

## **BAB II**

### **TINJAUAN PUSTAKA**

#### **2.1 PPKM (Pemberlakuan Pembatasan Kegiatan Masyarakat).**

Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) merupakan kebijakan Pemerintah Indonesia sejak awal tahun 2021 untuk menekan angka penularan Covid-19 di Indonesia. Sebelum pelaksanaan PPKM, pemerintah telah melaksanakan pembatasan sosial berskala besar (PSBB) yang berlangsung di sejumlah wilayah di Indonesia. PPKM berlangsung di beberapa wilayah yang menjadi titik penyebaran infeksi Covid-19, yakni di Pulau Jawa dan Bali.

Intruksi Menteri Dalam Negeri (Mendagri) Muhammad Tito Karnavian terkait Pemberlakuan Izin Kegiatan Masyarakat (PPKM) mikro nomor 03 tahun 2021. Dikeluarkannya Inmendagri sebagai upaya dalam menindaklanjuti arahan Presiden RI Joko Widodo agar di berlakukan perpanjangan terhadap kebijakan Pemberlakuan Kegiatan Masyarakat (PPKM) dengan berbasis mikro yang selanjutnya membentuk posko penanganan corona virus disease (covid-19) mulai dari tingkat desa dan kelurahan (Humas Kemendagri.2021).

Kebijakan tersebut di instruksikan kepada seluruh gubernur se-Jawa dan Bali beserta Bupati dan Walikota untuk mengatur PPKM berbasis mikro di wilayahnya asing-masing. “Mengatur PPKM yang berbasis mikro yang selanjutnya disebut PPKM Mikro sampai dengan tingkat Rukun Tetangga (RT)/Rukun Warga 16 (RW) yang berpotensi menimbulkan penularan COVID-19,” bunyi instruksi inmendagri diktum kesatu (Humas Kemendagri.2021).

“Pada Diktum ke-dua di tegaskan kembali, PPKM mikro di lakukan dengan melakukan pertimbangan kriteria zonasi yang digunakan dalam melakukan pengendalian wilayah hingga tingkat RT.” (Humas Kemendagri, 2021) :

a. Zona Hijau

Tidak ditemukan kasus covid-19 dalam satu RT, pengendaliannya dilakukan dengan surveilans aktif. Pemantauan dilakukan secara rutin dan berkala selanjutnya dilakukan pengecekan pada seluruh suspek.

b. Zona Kuning

Terkonfirmasi kasus positif 1 sampai 5 rumah dalam 1 RT selama 7 hari terakhir. Skenario pengendalian adalah “Menemukan kasus suspek dan pelacakan kontak erat, lalu melakukan isolasi mandiri untuk pasien positif dan kontak erat dengan pengawasan ketat,” instruksi Tito Karnavian.

c. Zona Oranye

Kasus konfirmasi positif ditemukan 6 sampai 10 rumah dalam 1 RT selama 7 hari terakhir, Skenario pengendalian yang dilakukan yaitu dengan menemukan kasus suspek dan pelacakan kontak erat, lalu melakukan isolasi mandiri untuk pasien positif dan kontak erat dengan pengawasan ketat, serta menutup rumah ibadah, tempat bermain anak, dan tempat umum lainnya kecuali sektor esensial.

d. Zona Merah

Terdapat kasus terkonfirmasi lebih dari 10 rumah dalam 1 RT selama 7 hari terakhir, maka skenario pengendalian adalah pemberlakuan PPKM tingkat RT yang mencakup 6 hal, yaitu, menemukan kasus suspek dan pelacakan kontak erat; melakukan isolasi mandiri/terpusat dengan pengawasan ketat; serta menutup rumah ibadah, tempat bermain anak dan tempat umum lainnya kecuali sektor esensial. Juga melarang kerumunan lebih dari tiga orang; membatasi keluar masuk wilayah RT maksimal hingga Pukul 20.00; serta meniadakan kegiatan sosial masyarakat di lingkungan RT yang menimbulkan kerumunan dan berpotensi menimbulkan penularan.

## 2.2 *Twitter*

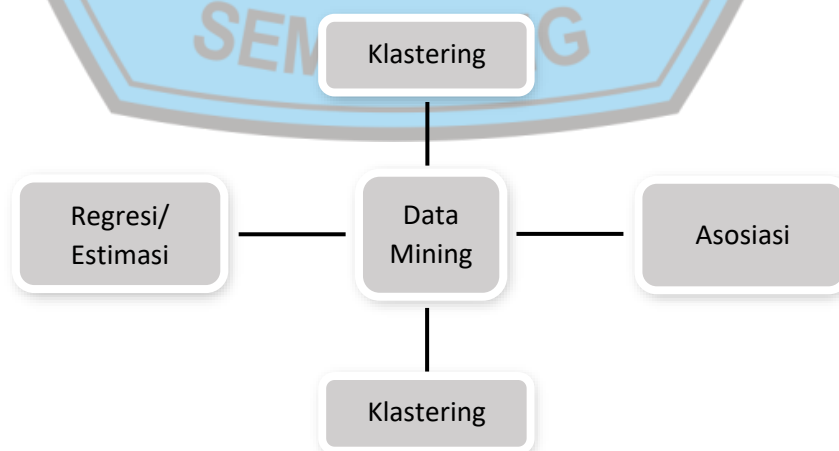
*Twitter* adalah layanan jejaring sosial dan mikroblog daring yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks dengan batas 140 karakter, akan tetapi pada tanggal 7 November 2017 bertambah hingga 280 karakter yang dikenal dengan sebuah kicauan (*tweet*).

*Twitter* didirikan pada bulan Maret 2006 oleh Jack Dorsey, dan situs *Twitter* mulai diluncurkan pada bulan Juli. *Twitter* Inc berbasis di San Fransisco dengan kantor tambahan di New York, Boston, dan San Antonio. Sejak diluncurkan *Twitter* telah menjadi salah satu dari sepuluh situs yang paling sering dikunjungi di internet.

Pengguna yang tidak terdaftar hanya bisa membaca kicauan, sedangkan pengguna terdaftar bisa menulis, menyukai, dan menyebarkan kicauan. Tingginya popularitas *Twitter* menyebabkan layanan ini telah dimanfaatkan untuk berbagai keperluan dalam berbagai aspek, misalnya sebagai sarana protes, kampanye politik, sarana pembelajaran, dan sebagai media komunikasi darurat. Semua pengguna dapat mengirim dan menerima *tweet* melalui situs *Twitter*, aplikasi eksternal yang kompatibel (telepon seluler), atau dengan pesan singkat (SMS) yang tersedia di negara-negara tertentu (*Twitter*, 2013). Pengguna dapat 13 menulis pesan berdasarkan topik dengan menggunakan tanda # (hashtag). Sedangkan untuk menyebutkan atau membalas pesan dari pengguna lain bisa menggunakan tanda @.

### 2.3 Data Mining

*Data mining* adalah kegiatan mengekstrak informasi atau pengetahuan (*knowledge*) penting dari suatu set data berukuran besar dengan menggunakan teknik tertentu informasi yang dihasilkan dari data mining ini bisa dipakai untuk guna pengambilan keputusan. Tugas-tugas yang biasa dilakukan oleh *data mining* adalah klustering, klasifikasi, estimasi dan Asosiasi (Santoso dan Umam, 2018).



Gambar 2.1. Lingkup Kajian *Data Mining*

## 2.4 Analisis Sentimen

Sentiment analysis atau analisis sentimen dalam bahasa Indonesia adalah sebuah teknik atau cara yang digunakan untuk mengidentifikasi bagaimana sebuah sentimen diekspresikan menggunakan teks dan bagaimana sentimen tersebut bisa dikategorikan sebagai sentimen positif maupun sentimen negatif. Hasil sistem prototipe mencapai tinggi presisi (75-95% tergantung pada data) dalam mencari sentimen pada halaman web dan artikel berita. (Nasukawa.T.& Yi, 2003) Menurut (Liu, 2011) sentiment analysis atau opinion mining mengacu pada bidang yang luas dari pengolahan bahasa alami, komputasi linguistik dan text mining yang memiliki tujuan menganalisa pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang apakah pembicara atau penulis berkenan dengan suatu topik, produk, layanan, organisasi, individu, ataupun kegiatan tertentu. Tugas analisis sentimen yaitu mengelompokkan teks ke dalam kalimat atau dokumen kemudian menentukan pendapat yang dikemukakan dalam kalimat atau dokumen yang dianalisis apakah bersifat positif, negatif, atau netral (Dehhaf, 2010).

Menurut (Akbari, Novianty, dan Setianingsih, 2012), Analisis sentimen merupakan salah satu cabang ilmu dari *text mining*, *natural language program*, dan *artificial intelligence*. Proses yang dilakukan oleh analisis sentimen untuk memahami, mengekstrak, dan mengolah data teks secara otomatis sehingga menjadi suatu informasi yang bermanfaat. Selain itu analisis sentimen merupakan bidang ilmu yang menganalisis pendapat, sikap, evaluasi, dan penilaian terhadap suatu peristiwa, topik, organisasi, maupun perseorangan.

Tugas dasar dalam analisis sentimen adalah mengelompokkan teks yang ada dalam sebuah kalimat atau dokumen kemudian menentukan pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif atau negatif. *Sentiment analysis* juga dapat menyatakan perasaan emosional sedih, gembira, atau marah. Kita dapat mencari pendapat tentang produk-produk, merek atau orang-orang dan menentukan apakah mereka dilihat positif atau negatif di web. Ekspresi atau *sentiment* mengacu pada fokus topik tertentu, pernyataan pada satu topik mungkin akan berbeda makna dengan pernyataan yang sama pada *subject*

yang berbeda. Oleh karena itu pada beberapa penelitian, pekerjaan didahului dengan menentukan elemen dari sebuah produk yang sedang dibicarakan sebelum memulai proses *opinion mining* (Kurniawan, T. 2017).

## 2.5 Preprocessing

Pre-processing adalah tahap awal yang digunakan untuk mendapatkan data yang akan digunakan dalam proses selanjutnya. Tahap dalam proses ini terdiri dari (Onantya, Indriati dan Adikara, 2019) :

### 4.2.1 Case Folding

*Case folding* adalah tahap awal *pre-processing* yang berfungsi mengubah huruf awal pada suatu kata masukan menjadi huruf kecil (*lower case*).

### 4.2.2 Tokenizing

*Tokenizing* merupakan proses pemisahan kalimat menjadi kata atau token yang merupakan penyusun dari suatu dokumen. Pada tahap ini, karakter-karakter pemisah kata akan dihilangkan karena tidak berpengaruh terhadap pemrosesan teks.

### 4.2.3 Filtering

*Filtering* merupakan proses menghapus kata yang tidak mengandung arti yang bermakna. Proses ini dilakukan dengan menggunakan *Stopword list* sebagai delimiter untuk menghilangkan kata yang tidak dibutuhkan.

### 4.2.4 Stemming

*Stemming* merupakan proses yang berfungsi untuk mengubah bentuk suatu kata menjadi kata dasar dan sesuai dengan struktur morfologi Bahasa Indonesia yang benar.

## 2.6 K-Nearest neighbor (K-NN)

Algoritma Nearest Neighbor adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama, yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada. (Kusrini dan Luthfi, 2009). Algoritma K-NN adalah suatu metode yang menggunakan algoritma supervised.



Perbedaan antara supervised learning dengan unsupervised learning adalah pada supervised learning bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang sudah ada dengan data yang baru. Sedangkan pada unsupervised learning, data belum memiliki pola apapun, dan tujuan unsupervised learning untuk menemukan pola dalam sebuah data. Tujuan 12 dari algoritma k-NN adalah untuk mengklasifikasi objek baru berdasarkan atribut dan training samples.

K-Nearest Neighbor (KNN) merupakan algoritma yang biasanya digunakan untuk proses klasifikasi data. KNN merupakan suatu tahapan yang menggunakan algoritma terawasi, dimana hasil dari query instance yang baru nantinya diklasifikasikan berdasarkan mayoritas dari kategori yang ada (Simanjuntak, Mahmudy dan Sutrisno, 2017). Algoritma ini digunakan untuk pengklasifikasian objek berdasarkan jarak yang paling dekat dengan objek tersebut. Untuk menghitung jarak antar objek biasanya dihitung berdasarkan *euclidean distance* dengan persamaan berikut :

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^n (a_r(x_i) - a_r(x_j))^2} \quad (2.1)$$

dimana :

$d(x_i, x_j)$  : jarak euclidean       $x_i$  : record ke-i  
 $x_j$  : record ke-j                       $a_r$  : data ke r  
 $i, j$  : 1,2,3,...n.

K-Nearest Neighbor digunakan untuk mengklasifikasi data yang tidak dilabeli. Karakteristik data didapatkan dari *training set* dan *test set*. Langkah-langkah klasifikasi data menggunakan K-Nearest Neighbor adalah sebagai berikut :

- a. Tentukan nilai K.
- b. Hitung jarak euclidean antara data baru ke setiap label data.

- c. Tentukan  $k$  labeled data yang mempunyai jarak yang paling minimal.
- d. Klasifikasikan data baru ke dalam label data yang mayoritas K-NN dipilih berdasarkan metrik jarak.

## 2.7 Naïve Bayes Classifier

Teorema Bayes merupakan teorema yang mengacu pada probabilitas bersyarat. Secara umum teorema Bayes dapat dinotasikan pada persamaan berikut.

$$P(A|B) = \frac{P(A|B) P(A)}{P(B)} \quad (2.2)$$

*Naive Bayes Classifier* merupakan algoritma yang digunakan untuk mencari nilai probabilitas tertinggi untuk mengklasifikasi data uji pada kategori yang paling tepat (Feldman & 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 (Falahah & Nur, 2015 dalam ). 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.

$$V_{MAP} = \arg \max P(V_j) \prod_i P(a_i|v_j) \quad (2.3)$$

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

$$P(V_j) = \frac{|doc|}{|training|} \quad (2.4)$$

dimana  $|doc|$  merupakan jumlah *tweet* yang memiliki kategori  $j$  dalam *training*. Sedangkan  $|training|$  merupakan jumlah *tweet* dalam contoh yang digunakan untuk *training*. Untuk setiap probabilitas kata  $a_i$  untuk setiap kategori  $P(a_i|v_j)$ , dihitung pada saat *training*.

$$P(a_i|v_j) = \frac{n_i + 1}{n + \text{kosakata}} \quad (2.5)$$

Dimana ini adalah jumlah kata  $a_i$  dalam *tweet* yang berkategori  $v_j$ , sedangkan  $n$  adalah banyaknya seluruh kata dalam *tweet* dengan kategori  $v_j$  dan  $|\text{kosakata}|$  adalah banyaknya kata dalam training, Langkah-langkah klasifikasi data menggunakan K-Nearest Neighbor adalah sebagai berikut :

- a. Menghitung probabilitas setiap kelas sentimen terhadap setiap dokumen.
- b. Lakukan perhitungan probabilitas kemunculan setiap kata pada masing-masing kategori dengan persamaan.
- c. Mencari probabilitas tertinggi dari *tweet* yang diujikan.
- d. Mengklasifikasikan *tweet* kedalam label.

## 2.8 Pembobotan TF-IDF

*Term Frequency-Inverse Document Frequency* (TF-IDF) adalah metode yang digunakan untuk menghitung bobot setiap kata yang telah diekstrak. Penggunaan metode ini umumnya dilakukan untuk menghitung kata umum yang ada pada *information retrieval*. Model pembobotan TF-IDF merupakan metode yang mengintegrasikan *model term frequency* ( $tf$ ) dan *inverse document frequency* ( $idf$ ). *Term frequency* ( $tf$ ) merupakan proses untuk menghitung jumlah kemunculan *term* dalam satu dokumen dan *inverse document frequency* ( $idf$ ) digunakan untuk menghitung *term* yang muncul di berbagai dokumen (komentar) yang dianggap sebagai *term* umum, yang dinilai tidak penting (Akbari, Novianty, dan Setianingsih, 2012).

Tahapan pembobotan dengan TF-IDF adalah :

- a. Hitung *term frequency* ( $tf_{t,d}$ )
- b. Hitung *weighting term frequency* ( $W_{tf}$ )

$$W_{tf_{t,d}} = \begin{cases} 1 + \log_{10} tf_{t,d}, & \text{if } tf_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (2.6)$$

- c. Hitung *document frequency* ( $df$ )



d. Hitung bobot *inverse document frequency* (*idf*)

$$idf_t = \log_{10} \frac{N}{df_t} \quad (2.7)$$

e. Hitung nilai bobot TF-IDF

$$W_{t,d} = W_{tf_{t,d}} \times idf_t \quad (2.8)$$

dimana :

$tf_{t,d}$	:	Frekuensi term
$W_{tf_{t,d}}$	:	Bobot frekuensi term
$df$	:	Jumlah frekuensi dokumen yang mengandung term
$N$	:	Jumlah total dokumen
$W_{t,d}$	:	Bobot TF-IDF

## 2.9 Evaluasi Performa Klasifikasi

### 2.9.1 Confusion Matrix

*Confusion matrix* adalah suatu metode yang biasa digunakan untuk melakukan perhitungan akurasi pada konsep data mining. *Confusion matrix* digambarkan dengan tabel yang menyatakan jumlah data uji yang benar diklasifikasikan dan jumlah data uji yang salah diklasifikasikan. (Wisdayani, Nur, dan Wasono, 2019).

**Tabel 2.1.** *Confusion Matrix*

		Clasif ied as	
		+	-
Correct Clasification	+	True Positive	False Positive
	-	False Negative	True Negative

### 2.9.2 Ukuran Performa Klasifikasi

Dalam pengukuran performa klasifikasi terdapat beberapa cara, namun cara yang paling sering digunakan adalah dengan menghitung akurasi, *precision*, *recall* dan *f-measure* (Witten dan Frank, 2005).

Akurasi merupakan persentase dari total sentimen yang benar dikenali. Perhitungan akurasi dilakukan dengan cara membagi jumlah data sentimen yang benar dengan total data dan data uji. Untuk menghitung nilai akurasinya dilakukan dengan menggunakan Persamaan (2.9).

$$akurasi = \frac{jumlah\ sentimen\ benar}{jumlah\ data\ tes} \times 100\% \quad (2.9)$$

Untuk pengukuran performa klasifikasi cara yang digunakan selain menghitung akurasi adalah menghitung *precision*, *recall* dan *f-measure*. *Precision* merupakan perbandingan jumlah data relevan yang ditemukan terhadap jumlah data yang ditemukan. Perhitungan *precision* dilakukan dengan cara membagi jumlah data benar yang bernilai positif dibagi dengan jumlah data benar yang bernilai positif ditambahkan dengan data salah yang bernilai positif. Nilai dari data salah bernilai positif diambil dari jumlah nilai selain *true positive* kolom yang sesuai tiap kelasnya. Untuk menghitung nilai *precision* dapat dilakukan dengan menggunakan Persamaan (2.10).

$$Presisi = \frac{true\ positive}{true\ positive + false\ positive} \quad (2.10)$$

*Recall* merupakan perbandingan jumlah materi relevan yang ditemukan terhadap jumlah materi yang relevan. Perhitungan *recall* dilakukan dengan cara membagi data benar bernilai positif dengan hasil penjumlahan dari data benar yang bernilai positif dan data salah yang bernilai negatif. Nilai dari data salah

yang bernilai negatif diambil dari jumlah nilai selain true positive baris yang sesuai tiap kelasnya. Perhitungan recall dapat menggunakan Persamaan (2.1).

$$recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (2.11)$$

*F-measure* merupakan parameter tunggal ukuran keberhasilan retrieval yang menggabungkan recall dan precision. Nilai *F-measure* didapat dari perhitungan hasil perkalian precision dan recall dibagi dengan hasil penjumlahan precision dan recall kemudian dikalikan dua dan perhitungan *f-measure* menggunakan Persamaan (2.12).

$$f - measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (2.12)$$

