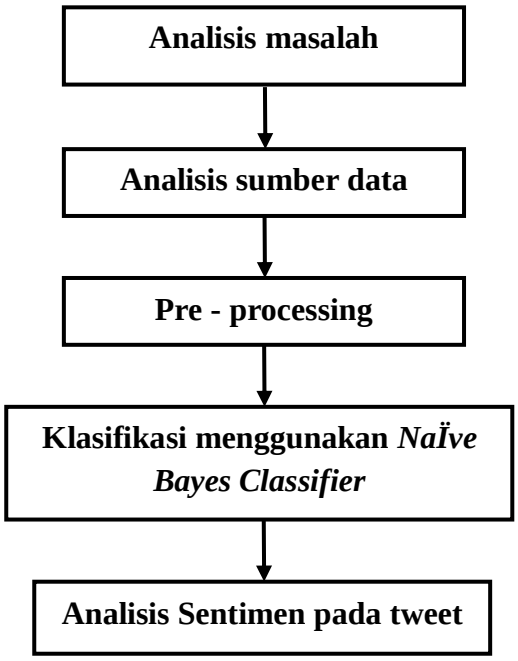
BAB III
METODOLOGI PENELITIAN

3.1 Analisis Penelitian

Pada bab ini membahas tentang metodologi penelitian tentang analisis sentimen dengan menggunakan algoritma *NaÏve Bayes Classifier.* Langkah – langkah yang dilakukan pada analisis ini sebagai berikut :

1. Analisis masalah.
2. Analisis sumber data.
3. Analisis preprocessing.
4. Klasifikasi dan penerapan*NaÏve Bayes Classifier.*
5. Analisis sentimen pada tweet.

Analisis diatas juga dapat dilihat pada tabel 3.1



Tabel3.1 Analisis Penelitian

3.1.1 Analisis Masalah

Pemilihan presiden yang akan dilaksanakan pada tahun 2019 sudah mulai dapat dirasakan pada saat sekarang. Hal tersebut terbukti dengan banyaknya tagar –tagar dimedia sosial khususnya Twitter yang bertuliskan "#2019GantiPresiden" maupun "#2019TetapJokowi". Fenomena ini membuat *netizen – netizen* khususnya *Twitter* peduli akan politik dan ini menjadi kesempatan bagi masyarakat untuk dapat menyuarakan pendapatnya atau bahkan dapat berinteraksi dengan para elite politik disosial media menjelang pilpres 2019. Hal ini juga dapat

menjadi kesempatan untuk Capres untuk dapat mengambil hati para pemilihnya nanti dengan menebarkan pencitraan didalam sosial media.

*Twitter*sendiri merupakan microblogging yang populer dikalangan masyarakat karena mudah diakses. Dalam penelitian ini semua data-data diambil dari *twitter* melalui aplikasi *tweetdeck*. Dan data-data yang diambil tersebut menyangkut tentang politik khususnya Pilpres 2019. *Tweet* tersebut mengandung sebuah opini-opini yang ditujukan untuk Pilpres 2019. Opini tersebut nantinya akan dilakukan analisis sentimen apakah termasuk opini positif atau opini negatif. Namun analalisis sentimen itu sendiri mendapatkan tantangan berupa model bahasa yang tidak formal yang digunakan di *Twitter*.

Maka dari itu sebelum melakukan analisis sentimen perlu diperhatikan tahap preprocessing pada data tweet yang sudah diambil. Hal ini berguna untuk mengatasi model bahasa yang tidak formal yang sering ditemukan pada *twitter*. Selain itu, pengklasifikasian sentimen ini dilakukan manual oleh manusia. Permasalahan yang terjadi adalah ketepatan dan kecepatan dalam menganalisis sentimen dalam jumlah data yang sangat banyak. Oleh karena itu, penggunaan aplikasi untuk dapat menganalisis sentimen secara otomatis merupakan solusi atas permasalahan tersebut.

3.1.2 Analisis Sumber Data

Data *tweet* yang digunakan adalah kumpulan *tweet* yang didalamnya mengandung tema "Pilpres" dalam bahasa Indonesia yang diambil menggunakan aplikasi Tweetdeck setiap hari pada pukul 12.00-15.00. Data *twitter* bertema pilpres ini menggunakan kata kunci tertentu agar *tweet* dapat diakui jika *tweet* tersebut benar benar bertema pilpres. Kata kunci yang digunakan untuk mendapatkat tema pilpres dapat dilihat pada tabel 3.1

Tabel 3.2 Daftar kata kunci untuk data tweet bertema "Pilpres"

No	Kata Kunci
1	Pilpres 2019
2	Jokowi Pilpres
3	Prabowo Pilpres
4	#GantiPresiden

Data *tweet* yang terkumpul nantinya akan melewati tahap *preprocessing* dan selanjutnya akan diklasifikasikan. Dalam sistem analisis sentimen ini, *tweet* akan diklasifikasikan ke dalam dua kelas yaitu kelas sentimen positif dan kelas sentimen negatif. Contoh dari tweets yang termasuk sentimen positif dapat dilihat pada tabel 3.2(a) untuk "Jokowi Pilpres" dan (b) untuk "Prabowo Pilpres"

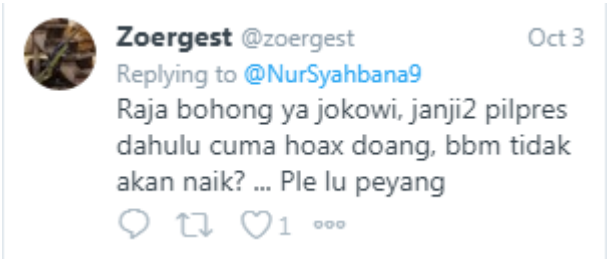
sedangkan *tweet* yang termasuk dalam sentimen negatif dapat dilihat pada gambar (c) untuk "Jokowi Pilpres" dan (d) untuk "Prabowo Pilpres".

(a)

Gambar 3.1Sentimen Positif "Jokowi Pilpres"

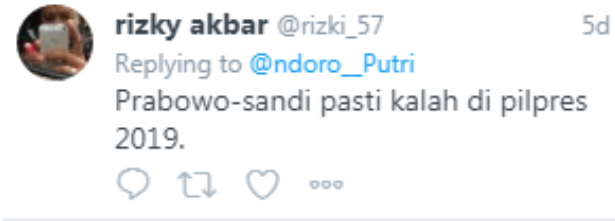
(b)

Gambar 3.1 Sentimen Positif "Prabowo Pilpres"



(c)

Gambar 3.1 Sentimen Negatif "Jokowi Pilpres"



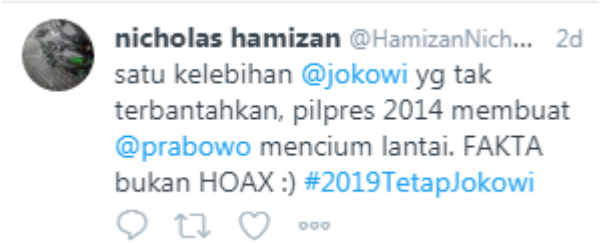
(d)

Gambar 3.1 Sentimen Negatif "Prabowo Pilpres"

Data yang dibutuhkan dalam penelitian ini terdiri dari dua jenis, yaitu data latih dan data uji. Data latih yang digunakan diambil dari sekumpulan tweet yang telah dilabeli dengan kelas sentimen secara manual. Data inilah yang digunakan untuk data latih membentuk model analisis sentimen. Model ini juga nantinya akan mengklasifikasikan *tweet* pada kelas sentimennya. Pada analisis ini metode yang digunakan adalah *NaÏve Bayes Classifier*. Sedangkan sebagian data yang sudah diambil akan dijadikan sebagai data uji. Data uji ini adalah kumpulan data *tweet* yang sudah memiliki label.

Setiap netizen memiliki ciri khasnya sendiri dalam menuliskan sebuah tweet. Dari hasil observasi dengan menuliskan "Jokowi Pilpres" dan "Prabowo Pilpres" dikolom search pada *Tweetdeck* terdapat beberapa karakteristik sebagai berikut :

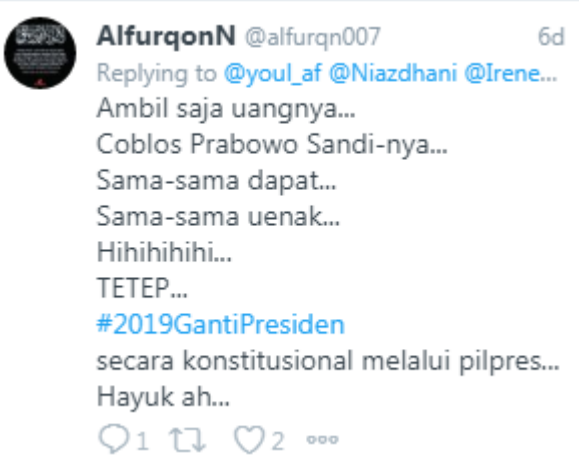
1. Penulisan kata yang disingkat.
   Keterbatasan karakter yang dapat ditulis pada suatu tweet yang hanya bisa menuliskan sebanyak 280 karakter saja membuat kata yang akan ditulis dapat disingkat. Contoh pada gambar 3.1(e).



(e)

Gambar 3.1 Contoh *Tweet* menggunakan "*singkatan*"

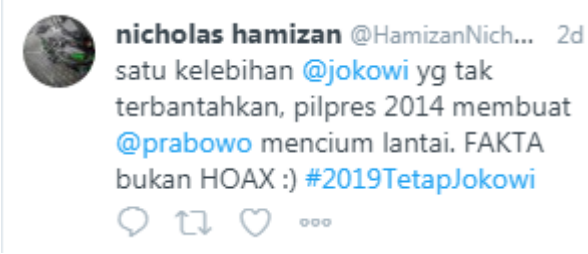
2. Penggunaan tanda baca titik (.) atau koma (,) pada akhiran *tweet*.
   Beberapa orang yang biasanya menggunakan banyak titik maupun koma pada akhiran *tweet* yang mereka tulis. Contoh pada Gambar 3.1(f).



(f)

Gambar 3.1 Contoh *Tweet* menggunakan banyak titik (.)

3. Penggunaan *emoticon*.
   Beberapa orang juga biasanya menambahkan sebuah *emoticon* supaya dapat mempertajam sentimen dari *tweet* tersebut. Contoh pada Gambar 3.1(g).

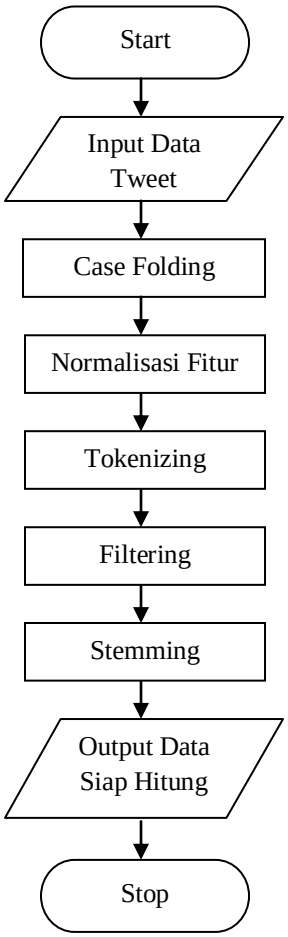


(g)

Gambar 3.1 Contoh *Tweet* menggunakan *emoticon* :)

3.2 Analisis Preprocessing

Preprocessing merupakan salah satu proses yang penting pada penelitian ini. Preprocessing disini yaitu proses untuk membuat data mentah menjadi data yang berkualitas dan siap diolah kepada proses berikutnya. Tahapan dari preprocessing itu sendiri dapat dilihat pada gambar flowchart 3.2



Gambar 3.2 Analisis Preprocessing

3.2.1 Case Folding

Case Folding adalah sebuah tahapan yang mengubah semua karakter pada tweet menjadi huruf kecil. Contoh (1): "@amossuryel: Setelah Pilpres 2019, semua akan kembali normal... Jangan terlalu ngegas ke kubu yang kontra dengan Anda, karena #KitaBersaudara @jokowi @prabowo". Contoh (2): "

1. Memeriksa ukuran setiap karakter dari awal sampai akhir tweet.
2. Jika ditemukan karakter yang menggunakan huruf kapital maka harus diubah menjadi huruf kecil.

Gambar dari *Case Folding* dapat dilihat pada Gambar 3.2.1

Input
@amossuryel: Setelah Pilpres 2019, semua akan kembali normal... Jangan terlalu ngegas ke kubu yang kontra dengan Anda, karena #KitaBersaudara @jokowi @prabowo :)

Output
@amossuryel: setelah pilpres 2019, semua akan kembali normal... jangan terlalu ngegas ke kubu yang kontra dengan anda, karena #kitabersaudara @jokowi @prabowo :)

Input
@IntanPermata: Jangan percaya koalisi Prabowo, kebanyakan bohong, penyebar hoax. Pilpres 2019 menjadi tidak menarik karena banyak kebohongan yang dilakukan @prabowo!!! #Prabohong

Output
@IntanPermata: jangan percaya koalisi prabowo, kebanyakan bohong, penyebar hoax. pilpres 2019 menjadi tidak menarik karena banyak kebohongan yang dilakukan @prabowo!!! #prabohong

Gambar 3.2.1 Contoh *Case Folding*

3.2.2 NormalisasiFitur

Tahapan ini berfungsi untuk menghilangkan komponen yang tidak mempengaruhi dari sentimen. Pada twitter terdapat fitur Hastag (#), Re-Tweet (RT), URL (Uniform Resource Locator), "@" dan username itu

sendiri. Beberapa fitur tersebut akan dibuang karena tidak berpengaruh pada sentimen itu sendiri. Berikut contoh dari tahap normalisasi fitur:

1. Data yang digunakan adalah data dari hasil case folding.
2. Hasil dari case folding diperiksa lagi, apakah masih ada kata yang mengandung username, Hastag (#), URL, @ atau RT.
3. Jika masih terdapat beberapa fitur diatas maka semua fitur tersebut akan dihapus.

Gambar dari normalisasi fitur dapat dilihat pada Gambar 3.2.2

Iutput
@amossuryel: setelah pilpres 2019, semua akan kembali normal... jangan terlalu ngegas ke kubu yang kontra dengan anda, karena #kitabersaudara @jokowi @prabowo :)

Output
setelah pilpres 2019, semua akan kembali normal... jangan terlalu ngegas ke kubu yang kontra dengan anda, karena:)

Input
@IntanPermata: jangan percaya koalisi prabowo, kebanyakan bohong, penyebar hoax. pilpres 2019 menjadi tidak menarik karena banyak kebohongan yang dilakukan @prabowo!!! #prabohong

Output
jangan percaya koalisi prabowo, kebanyakan bohong, penyebar hoax. pilpres 2019 menjadi tidak menarik karena banyak kebohongan yang dilakukan!!!

Gambar 3.2.2 Contoh Normalisasi Fitur

3.2.3 Tokenizing

Tahap ini adalah untuk mengecek kembali pada tweet dari karakter pertama hingga karakter terakhir dan memotong string input berdasarkan tiap kata yang menyusunnya. Apabila karakter ke-i bukan tanda pemisah kata seperti titik(.), koma(,), spasi atau lainnya maka akan digabungkan dengan karakter selanjutnya. Tahapan dari tokenizing adalah sebagai berikut :

1. Data yang digunakan ada data dari hasil convert emoticon.

2. Memotong setiap kata dalam teks berdasarkan pemisah kata seperti titik(.), koma(,), dan spasi.
3. Bagian yang hanya memiliki satu karakter non alfabet dan angka akan dibuang.

Gambar dari tokenizing dapat dilihat pada Gambar 3.2.3

Gambar 3.2.3Contoh Tokenizing

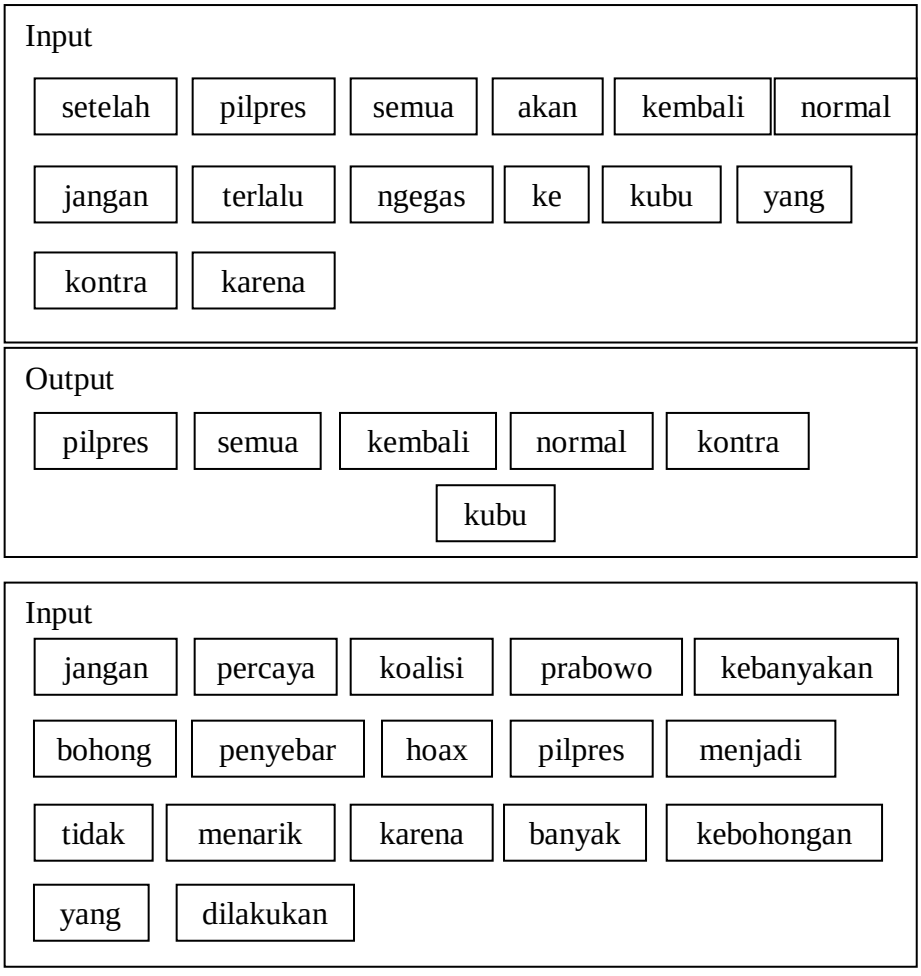
3.2.4 Filtering

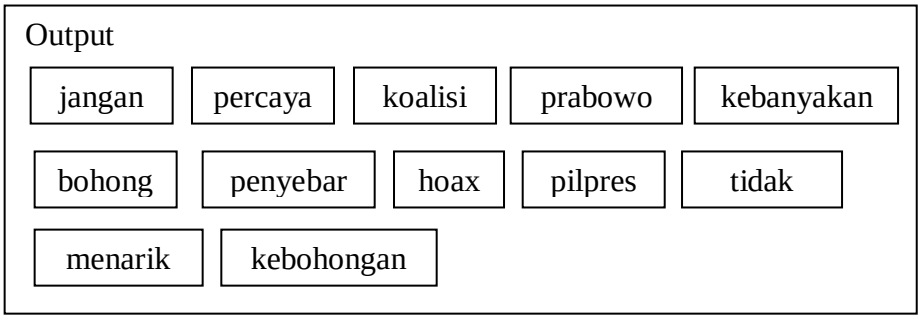
Tahap filtering adalah tahap mengambil kata-kata penting dari hasil tokenizing. Proses filtering dapat menggunakan algoritma stoplist (membuang kata yang kurang penting) atau wordlist (menyimpan kata penting). Jika ada kata sambung, kata depan, kata ganti atau kata yang

tidak ada hubungannya dengan analisis sentimen maka kata kata tersebut akan dibuang. Stoplist/stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang. Contoh dari stopword adalah “ke”, “di”, “yang”, “dan”, “dari” dan lain lain. Tahap dari filtering adalah sebagai berikut:

1. Data hasil dari tokenizing akan dibandingkan dengan daftar stopword.
2. Dilakukan pengecekan apakah kata sama dengan daftar stopword atau tidak.
3. Jika kata sama dengan yang apda daftar stopword maka kata tersebut akan dibuang.

Gambar dari filtering dapat dilihat pada Gambar 3.2.4





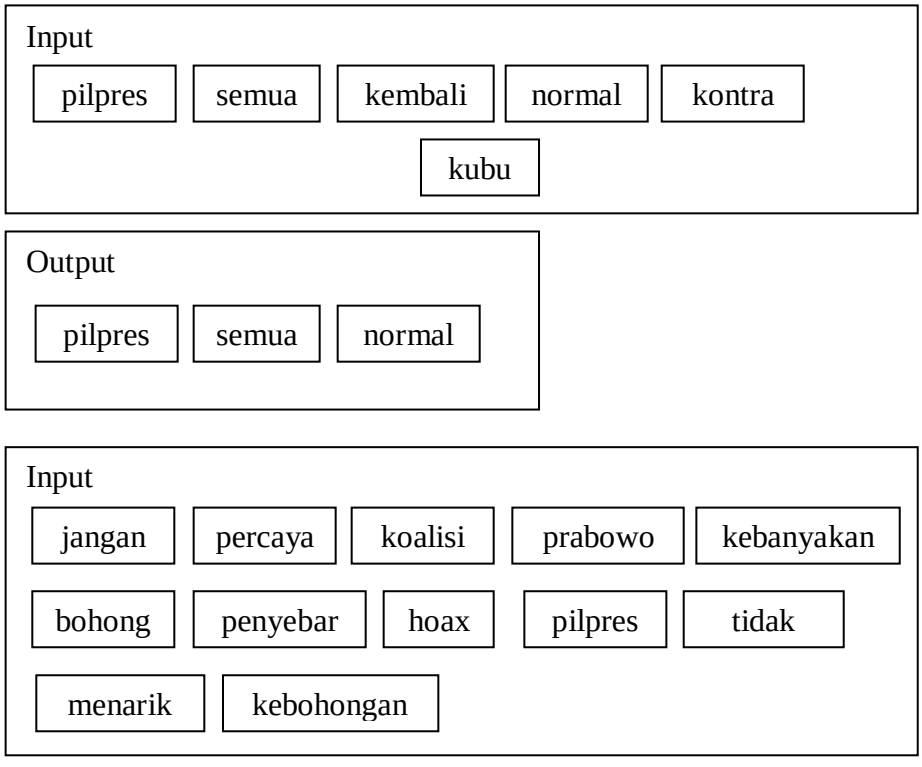
Gambar 3.2.4 Contoh Filtering

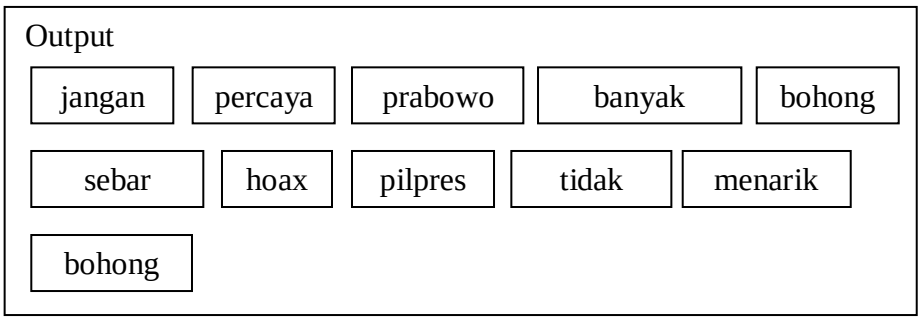
3.2.5 Stemming

Kata-kata yang muncul didalam tweet yang sering mempunyai banyak varian morfologik. Oleh karena itu, setiap kata direduksi ke bentuk stemmed word(term) yang cocok. Kata-kata tersebut diambil dari bentuk dasarnya tanpa menambahi awalan dan akhiran. Tahapan dalam stemming adalah sebagai berikut:

1. Data yang digunakan adalah hasil dari filtering
2. Setiap kata yang terdapat dalam tweet diperiksa dari awal hingga akhir
3. Jika terdapat kata imbuhan, maka imbuhan pada kata tersebut akan dihilangkan

Gambar dari stemming dapat dilihat pada gambar 3.2.5





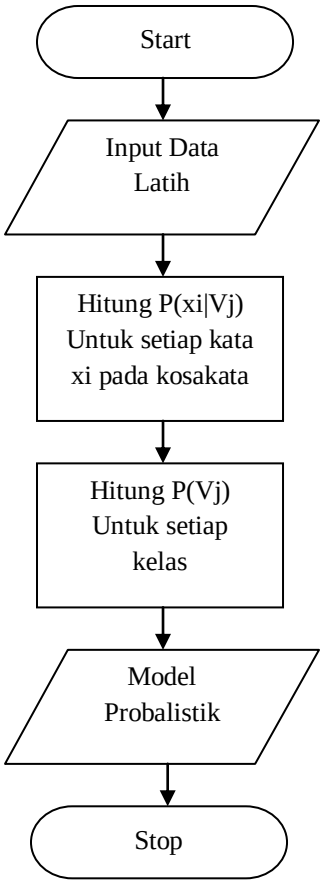
Gambar 3.2.5 Contoh Stemming

3.3 Klasifikasi dan Penerapan *NaÏve Bayes Classifier*.

Pada tahap ini, metode yang digunakan dalam pengklasifikasian sentimen adalah *NaÏve Bayes Classifier*. Metode ini terdiri dari dua proses, yaitu sebagai berikut :

1. Pelatihan *NaÏve Bayes Classifier.*

   Secara umum proses ini dibagi menjadi beberapa tahap. Tahapan tahapan tersebut dapat dilihat pada Gambar 3.3



Gambar 3.3 Flowchart pelatihan *NaÏve Bayes Classifier*

Ket:

- Input Data Latih
  Menginputkan data tweet yang akan dilatih.
- Hitung P(Vj) untuk setiap kelas
  Menghitung probabilitas (nilai kemunculan) dari Vj dengan persamaan (2.10)
- Hitung P(xi|Vj) untuk setiap kata xi pada kosakata
  Menghitung probabilitas xi pada kategori Vj dengan persamaan (2.11)
- Model Probabilistik
  Menentukan model dari probabilitas yang akan dicari.

Sebelum lebih jauh tentang *NaÏve Bayes Classifier*, hal hal yang harus diperhatikan adalah sebagai berikut:

Kosakata

|kosakata| adalah jumlah kata yang unik dalam semua data latih. Data latih adalah data-data tweets yang sudah dilakukan pengklasifikasian. Pada penelitian ini, kata dibagi menjadi dua kelas yaitu:

a. Data1 (D1) = kelas sentimen positif.
b. Data2 (D2) = kelas sentimen negatif.
c. Data3 (D3) = kelas sentimen negatif.

Contoh dari kumpulan kata latih terdapat pada Tabel 3.2.

Data	Keyword (Kemunculan)	Kelas Sentimen
D1	Pilpres(1), Semua(1), Normal(1)	Positif
D2	Purnawirawan(1), Dukung(2), Jokowi(1), Pilpres(1)	Positif
D3	Dukung(2), PKS(1), Optimis(1), Prabowo(1), Menang(1), Pilpres(1)	Positif
D4	Jokowi(1), Prabowo(1), Pilpres(2), Damai(1)	Positif
D5	Jangan(1), Bohong(2), Tidak(1), Hoax(1)	Negatif
D6	Kalah(3), Palsu(1), Provokasi(1)	Negatif
D7	Jokowi(1), Kalah(2), Pilpres(1)	Negatif

Perhitungan kelas sentimen positif

Tweet yang digunakan dalam perhitungan ini adalah tweet yang mempunyai kategori positif. Contoh tweet yang sudah terdapat pada tabel 3.2 adalah (Pilpres, Semua, Normal), (Dukung, Purnawirawan, Dukung, Jokowi, Pilpres), (Dukung, PKS, Optimis, Dukung, Prabowo, Menang, Pilpres), (Pilpres, Jokowi, Prabowo, Pilpres, Damai). Semua kata yang muncul akan menjadi kata positif. Pada Tabel 3.3 akan ditampilkan untuk kata sentimen positif dan frekuensinya.

Tabel 3.3 daftar kata sentimen positif

No	Kata	Frekuensi(n_k)
1	Pilpres	5
2	Semua	1
3	Normal	1
4	Purnawirawan	1
5	Dukung	4
6	Jokowi	2
7	PKS	1
8	Optimis	1
9	Prabowo	2
10	Menang	1
11	Damai	1
		20

Dari tabel diatas dapat diketahui:
Jumlah keseluruhan kata sentimen positif (n) = 20
Jumlah kata (kata) = 34
Hitung probabilitas setiap kata dengan persamaan (2.11)

$$P(xi|Vj) = \frac{nk+1}{n+|kosakata|}$$

Contoh: $P(\text{dukung}|\text{positif}) = \frac{4+1}{20+34} = 0.092$

$P(\text{normal}|\text{positif}) = \frac{1+1}{20+34} = 0.037$

$P(\text{dukung}|\text{negatif}) = \frac{0+1}{0+34} = 0.029$

Pada Tabel 3.4 ditampilkan daftar dari probabilitas setiap kata.

Tabel 3.4 daftar probabilitas sentimen positif

No	Kata	Probabilitas
1	Pilpres	0.111
2	Semua	0.037
3	Normal	0.037
4	Purnawirawan	0.037
5	Dukung	0.092
6	Jokowi	0.055
7	PKS	0.037
8	Optimis	0.037
9	Prabowo	0.055
10	Menang	0.037
11	Damai	0.037

Diketahui:
Jumlah tweet positif = 4
Jumlah tweet negatif = 3
Maka nilai dari P(Vj) = $\frac{|docs\ j|}{|all\ docs|}$

P(Positif) = 4/7 = 0.571
P(Negatif) = 3/7 = 0.428

2. Perhitungankelassentimennegatif

Tweet yang digunakan dalam perhitungan ini adalah tweet yang mempunyai kategori negatif. Contoh tweet yang sudah terdapat pada tabel 3.1 adalah(Jangan, Bohong, Bohong, Tidak, Hoax) dan (Kalah, Kalah, Kalah, Palsu, Provokasi), (Jokowi, Kalah, Kalah, Pilpres) Semua kata yang muncul akan menjadi kata negatif. Pada Tabel 3.5 akan ditampilkan untuk kata sentimen negatif dan frekuensinya.

Tabel 3.5 daftar kata sentimen negatif

No	Kata	Frekuensi(n_k)
1	Jangan	1
2	Bohong	2
3	Tidak	1
4	Hoax	1
5	Kalah	5
6	Palsu	1
7	Provokasi	1
8	Jokowi	1
9	Pilpres	1
		14

Dari tabel diatas dapat diketahui:
Jumlah keseluruhan kata sentimen negatif (n) = 14
Jumlah kata sentimen positif dan sentimen negatif = 34
Maka akan didapatkan nilai probabilitas dari sentimen negatif dengan persamaan (2.11) sebagai berikut:

Contoh: P(jangan|negatif) = $\frac{1+1}{14+34}$ = 0.041

P(bohong|negatif) = $\frac{2+1}{14+34}$ = 0.062

P(kalah|negatif) = $\frac{5+1}{14+34}$ = 0.125

P(jangan|positif) = $\frac{0+1}{0+34}$ = 0.029

P(bohong|positif) = $\frac{0+1}{0+34}$ = 0.029

P(kalah|positif) = $\frac{0+1}{0+34}$ = 0.029

Pada abel 3.6 akan ditampilan nilai dari probabilitas kata sentimen negatif

Tabel 3.6 daftar probabilitas sentimen negatif

No	Kata	Probabilitas
1	Jangan	0.041
2	Bohong	0.062
3	Tidak	0.041
4	Hoax	0.041
5	Kalah	0.125
6	Palsu	0.041
7	Provokasi	0.041
8	Jokowi	0.041
9	Pilpres	0.041

3. Klasifikasi

Input data dari klasifikasi ini adalah sebuah data tweet yang belum mempunyai kelas sentimen kategorinya. Sebagai contoh tweet dibawah akan dilakukan proses klasifikasi untuk menentukan kelas sentimen tweet tersebut. Contoh tweet:

“Dukungan untuk Jokowi di pilpres 2019 bertambah, jokowi yakin pilpres 2019 pasti menang”

Proses perhitungannya menjadi seperti ini:

Hitung posterior probability pada tweet dengan persamaan (9)

$P(V_j|x_1, \dots, x_n) = \underset{Vj \in V}{argmax}\ P(V) \prod_{i=1}^{n} P(xi|Vj)$

D8	dukung(1), jokowi(2), pilpres(2), menang(1)	?

P(Positif) = P(dukung|positif) P(jokowi|positif) P(jokowi|positif) P(pilpres|positif) P(pilpres|positif) P(menang|positif) P(positif)

= (0.037) x (0.037) x (0.037) x (0.037) x (0.037) x (0.037) x (0.57)

= 1.542

P(Negatif) = P(dukung|negatif) P(jokowi|negatif) P(jokowi|negatif) P(pilpres|negatif) P(pilpres|negatif) P(menang|negatif) P(negatif)

= (0.029) x (0.041) x (0.041) x (0.041)x (0.041) x (0.029) x (0.43)

= 1.021

Jadi kesimpulan dari perhitungan dengan data tweet “Dukungan untuk Jokowi di pilpres 2019 bertambah, jokowi yakin pilpres 2019 pasti menang” adalah Positif.

UNIVERSITAS SAHID SURAKARTA