1. Pendahuluan

1.1 Latar belakang

Pada saat ini, selain situs jejaring sosial, situs yang cukup terkenal dan banyak dikunjungi oleh sebagian besar orang adalah forum internet. Forum di internet ini merupakan tempat diskusi, ataupun tempat saling bertukar informasi antara sesama *user*. Salah satu cara untuk berdiskusi atau bertukar informasi di forum adalah dengan cara membuat suatu thread atau topik. Setiap *user* dapat membuat suatu thread atau topik yang nantinya dapat dikomentari oleh *user* - *user* lainnya.

Masalah utama dari sebuah forum adalah adanya *user* yang sering mengomentari suatu thread atau topik tetapi apa yang dikomentari tidak berguna atau tidak berhubungan sama sekali dengan isi thread atau topik tersebut. Komentar atau postingan seperti ini dikenal dengan istilah *junk post*. Banyak sekali kerugian akibat adanya *junk post* ini, diantaranya adalah menambah halaman topik tersebut yang mengakibatkan user lain sulit untuk membaca seluruh topik serta menambah waktu membaca forum, mengubah arah pembicaraan sehingga bisa berkembang junk post lainnya dari user lain, pokok topik yang penting bisa dikaburkan oleh *junk post* sehingga ada beberapa hal yang terlewat, dan berbagai macam masalah lainnya.

Untuk itulah pada penelitian ini, penulis ingin mengklasifikasikan suatu komentar atau postingan apakah termasuk *junk post* atau tidak, sehingga diharapkan kedepannya dapat mengurangi *junk post* pada forum internet. Metode yang akan digunakan untuk penelitian ini adalah dengan menggunakan metode *K-Nearest Neighbor* untuk pengklasifikasian teks. Dalam penerapannya, dilakukan juga pembobotan postingan dengan tf-idf dan untuk menghitung nilai *similarity* dari postingan yang ada dengan thread atau topik yang berkaitan menggunakan *Pearson Correlation Distance.* Alasan menggunakan metode K-Nearest Neighbor untuk penilitian ini adalah karena berdasarkan riset yang dilakukan sebelumnya oleh (Khamar, Khushbu, 2013) yang berjudul *Short Text Classification Using kNN Based on Distance Function*, menunjukkan bahwa pengklasifikasian teks pendek yang terdapat pada twitter, blog, rangkuman buku atau film, forum, dan di media yang lain menggunakan kNN menghasilkan akurasi yang baik dibandingkan dengan metode Naïve Bayes dan SVM. Sehingga berdasarkan riset tersebut, penulis ingin menggunakan metode KNN ini karena cocok digunakan untuk mengklasifikasi spam pada forum dan diharapkan akan menghasilkan akurasi yang sama baiknya pula.

1.2 Perumusan masalah

Rumusan masalah dalam tugas akhir ini adalah

1. Bagaimana mengidentifikasi *junk post* dengan metode *K-Nearest Neighbor (KNN)* ?
2. Bagaimana menganalisa hasil yang didapat dari mengidentifikasi *junk post* menggunakan *K-Nearest Neighbor (KNN)* dengan parameter K, jumlah data, dan *stop word removal* ?

1.3 Batasan masalah

Batasan masalah dalam Tugas Akhir ini yaitu :

1. Data yang digunakan sebagai dataset adalah dataset dari kaskus sebagai forum terbesar di Indonesia.
2. Hanya menganalisa mengenai 1 buah topik dan berbagai komentar mengenai topik tersebut.

1.4 Tujuan

Tujuan dari Tugas Akhir ini yaitu :

1 Mengidentifikasi *junk post* dengan metode *K-Nearest Neighbor (KNN)*.
2. Menganalisa hasil yang didapat dari mengidentifikasi *junk post* menggunakan *K-Nearest Neighbor (KNN)* dengan parameter K, jumlah data, dan *stop word removal*.

1.5 Metodologi penyelesaian masalah

Beberapa tahapan yang dilakukan dalam pembangunan Tugas Akhir ini, yaitu :

1. Studi Literatur
   a. Mengumpulkan dan mempelajari literature dan referensi yang menunjang penelitian terkait *K-Nearest Neighbor (KNN)*, *spam opinion, tf-idf, dan pearson correlation distance*.
   b. Mencari, memilah dan mengumpulkan data yang menjadi inputan.

2. Observasi Data
   a. Analisis Input
      Data yang diinputkan berupa 1 data thread dan beberapa data komentar atau posting tentang thread tersebut yang diambil dari kaskus dan akan dibagi menjadi data training dan data testing.
   b. Analisis proses
      Sistem yang dibangun melewati proses *preprocessing, term-weighting,* perhitungan *similarity,* dan melakukan proses klasifikasi dengan metode *K Nearest-Neighbor*.

c. Analisis Output
   Hasil keluaran dari sistem berupa tingkat f-measure serta validitas hasil pengukuran akurasi serta pengklasifikasian data apakah termasuk *junk post* atau tidak.

3. Pembangunan Aplikasi
   a. Implementasi
   - Data training dan testing didapat dari data pada forum kaskus, dimana diambil 1 data thread dan beberapa data komentar atau postingan pada thread tersebut.
   - Melakukan *preprocessing* data, yaitu *stopword removal*, dan tokenisasi.
   - Melakukan pembobotan terhadap data komentar atau postingan dengan metode tf-idf
   - Menghitung similarity antara data testing komentar atau postingan tesdenga topik yang ada dengan menggunakan *Pearson Correlation Distance*.
   - Melakukan proses klasifikasi dengan menggunakan metode *K-Nearest Neighbor (KNN)*.
   b. Pengujian dan Analisis Lanjutan
      Membuat dan menganalisis skenario pengujian.

4. Pengamatan dan Evaluasi Sistem
   Melakukan pengujian sistem yang telah dibangun menggunakan data latih dan uji kemudian melakukan analisis hasil dan parameternya

5. Penyusunan Laporan
   Melakukan penyusunan dokumentasi dari sistem yang telah dibuat, dari tahap awal pembangunan hingga tahap akhir. Dokumentasi ini menjelaskan detail sistem yang dibangun dilengkapi dengan jurnal dan poster tugas akhir untuk mendukung publikasi tugas akhir.