BAB I
PENDAHULUAN

1.1 Latar Belakang

Pada saat ini analisis sentimen menjadi tren untuk melihat berbagai pendapat, sikap dan emosi pada sebuah objek, dan media yang sering digunakan untuk mendapatkan data dalam jumlah yang besar untuk dianalisis adalah Twitter. [1] Twitter adalah sebuah paltform besar yang berisi sumber kumpulan data sentimen yang tidak terstruktur sehingga dapat digunakan untuk dianalisis dan dapat menghasilkan sebuah emosi. Untuk menganalisa sebuah *tweet* tentunya harus mengerti elemen yang terdapat dari *tweet* tersebut sehingga dapat mengklasifikasikan *tweet* positif, negatif atau netral dan menghasilkan tingkat akurasi yang tinggi. Dalam menghasilkan analisis sentimen dan akurasi yang tinggi dapat dilakukan dengan memanfaatkan teknik *machine learning* dan *text mining*. Algoritma pada *machine learning* dianggap lebih adaptif untuk mengubah input dan umumnya digunakan untuk klasfikasi biner dan prediksi sentimen positif, negatif atau netral [2].

Supervised learning dan *unsupervised learning* merupakan teknik dasar yang terdapat *machine learning*. Perbedaan dari kedua teknik tersebut terdaapat pada label dalam subset data latih. Metode klasifikasi pada teknik *supervised learning* telah terdapat atribut output pada kumpulan data tersebut dalam mengklasifikasikan kelasnya. Sehigga hasil akurasi dan eror pada klasifikasi tergantung pada proses kinerja dan perhitungan atribut.

Sedangkan *unsupervised learning* tidak melibatkan atribut dalam proses klasifikasi, dikarenakan dalam mengidentifikasi pengelompokan data tidak ada kebutuhan untuk pemberian label dan hasilnya tidak mengidentifikasi contoh pada kelas yang telah ditentukan [3].

Konsep yang mendasar dalam analisis sentimen adalah *text mining*. Dalam data mining terdapat *text mining* yang merupakan teknik yang digunakan untuk menemukan atau mencari pola dalam suatu teks. *Text mining* memiliki tujuan untuk mengubah teks yang tidak terstruktur menjadi semi terstruktur atau terstuktur dan mengkategorikasn teks dan pengelompokan teks, hal tersebut disebut sebagai *text preprocessing*. [4] Dalam *text preprocessing* terdapat tahap yang penting, salah satunya pembobotan kata. Pembobotan kata dilakukan untuk memberikan bobot atau nilai pada kata yang terdapat pada suatu dokumen. Setelah *text preprocessing* dan pembobotan kata dilakukan, maka proses klasifikasi menggunakan algoritma dapat dilakukan.

Dalam supervised learning terdapat beberapa metode yang popular digunakan untuk analisis sentiment, salah satunya adalah *Naive Bayes*. *Naive Bayes* memiliki beberapa bentuk pemodelan diantaranya *Multinomial Naive Bayes*, *Bernoulli Naive Bayes*, *gaussian Naive Bayes* dan sebagainya. Dari pemodelan tersebut memiliki hasil akurasi yang berbeda – beda seperti yang dilakukan dalam penelitian mengenai ulasan film yang menguji berbagai macam algoritma. Pada penelitian tersebut algoritma *Multinomial Naive Bayes* dan *Bernoulli Naive Bayes* menghasilkan tingkat akurasi dua

tertinggi yaitu 88.50% untuk *Multinomial Naive Bayes*, dan 87.50% untuk *Bernoulli Naive Bayes* daripada *support vector machine*, *decision tree* dan *maximum entropy* [5].

Dalam penelitian ini penulis mencoba membandingkan dua macam pemodelan dari algoritma *Naive Bayes* yang dikombinasikan dengan dua teknik pembobotan kata (*term weigthing*) yang berbeda untuk tiap modelnya. Model yang digunakan dalam penelitian ini adalah metode *Multinomial Naive Bayes* dan *Bernoulli Naive Bayes*. Penggunaan kedua model tersebut dikarenakan memiliki akurasi yang cukup tinggi dalam analisis sentimen. Dan untuk pembobotan kata menggunakan TF-RF (*Term Frequency Relevance Frequency*) dan TF-IDF (*Term Frequency Inverse Document Frequency*) dikarenakan kedua pembobotan kata tersebut populer dan mempengaruhi hasil akurasi menjadi lebih baik. Pengkombinasian model dan pembobotan kata ditujukan untuk mencari kombinasi terbaik dari pemodelan *Naive Bayes* dengan pembobotan kata.

Berdasarkan latar belakang diatas, penulis melakukan penelitian dengan judul "Perbandingan Kombinasi Model Algoritma *Naive Bayes* dengan Teknik Pembobotan Kata dalam Analisis Sentimen".

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan di atas, maka dapat diambil suatu rumusan masalah sebagai berikut:

a. Bagaimana penerapan model algoritma *Naive Bayes* dan teknik pembobotan kata dalam analisis sentimen?

b. Bagaimana hasil akurasi perbandingan dari kombinasi teknik pembobotan kata dengan model pada algoritma *Naive Bayes*?

1.3 Batasan Masalah

Dalam sebuah penelitian diperlukan batasan – batasan agar tujuan penelitian tercapai. Adapun batasan masalah dalam penelitian ini sebagai berikut:

a. Data diambil dari hasil *crawling* di Twitter berbahasa Indonesia dengan kata kunci vaksin covid, corona.
b. Model dari algoritma naïve bayes yang digunakan dalam penelitian ini adalah algoritma *Multinomial Naive Bayes* dan *Bernoulli Naive Bayes*.
c. Menggunakan dua pembobotan kata yaitu TF-IDF (*Term Frequency Inverse Document Frequency*) dan TF-RF (*Term Frequency Relevance Frequency*).

1.4 Tujuan Penelitian

Tujuan dalam penelitian ini sebagai berikut:

a. Menerapkan teknik pembobotan kata TF-IDF dan TF-RF pada model algoritma *Multinomial Naive Bayes* dan *Bernoulli Naive Bayes* dalam analisis sentimen.
b. Untuk mengetahui hasil kombinasi terbaik dari dua model algoritma *Naive Bayes* yaitu *Multinomial Naive Bayes* dan *Bernoulli Naive Bayes* dengan teknik pembobotan kata TF-IDF dan TF-RF.

c. Mengetahui hasil akurasi, presisi dan recall dari kombinasi antara model algoritma *Multinomial Naive Bayes* dan *Bernoulli Naive Bayes* dengan teknik pembobotan kata TF-IDF dan TF-RF.

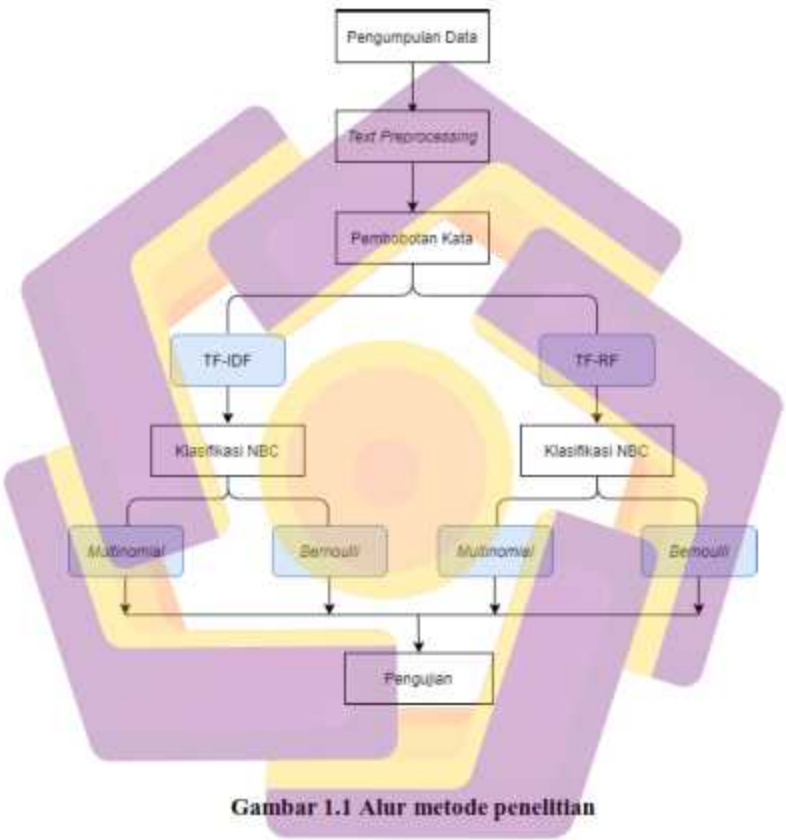
1.5 Manfaat Penelitian

Manfaat dalam penelitian ini sebagai berikut:

a. Memahami penerapan model algoritma *Naive Bayes* serta fitur pembobotan kata TF-IDF dan TF-RF dalam analisis sentimen.
b. Mengetahui performa (akurasi, presisi, recall, f1) tiap kombinasi antara model algoritma *Naive Bayes* dengan teknik pembobotan kata.
c. Memperoleh kombinasi terbaik dari dua model algoritma *Naive Bayes* dan pembobotan kata dari hasil performa yang didapat.

1.6 Metode Penelitian

Metode penelitian yang digunakan pada penelitian ini menggunakan metode penelitian eksperimen. Berikut tahapan – tahapan yang digunakan dalam penelitian ini.



Gambar 1.1 Alur metode penelitian

1.6.1 Pengumpulan Data

a. Studi Pustaka

Mencari literartur buku, paper, atau jurnal yang berkaitan dengan analisis sentimen dan klasifikasi menggunakan mode

algoritma *Naive Bayes*, pembobotan kata, serta penggunaan *k-fold cross validation* untuk mengukur hasil akurasi.

b. ***Text Mining***

Pada tahap ini peneliti melakukan pengumpulan data menggunakan API (*Application Programming Interface*) yang disediakan dari Twitter. Pengumpulan data *tweet* diambil berdasarkan kata kunci "corona, covid-19 dan vaksin".

1.6.2 Metode Analisis

a. ***Labeling***

Pada tahap ini data hasil *crawling* di beri label negatif, positif, dan netral. Pelabelan ini bertujuan untuk mempermudah dalam klasifikasi.

b. ***Text Preprocessing***

Tahap ini dilakukan dengan tujuan untuk membersihkan seluruh data yang telah diambil untuk mempermudah dalam tahapan proses pemberian bobot dan analisis. Dengan adanya tahap ini diharapkan hasil akurasi dalam klasifikasi yang didapatkan lebih akurat. Langkah – langkah dalam proses text processing antara lain *case folding dan cleaning* (pembersihan dokumen), *tokenization*, *filtering* atau *stopword removal*, dan *stemming*.

c. **Pembobotan Kata**

Pada tahap pembobotan kata, frekuensi kata terhadap dokumen akan dilakukan perhitungan nilai atau bobot. Tahap ini

bertujuan untuk mendapatkan nilai atau bobot pada tiap kata dasar yang telah diekstrak dan masing - masing nilai tersebut nantinya akan diinput pada metode klasifikasi. Pembobotan kata yang digunakan dalam penelitian ini adalah *term frequency inverse document frequency* (TF-IDF) dan *term frequency relevance frequency* (TF-RF).

d. Klasifikasi Algoritma *Naive Bayes*

Pada tahap ini akan dilakukan proses *training* dan *testing* pada dataset. Tahapan klasifikasi ini merupakan tahapan utama dalam analisis sentimen. Pada proses *training* data akan dilatih memahami mesin untuk klasifikasi. Sedangkan untuk proses *testing* untuk melihat performa atau akurasi dari proses klasifikasi dengan menggunakan nilai input dari masing – masing pembobotan kata, maka dari itu dalam penelitian ini akan dilakukan empat kali proses atau kombinasi yaitu TF-IDF dengan model *Multinomial Naive Bayes*, TF-RF dengan *Bernoulli Naive Bayes*, dan TF-IDF *dengan Bernoulli Naive Bayes*, serta TF-RF dengan *Multinomial Naive Bayes*.

1.6.3 Metode Pengujian

Pada tahap ini bertujuan untuk mengukur kinerja dan akurasi dari model klasifikasi yang dilakukan. Metode yang digunakan adalah metode *Confusion Matrix* atau sering disebut *eror matrix* dan *k-fold cross validation*. Dari hasil test data akan menghasilkan empat representasi

yaitu *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FP). Dari model klasifikasi tersebut juga akan dihitung nilai *precision*, *recall*, dan f1 *score*. Hasil akurasi yang didapat dihitung dengan mengkalkulasi semua nilai true dan dibagi dengan jumlah keseluruhan data. Sedangkan *k-fold cross validation* digunakan memvalidasi hasil akurasi model.

1.7 Sistematika Penulisan

Dalam penelitian ini terdapat 5 bab tahapan dalam proses penelitan, sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini menjelaskan mengenai latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitan, metode penelitian dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini berisi tinjauan pustaka dan landasan teori yang berisi dasar – dasar teori mengenai penelitan yang dilakukan.

BAB III METODE PENELITIAN

Bab ini berisi metode yang digunakan dalam penelitian dari metode pengumpulan data, *text preprocessing* yang berupa analisis data, pemberian bobot, tahapan klasifikasi hingga pengujian untuk mengukur akurasi dan kinerja dari algoritma yang digunakan.

BAB VI HASIL DAN PEMBAHASAN

Pada bab ini berisi hasil dan pembahasan perbandingan dari model algoritma naïve bayes dengan pembobotan kata.

BAB V PENUTUP

Pada bab ini berisi kesimpulan dan saran dari hasil penelitian.