

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

Pada saat ini perkembangan teknologi sangatlah pesat, masyarakat sudah tidak asing dengan media sosial, masyarakat menggunakan media sosial sebagai alat berjejaring dan berkomunikasi, seiring perkembangan media sosial yang semakin pesat, media sosial juga menyebabkan pertumbuhan data digital yang sangat besar. Media adalah satu contoh penghasil data internet terbesar (Dhawan, 2014). Di Indonesia pertumbuhan pengguna internet terus bertambah setiap tahun, pada Januari 2023 pengguna internet di Indonesia meningkat sekitar 3,85% dibandingkan tahun sebelumnya, pada tahun 2022 jumlah pengguna internet di Indonesia tercatat sebanyak 205 juta jiwa, dan tercatat pada Januari 2023 telah mencapai pada 212,9 juta jiwa, ini mengartikan bahwa sekitar 77% dari populasi Indonesia telah menggunakan internet (We Are Social, 2023).

*Youtube* adalah salah satu media sosial yang sangat cepat dalam merespon suatu isu, masyarakat sangat terbuka bebas dalam memberikan saran, isu-isu dan opini public. Penggunaan *Youtube* dalam berkomunikasi, berbagi informasi dan peristiwa akan menghasilkan data yang sangat besar yang *real time*. Dicitat oleh We Are Social pada Januari 2023, pengguna *Youtube* di Indonesia mencapai 139 Juta Jiwa, menempati peringkat ke-5 dari berbagai setelah Brazil. *Youtube* menyediakan suatu konten berupa video.

Video merupakan salah satu cara tercepat untuk terhubung dengan pelanggan dan membangun hubungan baik dengan mereka dan alasan bahwa video sangat kuat karena 90% dari komunikasi adalah non-verbal. (Evans W Wirga, 2017).

Para pengguna *Youtube* dapat memberikan opini dan pendapat melalui komentar, *Youtube* juga memberikan fitur *like* dan *dislike*, Salah satu topik yang ramai diperbincangkan di *Youtube* pada akhir tahun lalu adalah mengenai Resesi Indonesia pada tahun 2023.

Perbincangan masyarakat mengenai Resesi 2023 yang akan terjadi di Indonesia, berbagai influencer, artis, hingga ekonom menyatakan Indonesia akan mengalami resesi pada tahun 2023, beberapa hal yang menjadi alasan ramainya opini mengenai resesi yang akan terjadi di Indonesia pada tahun 2023 adalah menurunnya pertumbuhan ekonomi pada tahun 2020 sekitar -2,97%, penurunan drastis ini disebabkan oleh pergerakan ekonomi Indonesia yang kurang stabil. Munculnya *Trending Topic* mengenai resesi adalah kekhawatiran yang begitu besar terhadap perekonomian dunia yang baru pulih dari Pandemi Covid-19. Naiknya suku bunga yang dilakukan oleh bank sentral di seluruh dunia menyebabkan terjadinya resesi tahun 2023 yang diungkapkan oleh Bank Dunia.

Resesi Indonesia pada tahun 2023 memunculkan banyak opini positif dan negatif, terdapat pro dan kontra mengenai resesi yang akan terjadi di Indonesia pada tahun 2023, terdapat animo masyarakat yang takut dengan terjadinya resesi dan terdapat opini masyarakat yang optimis bahwa resesi tahun 2023 tidak akan terjadi. Melihat banyaknya tanggapan masyarakat di *Youtube* mengenai resesi yang akan terjadi pada tahun 2023, pendapat dan *sentiment* masyarakat hal tersebut menjadi sumber data yang dapat dianalisis, sehingga dengan data tersebut dapat diteliti untuk mengetahui seberapa besar persentase *sentiment* positif dan negative dari Tweet atau Cuitan. *Sentiment Analysis* atau bisa disebut dengan *textmining* bertujuan untuk menganalisis, memahami dan mengolah data tekstual yang berupa opini terhadap entitas seperti organisasi dan topik tertentu agar mendapatkan suatu informasi (Liu, 2012). *Textmining* merupakan teknik m dalam proses klasifikasi data dengan konten apapun. Teknik ini bekerja dalam memunculkan variasi dari kumpulan data yang tersedia dalam jumlah besar, kemudian menemukan pola yang sesuai dari kumpulan teks yang ada.

Dalam *Sentiment Analysis*, diperlukan algoritma *Classifier* & seleksi fitur yang mumpuni agar didapatkan hasil penelitian yang akurat, dalam melakukan *Sentiment Analysis* menggunakan algoritma *Naïve Bayes Classifier (NBC)*, dengan algoritma *Naïve Bayes* memprediksi peluang analisis

dengan akurasi yang tinggi, sehingga dapat memprediksi peluang terjadinya kejadian di masa depan berdasarkan data yang ada (Hakimi, 2018). Algoritma *Classifier Naïve Bayes* dirasa baik digunakan karena mampu mengklasifikasikan data dalam kategori positif dan negative.

Kemudian pada penelitian (Fajar Darwis Dzikril Hakimi, 2018), *sentiment analysis* digunakan untuk menganalisis Opini Pemilihan Kepala Daerah Jawa Timur 2018 Pada Dokumen Twitter. Data Twitter calon gubernur Jawa Timur nomor urut 1 dengan kata kunci “khofifah” & dengan kata kunci “gus ipul”. Data latih pada sistem pertama dengan kata kunci “Khofifah” sebanyak 2000 data, dan data latih kedua pada kata kunci “gus ipul” terdapat 1200 data. Kedua data latih tersebut diberikan label *sentiment* secara manual. *Sentiment analysis* pada penelitian ini menggunakan algoritma *Naïve Bayes Classifier* dengan seleksi fitur *Term frequency (TF) & Inverse Document Frequency (IDF)*. Pengujian pada data ini dilakukan sebanyak 2 kali, pada proses pertama menggunakan 487 data uji, dan proses kedua menggunakan 287 data uji. Hasil pada sistem pertama menghasilkan sebagai berikut: akurasi sebesar 98,99%, presisi sebesar 93,44%, recall sebesar 97,78% dan *f-measure* sebesar 95,56%. Sedangkan pada proses sistem kedua menghasilkan sebagai berikut: akurasi sebesar 98,95%, presisi sebesar 97,78%, recall sebesar 98,55% dan *f-measure* sebesar 98,17%. Berdasarkan penelitian sebelumnya, pada penelitian data ini menggunakan Algoritma *Classifier Naïve Bayes* dikarenakan terdapat fitur *Term Frequency (TF) & Inverse Document Frequency (IDF)*, kedua fitur seleksi tersebut dapat memprediksi *sentiment analysis* yang akurat. Dalam penelitian ini peneliti memutuskan berbagai hal berikut:

1. Menggunakan Algoritma *Naïve Bayes* dalam metode klasifikasi, dikarenakan dengan seleksi fiturnya mampu memprediksi hasil penelitian yang akurat, dengan rata-rata sekitar diatas 90%, sehingga memutuskan untuk menggunakan Algoritma *Naïve Bayes* untuk mengklasifikasikan data penelitian.
2. Dalam proses penelitian yang menggunakan Algoritma *Naïve Bayes*, proses penelitian ini menggunakan seleksi fitur *Term Frequency (TF) & Inverse*

*Document Frequency (IDF)*. Dikarenakan pada penelitian sebelumnya seleksi fitur ini mampu meningkatkan akurasi Algoritma *Naive Bayes*.

3. Untuk membantu menambah keakurasian data yang digunakan, peneliti menggunakan metode *Lexion Based* untuk mengklasifikasikan data, dengan menambah kamus kata, yang dilakukan secara manual.

## 1.2 Rumusan Masalah

Berdasarkan pernyataan dari latar belakang masalah diatas, maka penulis ingin memaparkan sebuah pertanyaan, sebagai berikut;

1. Bagaimana tingkat akurasi Algoritma *Naive Bayes Classifier (NBC)* dalam memprediksi data tanggapan masyarakat mengenai resesi Indonesia 2023 pada *Youtube*?
2. Bagaimana implementasi dari Algoritma *Naive Bayes Classifier (NBC)* dengan seleksi fitur *Term frequency (TF) & Inverse Document Frequency (IDF)* ?

## 1.3 Batasan Masalah

Adapun untuk memperoleh hasil dalam penelitian ini, terdapat Batasan-batasan masalah sebagai berikut :

1. Penelitian ini menggunakan data *Youtube* mengenai tanggapan masyarakat dengan topik Resesi Indonesia 2023.
2. Menggunakan metode Textmining & Algoritma *Naive Bayes Classifier* untuk memprediksi hasil *sentiment analysis* penelitian yang akurat.

## 1.4 Tujuan Penelitian

Adapun tujuan dilakukannya penelitian ini untuk mengetahui :

1. Penerapan metode *Naive Bayes Classifier*) dalam memprediksi data tanggapan masyarakat mengenai resesi Indonesia 2023 pada *Youtube*.
2. Mengetahui kelayakan dan tingkat akurasi hasil penelitian untuk penggunaan Algoritma *Naive Bayes Classifier*, dengan seleksi fitur *Term frequency (TF) & Inverse Document Frequency (IDF)*.

### 1.5 Manfaat Penelitian

Adapun manfaat dari penelitian ini, yaitu:

1. Mengetahui pengetahuan mengenai *Textmining*, *Sentiment Analysis* menggunakan metode *Algoritma Naïve Bayes Classifier* mengenai tanggapan masyarakat mengenai Resesi Indonesia 2023 pada *Youtube*.



## BAB II

### KAJIAN PUSTAKA DAN LANDASAN TEORI

#### 2.1 Penelitian Terdahulu

Dalam melaksanakan penelitian ini tinjauan Pustaka diperlukan sebagai kajian literatur, untuk mengetahui keterkaitan tentang penelitian sebelumnya, berikut ini adalah penelitian terdahulu yang digunakan sebagai tinjauan pustaka dalam penelitian ini :

**Tabel 2. 1** Penelitian Terdahulu

NO.	NAMA (TAHUN)	JUDUL	METODE	HASIL
1	Khofiya Mulya Rahmi (2021).	<i>Textmining Analysis</i> dan Sentiment Analysis dengan Menggunakan Metode <i>Naïve Bayes Classifier</i> (Studi Kasus: Mengenai Tanggapan Tokopedia Melalui Twitter)	- <i>Textmining &amp; Naïve Bayes Classifier</i>	Dari proses penggunaan <i>Sentiment Analysis</i> dari data Twitter sebesar 5055 terklasifikasi, sebanyak 69,8% atau sebanyak 3530 data tergolong sentimen negative, dan 30,2% atau sebesar 1525 data tergolong pada sentiment positif, dengan akurasi 91%, recall 91% dan precision 82%.
2.	Sheila Farah Diba (2020).	Implementasi Metode <i>Naïve Bayes Classifier</i> dalam Analisis Sentimen Pada Opini Masyarakat Tentang RUU KUHP.	- <i>Naïve Bayes Classifier</i>	Hasil penelitian senstimen analisis pada pasal kontroversial RUU KUHP ini menggunakan data sebesar 3561 <i>Tweet</i> , dengan jumlah 2483 data <i>Tweet</i> tergolong Negatif, dan 1078 data <i>tweet</i> tergolong positif, dengan hasil prediksi penelitian ini nilai akurasinya adalah 93,1%, <i>recall</i> sebesar 78,9% dan nilai <i>precision</i> sebesar 97,6%.
3.	Kurniawan, (2017)	Klasifikasi Berita Menggunakan	- <i>Naïve Bayes Classifier</i>	Pada penelitian klasifikasi berita ini menggunakan delapan katategori berita,

		Metode <i>Naïve Bayes</i>		yaitu: ekonomi, entertainment, olahraga, teknologi, travel, kesehatan, makanan, dan otomotif. Penelitian ini menghasilkan nilai <i>Preicision 0,96</i> , <i>Recall 0.78</i> dan <i>f-measure 0.86</i> .
4.	Enggar Prima Jati, (2019)	Implementasi Textmining dan Sentiment Analisis pada Jejaring Sosial Media Twitter dengan menggunakan metode <i>Naïve Bayes Classifier</i>	<ul style="list-style-type: none"> <li>- <i>Textmining</i></li> <li>- <i>Sentiment Analysis</i></li> <li>- <i>Naïve Bayes Classifier</i></li> </ul>	Pada proses <i>Sentiment Analysis</i> sebanyak 6011, data terklasifikasi, sebanyak 1735 data atau 29% tergolong positif, dan 71% lainnya, sebesar 4276 data tergolong negative. Untuk kasus banjir DKI Jakarta, 5527 terklasifikasi, 44% tergolong sentiment positif, dan 56% data lainnya tergolong sentiment negative.
6.	“Eko Budi Santoso, Aryo Nugroho (2019)	Analisis Sentimen Calon Presiden Indonesia 2019 Berdasarkan Komentar Publik di <i>facebook</i> .	<ul style="list-style-type: none"> <li>- <i>Textmining</i></li> <li>- <i>Sentiment Analysis</i></li> <li>- <i>Naïve Bayes Classifiier</i></li> </ul>	Popularitas dari Joko Widodo dan Prabowo Subiyanto dari data acak sebesar 5000 komentar. Pada tahap <i>preProcessing</i> menghasilkan polaritas sentiment. Joko Widodo diperoleh 85% sentiment positif dan 15% Sentimen negative. Dan pada Prabowo Subianto dengan hasil 76% sentiment positif, 24% sentiment negative. Dalam metode <i>Naïve Bayes Classifier</i> dengan <i>Accuration 86,4%</i> .

Berdasarkan **Tabel 2.1** penelitian terdahulu yang menggunakan metode *Naïve Bayes Classifier* telah menghasilkan prediksi yang akurat.

## 2.2 Landasan Teori

### 2.2.1 Resesi Ekonomi

Pasca pandemic Covid-19 berdampak bagi beberapa segmen negara seperti ekonomi, sosial, dan Kesehatan. Keadaan ini juga berpengaruh pada laju ekonomi negara, resesi ekonomi menjadi ancaman beberapa negara yang ditanggapi secara serius, diantaranya adalah negara ASEAN, tiga negara diantara negara-negara ASEAN terancam resesi ekonomi, antaranya adalah Indonesia, Filipina, dan Singapura. Resesi ekonomi merupakan kondisi ekonomi suatu negara yang disebabkan beberapa kebijakan fiscal dalam mengatur ekonomi global, resesi ekonomi ini terjadi oleh beberapa hal, salah satunya adalah perubahan harga input yang digunakan dalam memproduksi barang dan jasa (Santi Dharmastuti, 2021).

Kebijakan PPKM memperbaiki gerak laju ekonomi Indonesia selama 2 tahun dilanda Covid-19, tak dapat dipungkiri bahwa Covid-19 permasalahan kesehatan, krisis kesehatan ini selanjutnya memberikan dampak penurunan aktivitas ekonomi yang diwarnai dengan sejumlah persoalan yang memiliki karakteristik seperti pada yang terjadi dalam sebuah resesi ekonomi (Shanti Darmastuti, 2022).

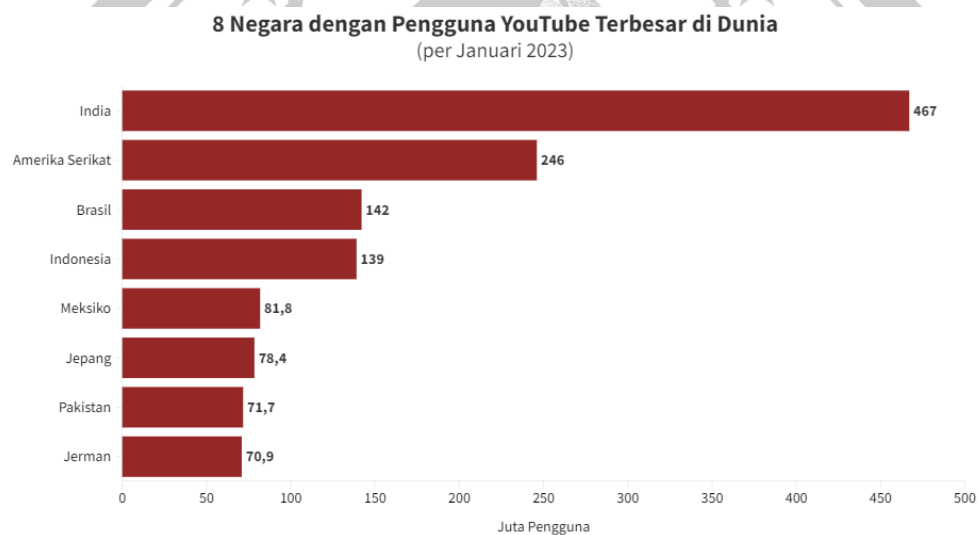
Pada tahun 2022 pemerintah melaporkan Kuartal I dan Kuartal II meningkat secara baik, Pada kuartal III terdengar isu resesi yang memunculkan ketakutan ekonomi di beberapa negara, ditambah dengan kondisi ekonomi yang disebabkan oleh perang Ukraina dan Rusia. Indonesia salah satunya negara yang terdampak, hal ini disebabkan oleh kebijakan moneter Bank Dunia menaikkan suku bunga dalam meresponnya lonjakan inflasi, hal ini menjadi potensi terjadinya resesi, kebijakan moneter ini berpotensi berdampak pada pada krisis keuangan dan ekonomi negara berkembang. Pada Oktober 2022, melalui Rapat Dewan Gubernur Bank Indonesia, memutuskan untuk menaikkan suku bunga acuan sebesar 50 basis poin menjadi 4,75%, alasan ini adalah untuk meredam lonjakan inflasi. Naiknya suku bunga yang Semakin besar bunga yang ditawarkan pada produk tabungan dan deposito, maka semakin besar pula minat masyarakat menjadi nasabah bank atau sebagai penanam modal (investor), hal ini dapat mengurangi terjadinya uang beredar yang mengakibatkan terjadinya inflasi (Yenni Samri Juliati Nasution dkk, 2022).



Dampak positif dari inflasi yaitu dapat meningkatkan perekonomian ke arah yang lebih baik seperti membuat masyarakat lebih giat dalam bekerja dan menabung, dan meningkatnya pendapatan nasional. Selain berdampak positif, inflasi juga memiliki dampak negatif yaitu dapat menyebabkan perekonomian suatu negara kacau dan perekonomian menjadi lesu (Laura Nur Aini, 2022).

### 2.2.2 Youtube API

*Youtube* merupakan suatu layanan platform web yang dapat digunakan dalam berbagi video, dimana pengguna dapat memberikan komentar, like, berbagi dan mengunggah video, layanan ini digunakan oleh individu, organisasi dan perusahaan-perusahaan media. Di Indonesia pengguna *Youtube* terus meningkat, per Januari 2023 Indonesia menempati peringkat ke-empat dari urutan berbagai negara lainnya.



**Gambar 2.1 Data Pengguna Youtube**

*Application Programming Interface (API)* sebagai fungsi, kumpulan perintah dan juga protokol yang berguna untuk membantu programmer dalam membangun perangkat lunak dalam sistem operasi tertentu. Khususnya dalam *Youtube* memungkinkan pengembang mengakses statistik video dan data saluran *Youtube* melalui dua jenis panggilan, REST, dan XML-RPC. Google sendiri mendeskripsikan Sumber Daya API *Youtube* sebagai API dan Alat yang memungkinkan Anda membawa pengalaman *Youtube* ke laman web, aplikasi, atau perangkat Anda. (Pramana Yoga S, 2019). Pada dasarnya fungsi

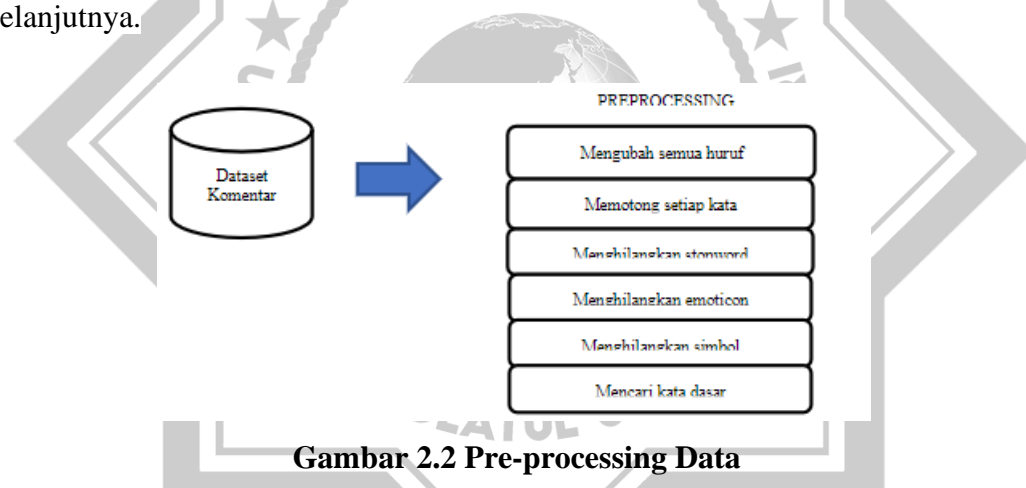
serta perintah pada *Youtube* API merupakan alat pemanggil sistem calls yang mana berhubungan langsung ke sistem operasinya.

### 2.2.3 Text Mining

Text mining merupakan ilmu yang bertujuan untuk memproses teks agar menjadi informasi, menambang suatu data yang berupa teks yang bersumber dari data tersebut (Luqyana, dkk. 2018). Data informasi yang berupa dokumen ini, dapat dicari kata-kata atau *keyword* untuk mewakili dari masing-masing data, sehingga dapat disimpulkan dan dianalisa hubungan antar data-data tersebut.

### 2.2.4 Text Processing

*Text Processing* merupakan tahapan awal yang dilakukan dalam *Text mining*, pada proses ini mempersiapkan data teks sehingga dapat diolah ke proses selanjutnya.



**Gambar 2.2 Pre-processing Data**

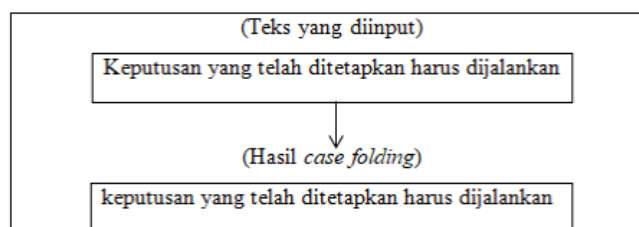
Pra-pemrosesan Data adalah langkah pertama dalam pembelajaran Mesin di mana data diubah / Dikodekan sehingga dapat dibawa dalam keadaan sedemikian rupa sehingga sekarang mesin dapat dengan cepat melewati atau mem-parsing Data tersebut. Dengan kata lain, dapat juga diartikan bahwa algoritma model dapat dengan cepat menganalisis fitur-fitur data. (Kiran Maharana, 2022). Adapun tahapan dalam *pre-processing* data sebagai berikut:

a) **Cleaning Data**

*Cleaning Data* tahapan memproses data tanggapan dari *emoticon*, tanda baca, hastag, *link*, dan karakter lainnya seperti tanda titik(.), koma(,), titik koma (;), mengurangi *noise* data dokumen untuk oleh sistem dan dengan baik

b) **Case Folding**

Pada proses *case folding* ini, data dokumen yang berupa huruf yang tidak teratur, tahap ini akan merubah data menjadi huruf kecil (*lowercase*) secara total. Huruf yang diubah mulai dari 'a' hingga 'z' (Damanik, 2014).



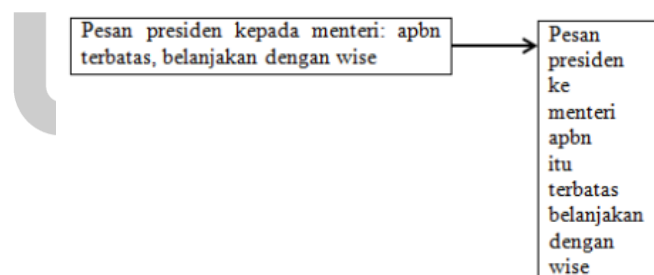
**Gambar 2.3** Proses *Case Folding*

c) **Tokenizing**

Tokenizing merupakan proses yang digunakan untuk memotong dokumen menjadi pecahan kecil yang dapat berupa kalimat, bab, dan kata pada proses ini akan menghilangkan whitespace (Luqyana, Dkk. 2018). Hasil pemotongan data pada proses ini disebut dengan *Token*.



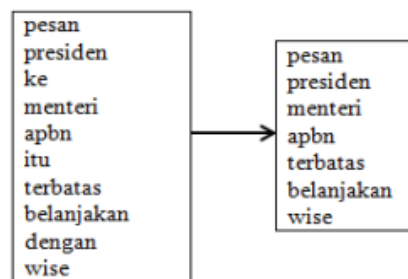
**Gambar 2.4** Proses Tokenizing



**Gambar 2.5** Proses *Tokenizing*

d) **Filtering,**

Pada tahapan *filtering* adalah tahapan penghapusan kata yang kurang penting, seperti kata ganti, kata keterangan, kata hubung dan berbagai lainnya menggunakan stopword, yakni 758 daftar kata yang akan dihapus pada data tanggapan terdapat pada lembar lampiran penelitian.



**Gambar 2.6** Proses *Filtering*

### 2.2.5 Term Weighting

*Term Weighting* ini adalah tahapan pembobotan kata, dalam tahapan ini menggunakan 2 tahapan, *Term Frequency (TF)* & *Inverse Document Frequency (IDF)* pada kata dari data tanggapan yang telah melalui tahapan sebelumnya. Pada tahapan ini adalah proses perhitungan pada kata dalam sebuah data, dan dapat menjadi sebuah acuan secara statistic, dan dapat menjadi ukuran seberapa penting sebuah kata dari data yang ada. *Term frequency (TF)* merupakan proses untuk menghitung jumlah kemunculan *term* dalam satu dokumen dan 17 *Inverse Document Frequency (IDF)* digunakan untuk menghitung *term* yang muncul diberbagai dokumen (komentar). (Farach&Nugraha, 2020).

1. Hitung *term frequency*  $tf_{t,j}$

$$tf_{t,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \quad (2.1)$$

$tf_t$ , = Frekuensi *term*

$n_t$ , = banyaknya kata i

## 2. Hitung *Document Frequency* (*df*)

$$P(X|Z) = \frac{P(Z|X)P(X)}{P(Z)} \quad (2.2)$$

*idf*, = *Inverse Document Freuquency*

*N* = jumlah dokumen

*df<sub>i</sub>* = jumlah frekuensi dokumen yang mengandung *term*

## 3. Perhitungan bobot TF-IDF

$$W_{i,j} = tf_{i,j} \times idf_i \quad (2.3)$$

Keterangan :

*W<sub>i,j</sub>* = Bobot TF-IDF

*idf<sub>i</sub>* = *Inverse Document Frequency*

*tf<sub>i,j</sub>* = Frekuensi suatu kata

### 2.2.6 Wordcloud

*Wordcloud* adalah metode data atau *Text Mining* yang menampilkan grafis data dari kata yang muncul tentang suatu objek yang sedang digali datanya (Limbong, 2022).



Gambar 2.7 Tampilan *Word Cloud*

### 2.2.7 Sentiment Analysis

*Sentiment Analysis* adalah proses mendeteksi kecenderungan seseorang melalui tulisan atau teks atau mengekstrak data teks untuk mendapatkan informasi tentang sentimen yang biasanya dituliskan didalam media sosial baik bermakna positif, negatif atau netral. Proses analisis ini biasanya dilakukan secara otomatis melalui beberapa aplikasi atau platform yang mendukung untuk dilakukannya analisis teks (Herwinsyah 2022). *Sentiment Analysis* dilakukan untuk melihat pendapat atau kecenderungan pendapat terhadap suatu masalah atau objek oleh seseorang, apakah cenderung berpendapat negatif atau positif. *Sentiment Analysis* biasa dilakukan untuk memantau perkembangan pasar atau menanggapi suatu permasalahan, salah satu contoh penggunaannya di dunia nyata adalah indentifikasi kecenderungan pasar 24 atau pendapat terhadap suatu objek. Analisis sentimen juga menganalisis sebagian data untuk mengetahui emosi manusia.

Contoh perhitungan skor sentimen sebagai berikut. (Santoso & Nugroho, 2019). Berdasarkan teks tanggapan “Jalan tol yang sudah dikerjakan sangat bagus dan indah, tapi untuk pembayaran tolnya sangat mahal”, terdapat 1 kata negatif dan 2 kata positif yang terdeteksi pada kamus *lexicon*, yaitu “bagus” dan “indah” sebagai kata positif, untuk kata negatif yaitu “mahal” sebagai kata negatif.

Dengan rumus perhitungan sebagai berikut:

$$\text{Skor} = (\sum \text{kata positif}) - (\sum \text{kata negatif})$$

**Tabel 2. 2** Perhitungan Kata

	Jumlah Kata Positif	Jumlah Kata Negatif
Jalan tol yang sudah dikerjakan sangat bagus dan indah, tapi untuk pembayaran tolnya sangat mahal	Bagus	Mahal
	Indah	
<b>Total Jumlah Kata</b>	2	1

Maka dari hasil perhitungan skor: Skor = (Jumlah kata positif) – (Jumlah Kata Negatif) Skor = 2-1 = 1, maka pada skor perhitungan menghasilkan skor 1, >0 maka dikategorikan pada positif.

### 2.2.8 Asosiasi Kata

Asosiasi kata dapat digunakan untuk mengetahui kata apa saja yang sering muncul pada sebuah dokumen. Asosiasi kata juga dapat mengetahui keterkaitan dan hubungan antar kata, misalnya antar dua kata atau lebih digunakan secara bersamaan dalam sebuah dokumen. Dalam asosiasi kata dapat juga dilihat dari nilai korelasi antar kata, dimana nilai korelasi berkisar antara -1 sampai 1. Jika nilai mendekati 1 atau -1 maka hubungan antar kata tersebut semakin kuat, sedangkan jika nilai mendekati 0 maka hubungan antar kata semakin lemah. Ada beberapa kategori nilai korelasi yang digunakan sebagai berikut. (Farach dan Nugraha, 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 $\Sigma xi =$

$$r_{xy} = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{(n\Sigma x_i^2 - (\Sigma x_i)^2)\{n\Sigma y_i^2 - (\Sigma y_i)^2\}}} \quad (2.4)$$

Dengan  $r_{xy}$  = nilai korelasi antar variabel x dan variabel y

n = banyaknya pasangan data x dan y

$\Sigma x_i$  = jumlah nilai pada variabel , i = 1,2,3,....., n

$\Sigma y_i$  = 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

### 2.2.9 Naïve Bayes Classifier

Naïve Bayes merupakan algoritma pembelajaran yang sering digunakan untuk mengatasi masalah klasifikasi teks, merupakan salah satu metode pembelajaran mesin yang menggunakan perhitungan probabilitas (P. Jurnal), 2020). Ciri utama

dari Naïve Bayes Classifier ini adalah asumsi yang sangat kuat (naif) terhadap tingkat independensi dari masing-masing kondisi atau kejadian. Terdapat dua tahap klasifikasi dokumen tweet pada penelitian ini. Tahap pertama adalah proses training terhadap dokumen yang sudah diketahui kategorinya. Sedangkan tahap kedua adalah proses testing yaitu mengklasifikasikan dokumen yang belum diketahui kategorinya. Terlepas dari kesederhanaannya, pengklasifikasi Naive Bayesian sering kali bekerja dengan sangat baik dan banyak digunakan karena sering mengungguli metode klasifikasi yang lebih canggih (Wagh & Shinde, 2016)

Metode Naïve Bayes Classifier dapat disimpulkan bahwa tingkat keakurasian dalam metode ini sangat stabil dan memiliki nilai akurasi yang sangat baik (Astari, Dkk, 2020)

Bayes merupakan teknik prediksi berbasis probabilitas sederhana yang berdasar pada penerapan teorema *Naïve Bayes* dengan asumsi independensi (ketidaktergantungan yang kuat (naif). Dengan kata lain, dalam *Naïve Bayes* yang digunakan adalah “model fitur independen” (Pramana Yoga Saputra Dkk, 2019).

$$P(C_j|W_i) = \frac{P(W_i|C_j)P(C_j)}{P(W_i)} \quad (2.5)$$

- $P(C_j|W_i)$  : peluang kategori  $j$ , Ketika terdapat kemunculan  $i$   
 $P(C_j|W_i)$  : Peluang kata  $i$  masuk ke kategori  $j$   
 $P(C_j)$  : Peluang kemunculan kategori  $j$   
 $P(W_i)$  : Peluang kemunculan kata

UNUGIRI



Cara yang paling sederhana dalam menghitung  $P(C_j|W_i)$  yakni dengan

$$P(W_i|C_j) = \frac{N_{ic}+1}{N_c+V} \quad (2.6)$$

$N_{ic}$  : Jumlah dokumen data latih dengan fitur  $W_i$  dengan kategori  $C_j$

$N_c$  : Jumlah dokumen dari kategori  $C_j$

$V$  : Jumlah kategori

