# BAB II

# LANDASAN TEORI

## *Text Mining*

### *Text mining* merupakan kegiatan eksplorasi dan analisis, dengan cara otomatis atau semi-otomatis pada data yang berupa teks dalam jumlah besar dan bersumber dari media sosial (Twitter, Facebook, Instagram, Youtube) atau sumber lain untuk menemukan pola yang bermakna (Tan & Steinbach, 2001). Menurut Fransiska Vina Sari dan Arief Wibowo dalam penelitianya (Vina & Wibowo, 2019), *text mining* merupakan proses penambangan teks melalui dokumen atau kalimat yang dilakukan melalui komputer yang bertujuan untuk mendapatkan sesuatu yang baru; yang belum diketahui sebelumnya; atau menemukan kembali informasi yang tersirat. Proses tersebut dilakukan dalam rangka penggalian, pengolahan, serta pengaturan pada informasi dengan menganalisa keterkaitan antara informasi satu dengan yang lainnya (Sudiantoro et al., 2018)

## Analisis Sentimen

Analisis sentimen merupakan sebuah metode yang digunakan untuk menentukan arah pandangan(sentimen) dan mengelompokkan polaritas teks dalam dokumen atau kalimat sehingga dapat ditentukan kategorinya sebagai sentimen positif, negatif, atau netral (Ramadhan & Setiawan, 2019) (Zuhdi dkk., 2019). Analisis sentimen juga dapat disamakan dengan *opinion mining* karena berfokus kepada pendapat yang mewakili pandangan positif(setuju) atau negatif(menolak) (Saputra et al., 2019). Saat ini, analisis sentimen banyak digunakan oleh peneliti sebagai salah satu cabang riset dalam ilmu komputer seiring dengan ledakan informasi di internet (Watrianthos, Ronal, 2020).

## Twitter

### Twitter merupakan jejaring sosial daring dan layanan *microblogging* yang memungkinkan pengguna terdaftar untuk membaca dan memposting pesan singkat yang dibatasi hingga 280 karakter, yang disebut dengan kicauan (*tweet*)(Nur Iksan, Helmi, 2019). Twitter juga merupakan media sosial yang populer dikalangan masyarakat Indonesia, tercatat negara Indonesia menempati peringkat ke-7 dengan 13.2 miliar pengguna pada Oktober 2020 menurut penelitian dan analisis oleh statista.com (Statista Research Departement, 2020). Pada umumnya *tweet* diunggah untuk menyampaikan sebuah berita atau informasi terkait peristiwa tertentu, isi *tweet* juga dapat mengekspresikan sebuah pendapat dari penggunanya. Karena hal tersebut, Twitter banyak digunakan sebagai objek penelitian. Hal ini karena tulisan-tulisan pada media sosial Twitter (*tweet*), memiliki struktur yang sangat cocok untuk digunakan pada analisis (Ferdiana, Jatmiko & dkk, 2019).

## *Crawling*

*Crawling* merupakan sebuah proses pengambilan dan pengimpulan data melalui komputer. Dalam penelitian ini proses *crawling* dilakukan menggunakan *standard* *search* *API* yang telah disediakan oleh Twitter dengan pustaka bantuan *tweepy*. Penggunaan pustaka *tweepy* bertujuan untuk menjembatani proses pengaksesan kicauan (*tweet)* melalui program yang dibuat dengan *standard* *search* *API* Twitter bedasarkan parameter seperti kata kunci dan waktu. Dalam penelitian ini kata kunci yang digunakan antara lain: X, Y dan Z dalam periode waktu 1 Desember 2020 hingga 31 Januari 2021.

## *Preprocessing*

*Preprocessing* atau pra-pemrosesan merupakan bagian dari *text mining* yang dilakukan untuk menghapus *noise* pada dokumen atau kalimat. Selain itu, proses ini bertujuan untuk menghindari data yang kurang sempurna; gangguan pada data; dan data yang tidak konsisten (Vina & Wibowo, 2019). Merujuk pada penelitian yang telah dilakukan (Watrianthos, Ronal, 2020) (Budi & Nugroho, 2020) maka pada penelitian ini akan dibahas beberapa tahapan pra-pemrosesan teks antara lain: *casefolding*, *cleansing*, mengubah *slang word,* menghapus *stop word,* dan *stemming*.

### *Casefolding*

*Case folding* merupakan proses yang bertujuan untuk mengubah seluruh huruf menjadi huruf kecil (*lowercase*) (Budi & Nugroho, 2020).

### *Cleansing*

*Cleansing* merupakan proses yang bertujuan untuk menghapus atribut yang tidak diperlukan untuk proses analisis (Watrianthos, Ronal, 2020) (Budi & Nugroho, 2020). *Cleansing* yang dilakukan pada penelitian ini antara lain: menghapus angka (0-9), *mention* (*@mention*), URL, *hastag* (*#hastag*), *unicode*, tanda baca, dan spasi berlebih.

### *Mengubah slang word*

*Slang word* merupakan kata yang tidak sesuai dengan ejaan bahasa Indonesia yang baku (EYD) baik berupa kata singkatan maupun kata gaul. *Slang word*  tersebut sebanyak mungkin akan ditampung ke dalam kamus *slang word*. Kamus tersebut kemudian digunakan sebagai pengetahuan untuk melakukan *replace* atau mengubah kata *slang* menjadi kata dengan bahasa Indonesia yang baku sesuai EYD.

### Menghapus *Stop word*

*Stopword* merupakan kata yang tidak berpengaruh atau kurang bermakna namun sering ditemui dalam dokumen atau kalimat, seperti kata ‘saya’, ‘dan’, ‘atau’ (Watrianthos, Ronal, 2020) (Budi & Nugroho, 2020). Dalam proses ini, kata yang tergolong ke dalam *stop word* akan ditampung ke dalam kamus *stop word.* Kamus tersebut kemudian digunakan sebagai pembanding untuk menghapus sebuah kata dalam dalam dokumen atau kalimat yang tergolong kedalam *stop word*.

### *Stemming*

*Stemming* merupakan proses mengubah kata berimbuhan menjadi kata dasar (Watrianthos, Ronal, 2020). Dalam penelitian lain *stemming* adalah proses pemotongan (pembuangan) imbuhan *(affix),* baik *prefix* maupun *suffix*, dari sebuah *term* untuk mendapatkan kata dasar *(root* atau *stem)* dari kata yang berimbuhan (Wahyudi, dkk, 2017). Dalam penelitian ini proses stemming dilakukan dengan menggunakan pustaka bantuan Sastrawi. Pada penelitian ini, tahap *stemming* merupakan tahap terakhir dalam proses pra-pemrosesan yang akan menghasilkan keluaran berupa teks bersih.

## *Labeling*

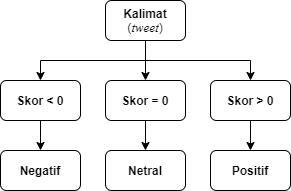
*Labeling* atau pelabelan merupakan proses pengklasifikasian berdasarkan ciri atau karakteristik yang terkandung dalam sebuah dokumen atau kalimat. Proses pelabelan pada dasarnya terbagi menjadi tiga kategori kelas sentimen, yakni sentimen positif, sentimen negatif dan sentimen netral (Budi & Nugroho, 2020). Dalam penelitian ini proses pelabelan dapat dilakukan dengan dua (2) cara antara lain: pelabelan manual dengan cara melabeli berdasarkan subjektifitas peneliti dan pelabelan dengan pendekatan kamus sentimen.

### Perhitungan skor sentimen

Perhitungan skor sentimen merupakan proses pelabelan dengan cara pendekatan kamus sentimen. Kamus tersebut berisikan kata opini positif dan kata opini negatif. Skor suatu kata dalam kalimat(*tweet*) akan bernilai +1 jika kata tersebut adalah kata opini positif, dan bernilai -1 jika kata tersebut adalah kata opini negatif (Liu, 2005). Adapun rumus yang digunakan dalam proses perhitungan skor sentimen adalah sebagai berikut:

### Pelabelan kelas sentimen

Setelah melakukan proses perhitungan skor sentimen dan diketahui nilai skor dari suatu kalimat. Proses selanjutnya dalam pelabelan dengan cara pendekatan kamus sentimen adalah mengklasifikasikan kalimat(*tweet*) berdasarkan skor kedalam kategori (sentimen). Jika dalam kalimat mempunyai skor > 0 akan diklasifikasikan dalam kelas positif, jika kalimat mempunyai skor = 0 akan diklasifikasikan dalam kelas netral, sedangkan jika kalimat mempunyai skor < 0 diklasifikasikan dalam kelas negatif (Asrofi Buntoro, Ghulam, 2017) (Budi & Nugroho, 2020). Adapun proses pelabelan dapat dilihat pada gambar X di bawah ini:



Gambar X. Pelabelan kelas sentimen

## *TF-IDF*

-- Vectorizer --

## *Naive Bayes Classifiier*

-- Classifier --

## Pengujian & Evaluasi

Pengujian merupakan hal penting untuk memastikan bahwa suatu algoritme yang telah dirancang dapat berjalan sesuai dengan harapan (Haryalesmana & Azhari, 2020). Dalam penelitian ini, pengujian dilakukan pada sebuah *model* terhadap data tes yang tersedia. *Model* tersebut dibangun menggunakan data latih yang telah diproses sebelumnya dengan algoritme Naive Bayes Classifier dan pembobotan TF-IDF.

Data latih yang dijadikan *model* dipilih dengan teknik sampling kuota (*quota sampling*). *Quota Sampling* merupakan teknik sampling yang menentukan jumlah sampel dari populasi yang memiliki ciri tertentu hingga jumlah kuota yang diinginkan tercapai (Vina & Wibowo, 2019). Dalam penelitian ini kuota data latih yang digunakan untuk membangun *model* adalah 1200 data *tweet,* terdiri dengan ciri 400 *tweet* positif, 400 *tweet* netral dan 400 *tweet* negatif. Hasil dari pengujian tersebut akan dievaluasi menggunakan *confused matrix* untuk mengukur tingkat akurasi.

## Studi Pustaka

Studi Pustaka