# BAB I
# PENDAHULUAN

## 1.1 Latar Belakang

Berita merupakan salah satu sarana untuk mendapatkan informasi mengenai suatu hal. Berita dapat disajikan dalam bentuk tulis atau lisan. Berita yang disajikan dalam bentuk tulis biasa disajikan dalam bentuk media cetak. Perkembangan teknologi yang pesat membuat berita tulis tidak hanya disajikan pada media cetak tetapi juga pada media elektronik.

Berita yang disajikan dalam bentuk tulis pada media elektronik atau media cetak, biasanya dikelompokkan berdasarkan isinya seperti berita olahraga, ekonomi, sains, dan lain sebagainya. Isi dari berita berupa teks, sedangkan berita pada media elektronik teks tersebut disimpan dan dikelompokkan sesuai dengan kelompoknya. Teks yang disimpan berjumlah sangat besar sehingga untuk pengambilan informasi dari teks tersebut membutuhkan waktu proses yang lama dikarenakan teks merupakan data yang tidak berstruktur dan tidak seperti data numerik yang terstruktur yang dapat dengan mudah untuk diolah menjadi informasi. Untuk memudahkan pengambilan informasi dari teks tersebut, maka diperlukan suatu metode pengorganisasian yang baik pada data tidak berstruktur seperti teks ini. Proses pengorganisasian teks tersebut adalah *Text Mining*.

*Text Mining (TM)* atau dikenal sebagai *Knowledge discovery from text (KDT)* merupakan proses penggalian pola yang menarik dari *database* teks yang sangat besar untuk kepentingan penggalian pengetahuan. Salah satu kegunaan dari *text mining* adalah pengklasifikasian dan pengorganisasian dokumen berdasarkan isinya atau disebut *text categorization*. [BEN-01].

Penelitian mengenai *text categorization* telah dilakukan oleh peneliti sebelumnya untuk menghasilkan performa yang baik dalam pengklasifikasian teks. Penelitian *text categorization* yang dilakukan oleh Retnani Latifah pada tahun 2012 dengan pendekatan klasifikasi menggunakan algoritma *Neighbor Weighted K-Nearest Neighbor* menunjukkan bahwa nilai akurasi dari algoritma ini dapat bertambah jika nilai *Document Threshold* yang dijadikan sebagai acuan untuk

membuang *term* (kata) dinaikkan menjadi 4 atau 5 [RET-12]. Pembuangan *term* ini dilakukan untuk mengurangi kompleksitas dari proses klasifikasi teks. Pendekatan *clustering* juga dapat digunakan untuk melakukan *text categorization*. Penelitian yang dilakukan oleh Munzir Umran pada tahun 2009 menggunakan algoritma K-Means dan *Singular Value Decomposition* untuk mengelompokkan teks atau dokumen. Hasil dari penelitian tersebut setelah dilakukan beberapa pengujian adalah jumlah anggota dokumen dalam setiap *cluster*-nya berubah-ubah. Hal ini dipengaruhi oleh penentuan nilai centroid awal pada algoritma K-Means yang juga berbeda-beda.

Beberapa algoritma yang digunakan dalam *text categorization* ini diantaranya adalah algoritma *K-Nearest Neighbor*, *Naïve Bayes Classifier*, dan ID3. Menuurut Yiming Yang, algoritma *K-Nearest Neighbor (KNN)* memiliki performa yang lebih baik dibandingkan dengan algoritma *decision tree* C4.5 dan algoritma *Rocchio*. Kelebihan dari algoritma KNN ini mampu untuk menangani klasifikasi teks dimana data yang digunakan berdimensi besar [YIM-02]. Implementasi dari algoritma-algoritma yang digunakan untuk *text categorization* ini adalah berdasarkan pada kemuculan kata atau morfologi kata seperti penelitian yang dilakukan sebelumnya oleh Retnani Latifah. Konsep dari klasifikasi teks adalah memasukkan teks baru yang belum diketahui kategorinya ke dalam kategori dengan melakukan pelatihan terhadap sekumpulan teks yang telah diketahui kategorinya. Proses pelatihan tersebut adalah menentukan kemiripan antara teks uji dengan setiap teks latih. Teks uji dan teks latih dikatakan mirip bila ada sekumpulan *term* yang muncul pada kedua dokumen tersebut. *Term* yang muncul tersebut adalah yang memiliki huruf penyusun yang sama. Semakin banyak term yang sama maka semakin mirip pula kedua teks tersebut. Proses penentuan kemiripan teks ini memiliki kelemahan karena apabila terdapat teks uji yang memiliki *term* yang berbeda secara morfologi dari *term* pada teks latih padahal kedua *term* tersebut memiliki makna yang sama maka kedua teks tersebut tidak dapat dikatakan mirip. Hal ini memungkinkan teks uji tersebut akan dikelompokkan ke dalam kelompok yang berbeda dari kelopok teks latih tersebut.

Untuk mengatasi masalah tersebut, algoritma yang digunakan harus mempertimbangkan kesamaan makna dari *term* tersebut yang merujuk pada konsep yang sama [ROS-04]. Untuk menentukan kesamaan makna antar *term*, teks dapat ditambahkan *knowledge background*. Salah satu *knowledge background* yang dapat digunakan adalah *lexical database WordNet* yang menghubungkan kata secara konseptual dan semantik [HOT-03].

Menurut [HOT-03], penambahan *knowledge background* pada teks yaitu dengan menggunakan *WordNet* dapat meningkatkan kinerja dari *text categorization*. Berdasarkan latar belakang yang telah dipaparkan, maka judul yang diambil dalam skripsi ini adalah **"Klasifikasi Berita Berbahasa Inggris Menggunakan Algoritma K-Nearest Neighbor Berbasis Ontologi".** Pemberian kategori pada teks berita adalah dengan melakukan pelatihan kepada setiap teks latih yang telah memiliki kategori dengan menggunakan algoritma KNN. Sebelum melalui proses pelatihan, teks akan melalui proses *preprocessing* yaitu proses perubahan representasi teks menjadi vektor teks. Proses *preprocessing* diantaranya adalah *case folding , tokenizing, stemming, ontology extraction, dictionary construction* dan pembobotan *term*. Setelah melalui proses *preprocessing,* proses selanjutnya adalah klasifikasi yaitu proses pemberian kategori pada teks berita uji. Dengan penambahan *knowledge background* pada teks diharapkan akurasi dari klasifikasi teks dapat meningkat.

## 1.2 Rumusan Masalah

Rumusan masalah dalam skripsi ini adalah :

1. Bagaimana implementasi dari algoritma *K-Nearest Neighbor* (KNN) berbasis ontologi dalam mengklasifikasikan berita berbahasa Inggris.
2. Berapa nilai evaluasi dari algoritma *K-Nearest Neighbor* berbasis ontologi dalam pengklasifikasian berita berbahasa inggris.

## 1.3 Batasan Masalah

Dari permasalahan yang yang dirumuskan di atas, maka batasan permasalahan yang digunakan untuk implementasi dari algoritma ini adalah:

1. Dokumen berita yang akan dipakai pada penelitian ini berupa dokumen dalam format **.txt*.
2. Berita yang digunakan adalah berita berbahasa Inggris dengan jumlah *corpus* yang digunakan seimbang untuk setiap kelas.
3. Berita terdiri dari 4 kategori yaitu *interest*, *money-fx*, *trade*, dan *crude*.
4. Proses klasifikasi dokumen hanya mengolah kata secara tunggal.
5. Setiap dokumen hanya berada pada satu kategori.
6. Hubungan semantik yang digunakan pada database *WordNet* adalah sinonim kata benda.
7. Evaluasi dari kinerja algoritma dihitung dengan menggunakan *precision*, *recall*,dan *f1-measure*.

## 1.4 Tujuan

Tujuan yang ingin dicapai dari penelitian dan penulisan skripsi ini adalah:

1. Mengimplementasikan algoritma *K-Nearest Neighbor* berbasis ontologi dalam pengklasifikasian berita berbahasa Inggris
2. Mengevaluasi kinerja dari algoritma tersebut berdasarkan nilai dari *precision, recall* dan *F-Measure*.

## 1.5 Manfaat

Adapun manfaat dari penelitian ini adalah berita teks yang merupakan data yang tidak berstruktur dapat diorganisasikan dengan baik sehingga diharapkan waktu proses pengambilan informasi dari teks dapat lebih efisien dan lebih terstruktur.

## 1.6 Sistematika Penulisan

Pembuatan tugas akhir ini dilakukan dengan sistematika penulisan sebagai berikut:

1. BAB I PENDAHULUAN

   Bab ini berisi pembahasan topik yang akan diangkat pada penelitian ini serta masalah yang dihadapi terkait penelitan ini. Bab I terdiri dari poin-poin sebagai berikut:

1.1 Latar Belakang
1.2 Rumusan masalah
1.3 Batasan Masalah
1.4 Tujuan
1.5 Manfaat
1.6 Sistematika Penulisan.

2. BAB II KAJIAN PUSTAKA

   Bab ini Berisi teori serta kajian pustaka tentang berita, *text mining*, metode *K-Nearest Neighbor*, ontologi dan teori-teori yang berhubungan dengan metode pengklasifikasian. Bab II terdiri dari poin-poin sebagai berikut:

   2.1 Klasifikasi Dokumen
   2.2 Berita
   2.3 *Data Mining*
   2.4 *Text Mining*
   2.5 Tahap *Preprocessing Text*
   2.6 *Classifier construction*
   2.7 Evaluasi

3. BAB III METODOLOGI PENELITIAN

   Bab ini berisi metode - metode yang digunakan dalam implementasi algoritma *K-Nearest-Neighbor* berbasis ontologi pada pengklasifikasian berita berbahasa Inggris. Bab III terdiri dari poin-poin sebagai berikut:

   3.1 Deskripsi Data
   3.2 Perancangan Sistem
   3.3 Perancangan Algoritma
   3.4 Perhitungan Manual
   3.5 Rancangan antar muka
   3.6 Rancangan Uji Coba

4. BAB IV HASIL DAN PEMBAHASAN

   Bab ini berisi tentang perancangan implementasi algoritma dan terdiri dari poin-poin sebagai berikut:

   4.1 Lingkungan Implementasi
   4.2 Implementasi Program

4.3 Implementasi Antar Muka

5. BAB V PENGUJIAN DAN ANALISIS

   Bab ini berisi pengujian dari algoritma yang telah diterapkan serta analisis hasil yang dihasilkan oleh algoritma. Bab ini berisi poin-poin sebagai berikut:

   5.1 Skenario Pengujian

   5.2 Hasil Pengujian

   5.3 Analisa Pengujian

6. BAB V PENUTUP

   Berisi Kesimpulan dari hasil penelitian serta saran untuk pengembangan penelitian. Bab ini berisi poin-poin sebagai berikut :

   6.1 Kesimpulan

   6.2 Saran