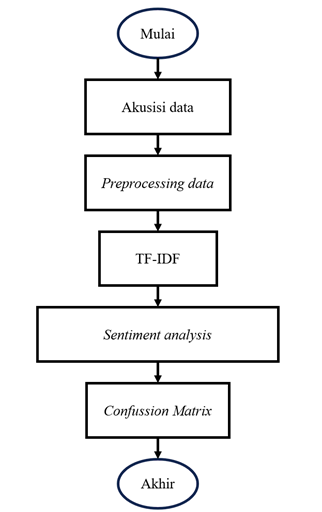
2. METODE PENELITIAN

Penelitian ini melalui beberapa alur tahapan pengerjaan seperti yang tertera pada gambar 1 dibawah ini,



Gambar 1. Alur tahapan pengerjaan

## 2.1 Akuisisi Data

Data yang diperlukan untuk melakukan sentimen analisis pada penelitian ini ialah data teks atau opini masyarakat yang didapatkan dari Kaggle pada tautan<https://www.kaggle.com/datasets/cttrhnn/cyberbullying-bahasa-indonesia/code>. Dataset ini merupakan dataset *cyberbullying* bahasa indonesia dengan sebanyak 650 dokumen yang terdiri dari dua label yakni *bullying* dan *non-bulyying*. Data latih yang digunakan 70% dari total data lalu sisanya digunakan sebagai data uji

## 2.2 Data Preprocessing

Data preprocessing merupakan tahapan awal untuk mengelola data agar data dapat dipelajari dengan mudah oleh mesin. Pada peneliian, tahapan *data prepocessing* antara lain:

## 2. 2.1 Case Folding

Case folding merupakan tahapan mengubah kata pada setiap dokumen menjadi huruf kecil. Case folding dilakukan karena mesin mengartikan huruf kecil dan huruf kapital memiliki arti yang berbeda, artinya mesin akan menganggap bahwa kata “Saya” dengan “saya” adalah dua kata yang berbeda, padahal kata tersebut merupakan satu kata yang sama.

## 2. 2.2 Data Cleanning

Data cleaning merupakan tahapan menghilangkan seluruh tanda baca seperti titik (.), koma (,) tanda seru (!), dan lain-lain karena tanda baca dianggap tidak memiliki arti yang begitu penting untuk mesin pada penelitian ini sehingga *noise* dapat dikurangi.

## 2. 2.3 Stopword Removal

Stopword removal merupakan tahapan untuk menghilangkan kata yang terlalu sering muncul dalam corpus. Stopword removal dilakukan dengan pertimbangan bahwa kata yang terlalu sering muncul bukanlah kata kunci atau kata yang memiliki arti yang penting untuk sentimen analisis. Contohnya, ‘dan’, ‘atau’, ‘ketika’, ‘kenapa’, dan lain-lain. Hal ini dilakukan agar fitur yang digunakan nantinya tidak akan terlalu besar dan tidak terlalu membebani sistem [8].

## 2. 2.4 Stemming

Stemming merupakan tahapan mengubah kata menjadi bentuk dasar dari kata tersebut. Contohnya, “berbahagia” berubah menjadi kata “bahagia” setelah melalui proses stemming. Dalam stemming data terdapat dua aturan yang bisa digunakan yaitu dengan pendekatan kamus atupun pendekatan aturan [9].

2. 2.5 Tokenizaiton

Tokenization merupakan tahapan pemecahan dokumen dalam korpus menjadi satu bagian kata. Tokenization dilakukan untuk mempermudah proses pembobotan kata menggunakan TF-IDF pada tahap selanjutnya.

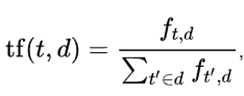
Hasil akhir dari tahap *preprocessing data* ialah data yang telah dipecah dari kalimat pada dataset menjadi unit-unit terkecil yakni kata dasar dan baku yang telah di *lowercase.* Tabel 2 merupakan contoh data yang telah melewati tahap seluruh *preprocessing data,*

## 2.3 Pembobotan Kata TF-IDF

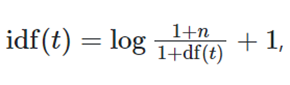
Mesin tidak dapat memproses data teks secara langsung, maka dari itu, dokumen yang telah melaui tahap tokenization saat data prepocessing perlu dikonversi tiap kata menjadi numerik. Tahap konversi kata menjadi numerik dilakukan dengan metode TF-IDF yang merupakan metode pembobotan kata dengan cara menghitung seberapa sering dan seberapa unik kata tersebut dalam korpus [10]. Secara matematis, TF-IDF dapat dituliskan sebagai berikut:

 (1)

Dimana tfij merupakan hasil perhitungan dari rumus Tearm Frequenty (TF) dan idfj merupakan hasil perhitungan dari rumus Inverse Document Frequency (IDF). Adapun rumus dari TF adalah sebagai berikut:

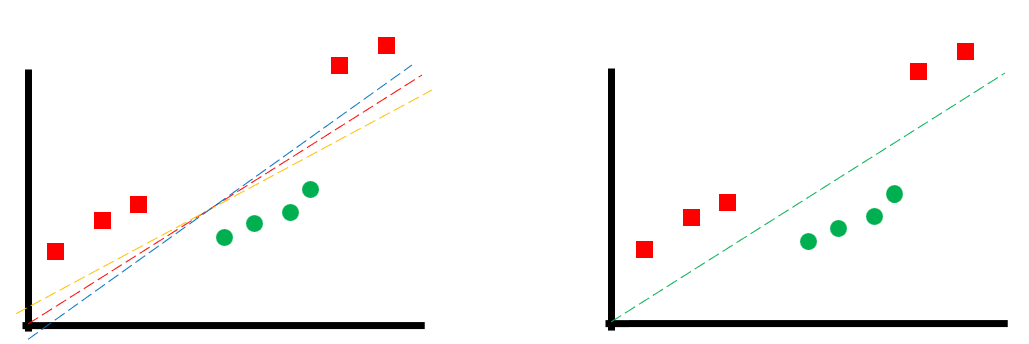
(2)

Sedangkan rumus dari IDF adalah sebagai berikut:

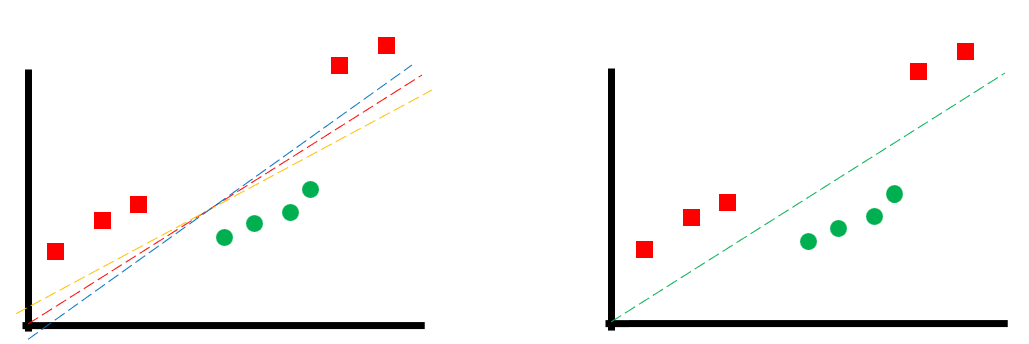
(3)

## 2.4 Support Vector Machine

Support Vector Machine (SVM) merupakan sebuah metode pembelajaran mesin untuk klasifikasi dengan pendekatan mencari *hyperplane* di ruang n-dimensi agar dapat menklasifikasikan data [11]. Seringkali, terdapat lebih dari satu *hyperplane* untuk menklasifikan data, maka dari itu, SVM akan mencari *hyperplane* terbaik seperti pada gambar dibawah ini,



Gambar 2. Contoh klasifikasi SVM dengan beberapa kemungkinan *hyperplane*



Gambar 3. Contoh klasifikasi SVM setelah diketahui *hyperplane* terbaik

*Hyperplane* dapat diketahui dari perhitungan *margin* terbesar*. Margin* merupakan selisih jarak terdekat antara hyperplane dengan data terdekat pada setiap kelas.

Pada beberapa kasus, kita tidak dapat menklasifikasikan data pada n-ruang dimensi karena tidak ada *hyperplane* yang dapat memisahkan date ke dalam kelas-kelas yang ada. Sehingga, kita perlu memetakan data ke dalam ruang yang lebih tinggi, hal ini disebut dengan fungsi karnel [12]. Ada beberapa fungsi karnel yang dapat digunakan antara lain:

## 2. 4.1 Kernel Linier

Umumnya digunakan pada data berdimensi satu atau dua.

## 2. 4.2 Kernel Polinomial

Karnel yang digunakan pada data berdimensi tinggi dengan cara mendefinisikan jarak antara dua titik data menggunakan polinomial.

## 2. 4.3 Kernel RBF

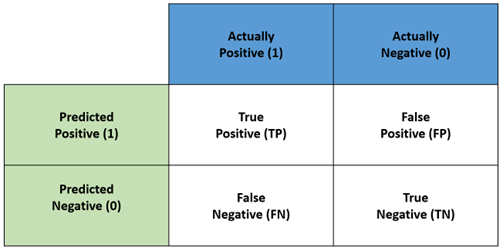
Karnel yang digunakan pada data berdimensi tinggi dengan cara mendefinisikan jarak antara dua titik data menggunakan fungsi radial basis.

## 2. 4.4 Kernel sigmoid

Karnel yang digunakan pada data berdimensi tinggi dengan cara mendefinisikan jarak antara dua titik data menggunakan fungsi sigmoid.

## 2.5 Confusion Matrix

Confusion matrix merupakan tabel yang digunakan untuk mengukur kinerja model klasifikasi. Confusion matrix dapat dilihat pada gambar 2 dibawah ini,



Gambar 2. Tabel Confusion matrix