BAB III

METODELOGI PENELITIAN

3.1 Metode Penelitian

Metode penelitian yang digunakan yaitu metode eksperimental dimana metode ini bekerja dengan memanipulasi dan melakukan kontrol pada objek penelitian [2]. Metode eksperimental bertujuan untuk menyelidiki hubungan sebab akibat dan seberapa besar hubungan sebab akibat tersebut dengan cara memberikan kontrol perbandingan. Berikut adalah beberapa kriteria umum pada metode eksperimantal:

a. Pemilihan masalah yang dipilih harus penting dan dapat dipecahkan
b. Mendefinisikan variable secara mendalam dalam suatu percobaan
c. Melakukan percobaan yang sesuai dengan desain percobaan yang cocok
d. Ketelitian saat observasi dan ketepatan pengukuran sangatlah diperlukan
e. Menjelaskan metode, material, dan referensi yang jelas
f. Analisis pengujian statistik
g. Interpretasi dan generalisasi

Syarat suatu percobaan yang baik adalah sebagai berikut :

a. Harus bebas dari bias
b. Mempunyai ukuran terhadap error atau kesalahan
c. Mempunyai ketepatan
d. Mendefinisikan tujuan dengan jelas
e. Mempunyai jangkauan percobaan yang cukup

3.2 Instrument Penelitian

Untuk melakukan tahapan proses *sentiment analysis* diperlukan adanya perangkat pendukung, diantaranya:

3.2.1 Hardware

Disini spesifikasi *hardware* yang diperlukan untuk melakukan penelitian ini sebagai berikut:

CPU	:	Intel Core i5
RAM	:	4 GB

Graphic Card : 256 MB

Connection : Internet Access

3.2.2 Software

Untuk spesifikasi *software* yang digunakan untuk penelitian ini dibagi menjadi 2 kategori:

Tabel 3.1 *Software Requirement*

Kategori	OS	Tools
Data Crawling and Modeling	Windows 8 32 BIT	• R GUI v3.2.2 • Microsoft Excel 2010
Data Preprocessing	Windows 8 32 BIT	• XAMPP v3.2.1 • Notepad ++ v6.3.2 • Mozila Firefox v42.0

3.3 Metode Pengumpulan Data

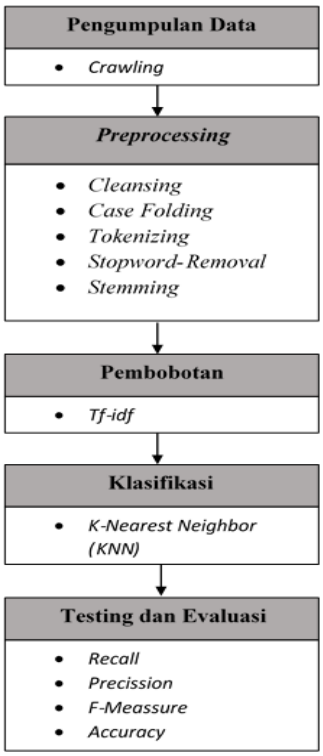
Pada penelitian ini, pengumpulan data dilakukan dengan cara melakukan *crawling* untuk mengambil *tweets* Berbahasa Indonesia tentang topik terkait melalui fasilitas *searching* yang disediakan oleh twitter dengan memanfaatkan API Twitter menggunakan *tools R GUI.*

3.4 Teknik Analisis Data

Data mentah yang telah diperoleh kemudian masuk ke tahapan *preprocessing*, dimana data tersebut akan melewati proses *cleansing*, *case folding* serta *tokenizing* untuk membersihkan data tersebut dari data yang tidak diperlukan sehingga dapat mengurangi resiko data *noise* yang tinggi.

3.5 Metode yang Diusulkan

Berikut ini adalah skema penelitian yang dilakukan pada penelitian ini:



Gambar 3.1 Skema Penelitian

3.5.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini diambil dari jejaring sosial *twitter*. Pengambilan data dengan memanfaatkan *tools R GUI* yang dihubungkan dengan API pencarian *twitter* yang berhubungan dengan topik terkait Universitas Dian Nuswantoro menggunakan kata kunci "udinus" dan "dinus". Di dalam satu data *tweet* memiliki maksimal 140 karakter. Setiap kali *request* pengambilan data, *API twitter* akan memberikan sampel *tweet* secara acak sebanyak jangka waktu seminggu kebelakang. Kita bisa menentukan batas maksimal data yang kita inginkan, namun data yang diberikan hanya sebatas berapa banyak *tweet* dengan kata kunci terkait dalam jangka waktu satu minggu sebelum tanggal pencarian. Sedangkan untuk seleksi bahasa digunakan library bawaan *TwitteR*; (lang='id') yang merupakan kode untuk teks Berbahasa Indonesia.

Berikut ini adalah contoh data twitter yang berhasil diambil:

no	text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID	statusSource	screenName	retweetCount	isRetweet	retweeted	longitude	latitude
1	Jika kita menetapkan ingin hidup ini seperti apa, lalu kerja ke	FALSE	0	NA	11/8/2015 0:16	FALSE	NA	6.63E+17	NA	<a href="http:	dinus_career	0	FALSE	FALSE	NA	NA
2	Hari Wayang Sedunia, Dalang Sukron Manggung bersama T	FALSE	0	NA	11/8/2015 0:27	FALSE	NA	6.63E+17	NA	<a href="http:	KotaSMG	0	FALSE	FALSE	NA	NA
3	Hari Wayang Sedunia, Dalang Sukron Manggung bersama T	FALSE	0	NA	11/8/2015 0:27	FALSE	NA	6.63E+17	NA	<a href="http:	KotaSMG	1	FALSE	FALSE	NA	NA
4	Hari Wayang Sedunia, Dalang Sukron Manggung bersama T	FALSE	0	NA	11/8/2015 0:27	FALSE	NA	6.63E+17	NA	<a href="http:	SeputariSemarang	0	FALSE	FALSE	NA	NA
5	Hari Wayang Sedunia, Dalang Sukron Manggung bersama T	FALSE	0	NA	11/8/2015 0:33	FALSE	NA	6.63E+17	NA	<a href="http:	ini_jateng	0	FALSE	FALSE	NA	NA
6	#BeritaSMG Hari Wayang Sedunia, Dalang Sukron Manggur	FALSE	0	NA	11/8/2015 0:59	FALSE	NA	6.63E+17	NA	<a href="http:	KotaSMG	4	FALSE	FALSE	NA	NA
7	RT @KotaSMG: #BeritaSMG Hari Wayang Sedunia, Dalang	FALSE	0	NA	11/8/2015 1:03	FALSE	NA	6.63E+17	NA	<a href="http:	MdhesthaSPA_	4	TRUE	FALSE	NA	NA
8	#InfoSemarang: Hari Wayang Sedunia, Dalang Sukron Many	FALSE	0	NA	11/8/2015 1:10	FALSE	NA	6.63E+17	NA	<a href="http:	AreaSemarangID	0	FALSE	FALSE	NA	NA
9	Hari Wayang Sedunia, Dalang Sukron Manggung bersama T	FALSE	0	NA	11/8/2015 1:10	FALSE	NA	6.63E+17	NA	<a href="http:	InfoSemarang468	0	FALSE	FALSE	NA	NA
10	RT @KotaSMG: #BeritaSMG Hari Wayang Sedunia, Dalang	FALSE	0	NA	11/8/2015 1:14	FALSE	NA	6.63E+17	NA	<a href="http:	MYGKSW	4	TRUE	FALSE	NA	NA
11	Follow @hmti_udinus buat kalian anak Teknik Informatika ya	FALSE	0	NA	11/8/2015 1:15	FALSE	NA	6.63E+17	NA	<a href="http:	anakdinusdotcom	2	FALSE	FALSE	NA	NA
12	RT @KotaSMG: Hari Wayang Sedunia, Dalang Sukron Many	FALSE	0	NA	11/8/2015 1:15	FALSE	NA	6.63E+17	NA	<a href="http:	MYGKSW	1	TRUE	FALSE	NA	NA
13	#BeritaSMG Hari Wayang Sedunia, Dalang Sukron Manggur	FALSE	0	NA	11/8/2015 1:19	FALSE	NA	6.63E+17	NA	<a href="http:	InfosemarangID	0	FALSE	FALSE	NA	NA
14	RT @anakdinusdotcom: Follow @hmti_udinus buat kalian ai	FALSE	0	NA	11/8/2015 1:19	FALSE	NA	6.63E+17	NA	<a href="https	hmti_udinus	2	TRUE	FALSE	NA	NA
15	RT @KotaSMG: #BeritaSMG Hari Wayang Sedunia, Dalang	FALSE	0	NA	11/8/2015 1:19	FALSE	NA	6.63E+17	NA	<a href="http:	sgci_semarang	4	TRUE	FALSE	NA	NA
16	ketika anak Udinus dan Unika disatukan, maka semua akan	FALSE	0	NA	11/8/2015 1:20	FALSE	NA	6.63E+17	NA	<a href="http:	akiramitsuda	0	FALSE	FALSE	NA	NA
17	Hari Wayang Sedunia, Dalang Sukron Manggung bersama T	FALSE	0	NA	11/8/2015 1:40	FALSE	NA	6.63E+17	NA	<a href="http:	tamkaRHidayanto	0	FALSE	FALSE	NA	NA
18	Hari Wayang Sedunia, Dalang Sukron Manggung bersama T	FALSE	0	NA	11/8/2015 1:40	FALSE	NA	6.63E+17	NA	<a href="http:	WelcomeSemarang	0	FALSE	FALSE	NA	NA
19	Untuk info #Lowongan selengkapnya, silahkan klik link yang	FALSE	0	NA	11/8/2015 1:45	FALSE	NA	6.63E+17	NA	<a href="http:	dinus_career	0	FALSE	FALSE	NA	NA
20	Di UDINUS ada komunitas yang bergerak di bidang Open So	FALSE	1	NA	11/8/2015 2:01	FALSE	NA	6.63E+17	NA	<a href="http:	anakdinusdotcom	2	FALSE	FALSE	NA	NA

Gambar 3.2 Data Utuh

Dari data utuh yang terkumpul kemudian akan dipilah dan nantinya yang akan digunakan adalah data pada kolom text yang berisi *tweets* dari berbagai *user* dengan topik mengenai Universitas Dian Nuswantoro Semarang. Berikut ini contoh *tweets* yang berhasil diperoleh:

no	text
1	Jika kita menetapkan ingin hidup ini seperti apa, lalu kerja keras untuk mencapai tujuan, kita tdk akan pern
2	Hari Wayang Sedunia, Dalang Sukron Manggung bersama Tim E-gamelankuÿUdinus https://t.co/hGgAQ4nC
3	Hari Wayang Sedunia, Dalang Sukron Manggung bersama Tim E-gamelanku Udinus https://t.co/hGgAQ4nC
4	Hari Wayang Sedunia, Dalang Sukron Manggung bersama Tim E-gamelankuÿUdinus https://t.co/k4MjNE5o
5	Hari Wayang Sedunia, Dalang Sukron Manggung bersama Tim E-gamelanku Udinus Tim e-gamelanku Udinu
6	#BeritaSMG Hari Wayang Sedunia, Dalang Sukron Manggung bersama Tim E-gamelanku Udinus https://t.co
7	RT @KotaSMG: #BeritaSMG Hari Wayang Sedunia, Dalang Sukron Manggung bersama Tim E-gamelanku Ud
8	#InfoSemarang: Hari Wayang Sedunia, Dalang Sukron Manggung bersama Tim E-gamelanku Udinus https://
9	Hari Wayang Sedunia, Dalang Sukron Manggung bersama Tim E-gamelanku Udinus https://t.co/YRSctfg8dK
10	RT @KotaSMG: #BeritaSMG Hari Wayang Sedunia, Dalang Sukron Manggung bersama Tim E-gamelanku Ud
11	Follow @hmti_udinus buat kalian anak Teknik Informatika yang mw dapet info terkini di jurusan Teknik Info
12	RT @KotaSMG: Hari Wayang Sedunia, Dalang Sukron Manggung bersama Tim E-gamelanku Udinus https://
13	#BeritaSMG Hari Wayang Sedunia, Dalang Sukron Manggung bersama Tim E-gamelanku Udinus https://t.co
14	RT @anakdinusdotcom: Follow @hmti_udinus buat kalian anak Teknik Informatika yang mw dapet info terk
15	RT @KotaSMG: #BeritaSMG Hari Wayang Sedunia, Dalang Sukron Manggung bersama Tim E-gamelanku Ud
16	ketika anak Udinus dan Unika disatukan, maka semua akan menjadi gilaaaa? https://t.co/yjeQuwrtZC
17	Hari Wayang Sedunia, Dalang Sukron Manggung bersama Tim E-gamelanku Udinus https://t.co/U0UxQ4lVP

Gambar 3.3 Data Tweets

3.5.2 Preprocessing

Pada proses preprocessing ini, data yang berhasil kita ambil dari proses *crawling* tadi diproses kedalam 3 tahapan yaitu [8]:

1. *Cleansing*, yaitu proses pembersihan dokumen dari kata yang tidak diperlukan untuk mengurangi data *noise*. Kata yang dihilangkan adalah karakter HTML,

kata kunci, emotikon, *hashtag* (#), *username* (@username), url (http://situs.com) dan email (email@situs.com).

2. *Case Folding*, yaitu proses penyeragaman bentuk huruf, dan penghapusan angka serta tanda baca. Pada kata lain data yang digunakan hanya karakter huruf a sampai z.
3. *Tokenizing*, yaitu proses memecah dokumen teks menjadi sebuah kata.

3.5.3 Pemilihan dan Fitur Ekstraksi

Proses ini dilakukan sebagai dasar proses klasifikasi yang nantinya akan dilakukan, proses ini terbagi menjadi 2 tahapan yaitu [7]:

1. *Stopword Removal*, yaitu proses penghilangan kata yang tidak mendeskripsikan sesuatu dalam Bahasa Indonesia seperti “di”, “ke”, “dari”, “yang”, “sedang”, “ini”, dan lain sebagainya. Namun didalam *text classification* keberadaan kata seperti “tidak”, “bukan”, “tanpa” tidak begitu penting sehingga kata ini biasanya tidak ikut dihilangkan.

   Berikut contoh kata dalam Bahasa Indonesia yang masuk dalam *stopword list* menurut KBBI (Kamus Besar Bahasa Indonesia):

Tabel 3.2 ***Stopwords List***

id	stoplist
1	ada
2	adalah
3	adanya
4	adapun
5	agak
6	agaknya
7	agar
8	akan
9	akankah
10	akhir
11	akhiri
12	akhirnya
13	aku
14	akulah
15	amat

2. *Stemming*, yaitu proses penghilangan imbuhan yang masih melekat sehingga diperoleh sebuah kata dasar, contoh: “membaca”, “dibaca”, “dibacakan” akan dikonversi menjadi kata dasar (*stem*) “baca”. Dalam proses ini terdapat 5 aturan yaitu:
    a. Menghilangkan partikel (*-lah, -kah, -tah,* dan *-pun*).
    b. Menghilangkan kata ganti kepemilikan (*-ku, -mu,* dan *-nya*) .
    c. Menghilangkan awalan tingkat pertama (*meng-, di-, ter-*, dan *ke-*).
    d. Menghilangkan awalan tingkat kedua (*per-*, dan *ber-*).
    e. Menghilangkan akhiran (*-i, -kan,* dan *-an*).

    Berikut ini adalah contoh daftar kata dasar dalam Bahasa Indonesia menurut KBBI (Kamus Besar Bahasa Indonesia):

Tabel 3.3 Tabel Kata Dasar

id_katadasar	katadasar	tipe_katadasar
1	a	Nomina
2	ab	Nomina
3	aba	Nomina
4	aba-aba	Nomina
5	abad	Nomina
6	abadi	Adjektiva
7	abadiah	Nomina
8	abah	Nomina
9	abai	Adjektiva
10	abaimana	Nomina
11	abaka	Nomina
12	abaktinal	Adjektiva
13	abakus	Nomina
14	abal-abal	Nomina
15	aban	Nomina
16	abang	Nomina
17	abangan	Nomina
18	abangga	Nomina
19	abar	Nomina
20	abatoar	Nomina

Dalam penelitian ini untuk proses *stemming* akan dilakukan dengan memanfaatkan *library* "Sastrawi" yang mana *library* ini memang dikhususkan untuk proses *stemming* dokumen teks Berbahasa Indonesia.

3.5.4 *Pembobotan (Term Weighting)*

Sebuah dokumen teks memgandung banyak kumpulan kata, sehingga sebuah transformasi kedalam bentuk yang dapat digunakan dalam proses klasifikasi dibutuhkan. Dengan memodelkannya kedalam bentuk vektor, setiap dokumen C akan diubah kedalam bentuk vektor *term-space* (sekumpulan kata yang muncul pada keseluruhan dokumen) [5].

$$C = (t_1, t_2, \ \dots, \ t_n) \qquad (7)$$

Dimana t_n adalah kemunculan kata ke-*n* dalam dokumen. Ada dua landasan dalam pembentukan vektor ini.

a. *Binary*, yaitu hanya berdasarkan keberadaan sebuah kata pada sebuah dokumen,
b. *Frequency*, yaitu dengan berdasarkan frekuensi kemunculan kata dalam dokumen tekstual.

Pembobotan dokumen teks ini dilakukan didalam bentuk vektor dengan menggunakan *term* yang dapat dikenali dengan perhitungan berdasarkan metode *TF-IDF*. Metode ini merupakan penggabungan metode *Term Frequency (TF)* yang dihubungkan dengan *Inverse Document Frequency (IDF)* dengan rumus sebagai berikut:

$$w_{(i,j)} = TFIDF\left(d_i,\ t_j\right) = Nd_i, t_j \,.\log\frac{|c|}{N_{t_j}} \qquad (8)$$

Dimana:

Nd_i, t_j = jumlah dari *term* t_j dalam dokumen *d*

N_{t_j} = jumlah dokumen didalam kumpulan *C*

3.5.5 *Metode Klasifikasi*

Sebuah dokumen *d* haruslah dapat diklasifikasikan kedalam kelas yang tepat. Proses klasifikasi ini meliputi dua tahapan. Pertama, sebuah model dibuat dengan menggambarkan sekumpulan kelas data atau konsep dari sebuah populasi data yang sudah ditentukan sebelumnya. Model ini dibuat dengan menganalisa data training yang dideskripsikan berdasarkan atribut yang dimilikinya. Setiap tupel diasumsikan dimiliki oleh kelas yang sudah didefinisikan, yang ditentukan dengan sebuah atribut, yang disebut *class label attribute*.

Tahapan kedua adalah pengujian model terhadap data untuk mengukur tingkat akurasi model atau performanya didalam mengklasifikasikan *data testing*. Setelah semuanya diukur, pengambilan keputusan dapat ditentukan untuk menggunakan model tersebut atau mengulangi proses pembentukan model menggunakan *data training*.

A. Metode K-Nearest Neighbor (KNN)

KNN akan memproses data yang dihasilkan dari proses preprocessing. Dalam penyelesaiannya metode ini akan mencari termasuk kategori manakah data tersebut. Berikut ini langkah-langkahnya [7]:

1. Menghitung *similarity* (tingkat kemiripan) antara dokumen sampel dengan dokumen test menggunakan rumus nomor (3).
2. Berdasarkan rumus nomor (3), dilakukan multiplikasi matriks ${B^T}_{n,k} \,.\, A_{k,m}$ yang kemudian menghasilkan matriks $C_{n,m}$. Baris ke-*i* dari matriks $C_{n,m}$ menunjukan kemiripan dari dokumen test ke-*i* dan seluruh kategori sampel.

3. Menghitung jarak dari J_i dan K_i dari masing-masing matriks vektor $A_{k,n}$ dan $B_{k,n}$.
4. Membuat matriks baru $D_{n,m}$ dengan nilai item dari $\frac{C_{(i,j)}}{J_i * K_i}$.
5. Pada setiap baris vektor $D_{n,m}$ diurutkan dari bawah untuk i = 1…n.
6. Berdasarkan nilai *k* yang diberikan. Ambil nilai *k* yang paling besar dari setiap baris vektor yang telah diurutkan. Pemilihan nilai *k* terbesar ini merepresentasikan nilai *k* dari tetangga terdekat. Masing-masing nilai *k* terbesar dievaluasi kedalam *term* dari anggota kelasnya dengan menggunakan rumus nomor (5).
7. Dengan menggunakan rumus nomor (4), hitung probabilitas dari tiap-tiap dokumen test pada masing-masing kelasnya dengan memangkatkan vektor ke-*k* dengan hasil dari proses f.
8. Tentukan probabilitas terbesar dan hasil kelasnya.

3.5.6Testing dan Evaluasi

Recall mengacu pada jumlah pengenalan entitas yang bernilai *true* atau benar yang dilakukan oleh sistem, dibagi dengan jumlah entitas yang seharusnya diproses oleh sistem; kemudian *Precision* dihitung dari jumlah pengenalan yang memiliki nilai *true,* dibagi dengan keseluruhan data yang berhasil dikenali oleh sistem. Berikut contoh tabel *confusion matrix*:

Tabel 3.4 *Confusion Matrix*

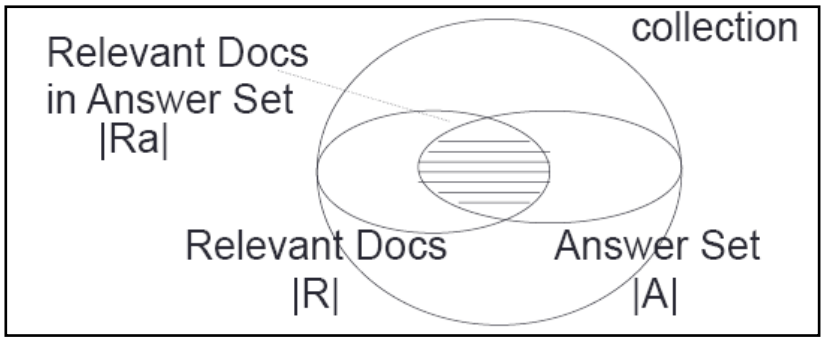
		Nilai Sebenarnya	
		TRUE	FALSE
Nilai Prediksi	TRUE	TP	FP
	FALSE	FN	TN

Sehingga dapat dirumuskan sebagai berikut [13]:

$$Precision = \frac{TP}{TP+FP} \qquad (9)$$

$$Recall = \frac{TP}{TP+FN} \qquad (10)$$

Berikut penggambaran *precision* dan *recall* kedalam sebuah diagram [13]:



Gambar 3.4 Diagram *Precission and Recall*

Precision dan *Recall* adalah ukuran himpunan. Dalam sebuah himpunan *rangked list*, kita dapat menghitung *precision* di setiap *recall point*. *Recall* meningkat ketika sebuah dokumen relevan terambil, menghitung *precision* di tiap dokumen relevan terambil, dari seluruh bagian dari retrieved set. Terdapat sebuah pertukaran pengaruh antara *precision* dan *recall*. Semakin banyak dokumen terambil, akan meningkatkan *recall*. Namun hal tersebut akan mengurangi *precision* [13].

F-measure merupakan hasil representasi keseluruhan sistem dan dimatematikakan dengan menggabungkan hasil dari *recall* dengan *precision* yang dapat dirumuskan sebagai berikut [6]:

$$Fmeasure = \frac{2*P*R}{P+R} \qquad (11)$$

Dimana *P* merupakan *precision* dan *R* adalah *recall.*

Accuracy merupakan tingkat keakuratan suatu metode yang diimplementasikan pada sebuah masalah, yang dapat dirumuskan sebagai berikut [8]:

$$Accuracy = \frac{\sum True\ positif + \sum True\ negatif}{\sum Data\ Testing} \qquad (12)$$