1. Pendahuluan

Hadis bagi umat islam merupakan sumber hukum kedua setelah Al-Qur'an yang dimana mengandung ajaran dari berbagai sunnah yang dilakukan Nabi Muhammad SAW. Setiap hadis terdiri dari dua bagian yaitu *Sanad* yang berartikan kumpulan nama para pembilang hadis yang menunjukkan keaslian dari hadis tersebut dan *Matan* yang berartikan penjelasan dari hadis [10]. Umumnya hadis dikumpulkan oleh beberapa imam besar, dan hadis Bukhari merupakan kumpulan hadis yang disusun oleh salah satu imam besar yaitu Imam Bukhari [2]. Maka dari itu, umat Islam dianjurkan untuk mengetahui dan mengamalkannya dengan mengenali hadis dari kumpulan buku hadis yang tersedia dalam versi cetak hingga versi digital. Di dalam hadis, terdapat beberapa jenis ajaran yang dapat diambil seperti jenis ajaran yang mengandung anjuran, larangan, dan informasi [1][13].

Atas dasar tersebut, dalam penelitian ini akan membangun suatu sistem yang dimana dapat mengelompokkan Hadis terjemahan Bahasa indonesia berdasarkan karateristik dari masing-masing kategori Hadis yaitu anjuran, larangan, dan informasi. Dalam pembangunannya, diperlukan sistem klasifikasi teks yang dapat mengelompokkan dan mengidentifikasi suatu hadis kedalam beberapa kategori. Proses klasifikasi ini dapat disebut dengan klasifikasi *multi label*. Untuk itu berbagai pendekatan klasifikasi yang dapat digunakan, salah satunya metode *Random Forest* (RF). *Random forest* merupakan metode klasifikasi yang memiliki performansi paling baik dalam melakukan klasifikasi teks. Hal ini disebabkan metode *Random Forest* memiliki kemampuan menggeneralisasi suatu data berdimensi tinggi[11]. *Random forest* memiliki kelemahan yaitu Ketika menghadapi jenis data yang mengandung jumlah kelas fitur yang sangat banyak[7][11]. Hal tersebut dapat menyebabkan metode RF memiliki nilai bias untuk banyak kelas data, sehingga dapat membuat terjadinya *overfitting* [11]. Untuk mencegah hal tersebut, dapat dilakukan dengan mereduksi dimensi vektor[2]. Proses yang dapat dilakukan yaitu proses seleksi fitur dengan tujuan menyeleksi kata atau fitur yang dianggap kurang berpengaruh dalam pembentukan suatu model untuk tahap klasifikasi[2][7]. Metode *Chi-Square* digunakan sebagai metode seleksi fitur karena memiliki kemampuan untuk memilih fitur-fitur yang berpengaruh dalam menghasilkan prediksi yang tepat dalam suatu kelas[1][2].

Topik dan Batasannya

Dalam penelitian akan dilakukan perancangan sistem klasifikasi *multi label* pada dataset hadis yang berjumlah 7000 data. Dataset hadis terlebih dahulu sudah dilakukan labeling dan divalidasi oleh pihak ulama. Sebelum dilakukan klasifikasi, dataset terlebih dahulu akan dilakukan *preprocessing* dan dilakukan *stemming* dengan bantuan library *sastrawi*.

Tujuan

Berdasarkan penjelasan dari latar belakang, maka akan dilakukan penelitian dengan klasifikasi *multi label* pada Hadis Bukhari terjemahan Bahasa Indonesia menggunakan *Random Forest* sebagai metode klasifikasi serta *Chi-Square* sebagai metode seleksi fitur. Setelah itu, Pada penelitian akan dilakukan pengujian dengan mengidentifikasi pengaruh dari penggunaan seleksi fitur, pengujian pengaruh penggunaan data *stemming* & tanpa *stemming*, dan pengujian pengaruh terhadap pengaturan nilai *threshold* yang digunakan untuk mendapatkan hasil optimum dari proses klasifikasi.