

BAB II

TINJAUAN PUSTAKA

2.1 Studi Pustaka

Penggunaan internet saat ini memudahkan kita dalam mencari dan mengakses informasi yang kita inginkan, mulai dari sosial, politik, kesehatan, jual beli dan masih banyak yang lainnya. Penggunaan Internet yang begitu mudah memudahkan kita dalam mengambil dan memilih informasi yang dibutuhkan, mulai dari persepsi, opini maupun ulasan banyak bertebaran di internet khususnya di media sosial. Media sosial menjadi sebuah wadah besar bagi para pengguna internet di seluruh dunia untuk saling berkomunikasi dan bertukar informasi. Dengan jumlah pengguna aktif media sosial yang begitu banyak, sehingga menghasilkan jumlah data yang besar. Munculnya teknologi *big data*, yaitu kumpulan data yang sangat besar, kompleks, dan tidak terstruktur, telah menjadi salah satu sumber daya utama yang dapat diproses untuk mendapatkan hasil emosional yang lebih akurat (Indrayuni, 2017).

Jumlah komentar dan opini sangat tinggi membuat pengguna internet sangat sulit untuk membaca teks komentar dan opini yang ada. Oleh karena itu, sebuah sistem dapat dirancang secara otomatis untuk mengklasifikasikan opini dan ulasan yang sesuai dengan kategorinya. Kategori sentimen dibagi menjadi positif, dan negatif. Sistem analisis sentimen dibangun menggunakan algoritma klasifikasi *Naive Bayes*. Fitur utama dari *Naive Bayes* yaitu untuk membuat asumsi yang sangat kuat tentang independensi setiap kondisi atau kejadian.

Kelebihan *Naive Bayes* yaitu untuk memproses dan mengklasifikasi sebuah data yang dapat disesuaikan dengan sifat dan kebutuhannya, hal tersebut sangat memudahkan para pelaku usaha untuk menganalisis permintaan pasar, untuk mengetahui kepuasan dari para pelanggan, serta untuk menjalin komunikasi baik antara pelaku usaha dengan para pelanggannya (Gunawan dkk., 2018). Metode analisis sentimen biasa digunakan untuk memudahkan para pelaku usaha untuk melihat tren permintaan pasar yang sedang terjadi, serta memudahkan untuk menjaga hubungan baik kepada para pelanggannya (Indrayuni, 2017).

Beberapa penelitian tentang analisis sentimen dengan menggunakan metode *Naive Bayes* telah banyak dipelajari secara ekstensif. Berbagai metode dan analisis yang berkaitan dengan analisis sentimen telah banyak digunakan untuk memecahkan permasalahan yang terjadi. Pada studi pustaka ini digunakan berbagai sumber jurnal internasional. Jurnal yang digunakan berasal dari *database Science Direct dan Google Scholar* (Magro dan Talamini, 2019).

Analisis sentimen merupakan proses untuk mengklasifikasikan dan menentukan polaritas dari sebuah teks dalam sebuah dokumen atau kalimat sehingga memudahkan seseorang dalam menentukan kategorinya baik sebagai positif, negatif, maupun netral (Dery dan Setiawan, Erwin Budi S.Si., 2019). Analisis sentimen banyak digunakan oleh para peneliti yang biasa mereka gunakan untuk mengambil dan menentukan kategori data yang mereka dapatkan termasuk kategori data yang positif, negatif atau netral, khususnya untuk penelitian yang berkaitan dengan ilmu komputer. Dalam jejaring sosial, seperti *Google, YouTube dan Twitter, Instagram, Facebook* dan masih banyak yang lainnya. Analisis sentimen sering digunakan untuk melihat dan menentukan persepsi yang dibuat oleh publik (Rifan dkk., 2019).

Analisis sentimen bisa disamakan dengan penambangan opini karena tugasnya yang berfokus untuk membandingkan antara opini positif atau negatif (Saputra dkk., 2019). Dalam analisis sentimen, pengambilan data dilakukan dengan cara menganalisis, memproses, dan mengekstrak data tekstual dalam suatu entitas, seperti layanan, produk, orang, fenomena, atau topik dari tren tertentu (Liu, 2012). Seperti fenomena bencana alam yang terjadi di Indonesia, pemerintah menggunakan analisis sentimen untuk memahami kebutuhan masyarakat segera setelah bencana, sehingga informasi yang diperoleh dapat membantu banyak pihak untuk mendukung pengambilan keputusan atau pilihan, seperti mengklasifikasikan berbagai kebutuhan masyarakat seperti obat-obatan, perawatan kesehatan, makanan, air dan listrik (Fahrur dkk., 2020).

Tokopedia menggunakan analisis sentimen untuk mengoptimalkan nilai informasi, mengolah pandangan dan pendapat yang diungkapkan pelanggan menjadi informasi yang dapat memberikan kritik, saran dan rekomendasi kepada

tokopedia, sehingga layanan yang mereka berikan lebih memuaskan pelanggan mereka (Pajri dkk., 2020). Analisis sentimen digunakan untuk membantu perusahaan mengetahui persepsi dan opini masyarakat dalam menilai sebuah produk yang mereka buat dan mereka jual melalui media sosial (Indrayuni, 2017).

Media sosial merupakan tempat para pelaku usaha memasarkan produknya, dan para pelaku usaha sering menggunakan jangkauan luasnya di media sosial untuk memasarkan produk yang mereka jual, antara lain elektronik, fashion, kosmetik, dan peralatan kecantikan yang banyak dipasarkan melalui media sosial (Kamhar dan Lestari, 2019). Analisis sentimen biasanya digunakan untuk menganalisis dan mengetahui opini maupun sentimen yang sering muncul dalam ulasan. Serta menentukan persepsi konsumen terhadap suatu barang maupun produk berdasarkan pengalaman yang pernah dilakukan (Salsabila dan Trianasari, 2021).

Media sosial memudahkan pemilik restoran untuk memasarkan produknya. Restoran adalah tempat bisnis yang paling banyak dikomentari dan didiskusikan di media sosial berdasarkan opini pengalaman pribadi. Ada banyak aspek yang dapat mempengaruhi keputusan seseorang untuk mengunjungi restoran hanya dengan membaca ulasan orang lain, layanan yang ditawarkan dan banyak lainnya. Pemilik restoran biasanya menggunakan metode analisis sentimen untuk menganalisis keinginan dan kepuasan pelanggan (Indrayuni, 2017). Pemilik rumah makan Bakso Kepresidenan menggunakan analisis sentimen untuk mengetahui kepuasan pelanggan terhadap produk dan pelayanan yang diberikannya kepada pelanggan, dengan tujuan untuk mengevaluasi dan meningkatkan kualitas pelayanan yang diberikan kepada pelanggan oleh pihak restoran (Parasati dkk., 2020).

Melalui penelitian-penelitian sebelumnya dimungkinkan untuk dilakukan penelitian mengenai “Analisis Sentimen Ketertarikan Masyarakat Indonesia Terhadap Makanan khas Korea Selatan”. Dalam penelitian ini berfokus pada analisis sentimen yang berkaitan dengan ulasan, persepsi maupun opini masyarakat Indonesia terhadap makanan khas Korea Selatan di media sosial *Twitter* dan *YouTube*. Data ulasan, persepsi dan opini tersebut akan dianalisis dengan menggunakan algoritma *Naive Bayes* selanjutnya penelitian ini untuk melihat dan

mengetahui respon masyarakat terhadap makanan khas Korea Selatan yang ada di Indonesia, serta untuk melihat peluang bisnis baru bagi pelaku usaha yang hendak bergerak dibidang bisnis makanan khas Korea Selatan.

2.2 Dasar Teori

2.2.1 Persepsi

Persepsi dalam Psikologi mengarahkan pembaca ke berbagai situs psikologis dan kognitif di mana kesadaran dan pemahaman informasi sensorik dibahas (McDonald, 2011). Persepsi dapat dikatakan sebagai indra keenam manusia karena hampir semua fungsi kognitif kehidupan bergantung padanya. Persepsi juga merupakan fungsi kognitif penting di tingkat bawah sadar yang menentukan kepribadian. Sebanyak 76% konsumen mengatakan pernah mendengar tentang sertifikat halal; persepsi tentang sertifikat halal berbeda-beda berdasarkan nilai agama; dan terakhir, faktor demografi dan sosial ekonomi (status pendidikan, pekerjaan, usia, jenis kelamin, dan status perkawinan) mempengaruhi tingkat kesadaran konsumen akan sertifikat halal. makanan halal (Susilawati dkk., 2019).

2.2.2 YouTube

YouTube Merupakan sebuah *website* maupun aplikasi yang berkaitan dengan berbagi video yang sangat populer yang didirikan oleh tiga orang pada tahun 2005. *YouTube* merupakan basis data video paling populer, paling lengkap dan beragam serta paling luas penyebarannya yang ada di dunia Internet saat ini. *YouTube* merupakan sebuah situs layanan penyedia video *online* yang terkemuka di Amerika dan dunia yang berhasil menguasai 43% pasar dunia (Fatty dkk., 2016). *YouTube* menjadi salah satu wadah terbesar untuk menyimpan data dalam bentuk video. Pengguna *YouTube* yang begitu banyak membuat banyak video yang diunggah ke *YouTube* setiap menitnya. *YouTube* telah banyak menyediakan berbagai macam kebutuhan untuk para penggunanya. kemajuan teknologi yang dimiliki oleh *YouTube* sangat membantu para pengguna dari berbagai aspek kebutuhan yang dibutuhkan oleh para penggunanya (Kiftiyah dkk., 2017)

2.2.3 Twitter

Twitter merupakan sebuah layanan jejaring sosial *microblogging* yang memungkinkan para pengguna yang terdaftar dapat menyiarkan pesan singkat yang

disebut *tweets*. Para pengguna *Twitter* dapat menyiarkan *tweet* dan mengikuti *tweet* pengguna lain dengan menggunakan beberapa *platform* dan perangkat *smartphone* yang telah terhubung ke jaringan internet. *Tweets* atau pesan singkat yang dibuat oleh para pengguna bisa dilihat secara publik, namun pengguna yang mengirim *tweets* juga dapat membatasi pengiriman pesan ke daftar teman-teman tertentu saja. Para pengguna juga dapat melihat *tweets* pengguna lain yang bisa dikenal dengan sebutan pengikut atau *follower*. *Twitter API (Application Programming Interface)* merupakan sejumlah fungsi yang dapat digunakan dalam pengembang perangkat lunak untuk mengolah data saat membangun perangkat lunak tersebut (Rahmi, 2021).

Tampilan antarmuka pemrograman aplikasi *Twitter* atau biasa disebut (*Twitter API*) memiliki fungsi untuk memudahkan para pengembang perangkat lunak dalam memanggil fungsi tertentu dari aplikasi yang mereka kembangkan. *Twitter API* menggunakan arsitektur sistem yaitu *REST (Representational State Transfer)*, sehingga *Twitter* dapat digunakan dalam berbagai format data seperti *XML* dan *JSON*. *Twitter API* terdiri dari *Twitter Search API* dan *Twitter Streaming API*. Perbedaan keduanya adalah *Twitter Search API* berfokus pada kemampuan pencarian masa lalu, sedangkan *Twitter Streaming API* berfokus pada kemampuan pencarian masa depan (Rustiana dan Rahayu, 2017).

Beberapa istilah yang biasa digunakan pada *Twitter* yaitu (Rahmi, 2021):

1. *Tweet*

Tweet merupakan sebuah pesan standar yang ada pada *Twitter* yang terdiri dari 140 karakter atau kurang.

2. *Retweet*

Merupakan *tweet* atau pesan yang telah dibagikan ulang oleh pengikut seseorang.

3. *Hashtag*

Merupakan simbol # yang biasa digunakan untuk menandai kata kunci atau topik pada *tweet* yang bertujuan untuk mempermudah pencarian.

4. *Mention*

Merupakan pesan atau *tweet* yang berisi balasan dari pengguna lainnya dengan memberi awalan @ pada *username* mereka.

5. *Handle*

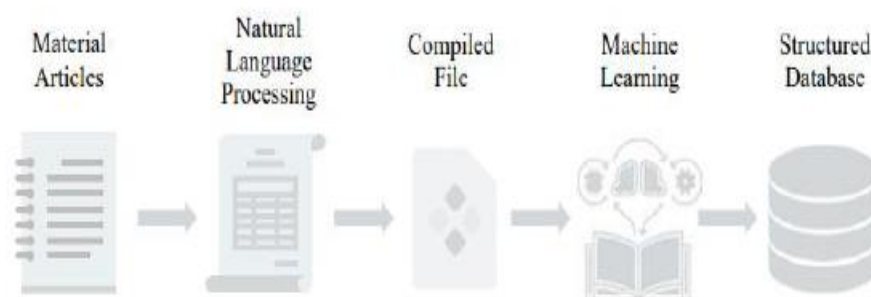
Merupakan fitur untuk menandakan *username* dan *URL* yang ada.

6. *Direct Message*

Fitur ini memberikan dukungan untuk pemberian pesan secara langsung antar pengguna

2.2.4 *Text Mining*

Penambangan teks biasa di sebut *text mining* merupakan sebuah proses penggalian teks dalam bentuk informasi dan pengetahuan dari sumber data tekstual yang sangat besar seperti dokumen *Word*, PDF, referensi teks, dan masih banyak yang lainnya. Jenis data yang diambil dalam melakukan *text mining* menggunakan data tidak terstruktur berbeda dengan halnya dengan *data mining* yang menggunakan data terstruktur atau *database* sebagai input datanya. Dalam proses penambangan teks dibagi menjadi dua tahap yang pertama menerapkan struktur ke sumber data tekstual, yang kedua mengekstrak informasi dan pengetahuan yang relevan dari data tekstual yang terstruktur menggunakan teknik dan alat yang sama seperti penambang data. Proses umum yang dilakukan oleh penambangan teks termasuk peringkasan otomatis, klasifikasi dokumen, pengelompokan teks (Fahrur dkk., 2020). Proses *text mining* dapat dilihat pada Gambar 2.1 (Samsir dkk., 2021).



Gambar 2.1 Proses *Text mining* (Samsir dkk., 2021)

2.2.5 Analisis Sentimen

Analisis sentimen merupakan proses untuk mengklasifikasikan dan menentukan polaritas dari sebuah teks dalam sebuah dokumen atau kalimat sehingga memudahkan seseorang dalam menentukan kategorinya baik sebagai positif, negatif, maupun netral (Dery dan Setiawan, 2019). Analisis sentimen banyak digunakan oleh para peneliti yang biasa mereka gunakan untuk mengambil dan menentukan kategori data yang mereka dapatkan termasuk kategori data yang positif, negatif atau netral, khususnya untuk penelitian yang berkaitan dengan ilmu komputer. Dalam jejaring sosial, seperti *YouTube dan Twitter*, analisis sentimen sering digunakan untuk melihat dan menentukan persepsi yang dibuat oleh publik (Rifan dkk., 2019).

Analisis sentimen bisa disamakan dengan penambangan opini karena tugasnya yang berfokus pada untuk membandingkan antara opini positif atau negatif (Saputra dkk., 2019). Dalam analisis sentimen, pengambilan data dilakukan dengan cara menganalisis, memproses, dan mengekstrak data tekstual dalam suatu entitas, seperti layanan, produk, orang, fenomena, atau topik dari tren tertentu (Liu, 2012). Proses analisis sentimen dapat mencakup teks komentar, forum, komentar, *tweet atau blog*, dan pra-pemrosesan data meliputi proses *tokenization*, *stopwords*, *penghapusan*, *stemming*, pengenalan sentimen, dan klasifikasi sentimen (Samsir dkk., 2021).

2.2.6 Pre-processing

Pre-processing merupakan proses untuk membersihkan dan mempersiapkan teks data untuk digunakan dalam klasifikasi. Data teks yang didapat dalam proses *Text mining* atau penambangan teks masih banyak mengandung *noise* dan banyak bagian tidak informatif seperti tag *HTML*, *script* dan iklan. Selain itu, tingkat kata-kata yang di dapat dari *text mining*, masih banyak kata-kata tidak sesuai pada dengan data yang kita butuhkan oleh karena itu proses *Pre-processing* sangat diperlukan untuk mendapatkan hasil data yang kita inginkan. 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 *online*, penghapusan ruang *spasi*, memperluas singkatan, kata dasar (*stemming*), penghapusan kata henti (*stopword removal*), penanganan negasi dan terakhir seleksi fitur (Indrayuni, 2017).

2.2.7 Labelling

Labelling atau pelabelan merupakan proses pemberian label pada data yang diperoleh berupa nilai atau simbol untuk dijadikan metrik dengan menggunakan metode *supervised learning*. Data yang digunakan terlebih dahulu harus memiliki sebuah indikator sehingga dalam proses pelatihan atau *training* dapat dilakukan pembelajaran berdasarkan label yang telah ditentukan (Noviantho dkk., 2021). *Labelling data* dapat dilihat pada Tabel 2.1 (Samsir dkk., 2021).

Tabel 2.1 *Labelling data*

<i>Label</i>	<i>Tweet</i>
Negatif	Espargaro: Sirkuit Mandalika Belum Siap Gelar MotoGP https://t.co/I96EkfpRjV
Negatif	Hey.. para pembalap MotoGP yang protes asphalt Mandalika. Gue kasi tau elu pada. Kalen itu balapan di Indonesia, https://t.co/AsQiGINIeu .
Positif	Sirkuit Mandalika sudah menggunakan aspal terbaik dari Inggris yang sudah digunakan di sirkuit besar dunia, tetapi kenapa bisa terlupas?
Negatif	Banyak Dikritik, Lintasan Sirkuit Mandalika Resmi Diaspal Ulang https://t.co/opBtM684Rs

2.2.8 Term Weighting

Term weighting merupakan metode yang dilakukan untuk memberikan bobot nilai pada setiap kata (*term*) di pada sebuah data yang telah didapat dalam sebuah dokumen. Salah satu metode yang dapat digunakan yaitu *TF-IDF* (*Term Frequency- Inverse Document Frequency*) rumus perhitungan dapat dilihat pada Persamaan (1), (2), dan (3) (Noviantho dkk., 2021).

$$TF_{i'j} = \frac{\text{Term } i \text{ frequency in document } j}{\text{Total words in document}} \dots\dots\dots(1)$$

$$IDF = \log \left(\frac{n}{df_i} \right) \dots\dots\dots(2)$$

$$W_{i'j} = tf_{i'j} * \log \left(\frac{n}{df_i} \right) \dots\dots\dots(3)$$

Keterangan:

$tf_{i'j}$ = jumlah kemunculan i di j

df_i = jumlah dokumen yang mengandung i

N = jumlah total dokumen

Term Frequency (TF) adalah nilai yang merepresentasikan seberapa sering sebuah kata muncul dalam sebuah dokumen. Semakin besar nilai TF suatu kata, semakin sering kata tersebut muncul dalam dokumen. Sedangkan *Inverse Document Frequency* (IDF) adalah nilai yang menunjukkan kebalikan dari jumlah dokumen yang mengandung sebuah kata. Semakin kecil jumlah dokumen yang mengandung kata, semakin besar nilainya.

2.2.9 Naive Bayes

Naive Bayes merupakan metode yang digunakan untuk mengklasifikasi data dalam melakukan analisis sentimen dalam *text mining*. 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 *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 (4) (Samsir dkk., 2021).

$$P(V|X) = \frac{P(X|V).P(V)}{P(X)} \dots\dots\dots(4)$$

Dari persamaan (4) 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 (5) (Samsir dkk., 2021).

$$P(V|X_1 \dots X_n) = \frac{P(V)P(X_1 \dots X_n|V)}{P(V|(X_1 \dots X_n))} \dots \dots \dots (5)$$

Dalam persamaan (5), variabel V mewakili kategori, dan variabel $X_1 \dots X_n$ menjelaskan instruksi yang diperlukan untuk klasifikasi. Persamaan (5) 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). Tahapan Penelitian Menggunakan Klasifikasi *Naive Bayes* dapat dilihat pada Gambar 2.2 (Samsir dkk., 2021).



Gambar 2.2 Tahapan Penelitian Menggunakan Klasifikasi *Naive Bayes*

(Samsir dkk., 2021)

Pada Gambar 2.2 merupakan tahapan alur proses yang dilakukan dalam penelitian ini yang diawali dengan proses pengambilan data dengan menggunakan metode *web scrapping* untuk mengambil data dari media sosial *YouTube*, sedangkan untuk pengambilan data *Twitter* metode pengambilan data yang digunakan yaitu *crawling data*.

Langkah berikutnya yaitu tahap *Pre-processing*. *Pre-processing* dilakukan untuk menyeleksi dan mengubah data agar menjadi lebih terstruktur. Pada tahap *Pre-processing* dilakukan proses *cleaning* yang berfungsi untuk mengurangi *noise*

dan *remove pada kata* maupun informasi yang didapat, proses *stopword* dilakukan untuk menghapus kata-kata yang tidak bermakna atau tidak penting seperti ‘saya’, ‘dan’, ‘atau’ dan seterusnya (Watrianthos, 2020). Proses *tokenization* dilakukan untuk mengidentifikasi kata-kata dari informasi yang didapat dan memecah kalimat menjadi istilah berdasarkan spasi dan tanda baca (Sholihin dkk., 2019). Tahap terakhir dalam proses *Pre-processing* yaitu *stemming*. *Stemming* merupakan langkah untuk merubah kata imbuhan menjadi kata dasar (Setian dan Seprina, 2019). Tahap ketiga dalam melakukan *opinion mining* yaitu mengekstrak informasi maupun fitur yang ada, hal tersebut dilakukan untuk mempermudah dalam klasifikasi *Naive Bayes* (Watrianthos, 2020).

2.2.10 Web Scrapping

Web Scrapping merupakan sebuah metode atau cara yang biasa digunakan untuk mendapatkan informasi yang dibutuhkan dari suatu situs web yang dilakukan secara otomatis dengan bantuan sistem komputerisasi. *Web Scrapping* biasa digunakan untuk menggali sebuah informasi penting dari sebuah halaman *website* tertentu yang berbeda, yang mana informasi yang berada di halaman *website* belum terstruktur dengan rapi lalu informasi yang di dapat disaring secara otomatis dengan bantuan sistem komputerisasi agar data dan informasi yang didapat lebih terlihat rapi dan terstruktur dan format data yang dihasilkan biasanya dalam format *spreadsheets*, maupun *comma separated values (CSV)* (Parasati dkk., 2020).

2.2.11 Crawling Data

Crawling Data merupakan metode yang digunakan untuk mengambil dan mengumpulkan data yang dibutuhkan dari sebuah *database*. Data yang dibutuhkan dalam penelitian ini merupakan data yang diambil dan diperoleh dari *database Twitter* berupa informasi berkaitan dengan *user* atau pengguna beserta *tweet* dan atribut lainnya. Pengambilan data *Twitter* menggunakan metode *Crawling Data* dapat dilakukan dengan dua cara yaitu dengan menggunakan *user* atau pengguna dan menggunakan sebuah kata kunci atau biasa disebut dengan *keyword* (Dery dan Setiawan, 2019). Pencarian data *Twitter* dengan menggunakan kata kunci merupakan pencarian data menggunakan potongan kata atau *hashtag*. Sedangkan pencarian berdasarkan pengguna merupakan pencarian data berdasarkan nama akun

pengguna *Twitter*. Ekstrak fitur data pengguna dari indeks *Twitter* berupa total *tweet*, total pengikut, total suka, situs web, sumber, profil, id, akun, nama, dan lokasi. Sedangkan ekstraksi fitur dari kata kunci atau keyword yang diperoleh dari indeks *Twitter* adalah opini atau ulasan, *url*, *mention*, *retweet*, *hashtag*, jumlah *likes* dan jumlah *retweet* (Samsir dkk., 2021).

2.2.12 *Word Cloud*

Word Cloud merupakan sebuah representasi visual dari sebuah informasi data tekstual. *Word Cloud* merupakan sebuah cara yang sering digunakan untuk menggambarkan atau memvisualisasikan metadata kata kunci dari sebuah situs *website* dalam bentuk teks bebas. *Word Cloud* biasanya menggunakan satu kata yang dihasilkan untuk membuat presentasi grafis dari kata tersebut berdasarkan seberapa sering kata tersebut digunakan dalam domain. Hal ini dapat diartikan sebagai keunggulan kata-kata dalam bidang tertentu. *Word Cloud* adalah cara untuk meringkas sejumlah besar data dengan cepat. Serta menggambarkan ideologi yang didukung oleh kumpulan data tertentu (dalam hal ini data *Twitter* dan *YouTube*) dan menunjukkan filosofi yang terkait dengan kumpulan ucapan tekstual tersebut. Jenis visualisasi ini dapat membantu penilai untuk melakukan analisis teks eksplanasi dengan mengidentifikasi kata-kata yang sering muncul dalam kumpulan *tweet* tertentu. Contoh *Word Cloud* biasa di lihat pada Gambar 2.3 (Ahuja dan Shakeel, 2017).



Gambar 2.3 *Word Cloud* (Ahuja dan Shakeel, 2017)

2.2.13 *Cleaning Data*

Cleaning Data merupakan tahapan yang digunakan untuk menghilangkan komponen khusus dari suatu komen maupun *tweet* yang ada di sebuah *website* maupun dalam sebuah *tweet* yakni, komponen khusus yang biasa ada dalam sebuah komen maupun *tweet* yaitu, *username*, *URL (Uniform Resource Locator)*, dan “RT” (tanda *retweet*). Hal tersebut dapat mempengaruhi hasil nilai sentimen yang didapatkan nantinya, maka komponen di atas harus dihilangkan maupun dibuang (Ghaniy dan Sihotang, 2019).

2.2.14 *Case Folding*

Case folding merupakan tahapan yang dilakukan untuk merubah huruf besar menjadi *lowercase* atau huruf kecil. Berikut salah satu contoh *case folding dalam sebuah tweets*: “Puas dengan layanan yang telah diberikan oleh @Nissan” diubah menjadi “puas dengan layanan yang telah diberikan oleh @nissan” (Ghaniy dan Sihotang, 2019). *Proses Case Folding* dapat dilihat pada Tabel 2.2 (Samsir dkk., 2021).

Tabel 2.2 *Proses Case Folding*

Sebelum	Sesudah
Sirkuit Mandalika sudah menggunakan aspal terbaik dari Inggris yang sudah digunakan di sirkuit besar dunia, tetapi kenapa bisa terlupas?	Sirkuit mandalika sudah menggunakan aspal terbaik dari inggris yang sudah digunakan di sirkuit besar dunia, tetapi kenapa bisa terlupas?

2.2.15 *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).

Proses *Tokenizing and Stopword Removal* dapat dilihat pada Tabel 2.3 (Ghaniy dan Sihotang, 2019).

Tabel 2.3 *Proses Tokenisasi and Stopword Removal*

Sebelum	Setelah
Sirkuit mandalika sudah menggunakan aspal terbaik dari inggris yang sudah digunakan di sirkuit besar dunia, tetapi kenapa bisa terlupas?	'sirkuit', 'mandalika', 'indonesia', 'pertamina', 'aspal',

2.2.16 *Stemming*

Stemming merupakan tahapan yang dilakukan untuk merubah 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).

2.2.17 *Stopword Removal*

Stopword Removal merupakan tahapan dari kumpulan komen dan *tweets* yang telah melewati tahap *cleaning data* yang mana data yang didapat telah dibersihkan mulai dari dihilangkan karakter, tanda baca, serta kata-kata umum yang tidak memiliki makna maupun informasi yang tidak kita dibutuhkan (Ghaniy dan Sihotang, 2019).

