks tersebut memiliki kelas perintah, larangan dan informasi.

Alasan penggunaan dataset tafsir Al-Quran karena isinya yang terkadang sulit untuk dibedakan bahkan oleh manusia. Ditemukan bahwa jumlah kata unik untuk setiap kelas sangat sedikit, namun jumlah kata *noise* cukup banyak. Tantangan lain adalah adanya ayat Al Quran yang maknanya bersifat implisit. Untuk menangani ketidakmampuan untuk mengenali makna implisit, digunakan kamus WordNet sebagai alat bantu untuk melakukan perhitungan kemiripan semantic. Di penelitian ini, dilakukan beberapa tahap untuk mengklasifikasikan sebuah dokumen, diantaranya *preprocessing,* ekstraksi fitur, pembobotan semantic, *classifier training*, dan evaluasi. Pada saat melakukan ekstraksi fitur, dihasilkan beberapa fitur diantaranya *Term Frequency* (TF), *Term Frequency–Inverse Document Frequency* (TF-IDF), *Part-of-Speech Tagging* (POSTAG), dan *Bigram*.

Metode yang diusulkan adalah melakukan perhitungan bobot yang dinamakan *Document-to-Class semantic similarity*. Metode baru yang digunakan untuk menghitung kemiripan semantic adalah gabungan dari metode Wu dan Palmer (WUP) dan metode jalur terpendek, lalu penulis sedikit memodifikasinya. Setelah itu, dilakukan *classifier training*, dimana *classifier* yang digunakan adalah Multinomial Naïve Bayes yang telah dimodifikasi. Metode yang diusulkan adalah memodifikasi *likelihood probability* dengan menggunakan nilai bobot yang telah didapatkan sebelumnya.

Pada saat proses evaluasi, dinilai performansi klasifier terhadap dataset Al-Quran yang telah dibuat. Untuk perbandingan, digunakan dataset ulasan Amazon, dataset ulasan Yelp dan dataset ulasan IMDB. Metode evaluasi yang digunakan adalah akurasi, *precission, recall* dan *F1-Measure.* Nilai F1 untuk klasifikasi Al-Quran dataset menggunakan kombinasi fitur POSTAG, BIGRAM dan TF adalah 60.5%. Nilai *F1-Measure* untuk kombinasi POSTAG, BIGRAM dan TFIDF adalah 58.6% dan nilai *F1-Measure* untuk kombinasi POSTAG, BIGRAM dan *Weighted TF* yang diusulkan adalah 66.4%

Kata Kunci: Klasifikasi teks, kemiripan semantik, ekstraksi fitur, tafsir Al-Quran