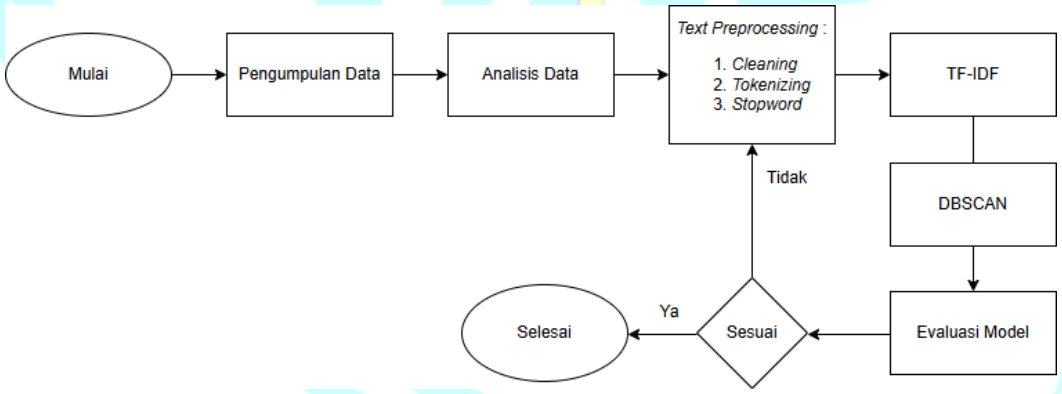
BAB III
METODOLOGI PENELITIAN

3.1 Bahan Penelitian

Bahan yang digunakan dalam penetilitian ini adalah dataset hasil penambangan data dari *website X* . Dataset yang digunakan berupa kalimat-kalimat tentang konflik sosial yang ditulis oleh pengguna *X*.

3.2 Prosedur Penelitian

Prosedur penelitian ini dapat diilustrasikan menggunakan bentuk Flowchart dalam Gambar 3.1. Penelitian dimulai dari Pengumpulan data studi literatur guna mendapatkan referensi dan *Crawling* data menggunakan *API X* lalu data di analisis. Data yang sudah terkumpul akan di proses pada bagian *text proscesing* yang terdiri dari dari proses *Cleaning, Tokenizing,* dan *Stopword* untuk membersihkan data. Setelah data bersih, lalu ke tahap *TF-IDF* dan dilanjut implementasikan algoritma *DBSCAN*. Kemudian dilanjutkan ke tahap pengujian model menggunakan *Sillhoute Coeffisien.*



Gambar 3. 1 *Flowchart* Prosedur Penelitian

3.3 Pengumpulan Data

Pengumpulan data pada penelitian ini dibagi menjadi dua bagian antara lain:

3.3.1 Studi Literatur

Studi literatur berguna sebagai pencarian referensi dan landasan teori. Referensi dari berbagai jenis jurnal, buku, dan bahkan internet sebagai sarana pencarian terhadap tinjauan pustaka, penelitian terdahulu, dengan metode yang dipakai.

3.3.2 *Crawling* Data

Proses mengumpulkan data *tweet* atau meng*crawling* melalui media sosial *X* dengan kata kunci "konflik sosial" dan mendapatkan data sebanyak 1.477.

3.4 Analisis Data

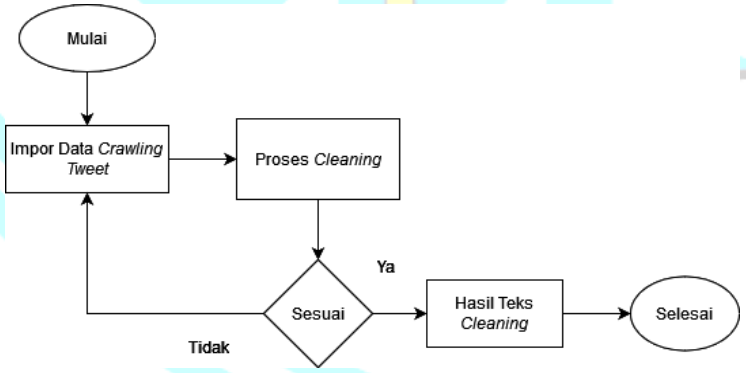
Analisis data berguna untuk mengidentifikasi topik dari data yang didapatkan dengan kata kunci "konflik sosial". Kegiatan dalam proses ini berfungsi untuk memilih data. Data yang telah dianalisis akan mendapatkan hasil, yaitu informasi terkait dengan konflik sosial yang sedang terjadi di masyarakat Indonesia.

3.5 *Text Preprocessing*

Text Preprocessing merupakan tahap untuk mengganti kata yang tidak terstruktur menjadi kata yang terstruktur sehingga mempermudah pemrosesan dataset kata tersebut. Tahap ini melakukan *Cleaning, Tokenizing,* dan *Stopword.*

3.5.1 *Cleaning*

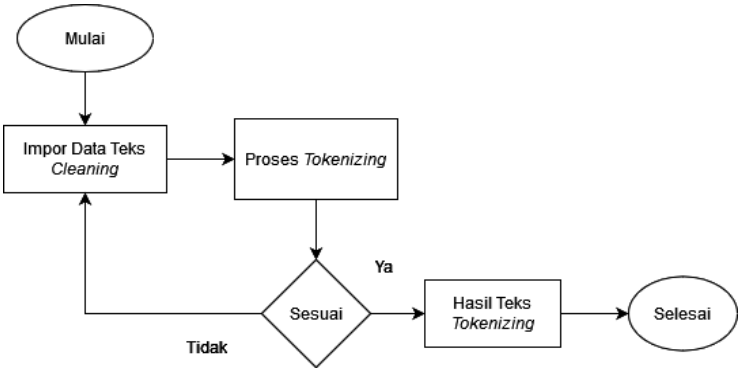
Pada tahap ini *Cleaning* berfungsi untuk menghapus karakter non-alfabetis dan menghilangkan tanda baca, simbol-simbol seperti tanda '@' untuk username, hashtag (#), emoticon dan url dari situs web.



Gambar 3. 2 *Flowchart Cleaning*

3.5.2 *Tokenizing*

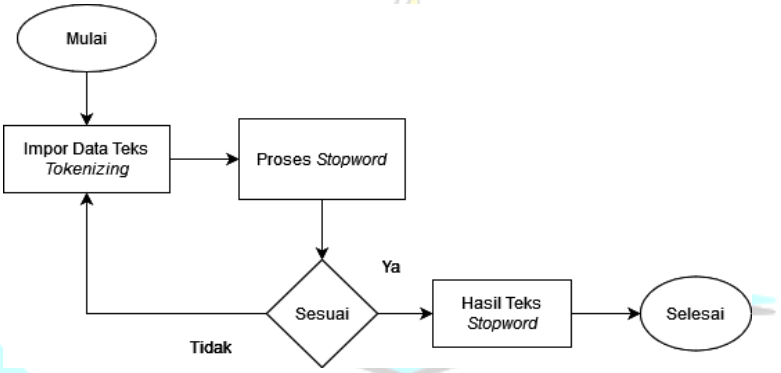
Tahap *Tokenizing* berfungsi memisahkan string atau kata terhadap suatu teks berdasarkan tiap kata pada kalimat kalimat yang terdapat dari tweet (menjadi potongan tunggal).



Gambar 3. 3 *Flowchart Tokenizing*

3.5.3 *Stopword*

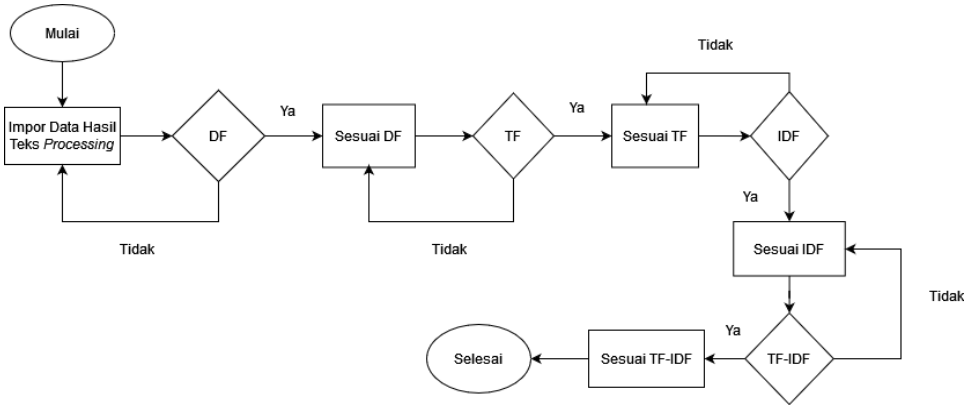
Stopword berfungsi untuk menyaring atau mengurangi kata yang tidak bermakna atau tidak berpengaruh dalam dokumen tersebut. Dengan menggunakan *stopword* bahasa Indonesia yang disusun dan menambahkan kata secara manual karena beberapa kata belum ada di *stopword*.



Gambar 3. 4 *Flowchart Stopword*

3.5.4 *TF-IDF*

Proses selanjutnya *TF-IDF* tahap yang berfungsi mengubah data teks menjadi data numerik, untuk mengetahui keseringan kata muncul dengan menghitung bobot dari setiap kata dan menampilkan hasil menggunakan fitur *wordcloud*. Pembobotan kata menggunakan *TF-IDF* merupakan perkalian dari nilai *TF* dan *IDF* yang akan mendapatkan bobot lebih kecil jika kata tersebut selalu muncul pada setiap dataset, sebaliknya bobot *TF-IDF* akan lebih besar jika kata tersebut tidak sering muncul pada setiap *dataset*. Gambar 3.6 merupakan alur proses dan metode *TF-IDF*.



Gambar 3. 5 *Flowchart TF-IDF*

3.6 Pengujian Model

Pada proses pengujian model ini, berguna untuk memastikan aktualitas dari proses pencarian *item* yang dicari. *Clustering* ini menggunakan formula TF-IDF dalam kemudian hasilnya akan diproses menggunakan *DBSCAN*. Setelah mendapatkan hasil dari proses sebelumnya, data akan dievaluasi menggunakan formula *Sum of Square Error* (SSE). *SSE* ini berguna untuk mengetahui jumlah *cluster* yang optimal (Mar'i & Supianto, 2018). Berikut persamaan (1) formula untuk menghitung *Sum of Square Error* (SSE) :

$$SSE = \sum_{K=1}^{K} \sum_{x_i \in S_k} \|X_i - C_j\|22 \quad (1)$$

Keterangan :

SSE = *Sum of Square Error*

K = Jumlah *cluster*

Cj = *Centroid* (Titik pusat *cluster*)

Xi = Data ke – i

Setelah melakukan proses evaluasi menggunakan *SSE* ini, nilai yang diperoleh akan dilakukan pemeriksaan *cluster* menggunakan *Silhoutte Coefficient*. Pemeriksaan ini berfungsi untuk menilai dan mengetahui kualitas *cluster* yang didapatkan (Wijaya et al., 2021).

Formula *Silhoutte Coefficient* untuk menghitung ketepatan pada *clustering* ditunjukkan pada Persamaan (2)

$$\boldsymbol{Si} = \frac{(\boldsymbol{bi} - \boldsymbol{ai})}{\boldsymbol{max}\,(\boldsymbol{bi} - \boldsymbol{ai})} \tag{2}$$

Keterangan :

Si = Hasil rata-rata *cluster*

bi = Nilai terkecil rata – rata objek dalam *cluster*

ai = Nilai rata-rata objek i dengan seluruh objek *cluster*