# 1. Pendahuluan

## 1.1 Latar belakang

Internet atau lebih tepatnya WWW(World Wide Web) merupakan sarana yang paling umum digunakan untuk menyebarluaskan informasi. Pesatnya penggunaan Internet mengakibatkan tersedianya informasi dalam jumlah yang sangat besar, beraneka ragam, dan pada umumnya tidak terstruktur. Hal ini mendorong meningkatnya kebutuhan untuk menemukan dan mengelola informasi dengan baik sehingga dihasilkan pengetahuan yang berguna.

Salah satu komponen yang penting dalam pengelolaan informasi adalah pengkategorisasian teks. Tugas utama dari pengkategorisasian teks ini adalah mengklasifikasikan dokumen ke dalam kategori tertentu yang telah didefinisikan sebelumnya. Hal ini dapat membantu pengguna dalam menemukan informasi yang dibutuhkan berdasarkan topik tertentu dengan cepat dan akurat.

Terdapat beberapa *classifier* yang dapat digunakan untuk mengkategorisasikan teks diantaranya yaitu *k-NN Feature Projection* dan *Feature Projection.* Ide utama dari kedua *classifier* ini berasal dari *k-Nearest Neighbor (k-NN). k-NN* merupakan *classifier* yang sering digunakan pada kategorisasi teks karena memiliki performansi yang tinggi, algoritma yang sederhana dan mudah untuk dipelajari. Tetapi *k-NN* memiliki kelemahan utama yaitu waktu yang dibutuhkan untuk mengklasifikasikan dokumen sangat lambat[5]. Hal ini karena *k-NN* harus menghitung nilai kesamaan antara setiap dokumen uji dengan keseluruhan data latih untuk mendapatkan sejumlah *k* tetangga terdekat dari dokumen uji tersebut.

Pada tugas akhir ini digunakan *k-NN Feature Projection dan Feature Projection* sebagai *classifier* untuk melakukan pengkategorisasian dokumen berita berbahasa Indonesia. Berdasarkan referensi 5, proses yang dilakukan pada kedua *classifier* ini yaitu menormalisasikan setiap data latih sebagai sebuah unit *vector*. Dokumen *vector* ini ditransformasikan ke dalam *feature projections* dan disimpan pada masing-masing dimensi *feature*-nya. Hasil dari *feature projection* pada sebuah *feature* dapat dilihat sebagai sekumpulan bobot dari dokumen untuk *feature* tersebut. Pada *k-NN Feature Projection* klasifikasi awal pada sebuah *feature* dilakukan berdasarkan pada sejumlah k tetangga terdekat dari data uji pada *feature* tersebut[2]. Sedangkan pada *Feature Projection* masing-masing *feature* dari data latih dilakukan *voting* awal untuk sebuah proses klasifikasi[5]. Klasifikasi akhir dari sebuah dokumen uji ditentukan oleh sebuah *majority voting* dari setiap *voting* awal yang dilakukan pada data latih untuk masing-masing *feature*.

Pada *k-NN Feature Projection* dan *Feature Projection*, setiap *feature* secara individual berkontribusi pada proses klasifikasi sehingga kedua *classifier* ini tidak terpengaruh oleh adanya *feature* yang tidak relevan. Selain itu pada *k-NNFP* perhitungan similarity hanya dilakukan pada masing-masing feature saja. Sedangkan pada *Feature Projection* pengklasifikasian dokumen uji hanya membutuhkan perhitungan yang sederhana yaitu berdasarkan pada hasil voting dari masing-masing *feature* untuk masing-masing kategori. Oleh karena itu kedua *classifier* ini mengatasi kekurangan dari *k-NN* dalam hal waktu klasifikasi dan

mampu mempertahankan poin kekuatan dari *k-NN* yaitu performansi yang tinggi[2].

## 1.2 Perumusan masalah

Pada tugas akhir ini masalah yang akan diselesaikan yaitu:

1. Bagaimana cara untuk mengatasi keberagaman nilai *document frequency* untuk masing-masing feature.
2. Bagaimana mengkategorisasikan dokumen berita berbahasa Indonesia dengan *Feature Projection.*
3. Bagaimana performansi dari penerapan *Feature Projection* pada pengkategorisasian dokumen berita berbahasa Indonesia.

Batasan masalah pada tugas akhir ini yaitu:

1. Dataset yang akan digunakan merupakan artikel berita berbahasa Indonesia yang bersifat *offline*.
2. Dataset yang digunakan adalah data yang telah berlabel.
3. Dokumen dikategorikan ke dalam *single label*.
4. Tidak menangani *preprocessing* data.
5. Klasifikasi dokumen berdasarkan pada kata tunggal bukan berdasarkan pada frasa.

## 1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah :

1. Mengimplementasikan *Feature Projection* dan *k-Nearest Neighbor (k-NNFP)* untuk mengklasifikasikan dokumen berita berbahasa Indonesia.
2. Mengevaluasi perbandingan performansi yang berupa nilai *precision, recall, f-measure* dan *micro average f-measure* serta waktu klasifikasi dari *Feature Projection* dan *k-Nearest Neighbor Feature Projection (k-NNFP)* serta dibandingkan dengan *k-Nearest Neighbour (k-NN)* .

## 1.4 Metodologi penyelesaian masalah

Metodologi yang digunakan untuk menyelesaikan masalah dalam Tugas Akhir ini adalah :

1. Studi literatur.
   Mencari, mengumpulkan informasi, memahami dan mempelajari konsep kategorisasi teks dengan menggunakan *Feature Projection* serta konsep *text mining* melalui literatur berupa makalah, buku, atau jurnal yang berhubungan dengan *text mining, text categorization* dan *Feature Projection*.
2. Pencarian dan pengumpulan data.
   Mengumpulkan data berupa dokumen berita berbahasa Indonesia yang dibutuhkan untuk keperluan proses implementasi dan pengujian metode yang digunakan.
3. Analisis kebutuhan dan perancangan aplikasi yang akan dibangun.

Tahapan ini dilakukan dengan menganalisa kebutuhan perangkat lunak dan merancang perangkat lunak untuk pengkategorisasian dokumen berita berbahasa Indonesia dengan *Feature Projection* dan *k-Nearest Neighbor (k-NNFP)*

4. Implementasi dan Pengujian
   Mengimplementasikan hasil analisis dan perancangan perangkat lunak serta melakukan pengujian terhadap dataset yang akan diklasifikasikan untuk mengukur performansi dari *classifier*.
5. Analisa hasil pengujian dan penarikan kesimpulan.
   Perangkat lunak yang dihasilkan dievaluasi berdasarkan data yang diperoleh dari hasil pengujian.
6. Penyusunan laporan tugas akhir.
   Pembuatan laporan tugas akhir yang mendokumentasikan tahap-tahap kegiatan dan hasil dalam tugas akhir ini.