BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi yang semakin pesat, hal ini tentunya sangat berpengaruh pada perkembangan internet. Data menunjukkan lebih dari tiga miliar masyarakat seluruh dunia sudah menggunakan internet dan bertambah 5% untuk setiap tahunnya. Hal ini juga berpengaruh terhadap meningkatnya kebutuhan masyarakat umum akan suatu informasi yang cepat dan mudah diakses sehingga juga berdampak pada meningkatnya pembaca dokumen berita secara *online* melalui internet.

Kebutuhan masyarakat umum akan suatu informasi yang cepat tidak lepas dari pengaruh gaya hidup masyarakat di era *modern* yang lebih cenderung membaca berita secara *online* daripada media cetak. Sehingga penumpukan data dan aliran informasi seperti dokumen teks meningkat setiap harinya. Selain itu juga berdampak pada sulitnya memperoleh informasi berita yang diinginkan dengan cepat. Oleh sebab itu diperlukan solusi manajemen data yang dikombinasikan dengan teknik pengelompokan dokumen yang dapat mengelompokan dokumen berita berdasarkan kesamaan topik bahasannya. Metode pengelompokan dokumen tersebut dinamakan klasifikasi dokumen.

Klasifikasi dokumen berita merupakan proses mengelompokan dokumen berita ke dalam kategori-kategori berita yang telah ditentukan. Frekuensi kata yang muncul di dalam dokumen tersebut menjadi acuan dasar yang akan digunakan untuk menentukan kategori berita yang sesuai. Salah satu algoritma klasifikasi yang populer adalah Naïve Bayes atau lebih dikenal *Naïve Bayes Classifer* (NBC) yang

dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class* (Andini, 2013; Natalius, 2011; Hamzah, 2012; Setiawan, et al., 2015). *Naïve Bayes Classifier* memiliki kemampuan serupa dengan *decision tree* dan *neural network*. Metode ini memiliki kelebihan pada tingkat akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam data yang besar (Kusrini dan Emha Taufiq Luthfi, 2009). Dengan klasifikasi dokumen, dokumen dikelompokokkan sesuai dengan kategorinya sehingga proses manajemen data dan proses pengklasifikasian dokumen berita dapat dilakukan lebih cepat.

Penelitian ini bertujuan melakukan klasifikasi dokumen berita yang diambil secara *online* melalui situs penyedia berita *online* menggunakan metode Naïve Bayes. Berdasarkan pemaparan yang telah diuraikan sebelumnya, penulis tertarik melakukan penelitian yang berjudul "*Text Classfication with Naïve Bayes*".

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan sebelumnya, penulis merumuskan permasalahan pada penelitian ini adalah bagaimana mengimplementasikan dan mengevaluasi metode *Naïve Bayes* dalam melakukan klasifikasi dokumen teks berita?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini adalah:

1. Penelitian hanya difokuskan pada dataset dokumen teks berita berbahasa Indonesia.

2. Penelitian hanya untuk melakukan klasifikasi dokumen dalam empat kelas yang telah ditentukan (olahraga, *lifestyle*, ekonomi dan politik).
3. Dataset yang digunakan di dalam penelitian ini hanya data yang diambil dari situs berita Indonesia yaitu www.detik.com, www.kompas.com dan www.okezone.com.
4. Pada penelitian ini tahap *text preprocessin*g tidak dilakukan proses *stemming*.

1.4 Tujuan Masalah

Tujuan dari penelitian ini adalah melakukan klasifikasi dokumen berita berbahasa indonesia ke dalam kelas atau kategori yang telah ditentukan menggunakan metode *Naïve Bayes Classifier*.

1.5 Manfaat Penelitian

Adapun manfaat yang diharapkan dari penelitian ini adalah:

1. Penelitian ini diharapkan dapat membantu *user* dalam mengelompokkan dokumen berita berdasarkan kelas yang telah ditentukan.
2. Hasil penelitian bermanfaat bagi Perguruan Tinggi Teknokrat sebagai informasi tentang penerapan metode *Naïve Bayes Classifier* dalam klasifikasi dokumen.

1.6 Keaslian Penelitian

Penelitian mengenai klasifikasi teks telah banyak dilakukan di Indonesia, antara lain oleh Natalius (2011) yang berjudul Metode *Naïve Bayes Classifier* dan Penggunaannya pada Klasifikasi Dokumen menemukan bahwa untuk mengetahui

suatu email merupakan suatu kelas spam atau non spam sangat sulit, apalagi jika sebuah email mengandung gambar hal ini semakin memperbesar probabilitas email tersebut masuk ke dalam kelas yang bukan sebenarnya hal ini dikarenakan *Naïve Bayes Classifier* dirancang untuk menganalisis kata-kata bukan gambar. Lalu Hamzah (2012) penelitiannya yang berjudul Klasifikasi Teks dengan *Naïve Bayes Classifier* (NBC) untuk Pengelompokan Teks Berita dan Abstrack Akademis mengemukakan akurasi suatu dokumen sebagai data uji sangat dipengaruhi oleh pemilihan fitur di dalam suatu koleksi dokumen latih. Filter kata akan semakin tinggi akurasi kemungkinan suatu dokumen masuk ke dalam kelas yang sebenarnya jika data latih menggunakan filter terhadap kata unik dan kinerja klasifikasi kurang optimal saat data latih tanpa menggunakan filter. Hal ini dikarenakan penggunaan filter kata menjadikan kata yang dijadikan fitur kategori lebih layak masuk ke dalam *vocabulary*. Kemudian penelitian Andini (2013) yang berjudul Klasifikasi Dokumen Teks Menggunakan Algoritma Naïve Bayes dengan Bahasa Pemrograman Java mengungkapkan bahwa klasifikasi membutuhkan data awal sebagai petunjuk akan digunakan untuk mendukung keputusan dalam klasifikasi dan data petunjuk akan membantu dalam mengetahui dokumen uji. Selanjutnya Samuel (2014) yang berjudul Metode K-Nearest Neighbor dengan *Decision Rule* untuk Klasifikasi Subtopik Berita, peneliti mengungkapkan bahwa dalam klasifikasi dokumen dapat ditinjau dari prosesnya yang mengambil aksi berdasarkan data-data yang telah ada sebelumnya. Lalu penelitian yang dilakukan oleh Setiawan, et al. (2015) yang berjudul Klasifikasi dan Pencarian Buku Refrensi Akademik Menggunakan Metode *Naïve Bayes Classifier* (NBC) (Studi Kasus: Perpustakaan Daerah Provinsi Kalimantan Timur) yang menggunakan Metode

Naïve Bayes Classifier dalam membangun sistem informasi berbasis pencarian buku refrensi.

Melihat penelitian-penelitian terdahulu seperti yang sudah dikemukakan penelitian ini akan mencoba menerapkan bagaimana metode *Naïve Bayes* dalam melakukan klasifikasi dokumen teks berita secara otomatis, sehingga peneliti menjamin keaslian penelitian ini dan dapat dipertanggung jawabkan.