BAB I

PENDAHULUAN

1.1 Latar Belakang

Meningkatnya perkembangan informasi pada dokumen online dari waktu ke waktu ini menyebabkan meningkatnya pula informasi yang berbentuk teks. Hal ini menimbulkan kesulitan bagi pembaca dalam mencerna sebuah informasi dimana bentuk informasi tersebut dalam format tidak terstruktur. Ketidakteraturan struktur berbentuk teks seperti ini memunculkan penganalisisan teks menjadi lebih terstruktur dengan *text mining. Text mining* mencoba untuk mengekstrasi pola berupa informasi dan pengetahuan yang berguna dari sejumlah besar sumber data melalui identifikasi dan eksplorasi dari suatu pola menarik dengan tujuan mendapatkan informasi yang berguna dari sekumpulan dokumen berbentuk teks yang memiliki format tidak terstruktur.

Dalam mengelola informasi dari sekumpulan dokumen dengan jumlah yang besar merupakan sebuah kesulitan untuk mengidentifikasi kata yang ada pada dokumen tersebut menurut masing masing kategori dari dokumen tersebut diperlukan suatu metode. Metode yang dapat mengorganisir dokumen teks secara otomatis diantaranya adalah klasifikasi. Adapun teknik yang banyak digunakan dalam klasifikasi dokumen diantaranya *Naive Bayes Classifier*. *Naive Bayes Classifier* merupakan salah satu metode *machine learning* yang menggunakan perhitungan probabilitas. Konsep dasar yang digunakan oleh *Naive Bayes Classifier* adalah *Teorema Bayes*, yaitu *teorema* yang digunakan dalam statistika untuk menghitung suatu peluang, *Bayes Optimal Classifier* menghitung peluang dari satu kelas dari masing-masing kelompok atribut yang ada, dan menentukan kelas mana yang paling optimal.

Dalam klasifikasi teks menggunakan *Naive B*ayes *classifier* ini ada salah satu model yang dapat membantu kita mengelompokkan dokumen yaitu *Bernoulli NB. Bernoulli* merupakan fungsi *Naive Bayes Classifier* yang

menggunakan unsur biner mengambil nilai 1 bila kata yang sesuai terdapat dalam dokumen dan 0 bila kata tersebut tidak ada.

Beberapa penelitian yang berkaitan dengan klasifikasi dokumen diantaranya klasifikasi otomatis dokumen berita kejadian berbahasa Indonesia menghasilkan keakurasian 86% (Musthafa, 2009), klasifikasi teks dengan *naive bayes classifier* (NBC) untuk pengelompokan teks berita dan akademis menghasilkan akurasi yang lebih tinggi pada teks berita dibandingkan dengan dokumen akademik (Hamzah, 2012), klasifikasi dokumen teks menggunakan metode *support vector machine* dengan pemilihan fitur *chi-square* menghasilkan akurasi 99,69% (Putri, 2013), Klasifikasi dokumen *naive bayes classifier* untuk mengetahui konten *E-goverment* menghasilkan akurasi 85% (Wijaya, 2015).

Berdasarkan penelitian yang ada tersebut, penulis mencoba melakukan penelitian bagaimana mengklasifikasi dokumen yang dilakukan dengan menggunakan berbagai metode pengklasifikasian tetapi pada penelitian kali ini menggunakan dokumen dengan berbagai kategori dan algoritma *naive bayes bernoulli* yang merupakan model penyederhanaan dari metode *naive bayes* yang cocok dalam pengklasifikasian teks atau dokumen. Algoritma ini diharapkan dapat menghasilkan akurasi yang tinggi melihat dari beberapa penelitian diatas penggunakan metode *naive bayes* yang memiliki akurasi yang tinggi. Dari uraian tersebut, penulis melalukan penelitian tugas akhir yang berjudul "Klasifikasi Dokumen Berkategori Menggunakan Algoritma *Naive Bayes* Berbasis *Bernoulli*" penulis ingin melakukan analisis akurasi algoritma *naive bayes* yang berbasis *bernoulli* untuk pengklasifikasian teks pada dokumen berdasarkan masing masing kategori. Objek penelitian adalah suatu teks dokumen dalam bentuk digital atau media massa dalam bentuk eletronik dengan berbagai kategori, diantaranya ekonomi, kesehatan, hiburan, dan teknologi.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang diuraikan sebelumnya, terdapat beberapa permasalahan yang akan diangkat dalam penelitian ini antara lain:

1. Bagaimana mengklasifikasikan dokumen berdasarkan masing-masing kategori menggunakan metode *Naive Bayes Classifier* berbasis *Bernoulli*
2. Bagaimana akurasi dari klasifikasi dokumen yang memiliki beberapa kategori dengan menggunakan N*aive Bayes Classifier* berbasis *Bernoulli*

1.3 Batasan Masalah

Adapun batasan masalah dari penelitian ini adalah sebagai berikut :

1. Data yang digunakan adalah dokumen online yang diambil dari situs berita.
2. Diklasifikasikan menjadi 4 kelas yaitu Ekonomi, Kesehatan, hiburan dan teknologi
3. Teknik klasifikasi data yang digunakan adalah Algoritma *Bernoulli Naive Bayes*.
4. Aplikasi yang dibuat adalah aplikasi dengan menggunakan bahasa PHP (PHP : *Hypertext Prepocessor*)

1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah:

1. Mengklasifikasi dokumen berkategori dengan menggunakan algoritma *Naive Bayes Classifier* berbasis *bernoulli.*
2. Pengimplementasian sistem untuk menganalisis keakurasian algoritma *Naive Bayes Classifier* berbasis *Bernoulli*.

1.5 Manfaat Penelitian

Adapun manfaat dari penelitian ini sebagai berikut :

1. Memberikan tambahan wawasan keilmuan serta memperdalam konsep dan teori teknik pengklasifikasi data khususnya *Bernoulli Naive bayes*.
2. Memahami penerapan *text mining* dalam pengklasifikasian dokumen.
3. Dapat mengklasifikasikan teks dokumen berdasarkan masing-masing kategori.