BAB 3
METODE PENELITIAN

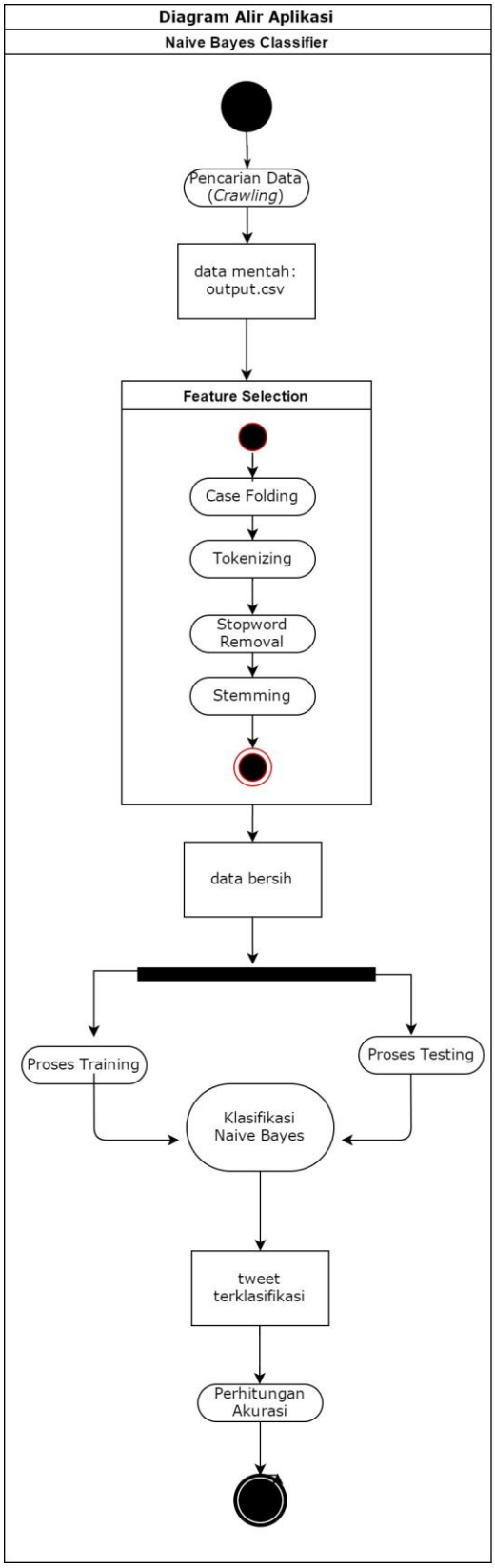
Penelitian ini dilakukan berdasarkan tahapan-tahapan yang telah disusun agar mencapai tujuan yang diharapkan. Pada bagian ini akan dipaparkan secara detail setiap tahapan penelitian, mulai dari metode pengumpulan data, metode analisa, perhitungan akurasi, hingga pembuatan prototype. Gambar 3.1 menggambarkan tahapan penelitian yang dilakukan.

Pada diagram alir tersebut proses dimulai dengan "Pencarian Data". Proses ini melakukan pencarian *tweet* dengan kata kunci tertentu pada *Twitter* menggunakan aplikasi berbasis *Java.* Hasil pencarian akan muncul menjadi data mentah dalam format .csv dengan nama "output.csv". Penjelasan lebih lanjut dipaparkan pada bab 3.1.

Feature Selection merupakan proses pembersihan *tweet* agar kata yang akan digunakan merupakan kata unik dan tidak memasukkan simbol atau karakter selain huruf. Proses ini terdiri dari proses-proses lain. Setelah proses ini dilakukan, data mentah akan menjadi data bersih yang siap dimasukkan dalam database. Penjelasan lebih lanjut dipaparkan pada bab 3.2.

Setelah data bersih, proses dapat dilanjutkan ke proses training atau proses testing. Kedua proses tersebut merupakan proses yang paralel, pengguna dapat memilih proses mana yang akan dilakukan. Pada tiap proses tersebut akan diberlakukan proses klasifikasi menggunakan algoritma Naïve Bayes. Hasil dari klasifikasi akan menjadi data tweet terklasifikasi yang dimasukkan ke dalam database. Penjelasan lebih lanjut dipaparkan pada bab 3.3 sampai bab 3.3.1.

Evaluasi terhadap hasil klasifikasi selanjutnya akan dihitung menggunakan *confussion matrix* pada proses "Perhitungan Akurasi" untuk mengetahui nilai akurasi yang telah dilakukan sistem. Penjelasan lebih lanjut dipaparkan pada bab 3.3.2. Perhitungan akurasi merupakan proses akhir dari rancangan sistem.



Gambar 3.1 Diagram Alir Proses Kerja Aplikasi

3.1 Metode Pengumpulan Data

Data yang digunakan pada penelitian ini bersifat kuantitatif dan didapat menggunakan metode observasi. Metode observasi disini dimaksudkan adalah penulis mencari data *tweet* yang berasal dari *Twitter* menggunakan tools tertentu.

Proses pengumpulan data *tweet* dilakukan dengan menggunakan aplikasi *GetOldTweets-java*. Pada aplikasi tersebut telah diberikan perintah-perintah yang dapat digunakan untuk melakukan kostumisasi pencarian.

Contoh perintah dan kostumisasi pencarian yang disediakan dalam laman web penyedia adalah sebagai berikut:

- Petunjuk pemakaian

```
java -jar got.jar -h
```

- Mencari *tweet* berdasarkan *username*

```
java -jar got.jar username=barackobama maxtweets=1
```

- Mencari *tweet* berdasarkan kata kunci

```
java -jar got.jar querysearch="europe refugees" maxtweets=1
```

- Mencari *tweet* berdasarkan *username* dan batasan waktu tertentu

```
java -jar got.jar querysearch="europe refugees" maxtweets=1
```

Contoh hasil pencarian menggunakan *GetOldTweets-java* dapat dilihat pada tabel 3.1.

Tabel 3.1 Contoh Hasil Pencarian

user name	date	retweets	favorites	text	geo	mentions	hash tags	id	permalink
DriverGojekPro	2/28/2017 21:15	4	0	Tarif minimum diturunin lagi sama @gojekindonesia Gua rasa demi promo tarif akan turun terus sampe 100 rupiah perkilometer.		@gojekindonesia		8E+17	https://twitter.com/DriverGojekPro/status/836580481055350784
titarahanip	2/28/2017 20:22	11	3	Semoga selalu dilindungi & dimudahkan rezekinya bapak-bapak gojek yg melayani gofood di kala hujan seperti ini @dramauber_id aamiin..		@dramauber_id		8E+17	https://twitter.com/titarahanip/status/836567170301620227
mngywnw	2/27/2017 17:26	13	4	@dramauber_id trip kemarin sore. sungguh abang gojek yg pantang menyerah pic.twitter.com/aEmfSxRQoB		@dramauber_id		8E+17	https://twitter.com/mngywnw/status/836160645217202178
kv_kahfi	2/26/2017 20:07	116	32	Cc: @Gemacan70 @RestySeterah RT @R_Yoyoy : bantu Viralkan kejahatan besar PT. Gojek yg membebankan biaya Administrasi kpd drivernya pic.twitter.com/7xcMN8aKCZ		@Gemacan70 @RestySeterah @R_Yoyoy		8E+17	https://twitter.com/kv_kahfi/status/835838697480187906
agunkdarmawan9	2/24/2017 2:40	0	0	Owh seperti ini yah gojek? ikut perpolitk https://twitter.com/condetwarrior/status/834779786207563777 …				8E+17	https://twitter.com/agunkdarmawan9/status/834850548037853184

Data yang didapat dari proses *crawling* tidak secara langsung dapat digunakan. Perlu adanya proses *data preparation (preprocessing)* yang meliputi pengolahan teks dan *feature selection.* Proses ini diperlukan untuk mempermudah proses berikutnya.

Berikut tahapan-tahapan *data preparation:*

1. *Case folding*
2. *Tokenizing*
3. *Stopword removal*
4. *Stemming*

Proses berikutnya adalah proses *tweet cleansing* yaitu menghilangkan data yang tidak memiliki sentiment agar proses *training* dapat dilakukan dengan optimal.

3.1.1 Cleansing Data

Dalam proses *crawling data* atau pengumpulan data, kerap ditemukan data-data yang tidak diperlukan. Proses ini merupakan prosses untuk menghapus record data tersebut atau memilih data mana yang diperlukan dan mana yang tidak. Pada penelitian ini hasil *crawling* berupa file berformat .csv, oleh karena itu proses *cleansing data* dapat dilakukan menggunakan Microsoft Excel.

3.1.2 Penghapusan Attribute

Hasil pencarian menampilkan 10 *attribute* yang terdiri atas: *username, date, reteets, favorites, text, geo, mentions, hashtags, id* dan *permalink.* Namun proses klasifikasi berfokus pada atribut "*text"* yang berisi *tweet* dari pengguna dan untuk melakukan analisa lebih jauh dibutuhkan pula *attribute* "*username*". Untuk itu perlu dilakukan penghapusan *attribute* yang tidak terpakai dengan menggunakan Microsoft Excel secara manual.

3.1.3 Pemberian Label

Sebelum memasukkan data training ke dalam database untuk dilakukan proses lebih lanjut, perlu dilakukan pemilahan terlebih dahulu sentimen dari data yang akan digunakan sebagai data training secara manual (subjektif), apakah positif ataukah

negatif. Proses ini merupakan proses yang penting karena data training akan sangat berpengaruh pada saat klasifikasi.

Data training yang digunakan berjumlah 500. Dengan 250 untuk sentimen positif dan 250 untuk sentimen negatif.

3.2 Preprocessing Text

Pada bagian ini akan dijelaskan lebih dalam mengenai persiapan data *(preprocessing)* yang digunakan. Persiapan data pada penelitian ini dibantu oleh *library* Sastrawi dalam bahasa pemrograman PHP.

Berikut adalah contoh 2 *tweet* yang akan digunakan sebagai masukan:

Semoga selalu dilindungi & dimudahkan rezekinya bapak-bapak gojek yg melayani gofood di kala hujan seperti ini @dramauber_id aamiin..

Contoh Tweet 1

Sudahkah anda menghapus aplikasi Gojek hari ini?!. #BoikotGojek

Contoh Tweet 2

3.2.1 *Case Folding*

Tahap *case folding* merubah seluruh huruf pada *tweet* masukan menjadi huruf kecil dan menghapus semua konten karakter selain huruf seperti simbol. Berikut merupakan hasil proses *case folding* dari contoh data diatas:

semoga selalu dilindungi dimudahkan rezekinya bapakbapak gojek yg melayani gofood di kala hujan seperti ini dramauberid aamiin

Contoh Tweet 1

sudahkah anda menghapus aplikasi gojek hari ini boikotgojek

Contoh Tweet 2

3.2.2 *Tokenizing*

Tahap *tokenizing* merubah struktur *tweet* yang berupa kalimat menjadi kata atau "*term*". Berikut contoh dokumen yang telah dilakukan tahap *tokenizing:*

semoga
selalu
dilindungi
dimudahkan
rezekinya
bapakbapak
gojek
yg
melayani
gofood
di
kala
hujan
seperti
ini
dramauberid
aamiin

Contoh Tweet 1

sudahkah
anda
menghapus
aplikasi
gojek
hari
ini
boikotgojek

Contoh Tweet 2

3.2.3 Stopword Removal

Tahap *stopword removal* melakukan penghapusan kata-kata yang terdapat pada daftar *stopword.* Daftar *stopword* yang digunakan pada *library* Sastrawi, adalah sebagai berikut:

Tabel 3.2 Daftar Stopword

```
'yang', 'untuk', 'pada', 'ke', 'para', 'namun', 'menurut', 'antara',
'dia', 'dua', 'ia', 'seperti', 'jika', 'jika', 'sehingga', 'kembali',
'dan', 'tidak', 'ini', 'karena', 'kepada', 'oleh', 'saat', 'harus',
'sementara', 'setelah', 'belum', 'kami', 'sekitar', 'bagi', 'serta',
'di', 'dari', 'telah', 'sebagai', 'masih', 'hal', 'ketika', 'adalah',
'itu', 'dalam', 'bisa', 'bahwa', 'atau', 'hanya', 'kita', 'dengan',
'akan', 'juga', 'ada', 'mereka', 'sudah', 'saya', 'terhadap', 'secara',
'agar', 'lain', 'anda', 'begitu', 'mengapa', 'kenapa', 'yaitu',
'yakni', 'daripada', 'itulah', 'lagi', 'maka', ‘tentang', 'demi',
'dimana', 'kemana', 'pula', 'sambil', 'sebelum', 'sesudah', 'supaya',
'guna', 'kah', 'pun', 'sampai', 'sedangkan', 'selagi', 'sementara',
'tetapi', 'apakah', 'kecuali', 'sebab', 'selain', 'seolah', 'seraya',
'seterusnya', 'tanpa', 'agak', 'boleh', 'dapat', 'dsb', 'dst', 'dll',
'dahulu', 'dulunya', 'anu', 'demikian', 'tapi', 'ingin', 'juga',
'nggak', 'mari', 'nanti', 'melainkan', 'oh', 'ok', 'seharusnya',
'sebetulnya', 'setiap', 'setidaknya', 'sesuatu', 'pasti', 'saja',
'toh', 'ya', 'walau', 'tolong', 'tentu', 'amat', 'apalagi',
'bagaimanapun', 'gojek',
```

Berikut contoh dokumen yang telah dilakukan tahap *stopword removal:*

semoga
dilindungi
dimudahkan
rezekinya
bapakbapak
melayani
gofood
kala
hujan
seperti
dramauberid
aamiin

Contoh Tweet 1

sudahkah
menghapus
aplikasi
hari
boikotgojek

Contoh Tweet 2

3.2.4 *Stemming*

Tahap *stemming* ini melakukan perubahan terhadap tiap kata yang ada untuk dijadikan input yaitu kata dasar. Sehingga imbuhan yang ada pada setiap kata akan dihilangkan. Berikut contoh dokumen yang telah melalui proses *stemming:*

semoga
lindung
mudah
rezeki
bapakbapak
layan
gofood
kala
hujan
seperti
dramauberid
aamiin

Contoh Tweet 1

sudah
hapus
aplikasi
hari
boikotgojek

Contoh Tweet 2

3.3 Metode Naïve Bayes Classifier

Perancangan model klasifikasi dokumen *tweet* dapat dijelaskan sebagai berikut:

3.3.1 Proses Klasifikasi

Pada penelitian ini proses klasifikasi terbagi atas dua tahap, algoritma pelatihan dan algoritma klasifikasi.

Algoritma pelatihan adalah sebagai berikut:

1. Melakukan penjumlahan semua token
2. Pada tiap kelas sentiment dilakukan:
   a. Jumlah record pada $kelas_j$
   b. Menghitung P($sentiment_j$) dengan persamaan

$$P(sentiment_j) = \frac{|docs\ j|}{|contoh|}$$

   c. Pada tiap kata w_k pada daftar semua token lakukan:
3. Menghitung P($kata_k|sentiment_j$) dengan persamaan

$$P(kata_k|sentiment_j) = \frac{n_k + 1}{n + |kosakata|}$$

   Keterangan:

   j = kategori *tweet*, dimana j_1 = tweet positif dan dimana j_2 = tweet positif

   |docs j| = jumlah dokumen pada tiap kategori j

   |contoh| = jumlah dokumen dari semua kategori

   n = jumlah frekuensi munculnya kata dari tiap kategori

   |kosakata| = jumlah semua kata dari semua kategori

Algoritma Klasifikasi:

1. Masukkan dokumen (tweet) yang akan diklasifikasikan
2. Hitung probabilitas untuk setiap kelas dengan persamaan di atas dengan menggunakan P($sentiment_j$) dan P($kata_k|sentiment_j$) yang telah diketahui dari tahap pelatihan

3. Probabilitas kelas maksimum merupakan kelas sentiment yang terpilih dari proses klasifikasi

3.3.2 Perhitungan Akurasi

Kinerja atau akurasi dari algoritma selanjutnya dapat dihitung dengan bantuan rumus perhitungan *confusion matrix:*

$$Sensitivity = \frac{TP}{P}$$

$$Specifity = \frac{TN}{N}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Accuracy = sensitivity\frac{P}{(P+N)} + specifity\frac{N}{(P+N)}$$

Keterangan :

TP = Jumah *True Positives*

TN = Jumah *True Negatives*

P = Jumlah Tupel Positif

N = Jumlah Tupel *Negatif*

FP = Jumlah *False Positives*

3.4 Prototype

Dari hasil proses text mining diatas, selanjutnya akan dikembangkan prototype aplikasinya. Adapun tahapan dari pengembangan prototype pada penelitian ini yaitu Pembacaan data, proses klasifikasi, dan pembuatan prototype.

3.4.1 Pembacaan data

Metode untuk memasukkan data latih akan terbagi atas dua, secara manual (satu per satu data latih) dan secara masal (menggunakan *file* Microsoft Excel). Data latih tersebut kemudian akan dilakukan preproses untuk mendapatkan kata-kata yang penting saja. Kata yang berhasil dimasukkan akan ditampilkan pada halaman Input Data Latih.

Data yang telah dilakukan preproses kemudian akan dimasukkan ke dalam *database MySql* dalam bentuk kata per kata (*term*). Pengulangan kata ditiadakan

agar tidak terjadi redudansi yang akan mengurangi keakuratan proses klasifikasi serta menambah beban data.

Data testing dapat dilakukan dengan memasukkan tweet satu per satu pada halaman Input Data Testing. Setelah pengujian dilakukan, akan ditampilkan nilai probabilitas tiap kata terhadap database pada tiap sentimen. Aplikasi akan menampilkan sentimen tweet dengan merujuk pada nilai probabilitas yang lebih besar.

3.4.2 Proses Klasifikasi

Proses klasifikasi menggunakan dasar rumus Naïve Bayes yang ditransformasikan ke dalam PHP dan JavaScript. Saat dimasukkan *tweet* pada proses *testing*, akan dilakukan perhitungan probabilitas kemunculan tiap kata pada database *training* baik pada sentimen positif maupun negatif. Setelah dihitung probabilitas kemunculan tiap kata, akan dibandingkan apakah *tweet* tersebut lebih banyak memiliki kemunculan pada *tweet* dengan sentimen positif atau negatif. Nilai terbesar dari probabilitas tersebut yang akan menentukan apakah *tweet* tersebut memiliki sentimen positif atau negatif. Hasil dari klasifikasi akan dimasukkan pada database *testing.*

3.4.3 Pembuatan Prototype

Prototype dibangun menggunakan bahasa pemrograman PHP dan bantuan JavaScript untuk melakukan beberapa perhitungan. Setiap proses akan ditulis pada file yang berbeda. Proses testing, proses training, proses perhitungan Naïve Bayes dan lainnya. Hal ini diharapkan mampu memudahkan dalam melakukan perbaikan jika terjadi kesalahan, selain itu juga memudahkan dalam penulisan untuk menghindari *coding* yang terlalu panjang dalam satu file.