# BAB III

# METODOLOGI PENELITIAN

## Data Penelitian

*Dataset* atau data yang digunakan dalam penelitian ini bersumber dari Twitter berupa data teks kicauan (*tweet)* yang diperoleh mulai tanggal 1 Desember 2020 hingga 31 Desember 2020 sejumlah 4.314 data. Data tersebut diperoleh menggunakan pustaka Tweepy melalui proses *crawling*. *Dataset* yang diperoleh dikumpulkan berdasarkan enam (6) parameter kata kunci yang terkait dengan sistem pembelajaran daring antara lain: ‘pembelajaran jarak jauh’, ‘belajar dari rumah’, ‘#belajaronline’, ‘#belajardarirumah’, ‘#belajardirumah’, dan ‘#kuliahonline’. Berikut beberapa contoh data *tweet* hasil dari proses *crawling* yang dapat dilihat pada Tabel 3.1 berikut:

**Tabel 13.1 Sampel data tweet**

| *Tweet ID* | *Username* | *Tweet* | Waktu *tweet* |
| --- | --- | --- | --- |
| 1335989354792103936 | LRomdani | Tetap memakai masker meski dirumah sendiri  Tetap semangat belajar dari rumah dimasa pandemi  #DiktiMengajarDariRumah  #DiktiDutaEdukasiPerubahanPrilaku https://t.co/c1WMa5SVSj | 2020-12-07 16:47:38 |
| 1336520460255724032 | kelaskitadotcom | Gunakan hak suara kamu dengan bijak, ya! Selamat memilih! #kelaskita #carabarubelajarseru #belajardirumah #elearning #belajaronline #dirumahaja #quotes https://t.co/1anyTiETlA | 2020-12-09 03:58:03 |
| 1336742494122340096 | fandimas16 | @collegemenfess 1. Jenuh banget di rumah 2. Gw dri dulu suka ama suasana kelas, dan suasana itu mendukung gw untuk belajar dan memahami suatu materi | 2020-12-09 18:40:20 |
| 1338003730587812096 | kumparan | Tanpa smartphone di masa pandemi, bisa berarti putus sekolah, karena kini dilakukan belajar online atau pembelajaran jarak jauh. https://t.co/rVW6xOgrfI | 2020-12-13 06:12:02 |

## Penerapan Metode

Dalam membangun aplikasi analisis sentimen yang dilakukan pada penelitian ini, terdapat beberapa tahapan yang dilakukan. Tahapan tersebut merepresentasikan setiap proses dan rancangan dalam penelitian, dari awal hingga akhir aplikasi berjalan. Tahapan yang dilakukan dapat dilihat pada Gambar 3.1 berikut:



**Gambar 13.1 Tahapan metode**

Pada Gambar 3.1, pengumpulan data dilakukan melalui proses *crawling* untuk mendapatkan *dataset* berupakicauan (*tweet*). Selanjutnya, *tweet* yang telah diperoleh dalam bentuk *excel*, kemudian dimasukkan kedalam basis data (*database*) untuk dilakukan proses *preprocessing*, pada proses *preprocessing* dilakukan penyaringan, pembuangan dan perbaikan kata. Hasil dari proses *preprocessing* menghasilkan kalimat yang lebih terstruktur (*clean text*) yang kemudian digunakan pada tahap selanjutnya. *Clean text* yang diperoleh dari proses *preprocessing* akan diproses dalam tahap *labeling* untuk menentukan kelas (*label*) berupa sentimen positif atau negatif, kemudian *tweet* yang telah berlabel akan dibagi menjadi dua (2) buah bagian antara lain: data uji dan data latih. Data latih merupakan data yang berfungsi sebagai pembangun pengetahuan untuk proses klasifikasi, proses pembangunan pengetahuan tersebut dilakukan melalui proses *modeling* dan menghasilkan sebuah model latih menggunakan data latih yang tersedia. Sementara data uji merupakan data yang disiapkan untuk menguji tingkat keakuratan model latih yang dihasilkan oleh proses *modeling*, proses pengujian tingkat akurasi tersebut dilakukan melalui proses pengujian dan evaluasi. Setelah dilakukan proses pengujiandan evaluasi, hasil penerapan metode akan divisualisasikan ke dalam bentuk persentase dan grafik.

### Pengumpulan data

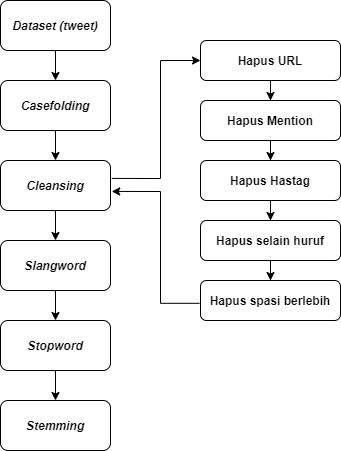
Pada tahapan pengumpulan data dilakukan melalui proses *crawling*. Proses tersebut meliputi: mendapatkan API *key* Twitter melalui akun *developer* Twitter (<https://developer.twitter.com/>). API *key* Twitter yang diperoleh antara lain: *Customer* API *key*, *Customer* API *Secret key*, *Access Token*, dan *Access Token Secret*. Proses selanjutnya adalah penambangan data yang bersumber pada media sosial Twitter menggunakan pustaka Tweepydengan akses dari API *key* yang telah didapatkan. Data *tweet* yang berhasil dikumpulkan akan disimpan ke dalam sebuah file *excel (.*xlsx*)*, yang kemudian dimasukkan ke dalam basis data (*database*) MySQL. Ilustrasi tahap pengumpulan data dapat dilihat pada Gambar 3.2 berikut:



**Gambar 23.2 Tahap pengumpulan data**

### *Preprocessing*

Pada tahapan *preprocessing* dilakukan penyaringan, pembuangan dan perbaikan kata melalui beberapa proses. Hal tersebut dimaksudkan untuk menghasilkan data kicauan (*tweet*) yang lebih terstruktur atau disebut dengan *clean text.* Berdasarkan teori yang telah dijelaskan pada sub bab (2.5), proses yang dilakukan dalam tahap *preprocessing* antara lain*: case folding, cleansing,* mengubah *slang word,* menghapus *stop word,* dan *stemming.* Ilustrasi proses tahap *preprocessing* data dilihat pada Gambar 3.3 berikut:



**Gambar 33.3 Tahap preprocessing**

#### *Case folding*

Proses *case folding* dilakukan penyetaraan teks menjadi huruf kecil (*lowercase*) secara keseluruhan, misalnya kata ‘BELAJAR’ atau ‘Belajar’, maka akan diubah menjadi ‘belajar’.

#### *Cleansing*

Pada proses *cleansing* dilakukan penyaringan dan pembuangan teks. Proses *cleansing* terdiri atas beberapa tahapan antara lain: menghapus URL, *mention* (*@mention*), *hastag* (*#hastag*), selain huruf (a-z) dan spasi berlebih.

##### Menghapus URL

Proses penghapusan URL akan menghapus semua teks yang diawali dengan kata ‘http’, misalnya: ‘https://google.com’, maka akan diubah menjadi ‘ ’.

##### Menghapus *mention* (*@mention*)

Proses penghapusan *mention* (*@mention*) akan menghapus semua teks yang diawali dengan ‘@’, misalnya: ‘@kominfo’, maka akan diubah menjadi ‘ ’.

##### Menghapus *hastag* (*#hastag*)

Pada Gambar 3.6 proses penghapusan tagar atau *hastag* (#hastag) akan menghapus semua teks yang diawali dengan ‘#’, misalnya: ‘#pjj’, maka akan diubah menjadi ‘ ‘.

##### Menghapus selain huruf

Proses penghapusan selain huruf akan menghapus karakter selain huruf a-z pada teks, misalnya terdapat teks: ‘tanggal: 01 desember 2020’, maka akan diubah menjadi ‘tanggal desember ‘.

##### Menghapus spasi berlebih

Proses penghapusan spasi atau (*whitespace)* berlebih akan menghapus *whitespace* berlebih pada teks, misalnya terdapat teks: ‘tanggal desember ‘, maka akan diubah menjadi ‘tanggal desember’.

#### Merubah *slang word*

Proses merubah *slang word* akan merubah setiap kata gaul, kata singkatan atau kata tidak baku ke bentuk bakunya, misalnya kata: ‘utk’ menjadi ‘untuk’, ‘yng’ menjadi ‘yang’, ‘apotik’ menjadi ‘apotek. Proses pengubahan tersebut melibatkan kamus *slang word* yang terdapat dalam basis data (*database*), diperoleh melalui penelitian sebelumnya yang dipublikasikan pada laman github (https://github.com/louisowen6/NLP\_bahasa\_resources).

#### Menghapus *stop word*

Proses menghapus *stop word* akan menghapus setiap kata yang kurang memiliki makna namun sering dijumpai dalam sebuah teks, misalnya penghapusan kata: ‘untuk’, ‘yang’, ‘apa’. Proses penghapusan tersebut melibatkan kamus *stop word* yang terdapat dalam basis data (*database*), diperoleh melalui penelitian sebelumnya yang dipublikasikan pada laman github (https://github.com/louisowen6/NLP\_bahasa\_resources).

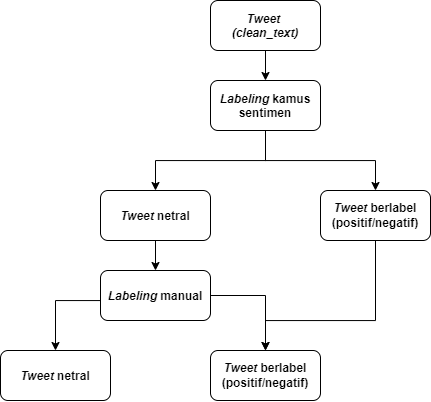
#### *Stemming*

Proses *stemming* akan mengubah kata berimbuhan menjadi kata dasar dengan melibatkan pustaka Sastrawi, misalnya kata: ‘membantu’ menjadi ‘bantu’, ‘pembelajaran’ menjadi ‘ajar’. Proses pengubahan tersebut dilakukan dengan melibatkan pustaka Sastrawi, menggunakan paket StemmerFactory.

### *Labeling*

Pada tahapan *labeling* dilakukan pemberian *label* (kelas) berdasarkan ciri atau karakteristik yang terkandung dalam sebuah dokumen atau kalimat. Pada tahap ini, t*weet* yang telah melalui proses *preprocessing* dan menghasilkan *clean text* akan diberikan kelas positif atau negatif. Kelas positif dimaksudkan untuk teks *tweet* tersebut yang mengandung pernyataan yang setuju, mendukung atau menerima proses berjalanya pembelajaran daring. Sedangkan kelas negatif dimaksudkan untuk teks *tweet* yang cenderung tidak setuju atau menolak proses berjalanya pembelajaran daring.

Berdasarkan teori yang telah dijelaskan dalam sub bab (2. 6), bahwa proses *labeling* dapat dilakukan dengan dua (2) buah cara, antara lain: *labeling* manual dan *labeling* dengan kamus sentimen. *Labeling* manual merupakan proses pemberian kelas berdasarkan subjektivitas beberapa orang. Sedangkan *labeling* dengan kamus sentimen merupakan proses pemberian kelas secara otomatis berdasarkan kamus sentimen, dimana prosesnya melibatkan kamus kata positif dan kamus kata negatif yang terdapat dalam basis data (*database*), diperoleh melalui penelitian sebelumnya yang dipublikasikan pada laman github (https://github.com/masdevid/ID-OpinionWords). Proses *labeling* utama dilakukan dengan cara *labeling* dengan kamus sentimen, kemudian untuk untuk *tweet* yang gagal mendapatkan *label* (*tweet* netral) akan dievaluasi ulang menggunakan cara *labeling* manual berdasarkan subjektivitas beberapa orang, dengan mengabaikan *tweet* yang dinilai sebagai netral. Ilustrasi tahap *labeling* dapat dilihat pada Gambar 3.4 berikut:



**Gambar 43.4 Tahap labeling**

### Pembagian data

Pada tahapan pembagian data, *tweet* yang telah berlabel akan dibagi menjadi dua (2) buah bagian antara lain: data uji dan data latih. Proses pembagian data dilakukan dengan membagi *dataset* menjadi 90% data latih dan 10% menjadi data uji. Ilustrasi tahap pembagian data dapat dilihat pada Gambar 3.5 berikut:



**Gambar 53.5 Tahap pembagian data**

#### Data Latih

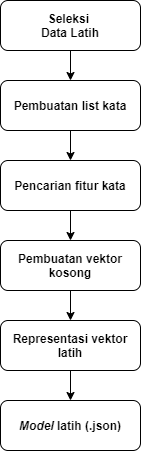
Data latih merupakan data yang berfungsi sebagai *model* latih, yaitu pembangun pengetahuan untuk proses klasifikasi.

#### Data Uji

Data uji merupakan data yang disiapkan untuk menguji tingkat keakuratan algoritme klasifikasi berdasarkan *model* latih.

### *Modeling*

Tahap modeling dilakukan untuk mengekstraksi *tweet* data latih menjadi representasi vektor menggunakan *CountVectorizer*. Pada tahap ini terdapat lima (5) proses utama yang dilalui untuk menghasilkan sebuah *model* latih, lima proses itu antara lain: seleksi data latih, pembuatan list kata, pencarian fitur kata, pembuatan vektor kosong dan membuat representasi vektor. Ilustrasi proses tersebut dapat dilihat pada Gambar 3.6 berikut:



**Gambar 63.6 Tahap modeling**

Pada Gambar 3.6, penyeleksian data latih dilakukan dengan menggunakan teknik sampling kuota (*quota sampling*), dimaksudkan untuk mendapatkan keseimbangan pada data latih. Tahapan seleksi tersebut dilakukan sebanyak kriteria tertentu. Data latih telah terseleksi akan diproses (ekstraksi fitur) menggunakan *CountVectorizer* sehingga dapat diperoleh daftar fitur dan representasi vektor angka untuk tiap data latih. Daftar fitur dan vektor angka tersebut kemudian akan disimpan dan dijadikan sebagai sebuah *model* pengetahuan (*model* latih) dalam bentuk file JSON (.json).

### Klasifikasi *K-Nearest Neighbor*

Tahap klasifikasi menggunakan *K-Nearest Neighbor* (KNN) merupakan tahapan yang dapat dilakukan setelah terdapat satu atau lebih *model* latih. *Model* latih tersebut merupakan data latih yang telah melalui tahap *modeling* yang dijelaskan pada sub-sub bab (3. 2. 5). Untuk menerapkan *model* klasifikasi menggunakan KNN, terdapat empat (4) buah proses utama yaitu: membuat representasi vektor uji, menghitung jarak antar data, mencari tetangga terdekat berdasarkan nilai K, dan menghitung nilai probabilitas *label* sentimen. Ilustrasi tahapan klasifikasi dapat dilihat pada Gambar 3.7 berikut:



**Gambar 73.7 Tahap klasifikasi**

Pada Gambar 3.7, pembuatan representasi vektor uji dilakukan menggunakan *model* latih yang dipilih, sehingga terbentuk representasi vektor uji yang sesuai dengan pengetahuan *model* latih. Hasil vektor uji tersebut akan dihitung tingkat kedekatannya (jarak) dengan vektor pada *model* latih, proses tersebut melibatkan perhitungan *euclidean distance*. Hasil perhitungan *euclidean distance* akan menghasilkan nilai jarak, yang kemudian akan disaring berdasarkan K tetangga terdekat. Selanjutnya dilakukan *voting* untuk menentukan *label* prediksi (positif atau negatif) berdasarkan dominasi *label* pada *K* tetangga terdekatnya.

## Rancangan Pengujian

Pengujian dilakukan untuk mengetahui nilai atau tingkat akurasi, presisi, dan *recall* dari *model* latih menggunakan algoritme yang diusulkan. Pada penelitian ini, pengujian dilakukan dengan cara membandingkan beberapa data hasil prediksi (data hasil tahap klasifikasi) dengan sekumpulan data aktual (data hasil tahap *labeling*). Adapun dimaksud dengan beberapa data hasil prediksi merupakan sekumpulan data yang telah diproses melalui algoritme *K-Nearest Neighbor* (KNN) dengan variasi nilai K, yaitu: K=3, K=5, K=7, K=9, dan K=11.

### Akurasi

Akurasi merupakan tingkat kedekatan antara nilai prediksi dengan nilai aktual persamaan (3. 1).

### Presisi

Presisi merupakan tingkat ketepatan antara informasi yang diminta dengan jawaban yang diberikan oleh sistem persamaan (3. 2).

### *Recall*

*Recall* merupakan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi persamaan (3. 3).

Berdasarkan pada sub bab (2. 10), pengukuran tingkat akurasi, presisi, dan *recall* dapat diketahui melalui *confusion matrix* dengan persamaan sebagai berikut:

|  |  |
| --- | --- |
|  | (3. 1) |
|  | (3. 2) |
|  | (3. 3) |

dengan,

1. *True Positive* (TP) merupakan data positif yang diprediksi benar. Misalnya: *tweet* 1 ber*label* positif dan dari *model* latih yang dibuat memprediksi *tweet* 1 bernilai positif juga.
2. *True Negative* (TN) merupakan data negatif yang diprediksi benar. Misalnya: *tweet* 1 ber*label* negatif dan dari *model* latih yang dibuat memprediksi *tweet* 1 bernilai negatif juga.
3. *False Positive* (FP) merupakan data negatif namun diprediksi sebagai data positif. Misalnya: *tweet* 1 ber*label* negatif namun dari *model* latih yang dibuat memprediksi *tweet* 1 bernilai positif.
4. *False Negative* (FN) merupakan data positif namun diprediksi sebagai data negatif. Misalnya: *tweet* 1 ber*label* positif namun dari *model* latih yang dibuat memprediksi *tweet* 1 bernilai negatif.