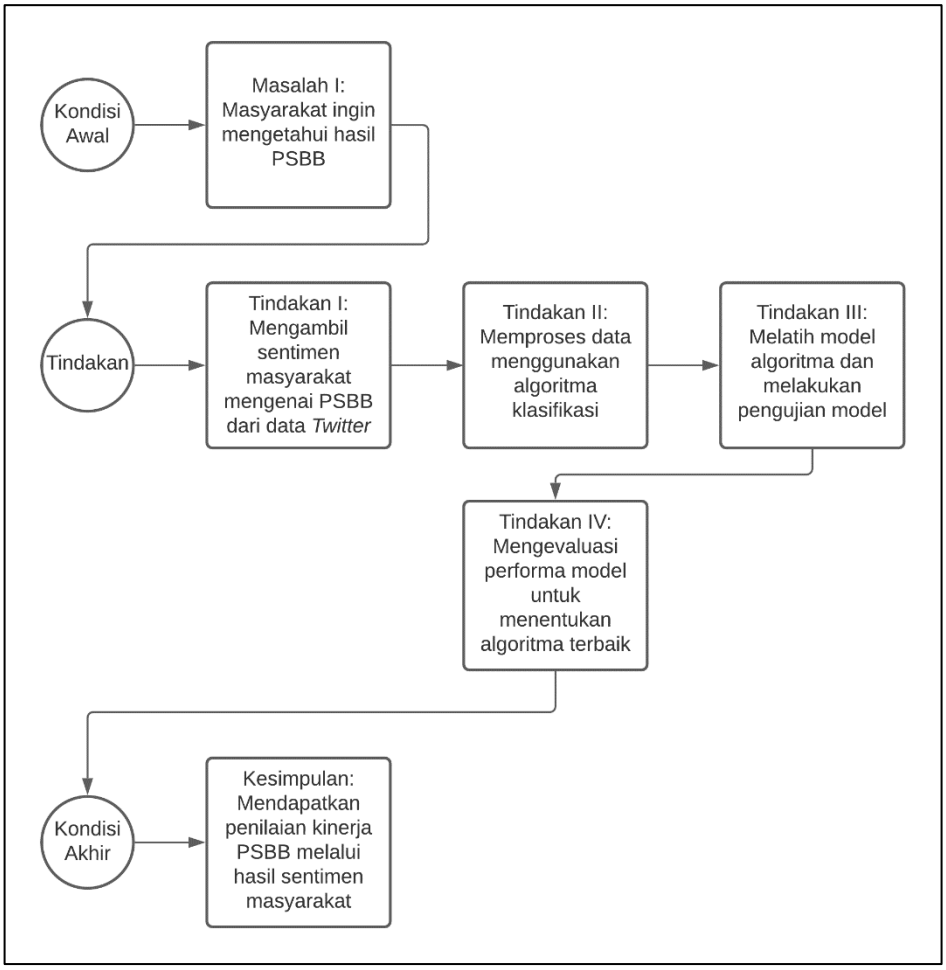
BAB III

METODOLOGI PENELITIAN

3.1. Gambaran Umum Objek Penelitian



Gambar 3.1. Objek Penelitian

Berdasarkan dari Gambar 3.1, objek penelitian dari skripsi ini adalah pendapat atau sentimen masyarakat Indonesia mengenai Pembatasan Sosial Berskala Besar (PSBB) yang diambil menggunakan sumber berupa media sosial

Twitter. Permasalahan pada objek penelitian ini adalah masyarakat ingin mengetahui hasil dari pelaksanaan PSBB. Dimana untuk menemukan hasil PSBB dapat dilakukan dengan menganalisis sentimen masyarakat. Melalui analisis sentimen, pendapat masyarakat digunakan untuk mencari tahu hasil sentimen melalui proses algoritma klasifikasi dalam menghasilkan model klasifikasi sentimen. Selanjutnya adalah menguji hasil model menggunakan *K-fold Cross Validation* untuk menemukan algortima terbaik dalam menangani kasus sentimen PSBB. Dengan begitu, sentimen dapat dibuat menjadi kesimpulan untuk mengetahui hasil PSBB.

Dalam penelitian ini, data yang digunakan untuk melakukan analisis sentimen adalah data *tweets* dalam bentuk teks. Selain itu, pengambilan data yang digunakan dalam penelitian ini terbatas pada periode 18 Februari 2021 dan 18 Maret 2021 dengan jumlah data sebesar 1400 *tweets*.

3.2. Metode Penelitian

Dalam memilih metode penelitian yang terbaik, penulis membuatkan tabel perbandingan antara tiga (3) algoritma terkenal yang umum digunakan dalam analisis sentimen. Ketiga algoritma tersebut adalah *Support Vector Machine*, *Naïve Bayes,* dan *Logistic Regression*. Perbandingan ini dilakukan dengan menyimpulkan hasil dari penelitian - penelitian terdahulu. Berikut adalah Tabel 3.1 mengenai perbandingan algoritma analisis sentimen:

Tabel 3.1. Perbandingan Algoritma SVM, NB, dan LR

No	Kategori	*Support Vector Machine (SVM)*	*Naïve Bayes (NB)*	*Logistic Regression (LR)*
1.	Tingkat Akurasi	SVM memiliki tingkat akurasi tertinggi.	NB memiliki tingkat akurasi yang cukup tinggi.	LR memiliki tingkat akurasi yang cukup tinggi
2.	Klasifikasi	SVM menghasilkan klasifikasi yang masuk ke dalam (*Excellent Classification*).	NB menghasilkan klasifikasi yang masih masuk ke dalam (*Poor Classification).*	LR masuk ke dalam (*Good Classification).*
3.	Tipe Algoritma	*Supervised Learning.*	*Supervised Learning.*	*Supervised Learning.*
4.	Kegunaan Algoritma	Membuat klasifikasi dan menyelesaikan masalah regresi.	Membuat klasifikasi.	Membuat klasifikasi.dan menyelesaikan masalah regresi.
5.	Cara Kerja	SVM menggunakan *vectors* untuk	NB menggunakan *features* untuk menentukan	LR menggunakan kurva S

No	Kategori	*Support Vector Machine (SVM)*	*Naïve Bayes (NB)*	*Logistic Regression (LR)*
		membentuk *hyperplane* (pembatas klasifikasi)	*variable* yang terikat atau tidak dalam melakukan klasifikasi.	(*sigmoid*) untuk melakukan klasifikasi.
6.	Ruang Kerja	Fungsi – fungsi linier berdimensi tinggi.	Dimensi peluang dari data.	Fungsi – fungsi sigmoid berdimensi tinggi.
7.	Tingkat Kesulitan	Sulit untuk diimplementasi.	Mudah untuk diimplementasi.	Mudah untuk diimplementasi
8.	Keterbatasan Algoritma	Semakin banyaknya data akan mempersulit penggunaan SVM.	Menggunakan asumsi variabel bebas dalam membuat klasifikasi.	Rentan terhadap variabel yang tidak seimbang menyebabkan pengurangan akurasi.

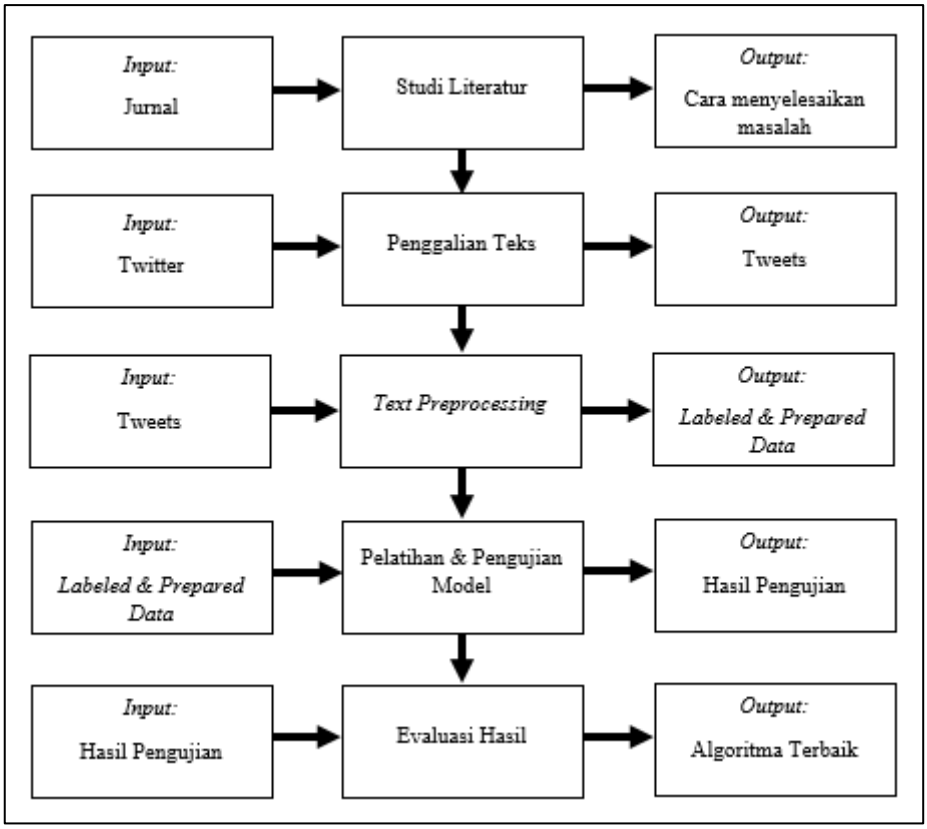
Penjelasan kategori untuk perbandingan algoritma yang berada pada Tabel 3.1:

1. Tingkat akurasi: Tingkat keakuratan yang dimiliki algoritma dalam menghasilkan klasifikasi.

2. Klasifikasi: Aspek penilaian tentang hasil klasifikasi yang dihasilkan algoritma.
3. Tipe algoritma: Tipe pembelajaran yang digunakan algoritma dalam membuat klasifikasi.
4. Kegunaan algoritma: Dasar tujuan atau kemampuan sebuah algoritma yang diciptakan terhadap pemecahan masalah.
5. Cara kerja: Metode yang digunakan algoritma dalam melakukan pembagian klasifikasi.
6. Ruang kerja: Perubahan data menjadi ruang kerja yang digunakan algoritma dalam mengolah proses klasifikasi.
7. Tingkat kesulitan: Tahapan atau proses yang dimiliki algoritma dalam melakukan klasifikasi.
8. Keterbatasan algoritma: Batasan yang dimiliki sebuah algoritma dalam melakukan klasifikasi.

Berdasarkan Tabel 3.1, perbedaan – perbedaan yang dimiliki setiap algoritma maka pada penelitian ini akan menggunakan ketiga algoritma tersebut untuk menemukan algoritma terbaik dalam menangani klasifikasi sentimen PSBB. Pemilihan teknik analisis pada penelitian ini didasarkan pada keunikan cara kerja algoritma dalam melakukan klasifikasi.

Dalam melakukan penelitian, penulis menggunakan kerangka pikir sebagai berikut (Gambar 3.2):



Gambar 3.2. Kerangka Pikir Penelitian

Pembuatan kerangka pikir (Gambar 3.2) ini dibuat menggunakan inspirasi dari kerangka pikir yang telah digunakan pada penelitian - penelitian terdahulu dengan rangkaian penyesuaian yang dilakukan. Berikut adalah penjelasan langkah - langkah penelitian yang ada dalam kerangka pikir penelitian:

1. **Studi Literatur**

   A. *Input*: Mengumpulkan jurnal - jurnal penelitian terdahulu.

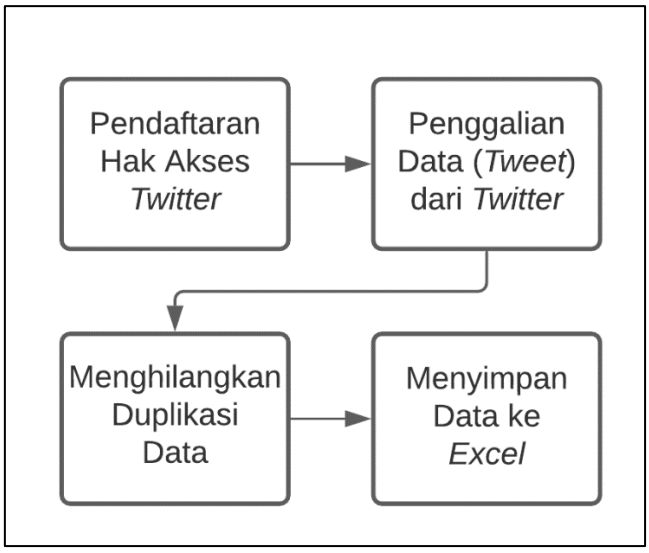
   B. *Process*: Mempelajari penemuan - penemuan yang ada pada jurnal peneliatian terdahulu.

   C. *Output*: Memperoleh metode - metode penyelesaian masalah dalam melakukan analisis sentimen.

2. **Penggalian Teks (*Text Mining*)**

   A. *Input*: Data yang digunakan berupa *tweets* yang berasal dari *Twitter*.

   B. *Process*:



Gambar 3.3. Proses Penggalian Teks

Pengambilan data teks dilakukan dengan menggunakan program *Rapidminer Studio*. Dalam pengambilan data terdiri dari alur proses yang dapat dilihat pada Gambar 3.3, yaitu:

1) Melakukan pendaftaran hak akses menggunakan akun *rapidminer* untuk dapat melakukan penggalian data dari *Twitter.*

2) Melakukan penggalian data (*tweet*) dengan kata kunci yang terdiri dari "PSBB" dan "Pembatasan Sosial Berskala Besar".

3) Sebelum menyimpan hasil penggalian data, dilakukan penghapusan data yang duplikat agar hasil data menjadi lebih bersih.

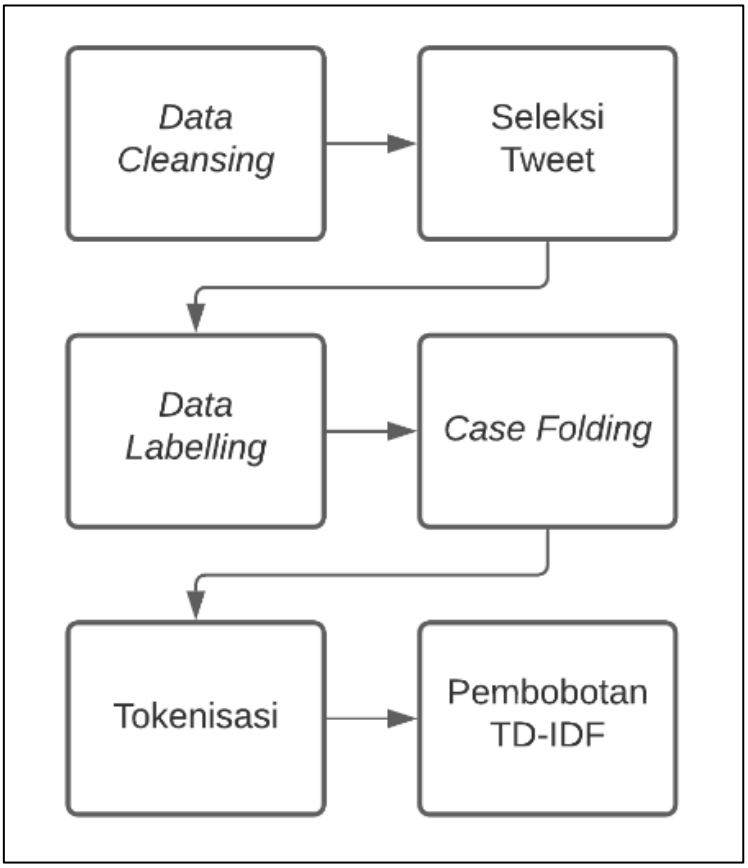
4) Hasil penggalian data disimpan ke dalam *file excel.*

C. *Output*: Data teks yang diperoleh ada *tweets* yang terdiri dari kata kunci "PSBB" dan "Pembatasan Sosial Berskala Besar".

3. *Text Preprocessing*

A. *Input*: Data yang *tweet* yang telah diperoleh dari penggalian data.

B. *Process*:



Gambar 3.4. Proses *Data Labelling*

Gambar 3.4 merupakan alur proses untuk melakukan *text preprocessing* yang terdiri dari:

1) Melakukan pembersihan data *tweets* sehingga hanya berisi teks yang murni dan juga menerapkan penghapusan data yang masih lolos pada saat penggalian data

2) Melakukan seleksi *tweets* secara manual untuk *tweets* yang tidak memiliki arti karena akan menjadi *noise* saat diterapkan pada klasifikasi.

3) Memberikan label pada setiap *tweet* untuk menandakan kategori sentimen yang dimiliki *tweet* tersebut.

4) Mengubah semua huruf besar pada *tweets* menjadi huruf kecil.

5) Membuat tokenisasi kata dari *tweets.*

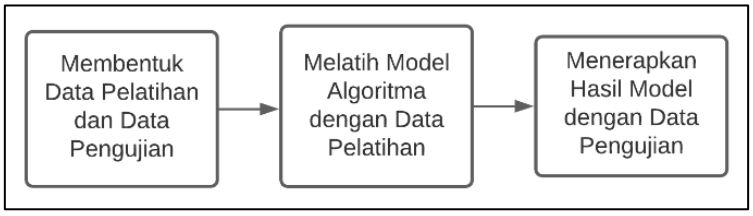
6) Memberikan pembobotan kata terhadap *tweets* menggunakan metode TD-IDF.

C. *Output*: Menghasilkan data yang siap untuk digunakan dalam analisis sentimen.

4. Pelatihan dan Pengujian Model

A. *Input*: Data yang sudah diproses sebelumnya dibagi menjadi dua untuk *data training* dan *data testing* menggunakan *K-fold Cross Validation*.

B. *Process*:



Gambar 3.5. Proses *Training Model*

Gambar 3.5 merupakan alur proses untuk melakukan pelatihan model algoritma yang terdiri dari:

1) Membentuk data pelatihan dan data pengujian dengan menentukan nilai K pada *cross validation*.
2) Data pelatihan digunakan untuk membentuk model.
3) Data pengujian digunakan untuk menguji model yang sudah terbentuk.

C. *Output*: Hasil pengujian setiap algoritma (*confusion matrix* dan AUC.

5. Evaluasi Hasil

A. *Input*: Hasil pengujian dari setiap algoritma.

B. *Process*: Melakukan seleksi hasil pengujian terbaik pada setiap algoritma untuk dijadikan perbandingan antar ketiga algoritma (SVM, NB, LR).

C. *Output*: Algoritma terbaik.

3.3. Variabel Penelitian

Berdasarkan penelitian yang akan dilakukan ini, variable penelitian adalah pendapat masyarakat mengenai PSBB di Indonesia. Dimana variabel bebas dari peneltian ini adalah opini masyarakat Indonesia terhadap Pembatasan Sosial Berskala Besar di Indonesia yang berada pada *Twitter*. Sehingga variabel terikat dari penelitian ini adalah data sentimen atau tanggapan dari masyarakat terhadap PSBB.

3.4. Teknik Pengumpulan Data

Teknik pengumpulan data berupa teks yang dibutuhkan dalam melakukan analisis sentimen ini diambil dari media sosial *Twitter* dengan menggunakan program *Rapidminer Studio.* Proses pengumpulan data ini dilakukan menggunakan dua (2) periode, yaitu 18 Februari 2021 dan 18 Maret 2021 menggunakan kata kunci berupa "PSBB" serta "Pembatasan Sosial Berskala Besar". Hasil dari pengumpulan data ini kemudian disimpan ke dalam format excel (.xlsx) untuk mempermudah akses dalam melihat hasil dan proses *labelling* data nantinya.

3.5. Teknik Pengambilan Sampel

Pengambilan sampel ini dilakukan melakukan seleksi hasil terhadap penggalian data *tweets* yang berisi sentimen mengenai PSBB dan memiliki nilai sentimen berupa positif atau negatif. Jumlah sampel data yang akan digunakan dalam membuat analisis sentimen adalah 1400 *tweets*.

3.6. Teknik Analisis Data

Berdasarkan seleksi penggunaan data yang mengandung nilai sentimen (positif dan/atau negatif), teknik analisis data yang digunakan dalam penelitian ini adalah kualitatif. Analisis data pada penelitian ini dilakukan dengan menggunakan program *Rapidminer Studio* dalam membuat model analisis sentimen menggunakan tiga (3) algoritma klasifikasi sebagai pembanding hasil. Ketiga algoritma tersebut adalah *Support Vector Machine*, *Naïve Bayes,* dan *Logistic Regression*. Berikut adalah langkah - langkah yang digunakan dalam menganalisis data:

3.6.1. Studi Literatur

Studi literatur merupakan langkah awal dalam melakukan penelitian ini. Pelaksanaan dari studi literatur ini adalah untuk mencari dan memahami langkah - langkah terbaik yang dapat digunakan dalam penelitian ini berdasarkan literatur - literatur yang ada.

3.6.2. *Text Preprocessing*

Langkah selanjutnya dalam analisis sentimen adalah mempersiapkan hasil pengumpulan data dari tahap sebelumnya agar menjadi data yang siap untuk diolah. Dalam tahap - tahap *preprocessing* ini terdiri dari:

3.6.2.1. *Data Cleansing*

Tahap awal dalam *text preprocessing* merupakan pembersihan data *tweets*. Berikut adalah rincian dari langkah – langkah yang digunakan untuk membersihkan data:

1. Membuka file excel yang berisi hasil penggalian data dari *Twitter*.
2. Menghapus URL yang terdapat dalam *tweets.*
3. Menghapus semua isi teks dalam *tweets* yang bukan merupakan alfabet sehingga menghasilkan teks yang hanya berupa alfabetis.
4. Menghapus semua kata "RT" yang terdapat dalam *tweets*.
5. Menghapus kembali data *tweets* yang terduplikat.
6. Menyimpan hasil pembersihan data ini pada file excel yang baru.

3.6.2.2. Seleksi *Tweet*

Pada seleksi *tweet*, dilakukan dengan menggunakan data yang sudah melalui tahap pembersihan data (memakai file excel yang baru). Proses seleksi *tweet* ini dilakukan secara manual oleh dua narasumber yang memiliki nilai mata kuliah Bahasa Indonesia yang bagus (nilai B keatas) untuk memilah *tweets* yang mempunyai sentimen (positif / negatif) dan akan digunakan sebagai bahan analisis sentimen. Penelitian ini tidak akan menggunakan sentimen yang tidak memiliki nilai atau bersifat netral untuk mendapatkan hasil akurasi terbaik.

3.6.2.3. *Data Labelling*

Dalam tahap ini, penulis membentuk kategori label untuk data *tweets* yang memiliki nilai atau tanggapan positif dan negatif. Kategori label tersebut digunakan untuk memasang label pada setiap data *tweet* yang sudah diperoleh dari penggalian data *tweets* dari *Twitter*. Pembentukan *labelling* ini akan digunakan sebagai dasar algoritma klasifikasi dalam membentuk model. Proses data labelling ini akan dilakukan oleh dua narasumber yang memiliki nilai mata kuliah Bahasa Indonesia yang bagus (nilai B keatas) untuk menentukan kategori klasifikasi teks sebagai berikut:

Tabel 3.2. Kategori Label

No.	Kategori Label	Ketentuan
1.	Positif	*Tweet* mengandung makna atau kata - kata publikasi positif mengenai PSBB.
2.	Negatif	*Tweet* mengandung makna atau kata - kata publikasi negatif mengenai PSBB.

3.6.2.4. *Case Folding*

Pada tahap ini, semua huruf pada data *tweets* yang telah memiliki label akan diubah menjadi huruf kecil. Proses ini dilakukan agar semua kata dalam kalimat memiliki konsistensi huruf kecil yang sama sehingga mempermudah proses tokenisasi nantinya dalam membedakan kata - kata.

3.6.2.5. Tokenisasi

Tahapan ini melibatkan proses pemecahan kata – kata pada setiap *tweets* yang kemudian dijadikan kata – kata yang terpisah (menjadi sebuah kumpulan token). Dengan memisahkan kata – kata menjadi individu akan mempermudah pelatihan model pengklasifikasian nantinya.

3.6.2.6. Pembobotan TD-IDF

Langkah terakhir dari *text preprocessing* adalah memberikan penilaian pembobotan pada setiap kata dari teks *tweets*. Pembobotan ini dilakukan dengan menggunakan metode TF-IDF yang tersedia pada fitur *process documents from data* yang disediakan dari *Rapidminer Studio*. Pembobotan menggunakan metode TD-IDF akan menghitung bobot dari setiap kata yang ada dalam data agar menambah nilai pada algoritma dalam membuat model.

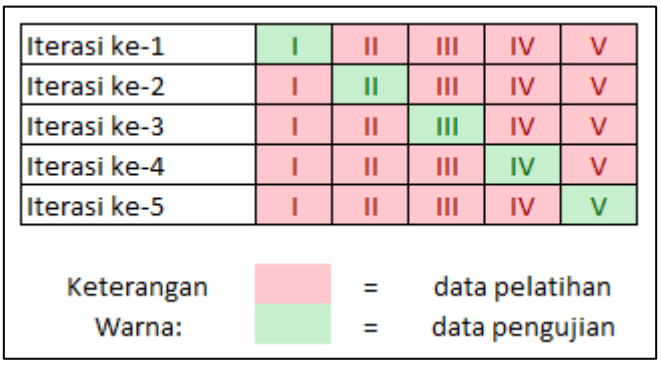
3.6.3. Pelatihan dan Pengujian Model

Tahap selanjutnya adalah melakukan pembagian data (*tweets*) yang sudah melewati *text preprocessing* menjadi dua bagian, yaitu: data pengujian dan data pelatihan. Proses pelatihan dan pengujian model ini dilakukan menggunakan metode *K-fold Cross Validation*. Penggunaan metode tersebut bertujuan untuk mengukur hasil akurasi dari model yang dihasilkan dan menemukan hasil yang terakurat yang dapat dihasilkan. Pengujian model menggunakan *K-fold Cross Validation* akan dilakukan menggunakan dua hingga sepuluh kali lipatan (K = 2 hingga 10) untuk menemukan hasil terbaik yang dapat dihasilkan oleh algoritma klasifikasi tersebut [22]. Proses pengujian akurasi bekerja dengan membagi data menjadi beberapa kelompok mengikuti jumlah nilai K yang telah ditentukan dan

menjalankan pelatihan serta pengujian secara iterasi. Dalam setiap setiap iterasi yang dilakukan kelompok data yang digunakan untuk pengujian akan bergantian dengan kelompok data (pelatihan) yang lain mengikuti jumlah proporsi nilai K. Untuk memperjelas cara kerja *K-fold Cross Validation* berikut adalah simulasinya untuk nilai K=5:

1. Data terdiri atas 1000 *tweets* akan dibagi menjadi 5 kelompok yang pada setiap kelompok terdiri atas 200 *tweets*.

2. Pada iterasi pertama, kelompok I akan digunakan sebagai data pengujian sementara 4 kelompok lainnya akan dipakai untuk data pelatihan. Setiap iterasi yang dilakukan akan menggunakan kelompok selanjutnya sebagai data pelatihan sehingga setiap kelompok akan kebagian giliran untuk menjadi data pelatihan. Gambar 3.6 merupakan visualisasi iterasi *cross validation*.



Gambar 3.6. Visualisasi Simulasi Cross Validation

3. Setiap iterasi yang dilakukan menghitung nilai akurasi yang dimiliki menggunakan rumus berikut:

$$Akurasi = \frac{Total\ data\ pengujian\ yang\ benar}{Total\ data\ pengujian} x100$$

Rumus 3.1. Akurasi *K-fold Cross Validation*

4. Setelah selesai melakukan lima kali iterasi pengujian maka rata - rata hasil akurasi akan dihasilkan menggunakan akurasi dari setiap iterasi.

3.6.4. Evaluasi Hasil

Setelah melakukan pelatihan dan pengujian terhadap model yang dihasilkan oleh setiap algoritma adalah memilih hasil terbaik pada setiap algoritma untuk dijadikan perwakilan dalam melakukan perbandingan algoritma. Dalam memlih hasil pengujian berdasarkan *fold* terbaik setiap algoritma dan algoritma terbaik ini akan menggunakan *confusion matrix* dalam menghitung akurasi dan kemudian membuat kurva ROC-AUC untuk melihat kemampuan algoritma dalam melakukan klasifikasi.