BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan situs berita online di internet telah terjadi dengan sangat cepat dan menghasilkan teks-teks berita dalam jumlah yang sangat besar, struktur penulisan berita pada masing-masing situs tersebut memiliki perbedaan satu dengan yang lainnya (di luar kasus plagiat). Ketika mengumpulkan berita-berita dari berbagai situs berita online yang berbeda, misalkan *kompas.com, detik.com, tempo.com* dan lain sebagainya, tentunya harus dikelompokan pada kategori beritanya masing-masing seperti kriminal, ekonomi, bencana alam atau kategori lainnya. Proses memperoleh data berita dari berbagai situs online dapat dilakukan dengan teknik perayapan web atau *web crawling.* Proses dimulai dengan menentukan URL dari situs web yang dipilih, kemudian perayap secara otomatis akan menemukan semua tautan yang ditemukan di laman HTML dan akan mengaturnya, dari URL yang diatur, perayap kemudian akan mengunduh konten laman(Sharma, 2014). Data teks berita yang telah didapatkan, tentunya dalam jumlah yang sangat besar dan tidak dapat dilakukan pengklasifikasian secara manual karena akan sulit dan perlu waktu yang cukup lama. Sebuah metode atau algoritma dengan kemampuan untuk mengklasifikasikan teks ke dalam kategori tertentu sangat membantu dalam menghadapi tantangan tersebut. Klasifikasi teks otomatis adalah untuk menetapkan label atau kelas ke teks yang diberikan menggunakan program komputer(Shahi, 2018). Data teks berita yang didapat berupa data teks dan dapat dilakukan klasifikasi otomatis dengan algoritma tertentu pada suatu program komputer.

Terdapat langkah yang harus dilakukan sebelum dilakukannya proses klasifikasi pada sebuah teks berita, yakni tahap preproses teks atau *text preprocessing*. Langkah awal dimulai dengan *lemmatization* atau *stemming*, yang bertujuan untuk mendapatkan bentuk dasar dari setiap kata. Kemudian dilanjutkan dengan penghapusan *stopword* yang berfungsi juga diterapkan untuk menghapus kata-kata yang kurang bermakna, seperti preposisi dan konjungsi (misal: apa, yang, dimana, kompilasi, di, saat dan lain-lain) dan langkah terakhir adalah untuk menentukan pemilihan fitur dan penggolong yang akan menjadi bekas(Wongso *et al.*, 2017). Selanjutnya adalah tahap *stemming* yang dapat menggunakan algoritma Porter dan algoritma Nazief-Adriani, dimana masing-masing algoritma memiliki rata-rata presisi 70,6% dan 84%(Agusta, 2009). Dari segi nilai presisinya, dapat disimpulkan bahwa algoritma Nazief-Adriani miliki kinerja lebih baik daripada algoritma Porter. Setelah selesai melakukan preproses teks, kemudian dilakukan pengklasifikasian data berita dengan algoritma Naive-Bayes sehingga dapat proses klasifikasi dapat dilakukan dengan cepat tanpa dibaca satu per satu oleh manusia.

Algoritma Naive Bayes adalah klasifikasi statistik yang melakukan prediksi sebuah kelas, selain itu algoritma ini juga memiliki tingkat akurasi yang cukup baik(Setiawan and Nursantika, 2017). Algoritma Naive Bayes dipilih dikarenakan pada penelitian sebelumnya(Wongso *et al.*, 2017), ditemukan bahwa algoritma Naive Bayes memiliki akurasi 98,4% dalam melakukan klasifikasi teks berita berbahasa Indonesia. Selain algortima tersebut, terdapat penelitian lain yang membandingkan algortima Naive Bayes dan *K-Nearest Neighbor* (KNN)(Devita, Herwanto and Wibawa, 2018). Pada penelitian terebut diperoleh akurasi untuk algoritma Naive Bayes adalah sebesar 70% sedangkan untuk algoritma KNN hanya

mendapatkan nilai akurasi cukup rendah yakni hanya sebesar 40% saja. Dengan melihat perbandingan tersebut, maka digunakan algoritma Naive Bayes untuk dicoba ditingkatkan lagi nilai akurasinya dengan beberapa metode tambahan. Data sumber beirta tidak didapatkan dari penelitian-penelitian sebelumnya, melainkan memggunakan data baru yang diambil dari situs berita online kompas.com.

Pada teks berita terdapat kata-kata berbahasa Inggris yang mana kata-kata tersebut sebenarnya sama dengan arti dalam bahasa Indonesia, misalnya kata "reading" dalam Inggris sama artinya dengan kata "membaca" atau "baca" dalam bahasa Indonesia. Sehingga pada tahap preproses teks klasifikasi berita menggunakan algoritma Naive Bayes dapat ditambahkan proses terjemahan bahasa Inggris ke dalam bahasa Indonesia dalam bentuk kata dasarnya.

Naskah berita dibuat atau diketik oleh manusia, masih ada kemungkinan untuk terjadi salah ketik kata dan perlu diperbaiki. Kesalahan penulisan kata dapat terjadi juga pada saat selesai dilakukannya proses *stemming*, karena algoritma *stemming* memiliki tingkat akurasi tidak sepenuhnya benar maka terdapat kata yang terbentuk tidak memiliki arti. Levenshtein Distance (*edit distance*) merupakan metode metode yang digunakan untuk mengatasi kesalahan ejaan (Antinasari, Perdana and Fauzi, 2017). Algoritma *edit distance* memiliki kompleksitas O(m.n) yang lebih baik daripada algoritma *exhaustive search* $O(3^{m+n})$ maupun *greedy*(Sempena, 2009). Diharapkan dengan menggunakan algoritma tesebut dapat memperbaiki kata-kata yang mengalami salah ejaan agar sesuai dengan kata dasar yang baku.

1.2 Rumusan Masalah

Rumusan masalah pada penelitian yang dilakukan adalah

1. Bagaimana perbandingan hasil akurasi klasifikasi teks menggunakan algoritma Naive Bayes biasa dengan klasifikasi teks algoritma Naive Bayes yang telah dimodifikasi
2. Bagaimana perbandingan hasil kecepatan proses klasifikasi teks menggunakan algoritma Naive Bayes biasa dengan klasifikasi teks algoritma Naive Bayes yang telah dimodifikasi

1.3 Batasan Masalah

Berdasarkan rumusan masalah yang telah diuraikan sebelumnya, maka dilakukan pembatasan terhadap ruang lingkup permasalahan. Berikut ini merupakan batasan dari pada penelitian

1. Terjemhan kata hanya dilakukan dari bahasa Inggris ke bahasa Indonesia
2. Teks berita yang digunakan berbahasa Indonesia dan bersumber dari situs berita online, yakni kompas.com
3. Kategori berita digunakan berjumlah 7 kategori, yakni kategori Fenomena, Film, Internet, Keuangan, Musik, Olahraga dan Otomotif

1.4 Tujuan Penelitian

Adapun tujuan yang terdapat dalam penelitian ini adalah sebagai berikut,

1. Untuk mengetahui akurasi terbaik antara algoritma Naive Bayes dan Naive Bayes yang melibatkan Levenshtein Distance dan terjemahan kata bahasa Inggris ke dalam bahasa Indonesia

2. Untuk kecepatan proses saat dilakukannya klasifikasi teks berita antara algoritma Naive Bayes dan Naive Bayes yang melibatkan Levenshtein Distance dan terjemahan kata bahasa Inggris ke dalam bahasa Indonesia

1.5 Manfaat Hasil Penelitian

Manfaat yang dapat diperoleh adalah ditemukannya langkah yang lebih optimal secara akurat dan kecepatan proses dalam melakukan klasifikasi teks berita berbahasa Indonesia.

Menemukan algoritma yang memiliki kinerja yang baik dalam melakukan klasifikasi teks berita dari berbagai sumber berita online secara otomatis.

UNIVERSITAS PENDIDIKAN GANESHA
UNDIKSHA