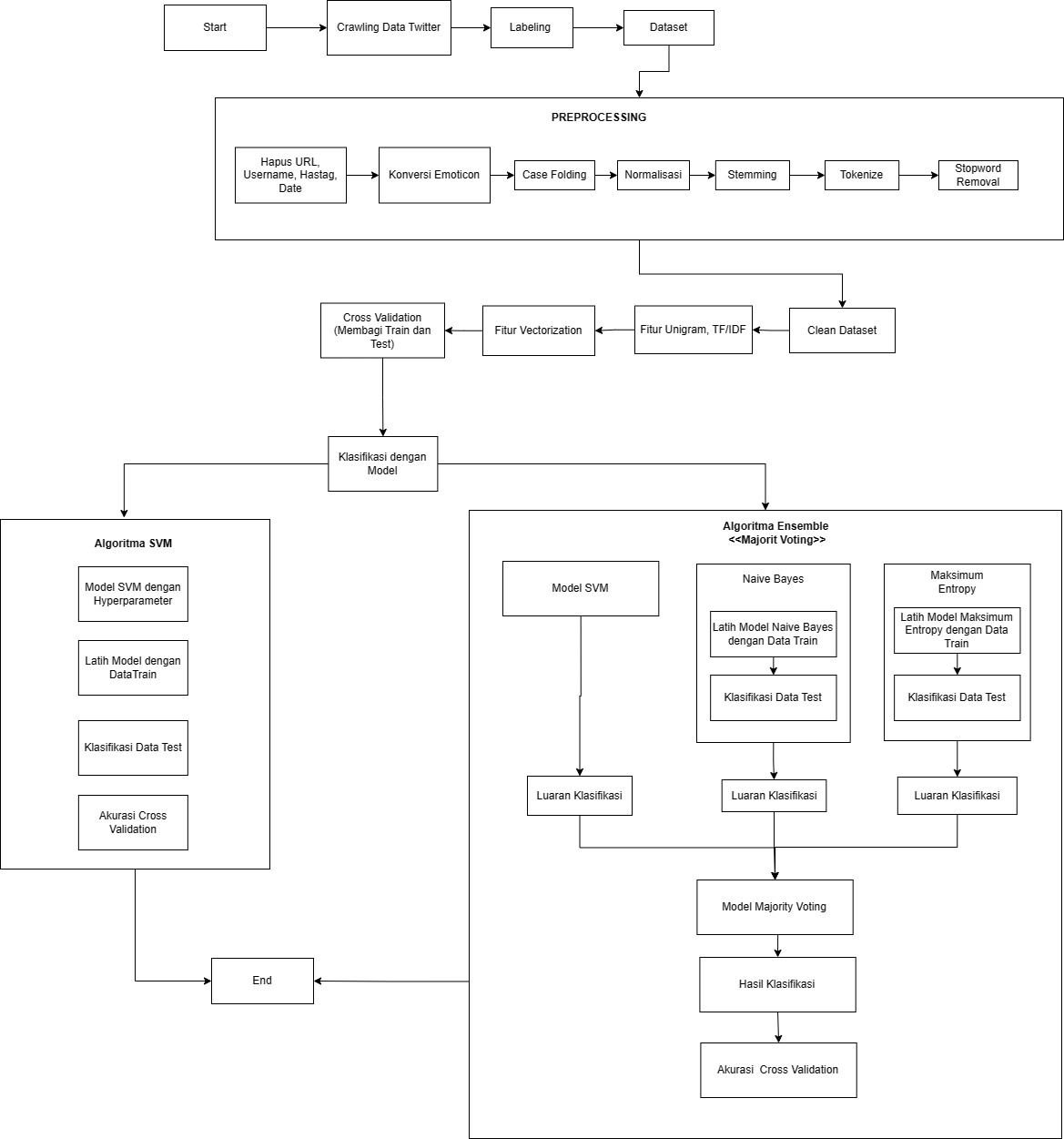
**BAB III**

**METODE PENELITIAN**

1. **Alur Penelitian**

****

**Gambar 1 Flowchart Penelitian**

1. **Pengumpulan Data**

Data yang digunakan dalam penelitian ini adalah tweet atau cuitan masyarakat di wilayah Indonesia yang diambil dari aplikasi media social twitter. Twitter

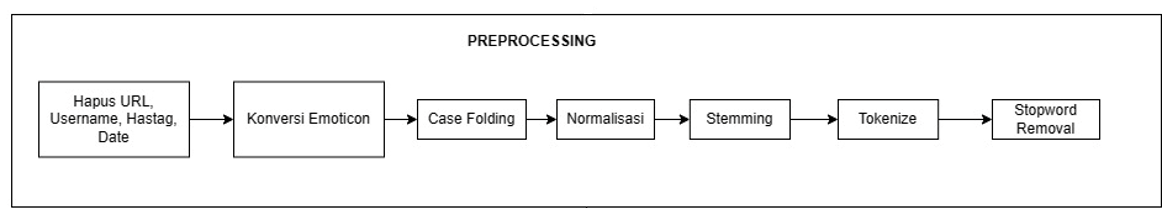
adalah sebuah situs microblogging, yang popular karena sebuah pesan singkat yang lebih dikenal dengan sebuta tweet. [1]. Tweet adalah ungkapan dari para pengguna yang bertujuan untuk mengungkapkan isi dari dalam pikiran, biasanya berkaitan dengan kondisi social masyarakat yang sedang terjadi.

Data yang diambil berjumlah 11.127 data tweet dari tanggal 24/03/2023 sampai 05/05/2023 yang didapat dengan menggunakan Teknik crawling. Data tweet yang didapat hanya bisa diambil dalam rentang waktu seminggu sesuai dengan kebijakan baru twitter API. Data dicari berdasarkan #ChatGPT dengan beberapa keyword pencarian yang termasuk seperti kata akademik, tugas, pr, skripsi, kurikulum, kuis, jurnal, paper, makalah, dll.

1. **Labeling**

Data yang didapat belum memiliki label, sehingga harus dilakukan pelabelan sentiment terlebih dahulu. Pelabelan ini menggunakan algoritma HuggingFace Transformers. HuggingFace Transformers adalah sebuah pustaka *open source* yang dirancang untuk memfasilitasi pengguna mengakses model berskala besar untuk membangun dan bereksperimen. [2]. Hugging Face Transformers biasa digunakan untuk melakukan berbagai tugas NLP seperti klasifikasi teks, menjawab pertanyaan, penerjemah mesin, dan label urutan. 11.127 data tweet akan diterjemahkan ke dalam Bahasa inggris, lalu dilakukan pelabelan dengan menggunakan Hugging Face Transformers, setelahnya akan diterjemahkan kembali ke dalam Bahasa Indonesia dengan bentuk yang sudah dilabeli.

1. **Preprocessing**



**Gambar 2 Alur Preprocessing Data**

1. ***Cleaning* : Hapus username, hastag, date, dan url**

Proses untuk menghilangkan semua karakter di dalam tweet yang tidak termasuk alfabet. Proses ini akan menghilangkan angka, tanda ttitik, koma, hastag, url dan symbol symbol yang yidak memiliki makna. Dalam proses ini juga akan dihapus username dan date karena dianggap tidak berpengaruh dalam pemrosesan teks lebih lanjut.

1. ***Emoticon conversion***

Proses untuk mengkonversi emoticon dari teks. Emoticon akan diubah ke dalam bentuk kalimat. Misalnya jika terdapat emot 😊 akan dikonversi menjadi “tersenyum”.

1. ***Case Folding***

Proses penyeragaman kata dengan cara mengubah semua kata menjadi huruf kecil (lowercase).

1. ***Normalization***

Proses untuk mengubah singkatan, akronim, atau kata-kata yang rancu dalam data teks. Dengan dilakukan normalisasi dapat menangani data yang tidak seimbang. Contohnya adalah :

|  |  |
| --- | --- |
| Original Word | Normalization Word |
| aj | saja |
| yg | yang |

1. ***Stemming***

Proses untuk menghilangkan awalan dan akhiran pada suatu akat untuk mendapatkan kata dasar dari suatu dokumen. Proses stemming ini menggunakan library sastrawi

1. ***Tokenize***

Proses pemotongan kalimat berdasarkan penyusunan katanya. Tokenize biasanya akan memenggal kata per kata dari sebuah kalimat tweet. Contohnya : “ChatGPT sangat membantu dalam menyelesaikan tugas kuliah”. Setelah melalui proses tokenize maka akan menjadi : ‘ChatGP’, ‘sangat’, ‘membantu’, ‘dalam’, ‘menyelesaikan’,’tugas’,’kuliah’.

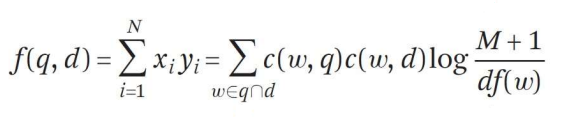
1. ***StopWord Removal***

Proses untuk mengambil kata kata penting dari hasil tokenizing. Proses ini juga bisa disebut dengan pengeliminasian kata kata sesuai dengan kaidahnya. Kata-kata yang akan dihilangkan adalah kata-kata yang dirasa tidak berpengaruh terhadap hasil sentiment, seperti aku, dan, yang, dll.

1. **Pembobotan Kata**

TF-IDF adalah metode pembobotan kata yang digunakan untuk mengukur signifikasi fitur dalam suatu dokumen. Pendekatan N-gram rentan terhadap overfit karena kapasitasnya untuk meningkatkan jumlah fitur. TF-IDF menangani ngram frekuensi tinggi dan rendah secara implisit. N-gram frekuensi tinggi tidak membantu membedakan tweet tapi N-gram dengan frekuensi rendah cenderung overfit.

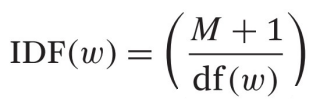
Secara matematis perhitungan TF-IDF dapat dituliskan sebagai berikut:

 (1)

Persamaan 1 diatas, terdiri dari perhitungan TF dan perhitungan IDF. Perhitungan TF dihitung dnegan mempertimbangkan jumlah atau frekuensi term/word dalam dokumen. Secara matermatis ditulis :

 (2)

Perhitungan IDF dihitung dengan memberi nilai lebih untuk kata/term yang tiddak banyak muncul dalam dokumen. Secara matematis ditulis :

 (3)

1. **Penggunaan Fitur**

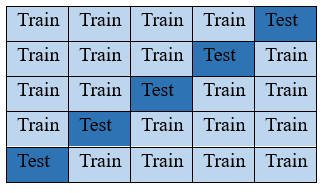
Fitur adalah atribut terukur yang berbeda di setiap sampel data inputan [3]. Dalam melakukan sentiment, perlu dilakukan ekstraksi fitur dan pemilihan fitur. Ekstraksi fitur melibatkan penentuan semua nilai input yang dapat menggambarkan objek yang akan diberikan, dengan kata lain ekstraksi fitur akan menentukan output dari inputan yakni labelnya. Fitur yang akan digunakan adalah Unigram. Fitur Unigram mengasumsikan bahwa urutan kata dihasilkan dari pembuatan kata secara mandiri atau terpisah, sehingga probabilits urutan kata akan sama dengan hasil kali probabiltas setiap kata [4].

Tweet adalah data yang tidak terstruktur sehingga tidak dapat langsung digunakan sebagai masukan ke algoritma untuk membangun model. Vektorisasi fitur adalah proses yang mengubah fitur teks yang diekstraksi menjadi matrix fitur numerik yang akan digunakan untuk estimasi dan prediksi model [3]. Setelah melalui proses ini, setiap dimensi angka dari vector itu sesuai dengan token tertentu dalam kumpulan data. Singkatnya, angka dalam vector matrix numerik akan merepresentasikan token tertentu dari data teks (tweet)

1. **Cross Validation**

Pembagian train dan test data dilakukan dengan menggunakan metode Cross Validation, yakni metode untuk mengevaluasi kinerja model pembelajaran mesin dengan meningkatkan beberapa model saat memprediksi data baru. Teknik validasi silang ini akan membagi data menjadi beberapa subset, kemudian melakukan pelatihan dan pengembangan model pada setiap subset secara terpisah. [5]

Dalam proses ini, data training dan test akan dipartisi dengan menggunakan metode k-fold cross validation dengan nilai k=5. Dari data tweet sebanyak 11.127, sebanyak 10.000 data akan digunakan sebagai pelatihan model (train dan validasi) lalu 1.127 akan digunakan untuk testing. 10.000 data akan dikenakan 5-fold cross validation. Penggambarannya sebagai berikut :



**Gambar 3 Ilustrasi Data Train dan Test menggunakan Cross Validation**

1. **SVM**

**….**

1. **Ensamble**

**….**