BAB III
METODE PENELITIAN

3.1 Proses Analisis Sentimen

Dalam penelitian ini ada beberapa tahap yang harus dilalui terlebih dahulu sebelum akhirnya didapat nilai akurasi dari masing-masing algoritma yang digunakan. Tahap pertama yang dilakukan yaitu pengumpulan data. Pada tahap pengumpulan data, ada dua proses yang dilakukan yaitu proses *crawling* dan *labeling.* Proses *crawling* yaitu proses dimana kita mengambil data dari media sosial *Twitter* untuk nantinya digunakan dalam penelitian. Kemudian peoses kedua yang dilakukan yaitu proses labeling. *Labeling* yakni proses melabeli data yang sudah diambil atau didapatkan dari proses pertama yang sudah dilakukan, *crawling*, dengan label positif dan negatif.

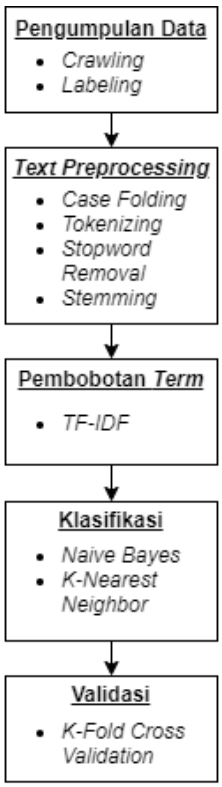
Pada tahap kedua, *text preprocessing*, terdapat empat proses yang akan dilakukan. Dimana pada tahap ini akan memproses atau mengolah data yang telah diambil atau didapat dan diberi label. Tahap ini dilakukan untuk mendapatkan data yang terstruktur serta memastikan untuk hasil yang baik dan konsisten. Proses pertama yang dilakukan yaitu *case folding*. Pada proses ini akan dilakukan pengubahan semua huruf yang terdapat dalam dokumen menjadi huruf kecil atau *lowercase* dan menghilangkan karakter selain huruf. Proses ini dilakukan untuk mendapatkan data yang tidak *redundant*. Selanjutnya akan dilakukan proses *tokenizing,* dimana pada proses ini akan dilakukan pemotongan atau pemisahan setiap kata yang terdapat dalam dokumen yang kemudian disebut dengan token. Setelah dilakukan proses *tokenizing*, proses selanjutnya yang dilakukan yaitu *stopword removal*. Pada proses ini semua kosakata yang tidak memiliki makna seperti kata ‘dan’, ‘di’, ‘oleh’, akan dihilangkan dari dalam dokumen sehingga menyisakan kata yang bermakna saja di dalam dokumen. Selanjutnya, proses terakhir dalam tahapan *text preprocessing* yaitu proses *stemming*. Semua kata yang terdapat dalam dokumen akan diubah menjadi bentuk kata dasar dengan menghapus atau menghilangkan imbuhan yang

terdepat pada kata tersebut. *Stemming* dilakukan untuk mengatasi adanya kata yang tidak biasa, memperkecil jumlah indeks yang berbeda dari suatu dokumen, serta untuk pengelompokkan kata lain yang memiliki bentuk kata dasar dan makna yang sama akan tetapi memiliki bentuk yang berbeda dikarenakan mendapat imbuhan yang berbeda.

Setelah melalui seluruh proses pada tahap kedua, akan didapatkan data atau dokumen yang sudah siap diolah serta diproses. Data yang sudah siap diproses itu kemudian dihitung seberapa banyak kemunculan atau frekuensi kemunculan setiap katanya di dalam dokumen. Tahap ini dinamakan tahap pembobotan *term.* Dalam prosesnya, akan digunakan metode pembobotan TF-IDF.

Tahap keempat yaitu tahap klasifikasi. Pada tahap ini data yang sudah melewati proses *labeling*, tahap *preprocessing* serta sudah dilakukan pembobotan dengan metode TF-IDF, akan diproses dengan algoritma yang telah dipilih yaitu algoritma *Naïve Bayes* dan *K-Nearest Neighbor*. Dalam tahapan ini mesin akan diajari untuk mengenal pola data atau dokumen yang ada untuk kemudian dapat mengklasifikasi sebuah data ke dalam dua kelas, yaitu kelas positif dan kelas negatif.

Selanjutnya tahap validasi. Pada tahap ini, proses validasi dilakukan dengan menggunakan *K-Fold Cross Validation*. Proses ini dilakukan untuk pengujian serta penilaian kinerja proses sebuah algoritma. Nantinya dari proses yang dilakukan pada tahapan ini akan didapatkan nilai akurasi dari masing-masing algoritma yang digunakan. Berikut merupakan alur proses analisis sentimen yang digunakan untuk penelitian ini:



Gambar 3.1 Proses Analisis Sentimen

3.1.1 Pengumpulan Data

Penelitian dilakukan terhadap cuitan pengguna media sosial *Twitter* tentang pandemi COVID-19. Data dibagi atas sentimen positif dan sentimen negatif. Sebanyak 1000 data akan digunakan untuk penelitian menggunakan metode *Naïve Bayes* dan *K-Nearest Neighbor*. Data yang diambil dari media sosial *Twitter* dan digunakan untuk penelitian ini merupakan data yang berbahasa Indonesia. Data diambil menggunakan *Twitter* API.

3.1.2 *Text Preprocessing*

Tahapan yang dilakukan terhadap dokumen selanjutnya adalah *text preprocessing.* Hal ini dilakukan guna menormalisasi data yang digunakan dalam proses analisis sentimen. Proses pertama yang dilakukan yaitu proses *case folding.* Pada proses ini akan dilakukan penghapusan atau penghilangan karakter-karakter pada dokumen yang tidak dibutuhkan, yang

mana hal itu dapat menimbulkan *noise* seperti misalnya emotikon, tanda baca dan lain sebagainya. Selain itu pada proses ini akan dilakukan pengubahan huruf dalam dokumen menjadi huruf kecil. Seperti misalnya "Gara gara COVID19 segala urusan ngantor jadi dilakuin di rumah WFH Work From Home gitu" menjadi "gara gara covid19 segala urusan ngantor dilakuin di rumah wfh work from home gitu". Proses selanjutnya yang akan dilakukan adalah proses *Tokenizing*. Pada proses ini dilakukan pemotongan/pemisahan setiap kata dalam teks yang disebut sebagai token. Selanjutnya akan dilakukan tahap *Stopword Removal* dimana pada tahap ini akan dihilangkan kata-kata yang tidak penting seperti kata "di", "dan", "karena" "oleh" dan lain sebagainya. Tahap ini dilakukan agar dapat memperbesar nilai akurasi. Proses selanjutnya yang akan dilakukan yaitu *Stemming*. Pada proses ini akan dilakukan pengubahan kata yang ada dalam dokumen menjadi bentuk kata dasar. Sebagai contoh, pengubahan kata "himbauan" menjadi "himbau", "larangan" menjadi "larang", "penyebaran" menjadi "sebar" dan lain sebagainya.

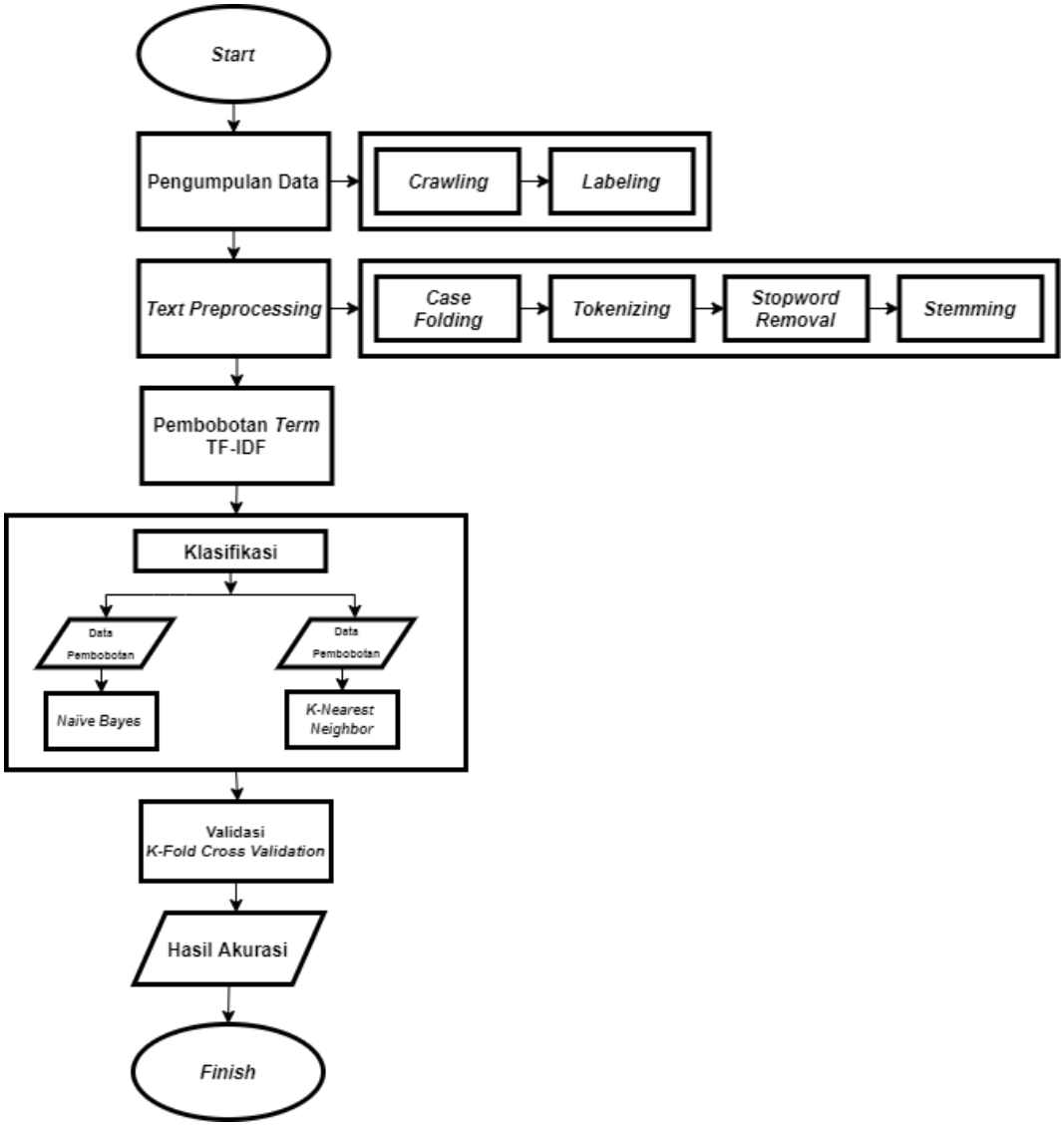
3.1.3 Pembobotan *Term*

Tahap selanjutnya yang dilakukan yaitu pembobotan term. Hal ini dilakukan dengan tujuan untuk menghitung seberapa banyak kemunculan kata pada dokumen. Dalam penelitian yang akan dilakukan ini metode yang digunakan yaitu TF-IDF sebagai proses pembobotan.

3.1.4 Klasifikasi

Setelah melalui tahap pengumpulan data, *text preprocessing* dan pembobotan *term* tahap selanjutnya yaitu tahapan klasifikasi. Dalam tahap ini terdapat dua jenis data yaitu data latih dan data uji. Data latih yaitu data yang sudah dilabeli dan memiliki label "positif" dan "negatif", sementara data uji merupakan data yang belum dilabeli dan belum memiliki label "positif" dan "negatif". Pada penelitian ini hanya dibuat klasifikasi untuk komentar positif dan negatif dikarenakan ingin menguji apakah kedua

algoritma yang digunakan dapat bekerja efektif dengan data yang tergolong sedikit, yaitu sebesar 1000 data. Untuk mendapatkan label-label tersebut, perlu dilakukannya pembelajaran. Jumlah total data yaitu 1000 dengan 500 data kelas positif dan 500 data kelas negatif.



Gambar 3.2 Diagram Alur Naive Bayes dan K-Nearest Neighbor

3.1.5 Validasi dengan *K-Fold Cross Validation*

Tahap validasi ini menggunakan *K-Fold Cross Validation*. Dokumen akan dibagi menjadi 10 bagian. Akan dilakukan percobaan

sebanyak 10 kali percobaan klasifikasi dokumen dan setiap percobaan data akan diacak. Kumpulan dokumen yang ada akan diacak urutannya terlebih dahulu sebelum akhirnya dimasukan dalam *fold*. Tujuan hal ini dilakukan adalah guna menghidari pengelompokkan dokumen dari satu kategori tertentu pada suatu *fold*.