BAB 1
PENDAHULUAN

1.1 Latar Belakang

Klasifikasi dokumen berbasis teks menjadi suatu bidang yang muncul dalam ranah penelitian *Text Mining*. Sebagai konsekuensi dari itu, telah terdapat banyak pendekatan yang dikembangkan untuk menerapkan hal tersebut, terutama dengan dikembangkannya berbagai macam *classifier* berbasis pembelajaran mesin seperti *Support Vector Machine* [1], *K-Nearest Neighbour* [2], Naive-Bayes [3], *Decision Tree* [4], *Neural Network* [5] dan *Maximum Entropy* [6]. Dalam mempertimbangkan pendekatan-pendekatan yang menggunakan algoritma berbasis pembelajaran mesin ini, varian Naïve-Bayes berupa *Multinomial Naive-Bayes* telah secara luas diterapkan karena kesederhanaan penerapannya di tahap pelatihan maupun di tahap pengujian yang telah dibuktikan akan keefektifannya dalam mengklasifikasi dokumen-dokumen berbasis teks yang tidak terstruktur di berberapa kasus penerapan yang berbeda-beda [3].

Beberapa penelitian juga telah menekankan masalah dari atribut-atribut yang redundan [7] terutama masalah *Curse of Dimensionality* yang mengakibatkan peningkatan kompleksitas komputasi dan kecenderungan untuk menghasilkan *error* (kesalahan) klasifikasi pada penerapan klasifikasi yang melibatkan dimensi fitur tinggi [8]. Hal ini secara umum disebabkan oleh jumlah fitur yang digunakan untuk menerapkan metode pembelajaran mesin untuk klasifikasi jauh melebihi baris data yang ada dalam dataset pelatihan awal sehingga terjadinya masalah *overfitting* [9], sebagaimana telah ditunjukkan dalam penelitian pada kasus penerapan klasifikasi berita bahasa Indonesia oleh Fanny dkk. (2018) yang hanya menghasilkan akurasi sebesar 53.1 % [10], dimana *classifier Multinomial Naive-Bayes* diterapkan tanpa implementasi seleksi fitur.

Seleksi fitur yang sering dianggap sebagai masalah pencarian dalam ruang sub himpunan fitur-fitur yang membutuhkan kondisi awal, strategi untuk me-*traverse* ruang sub himpunan fitur-fitur, fungsi evaluasi dan kondisi berhenti [11]

yang memiliki metode-metode beragam namun secara garis besar dapat dibagi menjadi dua metode utama yaitu *filter* dan *wrapper* [9]. Penerapan seleksi fitur Statistika *Chi-Squared* dengan *classifier Multinomial Naive-Bayes* dalam kasus penerapan klasifikasi berita seperti penelitian yang dilakukan oleh Mowafy dkk. (2018) [12] dan secara khusus untuk kasus penerapan berita bahasa Indonesia oleh Rahmad & Pribadi (2015) [13] merupakan contoh penerapan seleksi fitur dengan metode *filter*. Sedangkan penerapan seleksi fitur menggunakan metode *feature clustering* (klusterisasi fitur) dengan *classifier Multinomial Naive-Bayes* dalam penelitian Ismi dkk. (2016) [14] merupakan contoh penerapan dilakukan seleksi fitur dengan metode *wrapper*.

Sarkar dkk. (2014) [15] mengusulkan penerapan seleksi fitur dengan pendekatan baru bernama FS-CHICLUST yang melibatkan tahap *filter* dengan implementasi Statistika *Chi-Square* dan tahap *wrapper* yang menerapkan *feature clustering* dalam rangka memperbaiki performa akurasi dari *classifier Multinomial Naïve-Bayes* dengan cara mereduksi kompleksitas ruang fitur yang tinggi. Penelitian ini mengimplementasikan FS-CHICLUST sebagai seleksi fitur dan *classifier* Naïve-Bayes Multinomial yang dilibatkan dalam serangkaian proses dengan tujuan mengklasifikasi teks topik berita bahasa Indonesia sebagai upaya untuk mereduksi fitur-fitur redundan dan menguji pengaruhnya terhadap performansi akurasi klasifikasinya. Pengelompokan topik berita yang dijadikan acuan dalam penelitian ini mengacu pada 5 kelompok topik berita yaitu Ekonomi, Kesehatan, Pendidikan, Politik, dan Teknologi.

1.2 Rumusan Masalah

Berdasarkan apa yang telah diuraikan sebelumnya pada bagian latar belakang, masalah yang dapat dirumuskan adalah bagaimana akurasi dari metode Naive-Bayes Multinomial dalam kasus klasifikasi teks topik berita dapat ditingkatkan menggunakan seleksi fitur FS-CHICLUST

1.3 Maksud dan Tujuan

Maksud dari penelitian ini adalah untuk menerapkan metode Naïve-Bayes Multinomial dengan Seleksi Fitur FS-CHICLUST pada kasus penerapan klasifikasi teks berdasarkan topik berita.

Tujuan dari penelitian ini adalah untuk mengukur akurasi Naïve-Bayes Multinomial yang menggunakan seleksi fitur FS-CHICLUST dalam klasifikasi teks.

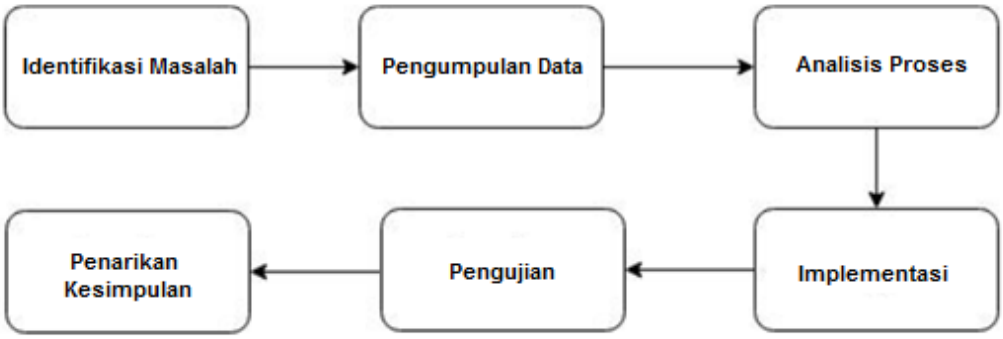
1.4 Batasan Masalah

Adapun batasan masalah yang dapat dijelaskan adalah sebagai berikut:

1) Data yang digunakan merupakan artikel-artikel berita dalam bahasa Indonesia yang diperoleh dari penelitian Irfa (2018)
2) Data yang berupa artikel-artikel berita daring tersebut dikategorikan menjadi 5 topik yaitu Ekonomi, Kesehatan, Pendidikan, Politik dan Teknologi
3) Banyak data yang digunakan adalah sebanyak 200 data (40 untuk masing-masing topik berita)

1.5 Metodologi Penelitian

Tahapan yang dilangsungkan dalam penelitian ini dapat dirincikan sebagai berikut:



Gambar 3.1 Alur Metodologi Penelitian

1.5.1 Identifikasi Masalah

Langkah pertama dalam metodologi penelitian adalah untuk menentukan masalah apa yang ingin dikaji. Pada tahap identifikasi masalah dilakukan analisis mengenai masalah yang dihadapi pada penelitian yang dilakukan.

1.5.2 Pengumpulan Data

Pengumpulan data bertujuan untuk mengumpulkan data – data yang memiliki hubungan dengan penelitian yang dilakukan, seperti data artikel-artikel berita berbahasa Indonesia yang diambil dari portal dataset daring. Selain itu, data yang dikumpulkan juga perlu diseleksi dan dipersiapkan lebih lanjut agar sesuai dengan kebutuhan penelitian sebagaimana telah dirumuskan dalam batasan masalah sebelumnya. Dataset artikel-artikel berita diambil dari penelitian serupa oleh Irfa (2018) [16] yang menerapkan metode *K-Nearest Neighbour* untuk mengklasifikasi teks berita Bahasa Indonesia. data berupa referensi literatur terkait penelitian klasifikasi teks terutama yang menerapkan varian dari metode Naive-Bayes dan berbagai macam seleksi fitur yang dapat diterapkan dengan metode klasifikasi tersebut juga dikaji.

1.5.3 Analisis Proses

Analisis proses bertujuan agar memperoleh gambaran rinci terkait tahap-tahap klasifikasi topik berita yang dilangsungkan. Tahap analisis secara garis besar terbagi menjadi analisis masalah, analisis pengumpulan dan persiapan data, analisis tahap *preprocessing*, analisis tahap seleksi fitur FS-CHICLUST dan analisis penerapan metode Naive-Bayes Multinomial.

1.5.4 Implementasi

Implementasi adalah tahap dimana serangkaian tahapan yang telah dispesifikasi dalam analsis proses dilangsungkan untuk sebelum hasilnya akan dikaji pada tahap pengujian.

1.5.5 Pengujian

Pengujian merupakan tahap dimana akurasi dari hasil klasifikasi metode Naïve-Bayes dengan penerapan seleksi fitur FS-CHICLUST dalam mengklasifikasi teks dokumen terhadap topik berita yang sesuai diukur. Untuk mencapai hal ini, hasil klasifikasi diuji menggunakan *Confusion Matriks* dan parameter-parameter yang diperoleh darinya berupa *Precision, F1-Score* dan Akurasi

1.5.6 Penarikan Kesimpulan

Fase ini akan menyajikan dan menguraikan hasil penelitian dan nilai keakuratan dari seraingkaian tahap klasifikasi teks yang telah dilangsungkan beserta kesimpulan terkait hasil pengujiannya.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun untuk memberikan gambaran umum tentang penelitian yang dijalankan. Sistematika penulisan tugas akhir ini adalah sebagai berikut:

BAB 1 PENDAHULUAN

Bab ini berisi latar belakang, rumusan masalah, maksud dan tujuan, batasan masalah, metodologi penelitian dan sistematika penulisan.

BAB 2 LANDASAN TEORI

Bab ini berisi uraian tentang *text preprocessing*, *Document-Term Matrix*, Pembobotan TF-IDF, Statistik Chi-Squared (Statistik X^2), *K-Means Clustering*, Algoritma FS-CHICLUST, Naïve-Bayes Multinomial, *Confusion Matrix*, *Precision, Recall, F1-Score* dan Akurasi

BAB 3 ANALISIS DAN PERANCANGAN

Bab ini membahas mengenai analisis masalah, analisis metode yang digunakan, analisis penyelesaian masalah, analisis simulasi (penerapan metode Naïve-Bayes Multinomial menggunakan seleksi fitur FS-CHICLUST) klasifikasi dan analisis perancangan sistem.

BAB 4 IMPLEMENTASI DAN PENGUJIAN

Bab ini membahas tentang implementasi antarmuka yang telah dirancang dan pengujian akurasi metode Naïve-Bayes pada klasifikasi berita dengan menggunakan program yang telah dibangun.

BAB 5 KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan yang dihasilkan dari pembahasan penelitian dan beberapa saran sebagai hasil akhir dari penelitian yang telah dilakukan dan pengembangan lebih lanjut.


UNIVERSITAS KOMPUTER INDONESIA
UNIKOM