BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Analisis sentimen adalah bidang studi untuk menganalisis pendapat, sentimen, evaluasi, penilaian sikap dan emosi terhadap entitas seperti produk, jasa, organisasi, individu, peristiwa dan atribut lainnya [1]. Analisis sentimen berfokus kepada opini yang mengekspresikan sentimen positif atau negatif [1]. Terdapat tiga tahapan umum dalam proses analisis sentimen, yaitu pengambilan data dari sumber data atau yang biasa disebut *crawling*, selanjutnya tahap *training* dan *testing* dengan menggunakan algoritma klasifikasi tertentu. Sumber data yang dapat digunakan untuk proses analisis sentimen salah satunya adalah media sosial Twitter. Twitter sebagai salah satu jejaring sosial yang interaktif memungkinkan penggunanya untuk memberikan suatu pendapat terhadap suatu hal secara *real time*. Data yang dapat dianalisis di Twitter adalah *tweet* yang dibuat oleh pengguna, dimana *tweet-tweet* tersebut nantinya akan diklasifikasikan ke dalam dua kelas sentimen yaitu sentimen positif atau sentimen negatif.

Data *tweet* yang akan diproses, terlebih dahulu dilakukan tahap *preprocessing*. Tahap *preprocessing* diperlukan untuk membersihkan data dari elemen-elemen yang tidak diperlukan dan memproses data yang sebetulnya dibutuhkan dengan tujuan saat masuk pada tahap klasifikasi akan lebih optimal dalam penghitungannya. Pada Twitter terdapat banyak penggunaan kata tidak baku dengan variasi penulisan kata. Berdasarkan hasil pengamatan terhadap *tweet-tweet* kepada akun Twitter salah satu acara pencarian bakat di Indonesia, ditemukan beberapa penggunaan kata tidak baku dengan variasi penulisan kata seperti “kereeeen”, “kren”, “seruuuu”, ”kerenzz”, ”baguss”, ”keereennnn”, ”keceee”, dan sebagainya. Keberadaan variasi kata yang tidak baku yang sebetulnya merujuk kepada suatu kata baku dapat mempengaruhi akurasi hasil penghitungan dengan Naive Bayes [2], karena setiap kata yang ada pada *tweet* memiliki peran terhadap

proses klasifikasi. Jika dilihat kata-kata tersebut merepresentasikan sebuah kata baku, namun ditulis dengan berbagai variasi penulisan.

Oleh sebab itu maka diperlukan suatu proses untuk mengakomodir penggunaan kata tersebut, yaitu dengan menggunakan metode n-gram. N-gram adalah potongan n-karakter yang dibangkitkan dari suatu string [3]. Keuntungan N-gram adalah berdasarkan karakteristik n-gram sebagai bagian dari suatu string, sehingga kesalahan pada sebagian string hanya akan berakibat perbedaan pada sebagian n-gram [3]. Jika dilihat kata "kren" dan "kereeeen" merupakan bentuk tidak baku dengan variasi penulisan dari kata "keren". Ketika dihitung jumlah kesamaan n-gram karakter kata pada kata-kata tersebut akan terdapat kedekatan.

Berdasarkan uraian di atas, pada penelitian ini akan diterapkan metode n-gram untuk optimasi akurasi analisis sentimen pada Twitter. Metode n-gram digunakan untuk mengambil fitur pada *tweet* yang selanjutnya akan diklasifikasi untuk menentukan kategori kelas sentimen positif atau negatif.

1.2 Rumusan Masalah

Berdasarkan penjelasan dari latar belakang ditemukan permasalahan yang dirumuskan ke dalam suatu rumusan masalah yaitu bagaimana menerapkan metode n-gram untuk mengambil fitur pada *tweet* supaya bisa mengakomodir kata tidak baku dengan variasi penulisan kata yang terdapat pada suatu *tweet*.

1.3 Maksud dan Tujuan

Maksud dari penelitian yang akan dilakukan adalah menerapkan metode n-gram untuk optimasi akurasi analisis sentimen pada Twitter. Sedangkan tujuan yang ingin dicapai adalah mengetahui pengaruh penerapan metode n-gram terhadap tingkat persentase akurasi yang akan digunakan untuk analisis sentimen pada Twitter.

1.4 Batasan Masalah

Agar penelitian yang dilakukan lebih terarah dan mencapai sasaran yang ditentukan, maka diperlukan sebuah pembatasan masalah atau ruang lingkup kajian, yaitu sebagai berikut:

1. Data *tweet* diambil dari Twitter The Dance Icon Indonesia (@TheDanceIconIND) atau mengandung keyword "TheDanceIconIND".
2. Tahap *preprocessing* yang dilakukan yaitu *convert emoticon*, *cleansing*, *case folding*, *tokenizing*, *convert negation*, *remove stopword* dan *stemming*.
3. Proses *stemming* menggunakan algoritma porter bahasa Indonesia.
4. Klasifikasi *tweet* ke dalam kelas sentimen positif dan negatif.
5. Menggunakan metode *Naive Bayes* untuk klasifikasi *tweet*.
6. Menggunakan metode n-gram karakter untuk pengambilan fitur, dengan nilai n=2, 3, 4, 5, 6, 7, 8, 9, 10.
7. Analisis pembangunan perangkat lunak yang digunakan pada penelitian ini menggunakan pendekatan berorientasi proses (terstruktur).
8. Aplikasi yang dibangun berupa *scientific*.

1.5 Metodologi Penelitian

Metodologi penelitian yang akan dilakukan pada penelitian ini yaitu sebagai berikut:

1. Pengumpulan Data

Melakukan pengumpulan data dengan cara mengumpulkan referensi seperti jurnal, paper, dan bacaan-bacaan yang berkaitan dengan topik dari penelitian yang akan dilakukan. Pengumpulan data juga yaitu mengumpulkan data *tweet* dengan cara mengambil langsung dari Twitter dengan menggunakan fasilitas API Twitter.

2. Analisis dan Perancangan Model Aplikasi

Melakukan beberapa analisis dalam penelitian ini, yaitu analisis masalah yang diangkat, analisis sistem atau aplikasi yang akan dibangun, analisis tahap *preprocessing*, analisis klasifikasi dengan melakukan penghitungan manual dari metode yang digunakan, serta analisis kebutuhan fungsional dari aplikasi yang akan dibangun. Setelah itu, dibuat suatu perancangan model aplikasi yang sebelumnya telah dianalisis kebutuhan fungsionalnya, dimana aplikasi yang dibangun untuk mensimulasikan metode yang diterapkan.

3. Pembangunan Aplikasi

Aplikasi yang telah dirancang, selanjutnya diimplementasikan. Implementasi dilakukan bertahap, dilakukan per-modul dari mulai dari tahap *preprocessing*, tahap *training*, tahap *testing*, hingga ke tahap pengujian akurasi hasil *testing*. Setiap modulnya dilakukan pengujian apakah sudah bekerja dengan baik (tidak ada *error*).

4. Uji coba dan evaluasi

Melakukan uji coba dengan menggunakan aplikasi yang sudah dibuat. Uji coba yang dilakukan adalah dengan melakukan proses *testing* dengan jenis fitur yang berbeda (dalam hal ini penggunaan fitur kata, fitur dengan n-gram dengan nilai n yang berbeda) dengan menggunakan metode *k-fold cross validation*, melihat akurasi dari setiap percobaan, serta membuat kesimpulan terhadap hasil percobaan tersebut apakah metode n-gram yang digunakan dalam penelitian ini dapat menambah tingkat akurasi dari klasifikasi *tweet* atau tidak.

1.6 Sistematika Penulisan

Sistematika penulisan penelitian ini disusun untuk memberikan gambaran umum mengenai penelitian yang dikerjakan. Sistematika penulisan penelitian sebagai berikut:

BAB 1 PENDAHULUAN

Pada bab ini berisi penjelasan mengenai latar belakang permasalahan, rumusan masalah, maksud dan tujuan, batasan masalah, metodologi penelitian, serta sistematika penulisan dalam penelitian tentang optimasi akurasi analisis sentimen pada Twitter.

BAB 2 LANDASAN TEORI

Pada bab ini berisi teori-teori atau tinjauan dari literatur pendukung yang akan digunakan untuk optimasi akurasi analisis sentimen pada Twitter. Adapun teori-teori yang akan ditulis pada Bab ini yaitu teori analisis sentimen, preprocessing (*convert emoticon*, *cleansing*, *case folding*, *tokenizing*, *convert negation*, *remove stopword* dan *stemming*), metode n-gram, metode *naive bayes*, *confussion matrix*. Teori analisis perancangan aplikasi yang akan dibuat seperti diagram konteks, data flow diagram, spesifikasi proses dan kamus data, serta teori tentang *tools* yang digunakan dalam pembangunan perangkat lunak atau aplikasi.

BAB 3 ANALISIS DAN PERANCANGAN

Pada bab ini dijelaskan secara rinci mulai dari analisis masalah, analisis sistem, analisis data masukan, analisis tahap *preprocessing*, analisis klasifikasi dengan menggunakan algoritma naive bayes dengan fitur kata dan fitur dengan n-gram. Selain itu juga dijelaskan analisis kebutuhan fungsional dan non-fungsional serta perancangan menu, antarmuka, pesan semantik dan prosedural dari aplikasi yang akan digunakan untuk penerapan dan pengujian akurasi klasifikasi dari metode yang digunakan.

BAB 4 IMPLEMENTASI DAN PENGUJIAN SISTEM

Pada bab ini ditampilkan implementasi dari hasil analisis dan perancangan yang sudah dilakukan sebelumnya menjadi sebuah aplikasi. Aplikasi digunakan untuk melakukan pengujian akurasi analisis sentimen terhadap *tweet* yang diatur dalam sebuah skenario pengujian. Adapun skenario pengujian menggunakan metode k-*fold cross validation* yang dilakukan kepada masing-masing nilai n pada metode n-gram. Informasi yang ditampilkan sebagai hasil dari pengujian yaitu persentase akurasi dari masing-masing pengujian pada nilai n yang digunakan.

BAB 5 KESIMPULAN DAN SARAN

Pada bab ini berisi kesimpulan dan saran tentang penelitian yang dilakukan. Pada bagian kesimpulan, berisi tentang apakah penerapan metode n-gram berpengaruh terhadap tingkat persentase akurasi dalam klasifikasi untuk analisis sentimen *tweet* pada Twitter atau tidak. Sedangkan pada bagian saran, berisi tentang hal-hal yang belum bisa terselesaikan pada penelitian yang telah dilakukan serta hal-hal yang bisa dilakukan untuk pengembangan selanjutnya.