

## BAB II

### TINJAUAN PUSTAKA

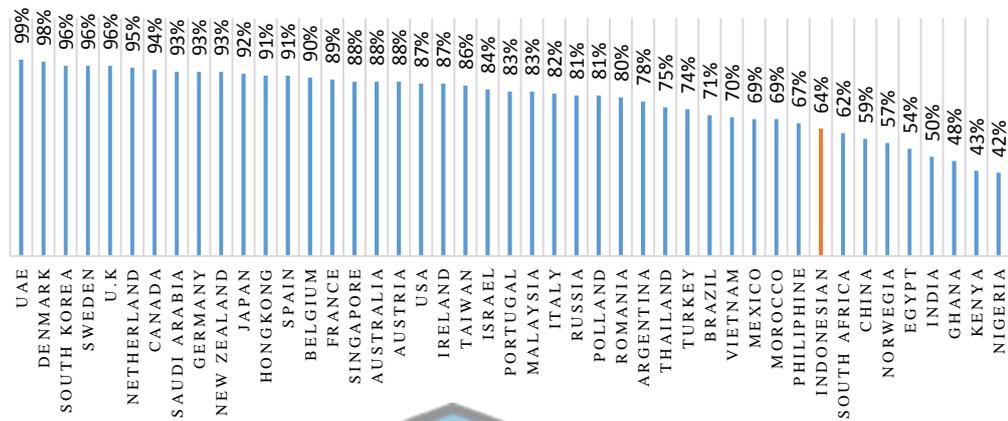
#### 2.1 Tinjauan Teoritik

##### 2.1.1 Media Sosial

Media sosial merupakan sebuah platform yang memungkinkan orang melakukan interaksi secara online di dunia maya. Menurut Anugratami, Maylanny dan Berlian (2015), media sosial merupakan sebuah media online dengan para penggunanya dapat dengan mudah berpartisipasi, berbagi, bahkan menciptakan isi meliputi blog, jejaring sosial, wiki, forum dan dunia virtual.

Data dari *We are social and Hootsuite* tentang lanskap digital tahun 2020, Indonesia sebagai negara yang terletak di wilayah Asia Pasifik menjadi bagian dari 4.3 milyar total penduduk di wilayah tersebut, sekitar 56% atau 2.42 milyar diantaranya sudah mendapatkan akses internet dan tepat 2.14 milyar penduduk telah menggunakan sosial media. Dibawah ini persentase pengguna internet di dunia pada tahun 2020 yang dikutip dari *We are social and Hootsuite* bulan Januari 2020, yang ditunjukkan pada Gambar 2.1.

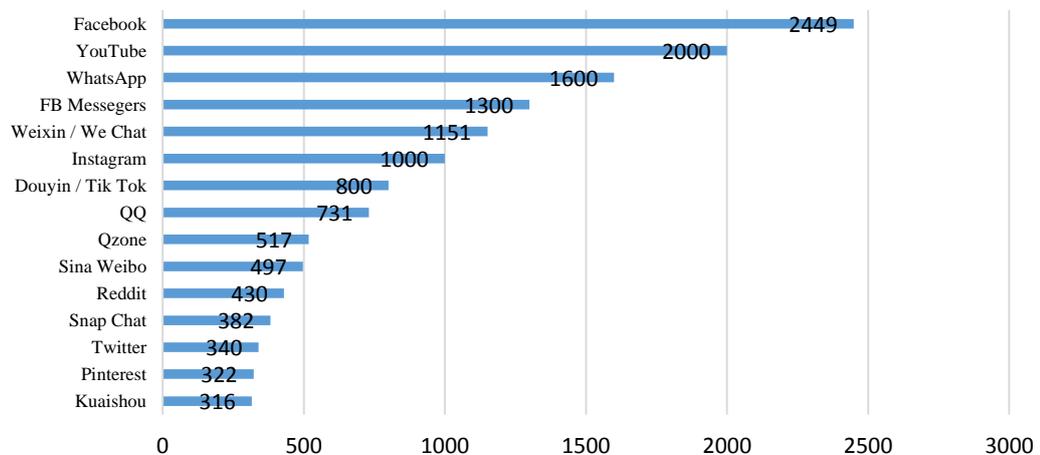
### Persentase Pengguna Internet



**Gambar 2.1 Persentase Pengguna Internet**

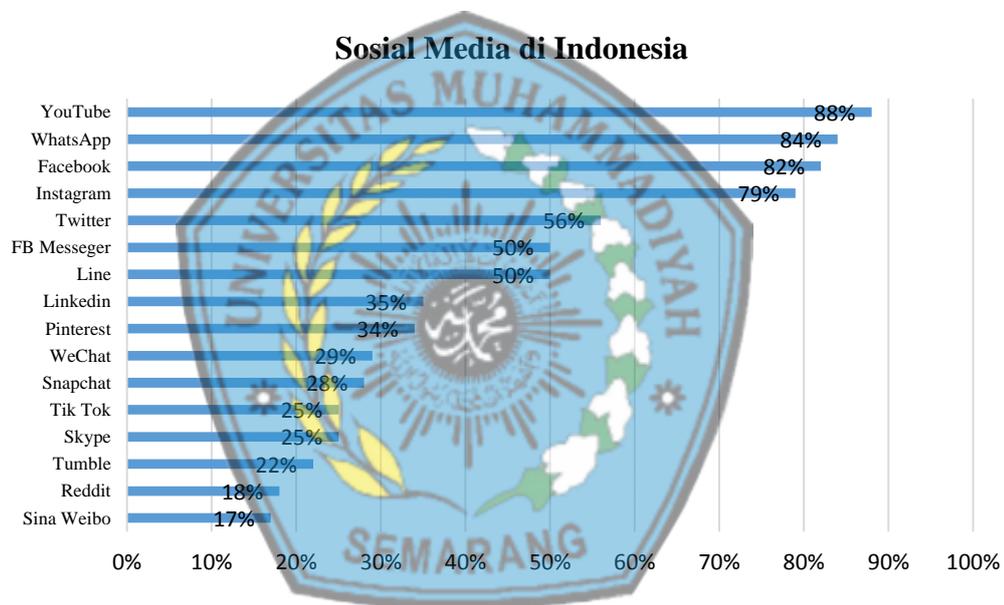
Berdasarkan grafik persentase pengguna internet di dunia tahun 2020, Indonesia merupakan salah satu negara yang menggunakan fasilitas jejaring internet yang signifikan sebesar 64%. Sedangkan untuk persentase pengguna internet tertinggi yaitu U.E.A (Uni Emirat Arab) sebesar 99%. Dengan adanya internet tentunya berkaitan dengan penggunaan sosial media.

### Sosial Media di Dunia



**Gambar 2.2 Sosial Media di Dunia**

Pada Gambar 2.2 sosial media paling banyak digunakan di dunia dikutip oleh *We are social and Hootsuite* bulan Januari 2020 dan dapat dilihat bahwasannya secara peningkatan *Facebook* masih menjadi sosial media yang paling populer di dunia dengan 2449 milyar akun, disusul oleh pengguna *YouTube*, sedangkan sosial media *Twitter* masuk dalam 15 sosial media terpopuler di dunia. Sedangkan data sosial media terpopuler di Indonesia dilansir oleh *We are social and Hootsuite* sebagai berikut:



**Gambar 2.3 Sosial Media di Indonesia**

Gambar 2.3 menjelaskan bahwa sosial media di Indonesia yang terpopuler saat ini adalah 88% masyarakat Indonesia menggunakan *YouTube*, yang kedua pengguna *WhatsApp* 84% masyarakat Indonesia. Sedangkan pengguna *Twitter* masuk dalam 5 besar pengguna sosial media di Indonesia dengan 56%.

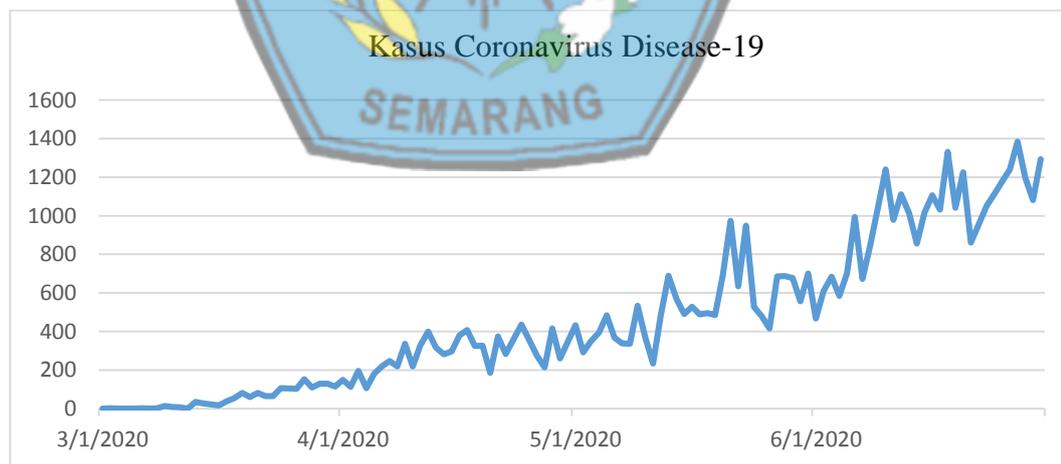
Dari data di atas, menggambarkan bagaimana sosial media saat ini menjadi layanan yang cukup dominan didunia maya. Kekuatan sosial media memang

terjadi tidak bisa dihindarkan serta tidak heran jika sosial media begitu intensif dan menyita banyak waktu.

### 2.1.2 *Coronavirus Disease (Covid-19)*

*Coronavirus Disease* 2019 atau disingkat dengan Covid-19, penyakit ini disebabkan oleh koronavirus jenis baru yang diberi nama SARS-Cov-2. Wabah covid-19 pertama kali dideteksi di Kota Wuhan, China pada bulan Desember 2019, dan ditetapkan sebagai pandemik oleh Organisasi Kesehatan Dunia (WHO) pada tanggal 11 Maret 2020.

Hingga 30 Juni 2020 lebih dari 216 negara atau wilayah kasus covid-19 telah dilaporkan, mengakibatkan lebih dari 500 ribu orang meninggal dunia (WHO,2020). Indonesia merupakan salah satu negara yang terkonfirmasi terpaparnya covid-19 yang masuk pada awal Maret 2020. Berikut ini data kasus terpaparnya covid-19 di Indonesia.



Sumber Data: Gugus Percepatan Penanganan Covid-19

**Gambar 2.4 Kasus Covid-19 di Indonesia**

Berdasarkan Gambar 2.4 bahwa jumlah kasus covid-19 di Indonesia dari awal kemunculannya bulan Maret 2020 sampai 30 Juni 2020 *update* data yang

dikeluarkan oleh Gugus Percepatan Penanganan Covid-19 sudah 56.325 terkonfirmasi terpapar virus tersebut. Dengan melihat grafik tersebut kasus terkonfirmasi covid-19 di Indonesia masih bersifat fluktuatif.

### **2.1.3 Kebijakan Pemerintah**

Dengan data terkonfirmasi kasus covid-19 di Indonesia yang masih bersifat fluktuatif. Pemerintah mengeluarkan sejumlah kebijakan untuk menangani wabah covid-19 yang saat ini melanda Indonesia. Penyakit yang disebabkan ini tidak hanya mendatangkan masalah dalam aspek kesehatan masyarakat, namun juga geliat perekonomian dari ranah mikro hingga makro. Untuk itu sejumlah kebijakan dicetuskan oleh pemerintah seperti, keringanan biaya listrik, pembatasan sosial berskala besar (PSBB), larangan mudik, pembebasan nabi, kartu pra kerja, keringan kredit, dan lain sebagainya (Azanella, 2020). Kebijakan yang paling disoroti yaitu dengan adanya pemberlakuan Pembatasan Sosial Berskala Besar (PSBB).

Pembatasan Sosial Berskala Besar (PSBB) sudah tercatat di dalam peraturan Menteri Kesehatan Nomor 9 Tahun 2020 tentang pedoman pembatasan sosial berskala besar dalam rangka percepatan penanganan corona virus disease 2019 (COVID-19) tanggal 3 April 2020. PSBB merupakan pembatasan dalam kegiatan-kegiatan di tempat umum dan melakukan karantina di rumah. Tujuan dari PSBB itu sendiri untuk mencegah meluasnya penyebaran covid-19 yang sedang terjadi saat ini. Pembatasan kegiatan yang dilakukan diantaranya liburan sekolah dan tempat kerja, pembatasan kegiatan keagamaan dan

pembatasan kegiatan-kegiatan umum. Tentunya dengan kebijakan pemberlakuan PSBB tersebut ada dampak positif maupun negatif.

Menurut Putri (2020) dilansir oleh Kompas.com manfaat dan dampak positif pemberlakuan PSBB selain memutus penyebaran covid-19, sebagai berikut:

1. Kualitas udara membaik
2. Solidaritas masyarakat meningkat
3. Munculnya relawan melawan covid-19
4. Kepedulian tinggi terhadap tenaga medis
5. Dan lain-lain

Menurut Armani (2020), dilansir oleh Kompasiana dampak yang timbul akibat diberlakukannya PSBB, sebagai berikut:

1. Banyak perusahaan yang bangkrut
2. Terjadi PHK besar-besaran
3. Bidang Industri tersendat
4. Bidang pariwisata sepi
5. Angka kemiskinan bertambah
6. Dan lain-lain

Seiring dengan kebijakan pemerintah yang sedang dilakukan saat ini tidak luput dari komentar publik dalam media sosial. Kebijakan yang menuai berbagai macam komentar, mulai dari pujian, kritik, saran bahkan sindiran. Berdasarkan hasil survei Indo Barometer yang diberitakan melalui laman kompas.com disebutkan sebanyak 40.3% responden yang menolak PSBB dengan alasan

kesulitan mencari nafkah. Secara keseluruhan, hasil survey tersebut menunjukkan bahwa 10% respon yang menganggap bahwa penerapan PSBB merupakan langkah yang tidak tepat, sementara 89.5% responden menilai bahwa penerapan PSBB sudah sanga tepat (Purnamasari, 2020).

#### **2.1.4 Twitter**

*Twitter* merupakan salah satu media jejaring sosial yang dapat digunakan dengan mendaftarkan diri sebagai pengguna twitter melalui website [www.twitter.com](http://www.twitter.com) dan memasukkan email untuk melakukan registrasi. Awalnya *twitter* membatasi penggunaannya untuk mengirim dan membaca teks hanya 140 karakter, namun pada akhir bulan September 2017 *twitter* memperbanyak jumlah karakternya menjadi 280 karakter. *Twitter* didirikan oleh Jack Dorsey pada tanggal 21 Maret 2006 di San Francisco, California, Amerika dan mulai diluncurkan situs jejaring sosialnya pada tanggal 15 Juli 2006. *Twitter* menyediakan wadah bagi penggunaannya untuk dapat membuat *Tweet* pada akun profil pengguna dan *Tweets* tersebut dapat ditampilkan pada beranda teman yang mengikutinya (*follower*). Selain membuat *Tweets*, pengguna *twitter* juga dapat melihat *Tweets* dari akun teman yang diikutinya (*following*). Pengguna *twitter* juga dapat mencari *Tweets* mengenai topik tertentu dengan cara menuliskan *keyword*, misalnya “PSBB”, maka akan muncul beberapa postingan pengguna-pengguna *twitter* yang memuat kata PSBB. Beberapa istilah atau simbol yang digunakan dalam *Twitter* antara lain :

1. *Tweet*, yaitu artikel atau tulisan atau postingan di dalam *Twitter*.

2. *DM/Direct Message*, yaitu pesan pribadi antara kedua pengguna/pemilik akun *Twitter*.
3. *Follow*, yaitu mengikuti pengguna *Twitter* lainnya. Pengguna akun *Twitter* yang mengikuti akan melihat semua *Tweet* terbaru dari pengguna akun *Twitter* yang diikutinya.
4. *Following*, yaitu memilih untuk melihat *Tweet* orang lain pada *timeline*.
5. *Handle*, yaitu akun *Twitter* disebut dalam pegangan dan biasanya diidentifikasi dengan simbol @.
6. *Lists*, yaitu cara untuk menggabungkan beberapa *Tweet* dari orang-orang atau perusahaan yang diikuti di satu tempat. *List* bisa berisikan *Tweet* dari teman dekat, kompetitor, orang-orang dari komunitas tertentu ataupun yang lainnya. *List* adalah cara yang bagus untuk fokus ke berita atau *Tweet* tertentu.
7. *RT/ReTweet*, yaitu melanjutkan (*forward*) *Tweet* yang diterima di *timeline Twitter*.
8. *Search*, yaitu kolom untuk mencari sesuatu yang dibutuhkan di dunia *Twitter*, karena banyak *Tweet* bersifat umum, maka *search* dapat digunakan untuk mencari apapun.
9. *Timeline*, yaitu daftar kronologi dari *Tweet*, orang yang diikuti atau *list*.
10. *Trending Topics*, yaitu *Tweet* yang sedang *tren* yang ada di *Twitter* berbasis web.
11. *Tweetup*, yaitu bahasa yang digunakan oleh pengguna *Twitter* jika mereka ingin bertemu di dunia nyata, yang biasanya pemimpin *Tweetup* membuat *hashtag* tertentu.

12. *Unfollow*, yaitu melakukan *unfollow* sehingga tidak akan menerima *Tweet* terbaru lagi di timeline dari pengguna *Twitter* yang sebelumnya diikuti.

13. *Hash Tag*, disimbolkan dengan # yang mana digunakan untuk menandai katakunci atau topik dalam sebuah *Tweet* dan diciptakan secara organik oleh pengguna *Twitter* sebagai cara untuk mengkategorikan pesan *Tweet*.

## 2.2 Tinjauan Statistika

### 2.2.1 Analisis Deskriptif

Statistik deskriptif adalah metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang bermakna. Tujuan dari statistik deskriptif untuk mendeskripsikan atau memberikan gambaran objek yang diteliti serta analisis deskriptif digunakan untuk menyajikan data dengan cara membuat tabulasi penyajian dalam bentuk grafik, diagram atau menyajikan dengan karakteristik-karakteristik dari ukuran pemusatan dan keragamannya (Sumanto, 2009)

### 2.2.2 Analisis Sentimen

Analisis sentimen *twitter* didefinisikan bagian dari pendapat pada media *twitter*. Pesan yang terdapat dari *twitter* yang pada kenyataannya, lebih mudah untuk menganalisis karena penulisannya dibatasi dibandingkan forum diskusi. Hal ini forum diskusi lebih sulit dikarenakan pengguna dapat mendiskusikan berbagai macam dan berinteraksi satu sama lain.

Pada dasarnya analisis sentimen yang merupakan tahapan klasifikasi, namun tahapan analisis sentimen pada *twitter* (tidak terstruktur) yang sedikit lebih sulit dibandingkan dengan klasifikasi dokumen yang sudah terstruktur. Menurut

Liu (2012), tahapan pertama dalam kasus analisis sentimen pada *twitter* adalah dengan mengklasifikasikan apakah kalimat yang mengungkapkan pendapat atau tidak. Selanjutnya tahapan kedua dengan mengklasifikasikan kalimat-kalimat pendapat menjadi positif ataupun negatif. Fitur-fitur *twitter* yang biasa digunakan dalam proses analisis sentimen yaitu *reTweets*, *hashtag* (#), link, *emoticon*, tanda seru ataupun tanda tanya (Barbosa dkk, 2010).

### 2.2.3 *Text Mining*

*Text mining* merupakan salah satu cabang ilmu dari *data mining* yang berguna untuk menganalisa data berupa dokumen teks (Prilianti dan Wijaya, 2014). Menurut Mariyanah (2018) yang dikutip dalam buku (Pramana dkk, 2018) menjelaskan bahwa *text mining* merupakan kumpulan dari proses-proses penambangan seperti pengumpulan, pemrosesan, analisis dan visualisasi guna mendapatkan informasi berharga dari teks tersebut. Menurut (Han dan Kamber, 2012), *text mining* adalah satu langkah dari analisis teks yang secara otomatis dilakukan oleh komputer dengan tujuan menggali informasi yang berkualitas dari suatu rangkaian teks yang terangkum dalam sebuah dokumen.

Perbedaan antara *data mining* dan *text mining* adalah pada data yang digunakan, *data mining* menggunakan data yang terstruktur (*structured data*) sedangkan *text mining* menggunakan data yang umumnya adalah *unstructured data* atau *semistructured*. Pada *unstructured data* atau *semistructured data* tentu akan sulit menggali suatu informasi, selain itu permasalahan yang terdapat dalam *text mining* adalah jumlah data yang besar, dimensi yang tinggi, serta adanya data

*noise* pada teks. Langkah awal sebelum dimulainya suatu analisis dalam data tekstual perlu dilakukan *pre-processing* teks.

Tujuan dari *text mining* itu sendiri untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Adapun tugas khusus dari *text mining* itu sendiri yaitu pengkategorisasian teks (*text classification*) dan pengelompokan teks (*text clustering*) (Syadid, 2019). *Text clustering* berhubungan dengan proses struktur kelompok yang belum terlihat dari sekumpulan dokumen. Sedangkan *Text classification* yaitu suatu proses untuk membentuk kelas-kelas dari suatu dokumen yang didasarkan pada kelas yang sudah diketahui sebelumnya (Darujati dan Agustinus, 2012).

#### 2.2.4 Pre-Processing Text

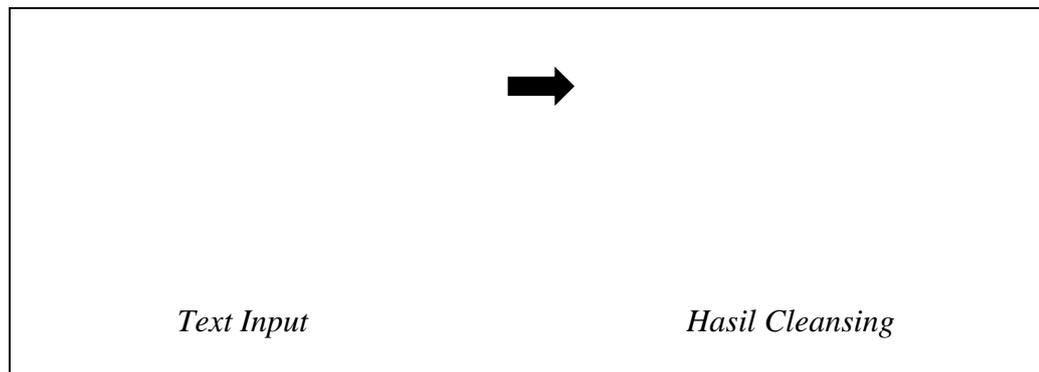
*Pre-processing* adalah suatu proses dalam mempersiapkan dokumen teks yang tidak terstruktur menjadi dokumen yang terstruktur dan selanjutnya proses klasifikasi dapat diproses dengan baik (Ipmawati, Kusriani, dan Taufik, 2017). Tahapan *text processing* yang akan dilakukan sebagai berikut:

##### a. *Cleansing*

Suatu proses untuk membersihkan *Tweet* yang tidak diperlukan untuk mengurangi *noise*. Penghilangan karakter HTML, kata kunci, ikon emosi, *hashtag* (#), *username* (@username), url (<http://situs.com>) ataupun e-mail ([nama@situs.com](mailto:nama@situs.com)) (Aditya, 2015).

#KarantinaWilayah dinilai lebih manusiawi ketimbang #PSBB terbukti di Pasal 55 ayat 1 dan 2 menyebutkan "Kebutuhan dasar hidup orang dan makanan hewan ternak menjadi tanggung jawab Pemerintah Pusat dengan melibatkan Pemerintah Daerah." #KarantinaWilayah or

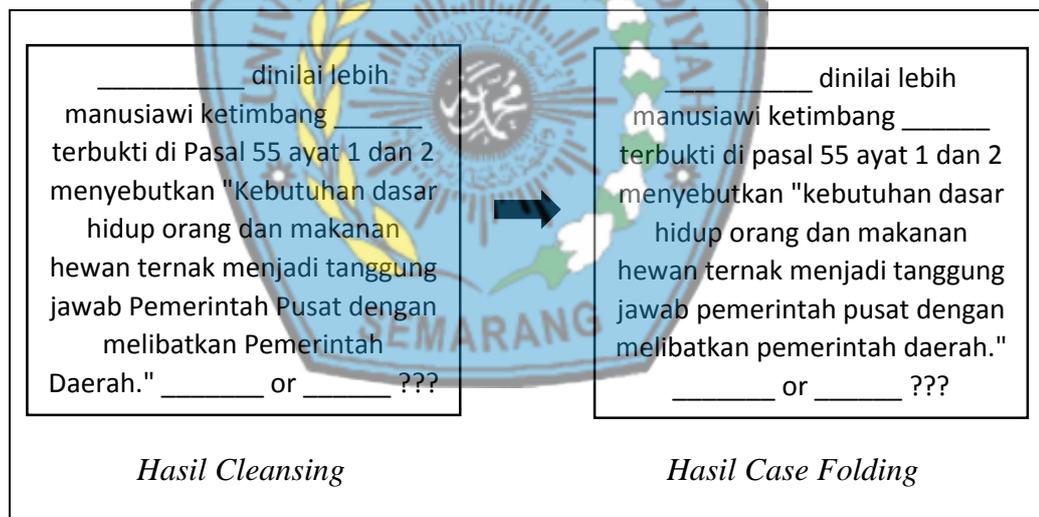
\_\_\_\_\_ dinilai lebih manusiawi ketimbang \_\_\_\_\_ terbukti di Pasal 55 ayat 1 dan 2 menyebutkan "Kebutuhan dasar hidup orang dan makanan hewan ternak menjadi tanggung jawab Pemerintah Pusat dengan melibatkan Pemerintah Daerah." \_\_\_\_\_ or \_\_\_\_\_ ???



**Gambar 2.5 Proses Cleansing**

**b. Case Folding**

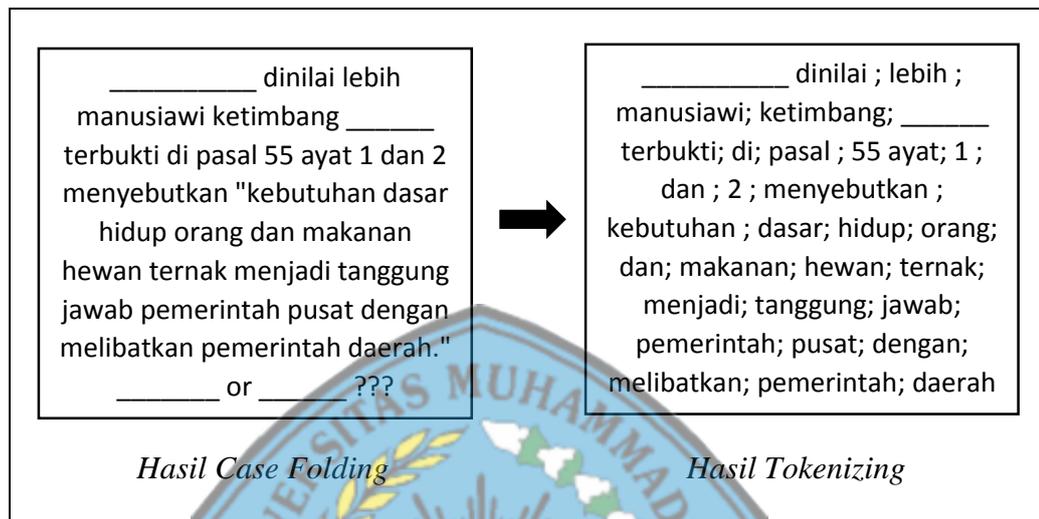
Proses yang digunakan untuk merubah setiap kata menjadi sama, misalnya merubah huruf besar (*uppercase*) didalam komentar menjadi huruf kecil (*lowcase*) (Salam, Junta, dan Rima, 2018).



**Gambar 2.6 Proses Case Folding**

**c. Tokenizing**

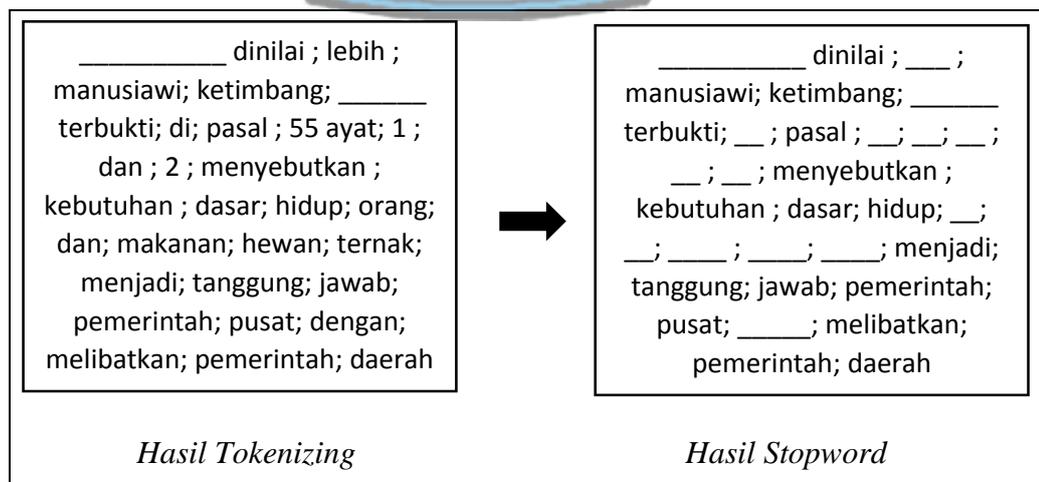
Proses untuk memenggal setiap kata pada suatu kalimat dan pada saat bersamaan membuang karakter tertentu yang dianggap sebagai tanda baca, misalkan, koma (,) ataupun titik (.) (Asiyah, 2016).



Gambar 2.8 Proses *Tokenizing*

#### d. *Stopword*

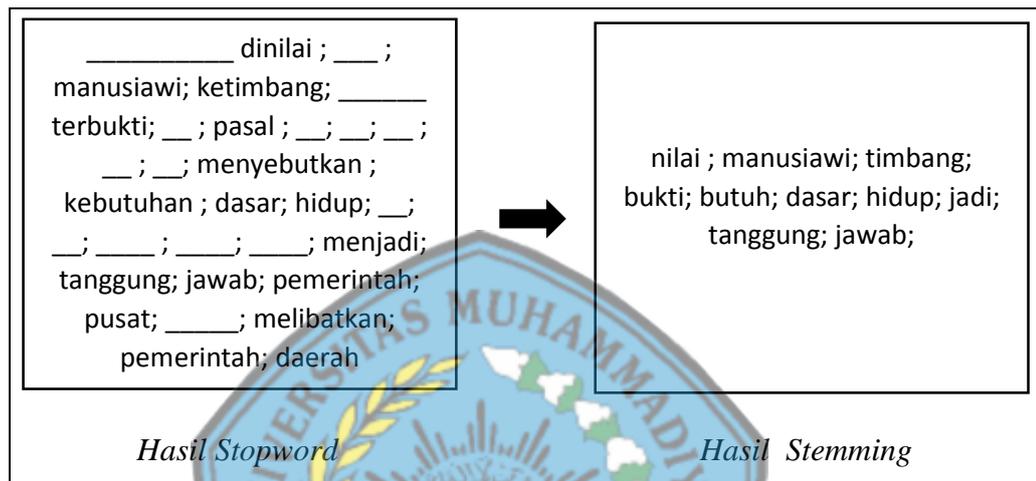
Suatu proses pengecekan setiap kata pada suatu komentar, selanjutnya penghapusan kata-kata yang terlalu umum dan kurang penting. Ciri-ciri pada kata ini biasanya frekuensi kemunculan yang jumlahnya cukup banyak.



Gambar 2.8 Proses *Stopword*

#### e. *Stemming*

Proses untuk mendapatkan kata dasar pada suatu kalimat dengan cara menghilangkan awalan, akhiran, sisipan, ataupun kombinasi ketiganya. Kata dasar disini yang mengandung sebuah makna dalam sebuah komentar sehingga lebih spesifik dalam pengkategorian.



**Gambar 2.9** Proses *Stemming*

### 2.2.5 Pembobot TF-IDF

*Term Frequency-Inverse Document Frequency* (TF-IDF) adalah metode yang digunakan untuk memberikan bobot pada setiap kata yang telah diekstrasi. Model pembobotan dari TF-IDF (2.1) merupakan model integrasi dari model *term frequency* dan *inverse document frequency*. *Term Frequency* (TF) adalah proses untuk menghitung jumlah kemunculan *term* dalam suatu dokumen. *Inverse Document Frequency* (IDF) berfungsi mengurangi bobot suatu *term* jika kemunculannya banyak tersebar diseluruh koleksi dokumen (Nurjannah, Hamdani, dan Astuti, 2013).

Tahapan pembobotan dengan TF-IDF adalah:

$$W_{at} = TF_{at} * IDF_{ft} \quad (2.1)$$

Dimana,

$d$  : dokumen ke- $d$

$t$  : kata ke- $t$  dari kata kunci

$W$  : bobot dokumen ke- $d$  terhadap kta ke- $t$

$tf$  : banyaknya kata yang muncul dalam dokumen

IDF : nilai idf dari hasil

$$IDF = \log \left( \frac{D}{df_t} \right) \quad (2.2)$$

$D$  : total dokumen

$df$  : banyaknya dokumen yang muncul berdasarkan kata kunci.

### 2.2.6 Klasifikasi

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia (Astuti dan Guntara, 2018). Klasifikasi memiliki sifat *Supervised Learning* yang artinya suatu proses pembelajaran adanya arahan (*supervised*), maksudnya metode ini diterapkan adanya latihan (*training*) dan ada guru (*teacher*) serta memerlukan target output. Dalam *Supervised Learning* ini adanya data *training* yang artinya data yang digunakan untuk membuat model, pola atau rumus. *Supervised Learning* juga mempunyai label, kelas atau yang lebih dikenal adanya variabel  $Y$  (Han, Kamber, dan Pei, 2012). Menurut Ibnu Santoso dan Budi Yuniarto dalam (Pramana, dkk, 2018) banyak metode yang termasuk dalam kategori *supervised learning* seperti, *K-Nearest Neighbour*, *Naïve Bayes*, *Support Vector Machine*, *Decision Tree*.

### 2.2.7 Naïve Bayes Classifier

Teorema Bayes merupakan teorema yang mengacu pada probabilitas bersyarat (Siang, 2005). Secara umum teorema Bayes dapat dinotasikan pada persamaan berikut.

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)} \quad (2.3)$$

Algoritma Naïve Bayes Classifier (NBC) merupakan algoritma yang digunakan untuk mencari nilai probabilitas tertinggi untuk mengklasifikasi data uji pada kategori yang paling tepat (Feldman dan Sanger, 2007). Metode Naive Bayes Classification merupakan salah satu metode yang dapat mengklasifikasikan teks. Kelebihan NBC adalah algoritmanya sederhana tetapi memiliki akurasi yang tinggi. Terdapat dua tahap dalam klasifikasi *tweet*. Tahap pertama adalah pelatihan terhadap *tweet* yang telah diketahui kategorinya. Sedangkan tahap kedua adalah proses klasifikasi *tweet* yang belum diketahui kategorinya (Falalah dan Nur, 2015). Dalam algoritma NBC setiap dokumen direpresentasikan dengan pasangan atribut “ $a_1, a_2, \dots, a_n$ ” dimana  $a_1$  adalah kata pertama,  $a_2$  adalah kata kedua dan seterusnya. Sedangkan  $V$  adalah himpunan kategori *tweet*. Pada saat klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan ( $V_{MAP}$ ). Adapun persamaan  $V_{MAP}$  adalah sebagai berikut

$$V_{MAP} = \underset{A_j}{\operatorname{argmax}} P(v_j) \prod_{i=1}^n P(a_i|v_j) \quad (2.4)$$

Nilai  $P(v_j)$  dihitung pada saat *training*, didapat dengan rumus sebagai berikut:

$$P(v_j) = \frac{|doc j|}{|training|} \quad (2.5)$$

$|doc j|$  merupakan jumlah *Tweet* pada kategori  $j$  dalam *training*. Sedangkan  $|training|$  merupakan jumlah *Tweet* dalam data yang digunakan untuk *training*.

Setiap probabilitas kata  $\mathbf{a}_i$  pada setiap kategori  $P(\mathbf{a}_i|\mathbf{v}_j)$ , dihitung pada saat *training*.

$$P(\mathbf{a}_i|\mathbf{v}_j) = \frac{n_i+1}{|n+kosakata|} \quad (2.6)$$

Di mana,

$n_i$  : jumlah kemunculan kata  $\mathbf{a}_i$  dalam *Tweet* yang berkategori  $\mathbf{v}_j$

$n$  : banyaknya seluruh kata *Tweet* dengan kategori  $\mathbf{v}_j$

$|kosakata|$  : banyaknya kata dalam data *training*

### 2.2.8 *K-Nearest Neighbour*

*K-Nearest Neighbour* merupakan salah satu metode yang digunakan dalam pengklasifikasian. *K-Nearest Neighbour* adalah algoritma pengklasifikasian yang didasarkan pada analogi, yaitu membandingkan data uji dengan data pelatihan (Tan, 2006). Sedangkan menurut (Krisandi, dkk. 2013) algoritma K-NN menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari sampel uji yang baru. Data latih akan dibangun dengan memperhatikan keseimbangan dokumen satu sama lain. Adapun algoritma K-NN dapat dijelaskan (Kurniniawan, 2012) sebagai berikut:

1. Hitung jarak antara data sampel (data uji) dengan data latih yang telah dibangun. Salah satu persamaan dalam perhitungan jarak kedekatan dapat menggunakan persamaan 2.8 *Euclidean Distance*
2. Menentukan parameter nilai  $k$  = jumlah tetangga terdekat.
3. Mengurutkan jarak terkecil dari data sampel
4. Pasangkan kategori sesuai dengan kesesuaian

5. Cari jumlah terbanyak dari tetangga terdekat dengan persamaan 2.8 kemudian tetapkan kategori.

Jarak yang digunakan dalam penelitian ini adalah *Euclidean Distance*

$$d(x_i, x_j) = \sum_{i=1}^p (x_{ip} - x_{jp})^2 \quad (2.8)$$

Keterangan:

$d(x_i, x_j)$  : Jarak Euclidean

$x_{ip}$  : Data *testing* ke-i pada variabel ke-p

$x_{jp}$  : Data *training* ke-i pada variabel ke-p

$p$  : Dimensi variabel bebas

### 2.2.9 Evaluasi Performa

Menurut Prasetyo (2012), klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi terdapat dua proses yang dilakukan yaitu dengan membangun model untuk disimpan sebagai memori dan menggunakan model tersebut untuk melakukan pengenalan atau klasifikasi atau prediksi pada suatu data lain supaya diketahui di kelas mana objek data tersebut dimasukkan berdasarkan model yang telah disimpan dalam memori. Sistem dalam klasifikasi diharapkan mampu melakukan klasifikasi semua set data dengan benar, namun tidak dapat dipungkiri bahwa kesalahan akan terjadi dalam proses pengklasifikasian tersebut sehingga perlunya dilakukan pengukuran kinerja dari sistem klasifikasi tersebut. Umumnya, pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi (confusion matrix). Matriks konfusi merupakan tabel

pencatat hasil kerja klasifikasi. Contoh dari matriks konfusi untuk dua kelas (biner) dapat dilihat **Tabel 2.1**

**Tabel 2.1 Confusion Matriks**

		Prediksi	
		Positif	Negatif
Aktual	Positif	PP	PN
	Negatif	NP	NN

$$Akurasi = \frac{PP+NN}{PP+PN+NP+NN} \times 100\% \quad (2.9)$$

Pengukuran performa dilakukan untuk menentukan bagus/tidaknya sebuah klasifikasi. Terdapat beberapa cara untuk mengukur performa, beberapa cara yang sering digunakan adalah dengan menghitung akurasi, *recall*, dan *precision*. Akurasi merupakan persentase dari total dokumen yang teridentifikasi secara tepat dalam proses klasifikasi. *Recall* mengindikasikan sebagian kecil dari dokumen yang relevan diambil. *Precision* mengkuantifikasi fraksi dokumen diambil yang sebenarnya relevan, dalam contoh milik kelas sasaran.

$$recall = \frac{PP}{(PP+PN)} \quad (2.10)$$

$$F \text{ Measure} = \frac{2 \times recall \times precision}{(recall+precision)} \quad (2.11)$$

APER (*Apparent Error Rate*) atau yang disebut laju error merupakan ukuran evaluasi yang digunakan untuk melihat peluang kesalahan klasifikasi yang dihasilkan oleh suatu fungsi klasifikasi. Semakin kecil nilai APER maka hasil

pengklasifikasian semakin baik (Prasetyo, 2012). Formulasi untuk menghitung APER (Johnson dan Wichern, 2007) yaitu:

$$APER = \frac{PN+NP}{PP+PN+NP+NN} \times 100\% \quad (2.12)$$



