BAB III
METODOLOGI PENELITIAN

3.1 Gambaran umum Objek Penelitian

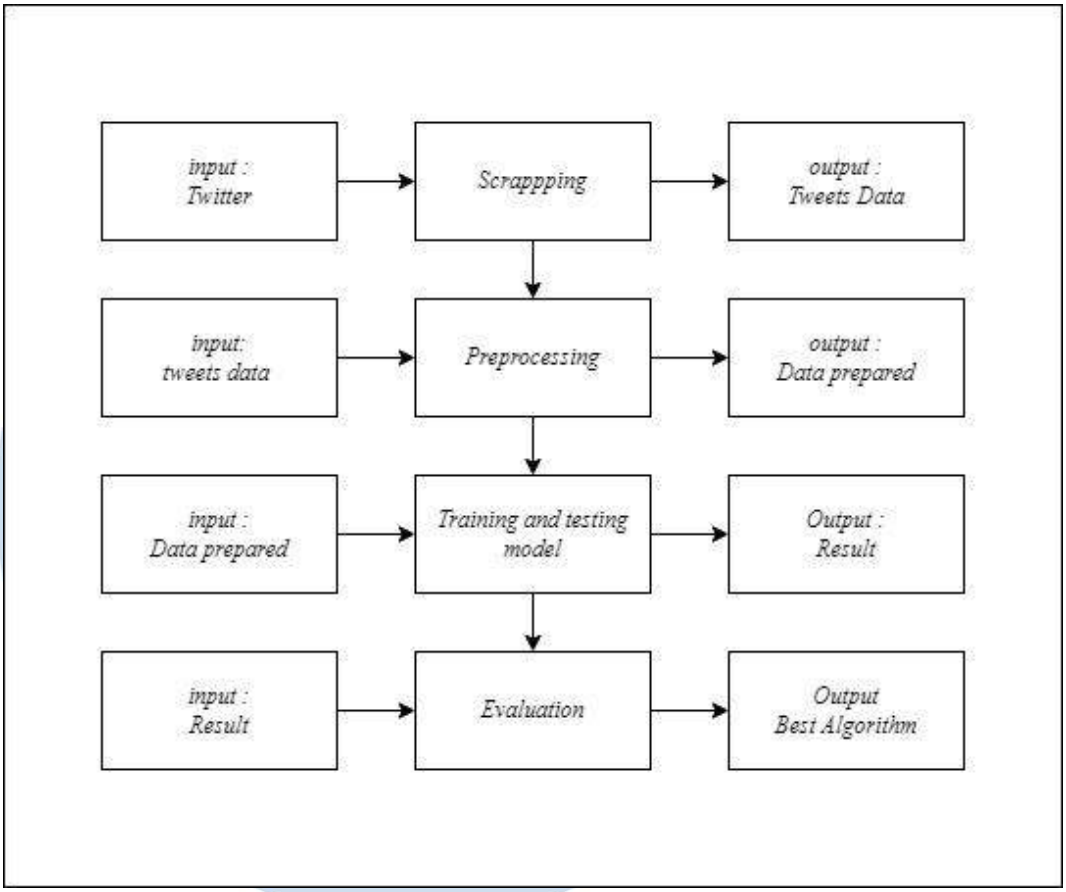
Objek penelitian dari skripsi ini adalah sentiment pengguna aplikasi lindungi yang diambil menggunakan berupa sumber *media social Twitter*. Melalui anlisis sentiment, pendapat masyarakat terhadap aplikasi peduli lindungi digunakan untuk mencari tahu hasil dari sentiment penguna aplikasi melalui proses algoritma klasifikasi *sentiment*. Tahapan berikutnya adalah menguji hasil *model K-Fold cross* validation untuk menemukan algoritma mana yang menghasilkan nilai yang paling baik dalam menangani kasus sentiment terhadap aplikasi PeduliLindungi. Dengan begitu Hasil dari penelitian diharapkan untuk mencari criteria yang ideal dalam menentukan algoritma yang digunakan untuk klasifikasi terhadap komentar pengguna aplikasi PeduliLindungi di *media social twitter*.

Maka dari itu penelitian ini bertujuan untuk membuat klasifikasi terhadap komentar dari *social media twitter* terhadap aplikasi PeduliLindungi dengan hasil yang lebih baik untuk dijadikan sebuah rekomendasi pengembangan terhadap aplikasi PeduliLindungi

3.2 Metode Penelitian

Pada penelitian ini *text mining* memiliki beberapa tahapan didalamnya yaitu D*ata Scraping, Data PreProcessing, Data Labeling, Data Sampling, Sentiment Classification,* Hasil Pembahasan. penelitian ini menggunakan refrensi dari penelitian terdahulu yang berjudul "*Analysis of user reviews for the PeduliLindungi application on google play using the Support Vector Machine and Naive Bayes algorithm based on particle swarm optimization*" dalam penelitian ini akan menggunakan tiga algoritma yang populer dipakai untuk sentimen analisis yaitu , *Naïve bayes, Support Vector Machine, K-NN*.

Dalam melakukan penelitian, penelitian ini memiliki kerangka pikir penelitian seperti gambar 3.1



Gambar 3. 1 Kerangka pikir penelitian.

kerangka pikir pada 3.1 dibuat dengan mengambil referensi kerangka pikir yang sebelumnya sdudah dibuat dari penelitian penelitian terdahulu, dengan rangkaian yang sudah dilkuakn untuk penyesuaian pada penelitian ini.

3.3 Variabel penelitian

Pada penelitian yang akan dilakukan ini, variabel penelitian yanga digunakan ialah pendapat masyarakat mengenai aplikasi *Contact tracing* yaitu PeduliLindungi. Dimana variabel independen dari penelitian ini merupakan komentar / opini masyarakat terhadap aplikasi PeduliLindungi yang berada pada media social twitter. variabel dependen dari penelitian ini adalah sentimen masarakat mengenai aplikasi PeduliLindungi, data dependen yang di pakai adalah sentimen negatif/ sentimen positif

3.4 Teknik Pengumpulan data

3.4.2 Pendaftaran akses twitter

Dalam melakukan penggalian data twitter penelitian ini menggunakan *Rapidminer studio*. Dibutuhkan sebuah izin akses authentifikasi dengan akun *twitter* yang ada. Akses *twitter* yang telah didapatkan akan digunakan dalam mencari koneksi *twitter*

3.4.1 Penggalian data twitter

Teknik pengumpulan data dilakukan dengan cara menggunakan *extension* yang berada didalam *Rapidminer* dengan cara menghubungkan dengan akun *twitter*, pengambilan sampel ini dilakukan dengan cara seleksi terhadap data *twitter* yang berisi sentimen mengenai aplikasi *Contact tracing* yaitu PeduliLindungi dan memiliki nilai sentimen berupa positif , dan negative. Jumlah sampel data yang akan digunakan berjumlah 7.587. hasil dari crawling data *twitter* akan diubah menjadi format xlsx.

3.4.1 *Twitter* data

Output yang didapat setelah melakukan data *Scraping* disimpan dalam format csv dan mempunyai bebarapa kolom atribut didalamnya, yang berisikan komentar komentar masyarakat mengenai aplikasi PeduliLindungi dalam kurun waktu 22 November 2021 hingga 27 Desember 2021, jangka waktu tersebut digunakan karena pemerintah sedang memaksimalkan penggunaan aplikasi pedulilndungi baik dari masyarakat dan pelaku bisnis dan di cabutnya Pemberlakuan Pembatasan Kegiatan Masyarakat tingkat 3. Data masih berbentuk *RAW* sebelum masuk kedalam tahapan *Pre-procecssing*.

3.4.3 Seleksi *tweet*

Pada tahapan ini, data twitter yang sudah di dapatkan akan dilakukan seleksi data twitter, beberapa tahapan yang dilakukan dalam proses seleksi yaitu penghapusan data twitter yang tidak berbahasa Indonesia, menghapus tweet yang sama dan menghapus tweet yang tidak berhubungan dengan aplikasi PeduliLindungi. Data yang sudah dibersihkan akan masuk kedalam tahapan data *Cleansing* dengan memakai file *excel* yang baru.

3.5 Teknik Analisa data

Teknik Analisa data dalam penulisan ini adalah kualitatif. Analisa data pada penelitian ini diproses dengan pengguanan progam Rapidminer studio dalam membuat model Analisa sentimen memakai tiga(3) algoitma yang sering digunakan dalam melakukan Analisa sentiment diantaranya ada Naïve bayes, *Support Vector Machine*, dan K-NN. sebelum melakukan tahapan pemodelan, text preprocessing dilakukaan guna mengolah data data twitter yang sudah didapatkan.

3.5.1 Text preprocessing

Tahapan tahapan yang selanjutnya dalam melakukan Analisa sentiment ialah menyiapkan hasil dari tahapan tahapan yang ada agar menjadi sebuah data yang sudah siap untuk dilakukan tahap pengolahan , dalam tahapan tahapan *pre-processing* terdiri dari beberapa proses yang dibutuhkan.

3.5.1.1 Cleansing

Tujuannya dari tahapan ini adalah membersihkan data *tweet* dan mengurangi *noise* yang ada didalam *data set*, beberapa contoh *noise* yang ada dalam sebuah teks adalah komponen-komponen yang tidak memiliki hubungan dengan informasi yang ada pada dokumen, seperti karakter atau simbol, angka, emoticon dan link URL.

3.5.1.2 Data *labeling*

Dalam tahapan kali ini, dalam penulisan ini akan membentuk kategori label untuk data tweets yang memiliki arti nilai atu tanggapan dengan makna positif , netral atau negative. Data yang sudah dibersihkan dan diseleksi akan di beri label secara manual yang dilakukan oleh tiga (3) narasumber yang memiliki nilai mata kuliah Bahasa Indonesia yang bagus (minimal B keatas) untuk memilah tweet sentiment postif, dan negatif. Pelabelan dilakukan dengan cara satu komentar tweet di labeli oleh 3 orang narasumber, pelabelan membutuhkan waktu 4 hari masa kerja setelah mengatahui hasil dari label mana yang paling dominan, maka hasil tersebutlah yang akan digunakan.

3.5.1.3 Case Folding

Case Folding merupakan tahapan awal pada *text pre-processing*, yang bertujuan utuk mengubah seluruh huruf kapital menjadi huruh kecil / *lower case*, yang bertujuan untuk mengubah kata menjadi sama.

3.5.1.4 Tokenization

Tokenization adalah langkah memecah sebuah dokumen, yang dapat berupa paragraf atau kalimat, menjadi bagian-bagian yang lebih kecil, dalam beberapa kasus menghilangkan tanda baca yang tidak perlu, sehingga proses tokenization juga dikatakan dilakukan.

3.5.1.5 Filtering Stopword

Pemfilteran adalah langkah untuk mengecualikan kata-kata yang umum tetapi tidak penting. Stopword adalah daftar kata-kata umum yang tidak terlalu penting. Dalam prosesnya, stopword dihilangkan sehingga pengguna dapat lebih fokus pada kata lain yang jauh lebih bermakna dan penting. Contoh stopword: juga, ini, itu, yang lain.

3.5.1.6 Pembobotan Td - Idf

Langkah pra-pemrosesan terakhir adalah menimbang setiap kata dalam teks tweet. Penimbangan ini dilakukan menggunakan metode Td-Idf yang tersedia dalam dokumentasi proses fungsi data yang disediakan oleh Rapidminer Studio. Penimbangan menggunakan metode Td-Idf menghitung bobot setiap kata dalam data untuk menambah nilai algoritma saat membangun model.

3.5.2 Perbandingan Algoritma

3.5.2.1 Naïve bayes

Pengklasifikasi *Naive Bayes* adalah pengklasifikasi paling sederhana dan paling umum digunakan. Model klasifikasi *Naive Bayes* menghitung probabilitas posterior suatu kelas berdasarkan distribusi kata dalam dokumen. Hal itu bergantung pada representasi dokumen yang sangat sederhana sebagai Bag of words. Model ini bekerja dengan mengekstraksi fitur bag of words yang mengabaikan posisi kata dalam dokumen. [11].

3.5.2.2 Suport vector machine

Support Vector Machine (SVM) adalah suatu teknik untuk melakukan suatu prediksi, baik dalam kasus klasifikasi atau regresi. Metode SVM memiliki prinsip dasar linier classifier yaitu kasus klasifikasi yang dapat dipisahkan seara linier, namun SVM yang dikembangkan dapat bekerja dengan problem nonlinier dengan memasukkan konsep kernel pada ruang berdimensi tinggi.[23]

3.5.2.3 K-NN

Algoritma KNN merupakan salah satu algoritma yang paling banyak digunakan. KNN ini termasuk dalam grup Pembelajaran Berbasis Instans. Metode KNN merupakan metode pembelajaran malas. Sebenarnya, metode ini digunakan untuk mengklasifikasikan data yang berjarak sempit. Disebutkan pula bahwa algoritma KNN merupakan algoritma pembelajaran yang banyak digunakan dalam *Cyber Physical Social Systems* (CPSS) untuk analisis dan pengumpulan data.[24]

Berikut akan ditampilkan tabel perbandingan dari ke-tiga algoritma yang telah dipilih, berikut adalah perbandingan tabel algoritma yang akan digunakan

Tabel 3. 1 Perbandingan algoritma

No.	Kategori	Naïve Bayes	*Support Vector Machine* (SVM)	K nearest neighbor (K – NN)
1	Kelebihan algoritma	NB hanya membutuhkan jumlah data pelatihan (*Training* Data) yang kecil untuk menentukan estimasi	SVM dapat menentukan hyperplane atau bidang pemisah dengan memilih bidang dengan optimal margin maka generalisasi pada SVM dapat terjaga dengan	pelatihan sangat cepat, sederhana dan mudah dipelajari, tahan terhadap data pelatihan yang memiliki derau, dan efektif jika data pelatihan besar

		paremeter yang diperlukan dalam proses pengklasifikasian	sendirinya	
2	Cara kerja	NB menggunakan features untuk menentukan variable yang terikat atau tidak dalam melakukan klasifikasi	SVM menggunakan vectors untuk membentuk hyperplane (pembatas klasifikasi)	klasifikasi Terdekat (Nearest Neighbor Classification)
3	Tipe algoritma	Supervised learning	Supervised learning	Supervised learning
4	Kegunaan algoritma	Membuat klasifikasi	Membuat klasifikasi dan menyelesaikan masalah regresi	Melakukan klasifikasi dengan data yang paling dekat

3.6 Pelatihan dan pengujian

Pada tahap selanjutnya adalah melakukan tahapan pelatihan dan pengujian pada data Twitter yang didapatkan. tahap yang dilakukan adalah tahapan pembagian data tweets, dibagi menajdi dua bagian yaitu pengujian dan pelatihan data twitter. data twitter yang dilakukan pada tahapan pengjuian berguna untuk melihat fungsi dari algoritma yang nantianya akan digunakan dan membuat hasil prediksi, sedangkan data twitter yang digunakan sebagai data test berguna untuk

melihat keakuratan atau performanya. proses validasi dari penelitian ini menggunakan K- fold Cross validation untuk dapat meningkatkan tingkat akurasi.

Maka dari itu penelitian ini memanfaatkan nilai lipatan atau nilai k untuk menentukan angka fold terbaik sesuai dengan akurasi yang dihasilkan dari masing-masing algoritma klasifikasi , dalam penelitian ini akan melakukan perbandingan lipatan atau nilai K dua hingga 10 kali lipatan. Dalam melakukan perbandingan nilai lipatan nantinya akan dapat dilihat algoritma manakah yang memiliki nilai akurasi dan auc yang terbaik sesuai dengan nilai lipatan dua hingga sepuluh.

proses pembagian data menjadi beberapa kelompok dengan menggunakan jumlah nilai K yang akan digunakan dan akan dijalankan dalam tahap pelatihan serta tahapan pengujian secara berulang ulang sesuai dengan nilai lipatan, data yang digunakan dalam tahap pengujian akan pergantian dengan data pelatihan yang lainnya. berikut adalah sekenario dalam tahapan pembagian data training dan testing dengan jumlah nilai K = 5

- Jumlah data yang dimiliki berjumlah 1000 data twitter lalu akan dibagi menjadi lima kelompok yang nantinya setiap kelompok memiliki jumlah data sebesar 200 data tweets
- Pada percobaan awal, kelompok pertama akan digunakan sebagai data *Testing* sementara kelompok lainnya akan digunakan sebagai data *training.* pada setiap percobaan akan menggunakan kelompok lainnya secara satu persatu sebagai data pelatihan, sehingga kelompok kelompok lainnya dapat bagian untuk menjadi sebuah data *testing*

  berikut adalah tabel skenario dalam tahapan pembagian dat *testing* dan *training* dengan nilai lipatan atau nilai K= 5.

Tabel 3. 2 Skenario Perbandingan Data Training & testing

Nilai K = 1	I	II	III	IV	V
Nilai K = 2	I	II	III	IV	V
Nilai K = 3	I	II	III	IV	V
Nilai K = 4	I	II	III	IV	V

Nilai K = 5	I	II	III	IV	V

Keterangan warna : = Data Training

= Data Testing

3.7 Evaluasi hasil

Setelah dilakukan proses analisis sentimen, selanjutnya dilakukan evaluasi hasil yang didapat dengan menghitung *Confusion Matrix*. Dalam melakukan pengujian dan pelatihan pada setiap model algortima lalu akan dipilih untuk dijadikan perwakilan dalam melakukan perbandingan algoritma mana yang terbaik dengan nilai K yang diuji.Perhitungan *Confusion Matrix* akan menghasilkan Akurasi, *Recall*, dan Presisi yang dalam memberikan pandangan bagaimana kinerja algoritma Naïve Bayes dan *Support Vector Machine* dan K- NN bekerja. setelah perhitungan *confusion Matrix* dalam melakukan perhitungan akurasi, lalu membuat grafik ROC-AUC untuk bisa dilihat kemampuan algoritma saat melakukan pemodelan klasifikasi. Setelah melalui semua tahapan pelatihan dan pengujian terhadap model yang di hasilkan oleh tiap tiap algoritma adalah hasil yang terbaik untuk dijadikannya acuan pada penelitaian dalam melakukan perbandingan algoritma .