BAB I

PENDAHULUAN

1.1. Latar Belakang

Seiring dengan semakin majunya perkembangan teknologi menyebabkan jumlah informasi yang tersedia juga semakin berlimpah. Sumber informasi dapat diperoleh melalui berbagai macam media informasi, seperti media elektronik dan media cetak. Dalam survey *online* yang dilakukan oleh Pew Internet Survey sejak tahun 2004, menemukan bahwa 92% dari penduduk Amerika Serikat menggunakan media internet yang merupakan salah satu dari media elektronik sebagai sumber untuk memperoleh informasi (Fallows 2004). Seiring dengan semakin berkembangnya pengguna dari media elektronik seperti internet, penyebaran informasi juga semakin pesat. Hal ini tentunya menyebabkan informasi-informasi seperti artikel ilmiah, jurnal, e-book, dan dokumen-dokumen elektronik lainnya semakin berlimpah.

Dengan semakin berkembangnya jumlah informasi mahasiswa pada saat ini mendapatkan keuntungan. Mahasiswa dengan mudahnya mendapatkan bahan sebagai dasar maupun landasan dalam melakukan penelitian atau sebagai dasar untuk mengerjakan tugas-tugas dalam masa perkuliahan. Dengan banyaknya jumlah dokumen yang tersedia, mulai dari bidang akademik sampai non-akademik menyebabkan mahasiswa sering merasa kesulitan dalam mendapatkan dokumen yang dibutuhkan sebagai dasar atau landasan literatur. Banyak kasus terjadi ketika mahasiswa tidak mendapatkan dokumen yang dibutuhkan karena isi dari dokumen tidak sesuai yang dibutuhkan. Oleh karena itu perlu adanya suatu sistem yang dapat

memberikan kemudahan dalam mengklasifikasikan dokumen berdasarkan dengan kategori yang sesuai.

Sistem temu kembali informasi adalah suatu sistem yang mampu dalam menyimpan, mendapatkan, dan me-*maintenance* informasi (Kowalski, 1999). Informasi yang dimaksud dapat berupa teks dalam dokumen, audio, video, dan objek multimedia lainnya. Pada dasarnya sistem temu kembali merupakan suatu sistem yang mencocokan *query* dari pengguna yang berisi statemen dari informasi yang diinginkan dengan dokumen-dokumen yang telah tersimpan dalam database (Frakes, 1984). Dalam proses sistem temu kembali informasi terdapat pengolahan data yang berupa *text mining*. Dalam *text mining* untuk pengolahan dokumen dilakukan dalam 4 tahapan utama. Pertama yaitu *text preprocessing* yang merupakan tahap awal dalam pengolahan dokumen untuk merubah dokumen menjadi bentuk paling sederhana dengan menggunakan tokenisasi. Kedua, dengan menggunakan *text transformasi* mengolah hasil dari tokenisasi yang berupa kumpulan kata menjadi bentuk dasar dari kata yang terkumpul dengan menggunakan metode *stemming*. Ketiga adalah *feature selection,* dalam tahap ini dibuat suatu *stoplist* yang berisi kumpulan kata-kata yang tidak relevan dengan isi dari tiap dokumen yang diolah. Kemudian dilakukan *stopword removal* untuk menghilangkan kata-kata yang terkandung didalam *stoplist* tersebut. Dan yang terakhir adalah *pattern discovery*. *Pattern discovery* merupakan tahap penting dalam pengolahan sistem temu kembali, karena dalam tahap ini ditentukan proses pengolahan yang akan dilakukan pada dokumen-dokumen yang telah terkumpul.

Dalam *pattern discovery* terdapat 2 jenis learning yang akan digunakan, yaitu *supervised* dan *unsupervised learning*.

Sistem temu kembali informasi yang sempurna adalah sistem yang mampu mendapatkan dokumen yang relevan dengan *query* yang dimasukan tanpa menampilkan dokumen-dokumen yang tidak relevan. Namun, sistem temu kembali informasi tidak akan ada yang sempurna, hal ini dikarenakan relevan tidaknya suatu dokumen yang didapatkan dalam sistem tersebut tidak semua pengguna menganggap dokumen tersebut relevan dengan queri yang dihasilkan (Hiemstra, 2009). Dalam penelitian ini, menggunakan sistem temu kembali dengan *supervised learning* yaitu metode learning atau pembelajaran pada dokumen yang dikumpulkan secara *supervised*. *Supervised* yang dimaksud adalah dengan memberikan label atau kategori kelas pada dokumen *training* yang menjadi dasar dalam melakukan klasifikasi dokumen selanjutnya atau dokumen baru. Sedangkan metode yang digunakan dalam penelitian ini adalah metode *K-Nearest Neighbor* (KNN) sebagai klasifikasi dokumen. Dengan menggunakan *cosine similarity*, klasifikasi dokumen akan lebih mudah dilakukan karena dengan menggunakan *cosine similarity* dokumen akan diolah dengan merubah dokumen tersebut menjadi bentuk vektor dan dengan melakukan perbandingan antara vektor dokumen tersebut dapat menghasilkan suatu nilai similaritas antar dokumen. Kemudian dengan menggunakan metode *K-Nearest Neighbor* dapat melakukan klasifikasi terhadap dokumen-dokumen yang telah menghasilkan nilai similaritas dengan melihat nilai similaritas terbesar dan mayoritas dari dokumen *training* sebagai pembanding yang muncul dalam lingkup nilai k yang sebelumnya telah di-inisialisasi oleh user.

1.2. Rumusan Masalah

Berdasarkan pada latar belakang diatas, maka rumusan masalah dalam penelitian ini adalah bagaimanakah implementasi yang dilakukan dalam pengembangan sistem klasifikasi jurnal dalam sistem temu kembali informasi dengan menggunakan *cosine similarity* dan *K-Nearest Neighbor* (KNN)

1.3. Tujuan

Tujuan yang ingin dicapai dari penelitian ini adalah untuk mengetahui bagaimana implementasi dari sistem temu kembali informasi dalam klasifikasi jurnal dengan menggunakan *cosine similarity* dan *K-Nearest Neighbor* (KNN).

1.4. Manfaat

Dalam pelaksanaan penelitian ini, manfaat yang didapatkan antara lain :

1. Mempermudah dalam pencarian dokumen yang dibutuhkan
2. Mempermudah dalam menyimpan dan memilah suatu dokumen dalam database berdasarkan jenis dokumen.

1.5. Batasan Masalah

Dalam penelitian ini batasan masalah yang digunakan adalah

1. Menggunakan jurnal sebagai dokumen yang akan diklasifikasikan.
2. Membagi kategori jenis dokumen berdasarkan isinya menjadi 4 kategori utama yaitu:
    1. *Physical Sciences and Engineering*
    2. *Life Sciences*
    3. *Health Sciences*
    4. *Social Sciences and Humanities*

3. Menggunakan *cosine similarity* dengan pembobotan TF-IDF dan K-*Nearest Neighbor* sebagai klasifikasi.