BAB III
METODE PENELITIAN

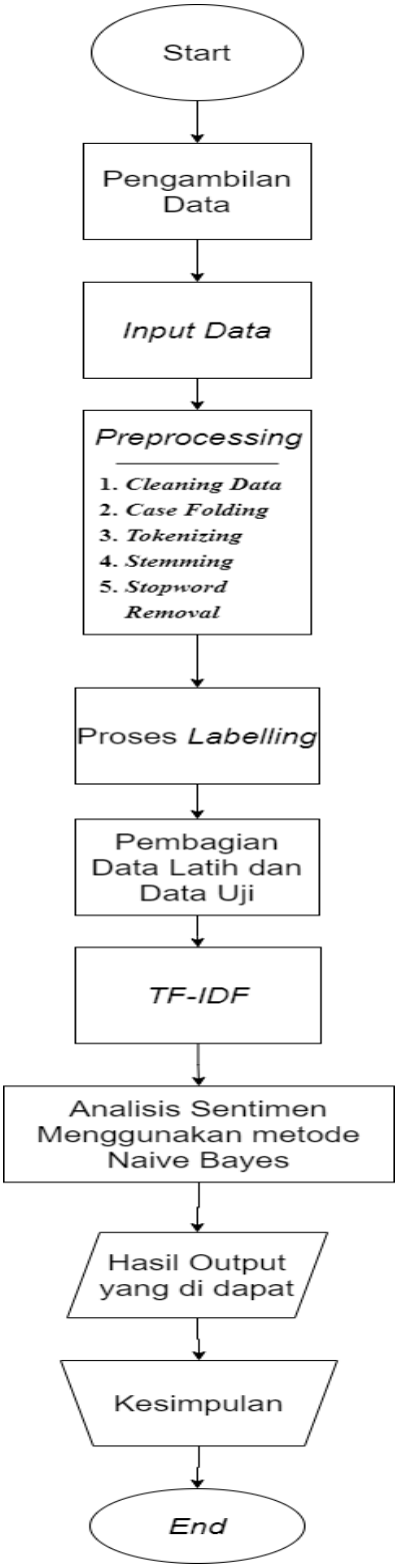
3.1 Objek dan Subjek Penelitian

Objek dari penelitian ini yaitu Analisis sentimen terhadap minat ketertarikan masyarakat Indonesia terhadap makanan khas Korea Selatan. Sedangkan untuk subjek dari penelitian yaitu seluruh masyarakat Indonesia khususnya para pengguna media sosial yang memberikan ulasan maupun komentar di *YouTube* dan *Twitter* yang berkaitan dengan makanan khas Korea Selatan.

1.2 Alur Penelitian

Penelitian ini dimulai dengan mengambil data di *Twitter dan YouTube,* di lanjutkan dengan *menginput* data yang telah didapat di *Microsoft Excel*, lalu dilanjutkan dengan tahap pembersihan data atau biasa disebut dengan *Pre-processing menggunakan software Rapidminer*, setelah data yang didapat bersih lalu masuk ke tahap seleksi data yang sering muncul atau biasa disebut dengan tahap *TF-IDF,* setelah didapatkan data yang sering muncul lalu dilakukan analisis sentimen untuk membedakan data yang bersifat positif atau negatif, setelah mendapatkan data positif dan negatif maupun netral lalu data dibedakan menjadi dua bagian data latih dan data uji.

Data latih yang digunakan sebanyak 595 opini maupun komentar, sedangkan untuk data uji yang digunakan sebanyak 105 opini maupun komentar dengan asumsi perbandingan data latih yang digunakan sebanyak 85% dan data uji sebanyak 15% dari total data yang digunakan dalam penelitian sebanyak 700 data opini *Twitter* maupun *YouTube*, setelah data dibagi menjadi dua data diolah menggunakan metode klasifikasi *Naive Bayes*, pengolahan data menggunakan metode klasifikasi *Naive Bayes* dilakukan untuk mendapatkan hasil atau *output* yang diinginkan. Hasil atau *output* yang telah didapat digunakan untuk menarik kesimpulan dari hasil penelitian yang dilakukan. Alur penelitian dapat dilihat pada Gambar 3.1.



Gambar 3.1 *Flow Chart Penelitian*

3.3 Teknik Pengumpulan Data

3.3.1 Metode Pengumpulan Data

Pengambilan data dilakukan lewat media sosial *Twitter* dan *YouTube* dengan menggunakan metode *web scrapping* untuk pengambilan data *YouTube* dan *crawling data* untuk pengambilan data *Twitter* dengan kata kunci yang digunakan yaitu " Makanan khas Korea*", "Korean food", "Street food Korea", "all you can eat korea", "Kimchi", "Ramyeon"," Jjajangmyeon"," Bibimbap"*. Total jumlah sampel yang digunakan dalam penelitian ini minimal 700 data komentar atau lebih. Sampel data yang akan diambil memiliki rentang waktu setahun yaitu pada tahun 2021.

3.3.2 Alat dan Bahan

Alat dan Bahan yang digunakan untuk mendukung penelitian ini dapat dilihat sebagai berikut:

a. Laptop

b. Media Sosial *Twitter dan You Tube*

c. Microsoft Excel

d. Software Rapid Miner

e. Alat Tulis

3.4 Teknik Analisa Data

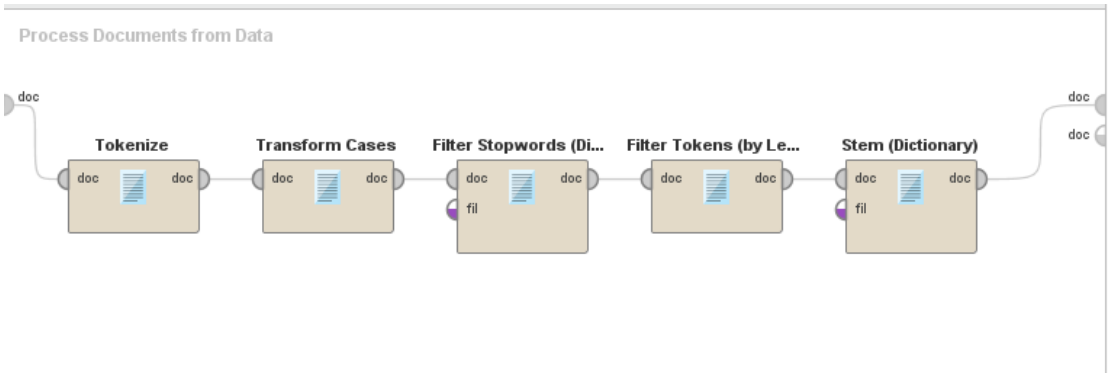
3.4.1 Metode Pemprosesan Data

Terdapat beberapa metode yang digunakan untuk memproses data yang telah di dapat dalam penelitian ini yaitu:

a. Pre-processing

Pre-processing merupakan tahap yang dimana data yang sudah didapat dibersuhkan dari karakter, simbol, spasi, maupun kata yang tidak penting sebelum dilakukan pembobotan maupun pemberian nilai pada setiap data yang dimiliki. Data teks yang didapat dalam proses *scrapping* masih banyak mengandung *noise* dan banyak bagian tidak penting seperti tag HTML, *script* dan iklan. Selain itu, tingkat tingkat kata-kata yang di dapat dari proses *scraping*, masih banyak kata-kata tidak sesuai dengan data yang butuhkan. *Pre-processing* sangat diperlukan untuk mendapatkan hasil data yang diinginkan. Mengurangi *noise* dalam teks dapat

membantu dalam meningkatkan kinerja *classifier* dan mempercepat proses klasifikasi sehingga memudahkan dalam menganalisis sentimen secara *real time*. Seluruh proses melibatkan beberapa langkah: mulai dari membersihkan teks, penghapusan ruang *spasi*, memperluas singkatan, kata dasar (*stemming)*, penghapusan kata ganti *(stopword removal)* (Indrayuni, 2017). Alur proses *pre-processing* dapat dilihat pada Gambar 3.2.



Gambar 3.2 *Alur proses Pre-processing*

a. Cleaning Data

Pada proses *cleaning data*, data yang sudah didapat dari proses *scraping* dibersihkan untuk menghilangkan komponen khusus yang ada di *website YouTube* maupun dalam sebuah *tweet.* Komponen khusus yang biasa ada dalam sebuah komen *website YouTube* maupun *tweet* yaitu, *username, URL (Uniform Resource Locator*), dan "RT" (tanda *retweet*). Komponen tersebut dibersihkan dan dihilangkan karena dapat mempengaruhi hasil nilai sentimen yang didapatkan nantinya (Ghaniy dan Sihotang, 2019).

b. Case Folding

Setelah data di bersihkan dari komponen khusus seperti *URL (Uniform Resource Locator*), dan "RT" (tanda *retweet*), lalu dilanjutkan dengan proses *case folding* yang mana data yang sudah didapat dan dibersihkan dari komponen khusus masih terdapat huruf besar di dalamnya sehingga harus diubah menjadi *lowercase* atau huruf kecil semua, agar data yang didapat mudah dibaca dan diproses oleh computer. (Ghaniy dan Sihotang, 2019).

c. Tokenizing

Tokenizing merupakan tahapan yang dilakukan untuk memisahkan kata agar tidak saling berpengaruh satu sama lain, sebelum dilakukan identifikasi menggunakan algoritma *Naïve Bayes* (Parasati dkk., 2020).

d. Stopword Removal

Setelah data dilakukan *tokenizing* tahap selanjutnya dilakukan proses *stopword removal* untuk menghilangkan kata-kata umum yang tidak memiliki makna maupun informasi yang kita dibutuhkan (Ghaniy dan Sihotang, 2019).

e. Stemming

Setelah data dilakukan *stopword removal* tahap selanjutnya dilakukan proses *stemming* untuk merubah dan menghilangkan kata yang masih berimbuhan pada data yang kita dapatkan menjadi kata dasar, dengan menghilangkan semua imbuhan yang ada pada kata dari data yang telah kita dapatkan (Ghaniy dan Sihotang, 2019).

3.4.2 Metode Pembobotan Kata

a. Term Frequency-Inverse Document Frequency (TF-IDF)

Penilaian dan pembobotan kata dari data yang sudah didapat dilakukan dengan menggunakan *Term Frequency-Inverse Document Frequency (TF-IDF)* merupakan metode yang digunakan untuk menghitung dan mengira berat setiap kata yang akan proses dan digunakan. *TF-IDF* digunakan sebagai alat ukur statistika yang digunakan untuk mengetahui seberapa penting kata yang didapat dalam dokumen. Pada proses pembobotan kata menggunakan dua model yang saling berkaitan yaitu model *term frequency* (tf) dan *inverse document frequency* (idf). Model *term frequency* (tf) digunakan untuk menghitung jumlah kemunculan kata dalam satu dokumen sedangkan *inverse document frequency* (idf) merupakan model yang digunakan untuk menghitung jumlah kemunculan kata diberbagai dokumen maupun komentar (Sheila, 2019), (Rahmi, 2021). Berikut Tahapan dalam proses pembobotan dengan *TF-IDF* yaitu:

1) Hitung *term frequency* tft, j

Rumus hitung *term frequency dapat di lihat pada persamaan (6)*(Sheila, 2019).

$tft, j = \frac{ni,j}{\Sigma kni,j}$.....(6)

2) Hitung *document frequency* (df) (Sheila, 2019).

3) Hitung bobot *inverse document frequency* (idf)

Rumus hitung bobot *inverse document frequency* (idf) *dapat di lihat pada persamaan (7)* (Sheila, 2019).

$IDF = \log(\frac{n}{dfi})$ .....(7)

4) Hitung nilai bobot *TF-IDF*

Rumus hitung nilai bobot *TF-IDF dapat di lihat pada persamaan (8)* (Sheila, 2019).

$Wi'j = tfi'j * idf$ .....(8)

Keterangan:

$tfi'j =$ jumlah kemunculan $i\ di\ j$

$df =$ jumlah dokumen yang mengandung term

ni, j $=$ banyak kata i dalam dokumen j

$N =$ jumlah total dokume

$Wi, j =$ bobot TF $-$ IDF

3.4.3 Metode Klasifikasi Data

Data yang sudah diberi bobot atau nilai lalu dilanjutkan ketahap proses klasifikasi data, dalam proses klasifikasi data terdapat beberapa metode yang digunakan untuk melakukan klasifikasi data yang telah diberi nilai dalam penelitian ini yaitu:

a. **Asosiasi Kata**

Asosiasi kata merupakan proses yang digunakan untuk mengetahui kata mana yang sering digunakan dalam komentar. Asosiasi kata juga dapat menemukan hubungan antar kata, misalnya hubungan antara dua kata atau lebih dapat digunakan

secara bersama dalam satu dokumen. Dalam asosiasi kata nilai korelasi antar kata berkisar antara -1 sampai 1. Jika mendekati 1 atau -1 hubungan antar kata semakin kuat, sedangkan Jika nilainya mendekati 0, hubungan antar kata menjadi lebih lemah. ada beberapa kategori nilai relevan yang digunakan adalah sebagai berikut (Sheila, 2019).

$0 : Tidak\ ada\ korelasi\ antar\ dua\ variabel$

$> 0 - 0{,}25 :$ Korelasi lemah

$0{,}25 - 0{,}5 :$ Korelasi cukup

$> 0{,}5\ 0{,}75 :$ Korelasi kuat

$1 :$ Korelasi sangat kuat

Dengan rumus perhitungan asosiasi kata yang dapat di lihat pada persamaan (9) (Rahmi, 2021), (Adawiyah, 2018).

$$r_{xy} = \frac{n\sum x_i y_{i-(\sum x_i)(\sum y_i)}}{\sqrt{n\sum x_i^2-(\sum x_i)^2\}\{n\sum y_i^2-(\sum x_i)^2\}}} \quad \text{.....(9)}$$

Keterangan:

r_{xy} : nilai korelasi antar variabel x dan variabel y

n : banyaknya pasangan data x dan y

Σx_i : jumlah nilai pada variabel , i $= 1,2,3,\ldots\ldots,$ n

Σy_i : jumlah nilai pada variabel y

$\Sigma_{x_i^2}$: kuadrat dari total nilai variabel x

$\Sigma_{y_i^2}$: kuadrat dari total nilai variabel y

$\Sigma x_i\ \Sigma y_i$: jumlah dari hasil perkalian antara nilai variabel x dan variabel y

b. **Analisis Sentimen**

Analisis sentimen dilakukan untuk melihat pendapat atau opini seseorang terhadap suatu isu atau objek tertentu, apakah cenderung berpendapat negatif atau positif. Analisis sentimen sering digunakan untuk melihat dan memantau perkembangan pasar atau untuk menanggapi suatu permasalah yang sedang terjadi

(Magro dan Talamini, 2019). Contoh penerapannya di dunia nyata adalah mengidentifikasi tren pasar atau persepsi dari suatu objek. Analisis sentimen juga digunakan untuk menganalisis beberapa data yang berfungsi untuk menentukan emosi manusia. Analisis sentimen dapat dibagi menjadi tiga tugas, yaitu untuk mendeteksi teks informatif, ekstraksi informasi, dan klasifikasi minat sentimenal. Klasifikasi sentimen terhadap suatu pandangan yang bersifat negatif maupun positif digunakan untuk memprediksi polaritas sentimen berdasarkan data sentimen pengguna (Santoso dan Nugroho, 2019).

c. Naive Bayes

Naive Bayes merupakan metode yang digunakan untuk mengklasifikasi data dalam melakukan klasifikasi data yang telah didapat. Metode ini memiliki potensi dan keunggulan dalam akurasi klasifikasi dan komputasi data. *Naive Bayes* merupakan teknik untuk melakukan klasifikasi data yang banyak digunakan pada saat penambangan data khususnya pada komentar *Twitter dan YouTube* (Samsir dkk., 2021). Metode *Naive Bayes* dapat memprediksi probabilitas keanggotaan dari suatu kelas data, setiap atribut dalam sebuah data bersifat terpisah atau *independent* (Rahmi, 2021). Dalam metode ini mengklasifikasikan kelas data berdasarkan pada probabilitas sederhana dengan menggunakan pendekatan algoritma bayes yang dapat dilihat pada persamaan (11) (Samsir dkk., 2021).

$$P\ (V|X) = \frac{P\ (X|V).P(V)}{P(X)} \quad \text{.....(11)}$$

Dari persamaan (11) di atas menunjukkan bahwa V merupakan suatu kelas yang spesifik, X merupakan suatu data dengan kelas yang belum diketahui, untuk P(V|X) merupakan probabilitas hipotesis berdasarkan kondisi, sedangkan untuk P(V) merupakan probabilitas V, sedangkan untuk P(X|V) merupakan probabilitas X berdasarkan kondisi V, sedangkan untuk P(X) merupakan probabilitas X, sedangkan yang terakhir untuk P(V|X) merupakan probabilitas hipotesis V berdasarkan kondisi X. Kemudian persamaan (4) dikembangkan lagi menjadi persamaan (12) (Samsir dkk., 2021).

$$P\ (V|X_1 \dots X_n) = \frac{P\ (V)P(X_1 \dots X_n|V)}{P\ (V|(X_1 \dots X_n)} \quad \text{.....(12)}$$

Dalam persamaan (5), variabel V mewakili kategori, dan variabel X1…Xn menjelaskan instruksi yang diperlukan untuk klasifikasi. Persamaan (12) menunjukkan bahwa peluang masuknya sampel dengan karakteristik tertentu masuk ke kelas V (*Posterior)* merupakan peluang munculnya kelas V atau biasa disebut *Prior*, dikali peluang kemunculan karakteristik sampel pada kelas V yang biasa disebut dengan *likelihood* (Susilawati dkk., 2019).

3.4.4 Metode Pengujian Data

Metode pengujian data yang digunakan pada penelitian ini yaitu dengan menggunakan data latih dan data uji dengan estimasi perbandingan data uji dan data latih yaitu 85%:15% yang dimana jumlah data latih harus lebih banyak dari data uji. Jika total data yang digunakan sebanyak 700 data opini, ulasan maupun komentar, maka perbandingan data latih dan data uji yang digunakan yaitu sebesar 595 data latih dan 105 data uji yang digunakan dalam penelitian ini. Data yang digunakan dalam proses pengujian harus data yang telah melewati proses pembersihan data atau *Pre-processing* (Zhafira dkk., 2021). Terdapat beberapa metode yang digunakan untuk menguji data yang telah di dapat dalam penelitian ini yaitu:

a. Confusion Matrix

Metode pengujian tingkat akurasi data yang digunakan pada penelitian ini yaitu dengan menggunakan metode *confusion matrix* yang biasa digunakan untuk mengukur tingkat akurasi baik buruknya *classifier* yang digunakan dalam mengenali *tuple* dari kelas yang berbeda,pengujian tingkat akurasi yaitu dengan menghitung *accuracy, precision, recall, f-measure* yang mengacu pada nilai *True Positive* (TP), T*rue Negative* (TN), *False Positive* (FP), dan *False Negative* (FN), rumus dapat dilihat pada Tabel 3.1 (Jiawei dkk., 2012)**.**

Tabel 3.1 *Confusion Matrix*

		True Class	
		Positive	*Negative*
Predicted Class	*Positive*	*True Positive Count (TP)*	*False positive Count (FP)*
	Negative	*False Negative Count (FN)*	*True negative Count (TN)*

1. *True Negative* (*TN*) adalah kelas yang dihasilkan dari prediksi klasifikasi negatif, meskipun kelas tersebut sebenarnya negatif.
2. *True Positive (TP*) adalah kelas yang dihasilkan dari prediksi klasifikasi positif, yang sebenarnya adalah kelas positif.
3. *False positive (FP*) adalah kelas yang memprediksi klasifikasi positif ketika kelas sebenarnya negatif.
4. *False Negatives (FN)* adalah kelas-kelas yang dihasilkan dari prediksi klasifikasi negatif, sedangkan kelas sebenarnya adalah kelas positif.

Dari Tabel 3.1 di atas dapat disimpulkan bahwa rumus-rumus perhitungan recall, presisi, akurasi, dll adalah sebagai berikut: (Rahmi, 2021). Akurasi merupakan suatu nilai ketepatan dari klasifikasi dalam bentuk persen dalam melakukan penghitungan tingkat akurasi rumus yang di gunakan dapat di lihat pada persamaan (13) (Fahrur dkk., 2020).

$$\text{Akurasi} = \frac{\Sigma\ data\ benar}{n\ dokumen} x100 \quad \text{.....(13)}$$

Presisi adalah nilai proporsi dari jumlah dokumen yang ditemukan dan dianggap relevan untuk mendapatkan suatu informasi yang di butuhkan. Perhitungan nilai Presisi dapat di lihat pada persamaan (14) (Fahrur dkk., 2020).

$$Presisi = \frac{\Sigma\ data\ positif\ atau\ negatif}{n\ dokumen\ positif\ atau\ negatif} x100 \quad \text{.....(14)}$$

Recall merupakan jumlah nilai proporsi yang dapat ditemukan kembali dalam proses pencarian. Perhitungan nilai Presisi dapat di lihat pada persamaan (15) (Fahrur dkk., 2020).

$$Recall = \frac{\Sigma\ doc\ relevan\ dan\ terambil}{\Sigma\ seluruh\ dokumen\ relevan} x100 \quad \text{.....(15)}$$

3.4.5 Jadwal Kegiatan

Berikut jadwal kegiatan pelaksanaan penelitian Tugas Akhir yang telah dijelaskan di metodologi penel, seperti pada Tabel 3.2.

Tabel 3.2 Jadwal Kegiatan

Kegiatan	Januari	Februari	Maret	April	Mei	Juni	Juli	Agustus	September
Studi Literatur									
Pengumpulan Data									
Pengolahan Data									
Analisis Data									
Pembuatan Proposal									