BAB III
METODOLOGI PENELITIAN

3.1. Metode Penelitian

Penelitian ini menggunakan jenis penilitian eksperimen, dengan tahapan penelitian sebagai berikut:

Gambar 3. 1 Alur Penelitian

Seperti yang ditunjukkan pada Gambar 3.1. Setelah melakukan pengumpulan data dan menentukan metode selanjutnya adalah tinggal melakukan uji coba dari hasil data yang sudah dikumpulkan. Selanjutnya data dari hasil penelitian akan dianalisis dengan cara melakukan verifikasi dan validasi data hingga akan menghasilkan output analisis yang memiliki tingkat akurasi yang tinggi.

3.1.1. Pengumpulan Data

Tahap ini dilakukan sebagai awal dari suatu penelitian. Studi literatur dilakukan pada literatur – literatur yang sesuai dengan penelitian ini. Dataset yang digunakan pada penelitian ini menggunakan dataset berupa postingan/unggahan Twitter yang diperoleh melalui internet. Pada gambar 3.2 menunjukan pengumpulan data yang dilakukan dengan cara menangkap tweet dengan cara mendapatkan data secara langsung (Crawl) Twitter menggunakan API (Application Interface) pada Twitter dengan program Pyhton. Data yang digunakan pada penelitian ini, sebanyak 51.250 tweet data dalam kurung waktu tanggal 10/01/2023 sampai 22/07/2023. Gambar 3.3 menunjukan bentuk data Twitter yang diperoleh melalui proses crawling.

```
# Crawl Data
filename = 'CrawData.csv'
search_keyword = 'indonesia until:2023-07-22 since:2023-01-10 lang:id'
limit = 1000

!npx --yes tweet-harvest@latest -o "{filename}" -s "{search_keyword}" -l {limit} --token ""
```

Gambar 3. 2 Pengumulan Data Penelitian dengan Python

	created_at	id_str	full_text	quote_count	reply_count	retweet_count	favorite_count	lang	user_id_str	conversation_id_str
0	Thu Jul 20 23:59:56 +0000 2023	1682178560524685312	Sejarah Singkat Bakso, Beginilah Asal-Usulnya ...	0	0	0	0	in	1180081130621394944	1682178560524685312
1	Thu Jul 20 23:59:54 +0000 2023	1682178549871181824	@Uki23 Elu kasih tau dong ke jokowi	0	0	0	0	in	1108606265284919296	1681677957537071105
2	Thu Jul 20 23:59:50 +0000 2023	1682178534046044160	@erickthohir memiliki 4 syarat jika dijadikan ...	0	0	7	8	in	326453024	1682178534046044160
3	Thu Jul 20 23:59:42 +0000 2023	1682178498637725697	Di Indonesia bisa begini gak yak? Ya minimal...	0	0	0	0	in	102019918	1682178498637725697
4	Thu Jul 20 23:59:39 +0000 2023	1682178487300546560	@KompasTV Artinya ada data double bosss Justr...	0	3	4	6	in	1088595784268824576	1681973959259209729
...	...	...	...	...	...	...	...	...	...	...
538	Thu Jul 20 22:49:56	1682160944947023872	ksian skli manusia manusia indonesia	0	0	0	0	in	1366062779606573057	1682160944947023872

Gambar 3. 3 Hasil Pengumulan Data Penelitian dengan Python

3.1.2. Preprocessing

Preprocessing merupakan suatu tahap untuk mempersiapkan data yang telah diperoleh dari tahap pengumpulan data sebelumnya sebelum data tersebut digunakan untuk tahap selanjutnya. Persiapan data yang dilakukan berupa membersihkan data dengan menghilangkan noise, menghapus data duplikat, memeriksa data untuk ketidakkonsistenan, dan memperbaiki kesalahan dalam data, seperti kesalahan ketik.

Pada tahap ini, data yang digunakan sebanyak 51.250 tweet berupa data text. Dan setelah melalui tahap tersebut, data menyusut menjadi 42.594 data tweet yang akan dibagi menjadi data testing dan data training dengan rasio 80% untuk data testing dan 20% data training dan didapatkan hasil. Seperti yang terlihat pada table 3.1.

Table 3. 1 Pembagian Data Testing dan Data Training

Jumlah Data (N)	Data Testing (N x 20%)	Data Training (N x 80%)
42594	8519	34075

a) Cleansing Data

Tahap Cleansing data dilakukan sebagai tahap awal yang sangat penting dalam penelitian ini. Hal ini dikarenakan data yang didapatkan dari Twitter masih dalam

bentuk teks yang tidak sesuai kaidah dan tidak lengkap. Normalisasi adalah proses penskalaan nilai atribut dari data sehingga terletak pada rentang tertentu[12].

Selain itu pada tahap normalisasi juga dilakukan secara Cleansing data menggunakan pemrograman Python, hal ini bertujuan untuk memastikan isi dari data tersebut. Cleaning data yang akan menghapus atribut-atribut yang tidak diperlukan di dalam data seperti, tanda baca, emot ikon, angka dan lain-lain. Atribut-atribut yang akan dihapus contohnya "@[A-Za-z0-9]+, [0-9]+, #, [^\w], '_', [\n]+, :, 'RT[\s]+, ^https?:\/\/.[\r\n], ^http?:\/\/.[\r\n]". Gambar 3.4 Cleansing Data dengan Python.

```
# Data Cleaning
def remove(tweet):
    tweet = re.sub(r'^RT[\s]+', '', tweet)
    tweet = re.sub(r'#', '', tweet)
    tweet = re.sub(r'http\S+', '', tweet)
    tweet = re.sub(r'[0-9]+', '', tweet)
    tweet = re.sub(r'\_', " ", tweet)
    tweet = re.sub("\W+", " ", tweet)
    tweet = re.sub(r'\\n', " ", tweet)
    tweet = emoji.get_emoji_regexp().sub(" ", tweet).strip()
    tweet = re.sub(r':', '', tweet)
    return tweet

def remove_whitesapce(tweet):
    return ' '.join(tweet.split())

def tokenizing(tweet):
    tokens = word_tokenize(str.remove(tweet))
    return ' '.join(tokens)
```

Gambar 3. 4 Cleansing Data dengan Python

Pada gambar 3.5 merupakan data yang dihasilkan setelah melalui tahan cleansing data.

	Text	clean_tweet	clean_bersih
0	? Panjat sosial\n? Panjat ranking FIFA\n\nSejak sanksi FIFA dicabut pada 2016, timnas Indonesia mencari lawan dari berbagai penjuru dunia untuk memperbaiki ranking FIFA. ??\n\n#PSSI #Indonesia #TahukahAnda #FIFAMatchday https://t.co/M3psaia90D	? Panjat sosial\n? Panjat ranking FIFA\n\nSejak sanksi FIFA dicabut pada 2016, timnas Indonesia mencari lawan dari berbagai penjuru dunia untuk memperbaiki ranking FIFA. ??\n\n#PSSI #Indonesia #TahukahAnda #FIFAMatchday https://t.co/M3psaia90D	Panjat sosial Panjat ranking FIFA Sejak sanksi FIFA dicabut pada timnas Indonesia mencari lawan dari berbagai penjuru dunia untuk memperbaiki ranking FIFA PSSI Indonesia TahukahAnda FIFAMatchday
1	PESAN BUYA YAHYA.\n.\nJika kita melihat kesalahan yang terjadi pada saudara kita, maka hendaklah kita melihat mereka dengan kasih sayang dan disertai dengan doa demi kebaikannya.\n.\n#buyayahya #ustazwadiannuar #kuliah #indonesia #malaysia #dakwah #islam #sunnah https://t.co/tRPDnNKxOj	PESAN BUYA YAHYA.\n.\nJika kita melihat kesalahan yang terjadi pada saudara kita, maka hendaklah kita melihat mereka dengan kasih sayang dan disertai dengan doa demi kebaikannya.\n.\n#buyayahya #ustazwadiannuar #kuliah #indonesia #malaysia #dakwah #islam #sunnah https://t.co/tRPDnNKxOj	PESAN BUYA YAHYA Jika kita melihat kesalahan yang terjadi pada saudara kita maka hendaklah kita melihat mereka dengan kasih sayang dan disertai dengan doa demi kebaikannya buyayahya ustazwadiannuar kuliah indonesia malaysia dakwah islam sunnah
2	Rakyat #Indonesia monitor ? https://t.co/gjczkWhrgV	Rakyat #Indonesia monitor ? https://t.co/gjczkWhrgV	Rakyat Indonesia monitor
3	Hanya dengan memiliki Akun MULTI PAY, anda bisa jalankan MULTI BISNIS ersebut diatas.\n#PeluangUsaha #PeluangBisnis #BisnisOnline #Usaha #UsahaOnline #Agent #Ticketing #Tour #Travel #Kurir #Ekspedisi #Umroh #Loket #IsiUlang #TopUp #BayarTagihan #UsahaMikro #UsahaKecil #Indonesia https://t.co/jvz8UuJLuL	Hanya dengan memiliki Akun MULTI PAY, anda bisa jalankan MULTI BISNIS ersebut diatas.\n#PeluangUsaha #PeluangBisnis #BisnisOnline #Usaha #UsahaOnline #Agent #Ticketing #Tour #Travel #Kurir #Ekspedisi #Umroh #Loket #IsiUlang #TopUp #BayarTagihan #UsahaMikro #UsahaKecil #Indonesia https://t.co/jvz8UuJLuL	Hanya dengan memiliki Akun MULTI PAY anda bisa jalankan MULTI BISNIS ersebut diatas PeluangUsaha PeluangBisnis BisnisOnline Usaha UsahaOnline Agent Ticketing Tour Travel Kurir Ekspedisi Umroh Loket IsiUlang TopUp BayarTagihan UsahaMikro UsahaKecil Indonesia

Gambar 3. 5 Hasil Cleansing Data dengan Python

b) Labeling Data

Labeling data adalah tahap memberikan label pada setiap data yang sudah dibersihkan atau telah melalui Cleansing data. Dalam penelitian ini data akan dilabeli melalui Natural Language Processing yang ada di Library Python yaitu Sastrawi. Proses Labeling melalui Sastrawi dilakukan dalam Bahasa Indonesia dimana data yang akan diberi label harus melalui tahap pemenggalan kata (stemming). Stemming sendiri merupakan proses menghilangkan imbuhan atau awalan kata sehingga hanya tersisa kata dasar. Penggunaan stemming ini dapat membantu dalam mengenali pola sentimen dalam kalimat tanpa harus memperhitungkan perbedaan kata berimbuhan. Proses Labeling juga dilakukan dengan berkonsultasi kepada ahli Bahasa yaitu Bapak Dr. Muhammad Sukirlan, M.A. selaku kepala UPT Bahasa di Unila, dan didapatkan hasil pada tabel 3.2.

Table 3. 2 Penjelasan Masing-Masing Label

No.	Sentimen	Penjelasan
1.	Negatif	Pernyataan yang sangat tidak mendukung terhadap konten yang dimaksudkan.
2.	Netral	Pernyataan yang tidak termasuk dalam mendukung dan setuju terhadap konten serta tidak termasuk kedalam pernyataan yang menentang atau tidak mendukung konten.
3.	Positif	Pernyataan yang sangat mendukung atau mengajak terhadap konten yang dimaksudkan.

Menurut Bapak Dr. Muhammad Sukirlan, M.A. labeling dilakukan dengan melihat kata kerja, kata bantu, kata sifat dan juga kata sambung di kalimatnya, contoh beberapa kata yang bisa diindikasikan sebagai kata yang negatif atau positif pada tabel 3.3.

Table 3. 3 Contoh Kata Negatif dan Kata Negatif

Positif	**Negatif**
Senang	Benci
Bahagia	Bodoh
Hebat	Catat
Sukses	Tolol
Terbaik	Gila
Suka	Buruk
Bagus	Jelek

Beliau juga mengatakan bahwa jika ada kata Positif namun diawali dengan kata “Tidak” maka kata tersebut akan berlabel Negatif, begitu juga sebaliknya jika ada kata Negatif yang diawali dengan kata “Tidak” maka kata tersebut akan berlabel Positif, sedangkan untuk kata Netral beliau mengatakan bahwa kalimatnya akan mengandung keduanya atau tidak menagandung keduanya dan juga bisa berupa pertaanyaan. Seperti pada tabel 3.4.

Table 3. 4 Cara Melabeli Kalimat

Kalimat	Sentimen
Senang sekali membaca tulisan dari tentang RKHUP dan perjalanannya dari rumah ke rumah seperti lagunya	Positif
Harkristuti Harkrisnowo berpendapat bahwa KUHP yang diterapkan sekarang sudah terlalu tua dan bukan merupakan buatan Indonesia	Negatif
Mana suaranya yang sepakat mempertemukan dan untuk debat RKUHP	Netral

c) Pembobotan Kata

Pada tahap pembobotan kata atau Terms Weight ini akan dilakukan perhitungan secara otomatis menggunakan python. Namun pada kali ini akan memberikan contoh dan Contoh akan dibagi menjadi 4 dimana 3 adalah data training dan 1 adalah data testing. Berikut adalah perhitungan manual dari pembobotan kata yang ada pada tabel 3.5.

Table 3. 5 Contoh Data Pembobotan Kata

Ket	Tweet	Sentimen
D1	gagalkan rkuhp uu kolonial gaya oligarki hentikan periode oligarki	Negatif
D2	hak publik untuk bersuara dan ikut terlibat dalam proses perumusan rkuhp ini	Positif
D3	senang sekali membaca tulisan dari tentang rkhup dan perjalanannya dari rumah ke rumah seperti lagunya	Positif
D4	rkuhp memudahkan masyarakat mencari pekerjaan tapi banyak pasal karet	?

Selanjutnya melakukan tokenisasi dan lakukan *stopwords*, tokenisasi dilakukan untuk memisahkan kalimat menjadi kata-kata dan *stopwords* digunakan untuk membuang kata-kata yang dirasa memiliki makna yang kurang berarti seperti: “yang”, “dan”, “atau” dan lain-lain. Berikut hasil dari pembobotan pada tabel 3.6.

Table 3. 6 Data Pembobotan Kata Tf-Idf

	Tf					Idf	Wt = Tf.Idf			
Term	**D1**	**D2**	**D3**	**D4**	**df**	**log(n/df)**	**D1**	**D2**	**D3**	**D4**
gagalkan	1	0	0	0	1	0,60205999	0,60205999	0	0	0
rkuhp	1	1	1	1	4	0	0	0	0	0
uu	1	0	0	0	1	0,60205999	0,60205999	0	0	0
kolonial	1	0	0	0	1	0,60205999	0,60205999	0	0	0
gaya	1	0	0	0	1	0,60205999	0,60205999	0	0	0
oligarki	2	0	0	0	2	0,30103	0,60205999	0	0	0
hentikan	1	0	0	0	1	0,60205999	0,60205999	0	0	0
periode	1	0	0	0	1	0,60205999	0,60205999	0	0	0
hak	0	1	0	0	1	0,60205999	0	0,60205999	0	0
publik	0	1	0	0	1	0,60205999	0	0,60205999	0	0
bersuara	0	1	0	0	1	0,60205999	0	0,60205999	0	0
ikut	0	1	0	0	1	0,60205999	0	0,60205999	0	0
terlibat	0	1	0	0	1	0,60205999	0	0,60205999	0	0
proses	0	1	0	0	1	0,60205999	0	0,60205999	0	0
perumusan	0	1	0	0	1	0,60205999	0	0,60205999	0	0
senang	0	0	1	0	1	0,60205999	0	0	0,60205999	0
membaca	0	0	1	0	1	0,60205999	0	0	0,60205999	0
tulisan	0	0	1	0	1	0,60205999	0	0	0,60205999	0
perjalanannya	0	0	1	0	1	0,60205999	0	0	0,60205999	0
rumah	0	0	1	0	1	0,60205999	0	0	0,60205999	0
lagunya	0	0	1	0	1	0,60205999	0	0	0,60205999	0
memudahkan	0	0	0	1	1	0,60205999	0	0	0	0,60205999
masyarakat	0	0	0	1	1	0,60205999	0	0	0	0,60205999
mencari	0	0	0	1	1	0,60205999	0	0	0	0,60205999
pekerjaan	0	0	0	1	1	0,60205999	0	0	0	0,60205999
pasal	0	0	0	1	1	0,60205999	0	0	0	0,60205999
karet	0	0	0	1	1	0,60205999	0	0	0	0,60205999

Kata “gagalkan” memiliki df (*document frequency)* sebanyak 1 dari 4 dokumen yang ada. Jadi dalam menentukan Idf (*inverse document frequency*) adalah:

Idf = log (4/1)

Idf = log (4)

Idf = 0,60205999

Setelah mendapatkan nilai Idf, dapat dilanjutkan ketahap selanjutnya dengan mengkalikan Idf dan Tf (*Terms frequency*), dengan nilai Tf “gagalkan” di masing-masing dokumen D1 = 1, D2 = 0, D3=0, D4=0, maka dapat dihasilkan Wt (*Weight terms*) sebagai berikut:

Wt = Tf . Idf

D1 =1

Wt = 1. 0,60205999

Wt = 0,60205999

D1 =2

Wt = 0. 0,60205999

Wt = 0

D1 =3

Wt = 0. 0,60205999

Wt = 0

D1 =5

Wt = 0. 0,60205999

Wt = 0

Data hasil dari pembobotan kata akan digunakan untuk mencari *Similiarity* atau persamaan menggunakan metode yang nanti digunakan.

3.1.3. Analisis Data

Tahap analisis merupakan suatu prosedur atau proses sistematis pengombinasian metode atau teknik untuk menentukan data yang sesui dan cara terbaik untuk memanfaatkannya. Tahap analisis data dilakukan dari awal sampai akhir adalah membandingkan hasil eksperimen. Data yang telah dikumpulkan akan dieksperimen dengan melakukan cleaning dan normalisasi. Selanjutnya tahapan yang dilakukan dengan transformasi data pengklusteran kemudian dibagi beberapa

kelompok data atau group. Setelah data sudah dikelompokkan maka selanjutnya akan dilakukan pemodelan data yang diubah menjadi nilai-nilai yang telah dipisahkan dengan bentuk pelebelan data, atau Comma Sparated Value (CSV) sebagai format data input ke dalam *Machine Learning*.

Setelah proses memasukkan data atau pemodelan data ke dalam suatu database selesai, selanjutnya proses yang akan dilakukan yaitu menerapkan model algoritma metode *Support Vector Machine, Logistic Regression, K-Nearest Neighbor* dan *Decision Tree*. Berdasarkan evaluasi dari model algoritma atau metode yang digunakan dapat dilakukan penarikan kesimpulan menggunakan pengukuran dari nilai performa akurasi, presisi, recall dan waktu. Hasil tersebut di jadikan sebagai acuan atau pedoman untuk pengukuran dalam klasifikasi pada penelitian ini.

3.1.4. Eksperimen dan Pengujian Metode

Tahap ini akan menjelaskan eksperimen dan teknik pengujian yang akan digunakan untuk menguji dan mengevaluasi suatu metode atau teknik dengan menggunakan data yang telah dikumpulkan dari observasi atau percobaan. Eksperimen dan pengujian metode merupakan bagian penting dari proses penelitian, karena dengan melakukannya maka akan tercipta suatu metode yang dapat digunakan secara efektif dan efisien untuk menyelesaikan suatu masalah.

Eksperimen dan pengujian metode juga bertujuan untuk mengetahui kelebihan dan kekurangan suatu metode atau teknik, serta untuk menemukan cara-cara baru yang lebih baik dalam menyelesaikan masalah. Oleh karena itu, eksperimen dan pengujian metode sangat penting dilakukan agar dapat memperoleh hasil yang bermutu dan dapat dipertanggungjawabkan.

a. Logistic Regression

Metode *Logistic Regression* menggunakan data historis untuk memprediksi probabilitas bahwa suatu peristiwa akan terjadi berdasarkan beberapa variabel input (fitur).

Table 3. 7 Kamus Kata (Representasi Biner)

Kata	terima	kasih	indonesia	rakyat	monitor	...	Bias
terima	1	0	0	0	0	...	1
kasih	0	1	0	0	0	...	1
indonesia	0	0	1	0	0	...	1
rakyat	0	0	0	1	1	...	1
monitor	0	0	0	0	1	...	1
mahfud	0	0	1	0	0	...	1
bidadari	0	0	1	0	0	...	1
masuk	0	0	1	0	0	...	1
sistem	0	0	1	0	0	...	1
setan	0	0	1	0	0	...	1
selamat	0	0	1	0	0	...	1
siang	0	0	1	0	0	...	1
ayo	0	0	1	0	0	...	1
tunjuk	0	0	1	0	0	...	1
senyum	0	0	1	0	0	...	1
indah	0	0	1	0	0	...	1
dunia	0	0	1	0	0	...	1
baru	0	0	1	0	0	...	1
news	0	0	1	0	0	...	1
korupsi	0	0	1	0	0	...	1
sejarah	0	0	1	0	0	...	1
nama	0	0	1	0	0	...	1

Representasi data biner pada tabel 3.7 berasal dari beberapa data text yang diubah dalam konsep biner. Hal guna memungkinkan model untuk dapat menghubungkan kata-kata dalam teks dengan bobot (koefisien) yang akan dihitung selama pelatihan. Untuk menghitung skor total (z), bobot akan dikalikan dengan representasi biner kata-kata dalam teks, dan akhirnya, kita dapat memprediksi sentimen berdasarkan probabilitas yang dihasilkan.

Table 3. 8 Inisialisasi Bobot (Koefisien)

Kata	Bobot
terima	-0.659
kasih	0.245
indonesia	0.772
rakyat	-0.123
monitor	0.418
mahfud	-0.512
bidadari	0.667
masuk	0.834
sistem	-0.398
setan	-0.943
selamat	0.032
siang	0.752
ayo	0.215
tunjuk	-0.709
senyum	0.579
indah	0.948
dunia	-0.279
baru	0.127
news	-0.016
korupsi	0.841
sejarah	-0.874
nama	0.355

Misalkan kita ingin menghitung skor total (z) untuk teks "terima kasih indonesia". Untuk suatu teks yang memiliki n fitur (kata-kata), skor total z dihitung sebagai jumlah dari perkalian bobot (koefisien) ω_i dengan representasi biner x_i dari setiap fitur. Contoh perhitungan manualnya:

Teks: terima kasih Indonesia

Tokenisasi: [terima, kasih, indonesia]

Skor Total (z) = (Bobot_terima * Biner_terima) + (Bobot_kasih *

Biner_kasih) + (Bobot_indonesia * Biner_indonesia)

= (-0.659 * 1) + (0.245 * 1) + (0.772 * 1)

= -0.659 + 0.245 + 0.772

= 0.358

Selanjutnya, kita hitung probabilitas (p) untuk setiap teks dalam dataset menggunakan fungsi sigmoid pada metode Regresi Logistik:

1) Untuk teks "terima kasih indonesia":

$$z = (-0{,}659\ x\ 1) + (0{,}245\ x\ 1) + (0{,}772\ x\ 1) = 0{,}358$$

$$p = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-0{,}358}} = 0{,}588$$

2) Untuk teks "rakyat indonesia monitor":

$$z = (-0{,}123\ x\ 1) + (0{,}418\ x\ 1) + (0{,}772\ x\ 1) = 1{,}067$$

$$p = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-0{,}358}} = 0{,}255$$

3) Untuk teks "mahfud bidadari masuk sistem indonesia setan":

$$z = (-0{,}512\ x\ 1) + (0{,}667\ x\ 1) + (0{,}834\ x\ 1)\ +$$
$$(-0{,}398\ x\ 1)\ + (-0{,}943\ x\ 1) = -0{,}352$$

$$p = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{0{,}352}} = 0{,}412$$

4) Untuk teks "selamat siang indonesia ayo tunjuk senyum indah dunia":

$$z = (\ 0{,}032\ x\ 1) + (0{,}752\ x\ 1) + (0{,}215\ x\ 1) +$$
$$(-0{,}709\ x\ 1)\ + (0{,}579\ x\ 1) + (0{,}948\ x\ 1) +$$
$$(-0{,}279\ x\ 1)\ = 2{,}538$$

$$p = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-2{,}538}} = 0{,}926$$

5) Untuk teks "baru news indonesia korupsi sejarah nama indonesia indonesia":

$$z = (-0{,}127\ x\ 1) + (-0{,}016\ x\ 1) + (0{,}772\ x\ 1)\ +$$
$$(0{,}841\ x\ 1)\ + (-0{,}874\ x\ 1)\ + (0{,}355\ x\ 1)\ +$$
$$(0{,}772\ x\ 1)\ + (0{,}841\ x\ 1) = 3{,}978$$

$$p = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-3{,}978}} = 0{,}981$$

Probabilitas ini mewakili kemungkinan sentimen positif (jika p > 0,5) atau sentimen negatif (jika $p \leq 0{,}5$) ntuk setiap teks dalam dataset.

b. Support Vector Machine

Metode *Support Vector Machine* adalah metode yang umumnya digunakan untuk klasifikasi biner, termasuk Sentimen Analisis, dengan membangun hyperplane yang memisahkan dua kelas data dengan margin maksimal. Berikut ini adalah hitungan manualnya:

Table 3. 9 Konversi Teks ke Fitur Numerik

Kalimat	Fitur 1 (terima)	Fitur 2 (kasih)	Fitur 3 (indonesia)	Fitur 4 (rakyat)	Fitur 5 (monitor)	Fitur 6 (mahfud)	Fitur 7 (bidadari)	Fitur 8 (masuk)	Fitur 9 (sistem)
terima kasih indonesia	1	1	1	0	0	0	0	0	0
rakyat indonesia monitor	0	0	1	1	1	0	0	0	0
mahfud bidadari masuk sistem indonesia setan	0	0	1	0	0	1	1	1	1

Dengan menggunakan representasi sederhana berdasarkan jumlah kata-kata dalam setiap kalimat. Kata-kata unik akan dihitung dalam seluruh dataset yang dapat dilihat pada table 3.9. Setelah data diubah kedalam bentuk biner, dapat kita lakukan penggantian label sentimen dengan nilai target numerik (Positif: 1, Netral: 0, Negatif: -1) seperti pada tabel 3.10.

Table 3. 10 Pelabelan Sentimen sebagai Nilai Target

Kalimat	Sentimen
terima kasih indonesia	1
rakyat indonesia monitor	0
mahfud bidadari masuk sistem indonesia setan	-1

Setelah didapatkan Nilai Target, Kita dapat menghitung matriks Hessian, yang berisi hasil perkalian dalam dari setiap data dari dataset. Setiap elemen dalam Hessian merupakan perkalian dot antara vektor fitur dari data yang ada. Berikut ini adalah hitungan manualnya:

- Produk Dot untuk terima kasih indonesia dan terima kasih indonesia:

(1 * 1) + (1 * 1) + (1 * 1) + (0 * 0) + (0 * 0) + (0 * 0) + (0 * 0) + (0 * 0) + (0 * 0) + (0 * 0) = 3

- Produk Dot untuk terima kasih indonesia dan rakyat indonesia monitor:
  (1 * 0) + (1 * 0) + (1 * 1) + (0 * 1) + (0 * 1) + (0 * 0) + (0 * 0) + (0 * 0) + (0 * 0) + (0 * 0) = 1
- Produk Dot untuk terima kasih indonesia dan mahfud bidadari masuk sistem indonesia setan:
  (1 * 0) + (1 * 0) + (1 * 1) + (0 * 0) + (0 * 0) + (0 * 1) + (0 * 1) + (0 * 1) + (0 * 1) + (0 * 1) = 1
- Produk Dot untuk rakyat indonesia monitor dan rakyat indonesia monitor:
  (0 * 0) + (0 * 0) + (1 * 1) + (1 * 1) + (1 * 1) + (0 * 0) + (0 * 0) + (0 * 0) + (0 * 0) + (0 * 0) = 3
- Produk Dot untuk rakyat indonesia monitor dan mahfud bidadari masuk sistem indonesia setan:
  (0 * 0) + (0 * 0) + (1 * 1) + (1 * 0) + (1 * 0) + (0 * 1) + (0 * 1) + (0 * 1) + (0 * 1) + (0 * 1) = 2
- Produk Dot untuk mahfud bidadari masuk sistem indonesia setan dan mahfud bidadari masuk sistem indonesia setan:
  (0 * 0) + (0 * 0) + (1 * 1) + (0 * 0) + (0 * 0) + (1 * 1) + (1 * 1) + (1 * 1) + (1 * 1) + (1 * 1) = 6

Dari seluruh perhitungan yang telah dilakukan dengan perkalian dot antara vektor fitur dari data yang ada. Akan menghasilkan data matriks seperti pada tabel 3.11.

Table 3. 11 Matriks Hessian

[3	1	2	0	0	0	0	0	0	0]
[1	4	3	3	3	1	1	1	1	1]
[2	3	10	0	0	3	3	3	3	3]
[0	3	0	6	3	0	0	0	0	0]
[0	3	0	3	6	0	0	0	0	0]
[0	1	3	0	0	6	3	3	3	3]
[0	1	3	0	0	3	6	3	3	3]
[0	1	3	0	0	3	3	6	3	3]
[0	1	3	0	0	3	3	3	6	3]
[0	1	3	0	0	3	3	3	3	6]

Untuk Menghitung Vektor Bobot (w) membuthkan invers dari matriks Hessian, supaya dapat menghitung vektor bobot (w) dengan mengalikan invers Hessian dengan vektor kolom yang berisi 1 (menandakan sentimen dari setiap contoh). Dan tabel 3.12 merupakan tebel invers dari matriks Hessian.

Table 3. 12 Matriks Invers Hessian

[10954	-0.2472	-0.0347	-0.0658	-0.0625	-0.0658	-0.0658	-0.0658	-0.0658	-0.0658]
[-0.2472	0.9035	-0.1963	0.0658	0.0625	0.0658	0.0658	0.0658	0.0658	0.0658]
[-0.0347	-0.1963	0.6422	0.1082	0.1028	0.1082	0.1082	0.1082	0.1082	0.1082]
[-0.0658	0.0658	0.1082	0.6098	-0.0417	-0.0434	-0.0434	-0.0434	-0.0434	-0.0434]
[-0.0625	0.0625	0.1028	-0.0417	0.6597	-0.0407	-0.0407	-0.0407	-0.0407	-0.0407]
[-0.0658	0.0658	0.1082	-0.0434	-0.0407	0.6711	-0.0434	-0.0434	-0.0434	-0.0434]
[-0.0658	0.0658	0.1082	-0.0434	-0.0407	-0.0434	0.6711	-0.0434	-0.0434	-0.0434]
[-0.0658	0.0658	0.1082	-0.0434	-0.0407	-0.0434	-0.0434	0.6711	-0.0434	-0.0434]
[-0.0658	0.0658	0.1082	-0.0434	-0.0407	-0.0434	-0.0434	-0.0434	0.6711	-0.0434]
[-0.0658	0.0658	0.1082	-0.0434	-0.0407	-0.0434	-0.0434	-0.0434	-0.0434	0.6711]

Menghitung nilai-nilai yang ada pada setiap elemen vektor bobot (w) menggunakan nilai-nilai matriks invers Hessian dan vektor sentimen. Berikut ini adalah hitungan manualnya:

w = Invers Hessian x Vektor Sentimen

w =

[(1.0954 * 1) + (-0.2472 * 0) + (-0.0347 * -1)]

[(-0.2472 * 1) + (0.9035 * 0) + (-0.1963 * -1)]

[(-0.0347 * 1) + (-0.1963 * 0) + (0.6422 * -1)]

[(-0.0658 * 1) + (0.0658 * 0) + (0.1082 * -1)]

[(-0.0625 * 1) + (0.0625 * 0) + (0.1028 * -1)]

[(-0.0658 * 1) + (0.0658 * 0) + (0.1082 * -1)]

[(-0.0658 * 1) + (0.0658 * 0) + (0.1082 * -1)]

[(-0.0658 * 1) + (0.0658 * 0) + (0.1082 * -1)]

[(-0.0658 * 1) + (0.0658 * 0) + (0.1082 * -1)]

[(-0.0658 * 1) + (0.0658 * 0) + (0.1082 * -1)]

w =

[-0.1919]

[0.3371]

[-0.4155]

[-0.1572]

[-0.1295]

[-0.2216]

[-0.2216]

[-0.2216]

[-0.2216]

[-0.2216]

w = [-0.1919, 0.3371, -0.4155, -0.1572, -0.1295, -0.2216, -0.2216, -0.2216, -0.2216, -0.2216]

Setelah mendapatkan nilai Vektor Bobot (w). Kita mencari nilai Konstanta (b) dimana dapat dihitung menggunakan salah satu dari titik data pelatihan yang berada pada batas margin (yaitu, memiliki nilai fungsi keputusan yang sama dengan 1 - jika positif - atau -1 - jika negatif).

Disini kita akan menggunakan data pelatihan pertama [terima kasih indonesia] dengan sentimen positif [1], kita dapat menghitung konstanta (b) dengan mengambil invers dari fungsi keputusan (f(x)) dan menggantikan dengan nilai vektor sentimen (y) dan vektor bobot (w):

Konstanta (b) = y - w * x

Dengan x = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0] (vectors sentimen pertama)

Konstanta (b) = 1 - [-0.1919, 0.3371, -0.4155, -0.1572, -0.1295, -0.2216, -0.2216, -0.2216, -0.2216, -0.2216] * [1, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Konstanta (b) = 1 - (-0.1919 * 1 + 0.3371 * 0 + -0.4155 * 0 + -0.1572 * 0 + -0.1295 * 0 + -0.2216 * 0 + -0.2216 * 0 + -0.2216 * 0 + -0.2216 * 0 + -0.2216 * 0)

Konstanta (b) = 1 + 0.1919 = 1.1919

Setelah mendapatkan nlai Vektor Bobot (w) dan Konstanta (b) Kita akan menghitung nilai prediksi (f) untuk setiap data yang ada. Berikut ini adalah hitungan manualnya:

- Data 1: [terima, kasih, indonesia, [Positif]]

  Nilai Prediksi = (terima * -0.1919) + (kasih * 0.3371) +
  (indonesia * -0.4155) + (-0.2216)
  = (1 * -0.1919) + (1 * 0.3371) + (1 * -0.4155) + (-0.2216)
  = -0.1919 + 0.3371 - 0.4155 - 0.2216

= -0.4919

- Data 2: [rakyat, indonesia, monitor, [Netral]]

  Nilai Prediksi = (rakyat * -0.1919) + (indonesia * 0.3371) +
  (monitor * -0.4155) + (-0.2216)
  = (1 * -0.1919) + (1 * 0.3371) + (1 * -0.4155) + (-0.2216)
  = -0.1919 + 0.3371 - 0.4155 - 0.2216
  = -0.4919

- Data 3: [mahfud, bidadari, masuk, sistem, indonesia, setan, [Negatif]]

  Nilai Prediksi = (mahfud * -0.1919) + (bidadari * 0.3371) +
  (masuk * -0.4155) + (sistem * -0.1572) +
  (indonesia * -0.1295) + (setan * -0.2216) + (-0.2216)
  = (1 * -0.1919) + (1 * 0.3371) + (1 * -0.4155) +
  (1 * -0.1572) + (1 * -0.1295) + (1 * -0.2216) + (-0.2216)
  = -0.1919 + 0.3371 - 0.4155 - 0.1572 - 0.1295 – 0.2216
  - 0.2216
  = -0.6852

Berdasarkan hasil nilai prediksi yang dihasilkan, semua data tampaknya cenderung memiliki sentimen yang lebih mendekati negatif.

c. K-Nearest Neighbor

Metode *K-Nearest Neighbor* menggunakan persamaan *Similiarity Vector* untuk mencari jarak antar data atau dokumen. Persamaan Similarity Vector dalam K-Nearest Neighbor (K-NN) digunakan untuk mengukur sejauh mana dua data atau dokumen mirip satu sama lain. Salah satu metode yang umum digunakan adalah Cosine Similarity, yang mengukur sudut antara dua vektor fitur yang mewakili data atau dokumen. Dengan menggunakan Cosine Similarity, kita dapat menentukan seberapa mirip atau berbedanya dua entitas dengan membandingkan orientasi vektor fitur mereka. Tabel 3.13 merupakan contoh hasil perkalian scalar vektor.

Table 3. 13 Perkalian Skalar Vektor

WD4.Wdi		
D1	**D2**	**D3**
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0

Perkalian Sklar yang dimaksudkan untuk mecari nilai $\sum_k(\boldsymbol{dik.dij})$ pada rumus similiarity, dimana dik adalah bobot data dokumen *Testing* (D1, D2, D3) dij adalah bobot data dokumen *Training* (D4). Perkalian ini pasangkan untuk masing-masing nilai *Weight Terms* atau bobot kata, contoh perhitungan manualnya:

D1: "gagalkan" = 0,60205999

D2: "gagalkan" = 0

D3: "gagalkan" = 0

D4: "gagalkan" = 0

$$\sum_{\boldsymbol{k}}(\boldsymbol{dik.dij})$$

D1 = bobot D1 . bobot D5

D1 = 0,60205999 . 0

D1 = 0

D2 = bobot D3 . bobot D5

D2 = 0 . 0

D2 = 0

D3 = bobot D3 . bobot D5

D3 = 0 . 0

D3 = 0

Selanjutnya adalah menghitung vector Panjang setiap dokumen dengan rumus:

$$\sqrt{\sum_k d^2ik}\ \sqrt{\sum_k d^2jk}$$

Table 3. 14 Perkalian Panjang Vektor

Panjang Vektor			
D1	**D2**	**D3**	**D4**
0,36247623	0	0	0
0	0	0	0
0,36247623	0	0	0
0,36247623	0	0	0

Panjang Vektor			
0,36247623	0	0	0
0,36247623	0	0	0
0,36247623	0	0	0
0,36247623	0	0	0
0	0,36247623	0	0
0	0,36247623	0	0
0	0,36247623	0	0
0	0,36247623	0	0
0	0,36247623	0	0
0	0,36247623	0	0
0	0,36247623	0	0
0	0,36247623	0	0
0	0	0,36247623	0
0	0	0,36247623	0
0	0	0,36247623	0
0	0	0,36247623	0
0	0	0,36247623	0
0	0	0,36247623	0
0	0	0,36247623	0
0	0	0	0,36247623
0	0	0	0,36247623
0	0	0	0,36247623
0	0	0	0,36247623
0	0	0	0,36247623
0	0	0	0,36247623
2,53733362	2,89980985	2,53733362	2,17485738
1,59290100	1,70288280	1,59290100	1,47473976

Panjang vektor:

D1 : $\sqrt{0{,}60205999^2 \; x \; 7}$

D1 : $\sqrt{2{,}53733362}$

D1 : 1,59290100

D2 : $\sqrt{0{,}60205999^2 \; x \; 8}$

D2 : $\sqrt{2{,}89980985}$

D2 : 1,70288280

D3 : $\sqrt{0{,}60205999^2 \; x \; 7}$

D3 : $\sqrt{2{,}53733362}$

D3 : 1,59290100

D4 : $\sqrt{0{,}60205999^2 \; x \; 6}$

D4 : $\sqrt{2{,}17485738}$

D4 : 1,47473976

Jika semua sudah didapatkan maka data tersebut dapat dimasukan kedalam rumus yang utuh:

Cos(D4,D1) = 0 / (1,47473976*1,59290100) = 0

Cos(D4,D2) = 0 / (1,47473976*1,70288280) = 0

Cos(D4,D3) = 0 / (1,47473976*1,59290100) = 0

Nilai kedekatan vector yang dimiliki oleh D1, D2, D3 terhadap D4 memiliki nilai yang sama yaitu 0, maka dapat disimpulkan bahwa D4 bukan termasuk dalam

sentiment “Negatif” maupun “Positif” atau D4 termasuk dalam sentiment “Negatif” maupun “Positif” yang mana itu berarti D4 adalah “Netral”.

d. Decision Tree

Metode *Decision Tree* menggunakan *Information Gain* (IG) untuk pemilihan fitur dalam analisis klasifikasi. IG mengukur seberapa banyak informasi yang dapat kita peroleh dari sebuah fitur dalam memprediksi atau mengklasifikasikan data. Berikut ini adalah hitungan manualnya:

Table 3. 15 Sempel Dataset

No	Teks Ulasan	Sentimen
1	terima kasih indonesia	Positif
2	rakyat indonesia monitor	Netral
3	mahfud bidadari masuk sistem indonesia setan	Negatif
4	selamat siang indonesia ayo tunjuk senyum indah dunia	Positif
5	baru berita korupsi sejarah nama	Negatif

Dari sempel dataset yang digunakan pada tebel 3.15. Kita dapat menghitung nilai Entropi awal dari dataset. Dimana kita dapat ketahui:

- Jumlah total data : 5
- Jumlah data positif : 2
- Jumlah data netral : 1
- Jumlah data negate : 2

Entropi awal (Entropy Before Split):

Entropy(S) = -p(positif) * log2(p(positif)) - p(netral) * log2(p(netral))

-p(negatif) * log2(p(negatif))

= -(2/5) * log2(2/5) - (1/5) * log2(1/5) - (2/5) * log2(2/5)

= 1.521

Setelah mendapakan nilai Entropi awal, kta akan menghitung information gain untuk setiap fitur. Dimana, kita akan mempertimbangkan fitur untuk membagi data menjadi dua berdasarkan "Teks Ulasan", yaitu: Teks Ulasan mengandung kata "indonesia" dan Teks Ulasan tidak mengandung kata "indonesia". Berikut ini adalah hitungan manualnya:

1) Kelompok 1 (Teks Ulasan mengandung kata "indonesia"):

Table 3. 16 Dataset Teks Ulasan Mengandung Kata "Indonesia"

No	Teks Ulasan	Sentimen
1	terima kasih indonesia	Positif
2	rakyat indonesia monitor	Netral
3	mahfud bidadari masuk sistem indonesia setan	Negatif
4	selamat siang indonesia ayo tunjuk senyum indah dunia	Positif

Bersarkan dataset yang digunakan pada tebel 3.16. Kita dapat menghitung nilai Entropinya. Dimana kita dapat ketahui:

- Jumlah data positif: 2, sehingga p(positif) = 2/4
- Jumlah data netral: 1, sehingga p(netral) = 1/4
- Jumlah data negatif: 1, sehingga p(negatif) = 1/4

Entropy (Kelompok 1)

Entropy = - (2/4) * log2(2/4) - (1/4) * log2(1/4) - (1/4) * log2(1/4)

= 1.5

2) Kelompok 2 (Teks Ulasan tidak mengandung kata "indonesia"):

Table 3. 17 Dataset Teks Ulasan Tidak Mengandung Kata "Indonesia"

No	Teks Ulasan	Sentimen
5	baru berita korupsi sejarah nama	Negatif

Dan dataset yang digunakan pada tebel 3.17. Kita dapat menghitung nilai Entropinya. Dimana kita dapat ketahui:

- Jumlah data negatif: 1, sehingga p(negatif) = 1/1 Entropy (Kelompok 2)
  Entropy = 0

Entropi kelompok 2 adalah 0, karena hanya terdapat satu kelas sentimen.

Setelah kita mendapatkan nilai entropi dari masing-masing kelompok, kita dapat melakukan perhitungan unutuk mencari nilai Information Gain. Dimana, Information Gain (IG) dihitung sebagai selisih antara entropi sebelum pemisahan dan jumlah weighted average entropi setiap kelompok setelah pemisahan. Berikut perhitungannya:

IG(Teks Ulasan) = Entropy(S) - [p(Kelompok 1) *

Entropy(Kelompok 1) + p(Kelompok 2) *

Entropy(Kelompok 2)

= 1.521 - [(4/5) * 1.5 + (1/5) * 0]

= 0.171

Dengan mengurangkan entropi awal dari jumlah rata-rata tertimbang dari entropi setelah pemisahan dalam kelompok-kelompok berdasarkan fitur. Dan menghitung Information Gain untuk fitur "Teks Ulasan" (mengandung kata "indonesia" atau tidak) dan menemukan IG(Teks Ulasan) didapatkan nilai IG sebesar 0.171.

3.1.5. Evaluasi dan Validasi Hasil

Tahap ini akan menyimpulkan dari hasil evaluasi dari eksperimen yang dilakukan dari dataset yang telah dikumpulkan. Evaluasi dan validasi hasil merupakan proses yang digunakan untuk mengevaluasi dan memvalidasi hasil suatu proyek atau kegiatan. Evaluasi adalah proses menilai keberhasilan suatu proyek atau kegiatan dengan mengukur sejauh mana tujuan yang telah ditetapkan tercapai. Validasi adalah proses menguji kebenaran atau keabsahan suatu hasil.

Evaluasi dan validasi hasil sangat penting dilakukan untuk memastikan bahwa proyek atau kegiatan yang dilakukan telah sesuai dengan tujuan yang telah ditetapkan dan hasil yang diperoleh dapat dipertanggungjawabkan. Dengan melakukan evaluasi dan validasi hasil, maka akan tercipta suatu proyek atau kegiatan yang berkualitas dan dapat memberikan manfaat yang optimal bagi yang bersangkutan. Adapun yang diterapkan pada kasus ini berupa;

1. Evaluasi Opini: Metode ini digunakan untuk mengevaluasi opini yang diberikan dalam sebuah teks. Ini melibatkan identifikasi pandangan atau pendapat yang dinyatakan dalam sebuah teks.
2. Evaluasi Performa: Metode ini digunakan untuk mengevaluasi kinerja model sentimen analisis. Ini melibatkan pengukuran akurasi, presisi, recall, F-measure, dan lain-lain.

F-measure adalah salah satu metode evaluasi performa untuk model klasifikasi, termasuk dalam sentimen analisis. Metode ini menggabungkan precision (presisi) dan recall (ingatan) ke dalam satu nilai yang merefleksikan performa keseluruhan model.

Precision mengukur seberapa banyak data yang diidentifikasi sebagai positif oleh model benar-benar positif. Recall mengukur seberapa banyak data positif yang diidentifikasi oleh model. F-measure menggabungkan precision dan recall menjadi satu nilai dengan memperhitungkan kekurangan masing-masing metrik.

F-measure adalah harmonic mean dari precision dan recall, dinyatakan sebagai:

$$F_{-measure} = 2\ x \frac{(Precision\ x\ Recall)}{(Precision + Recall)} \quad (1)$$

$$F1_{avg} = \frac{\sum_{i=1}^{n} \frac{2xPixRi}{Pi + Ri}}{n} \quad (2)$$

Nilai F-measure berkisar dari 0 hingga 1, di mana nilai 1 menunjukkan performa yang sempurna. Metode ini sering digunakan dalam penelitian sentimen analisis untuk mengevaluasi kinerja model dalam mengidentifikasi tipe sentimen dari sebuah teks[23].

3.2. Alat dan Bahan

Penelitian ini, menggunakan perangkat keras Laptop ASUS VivoBook dengan Intel® Core™ i5 8250U Processor (6M Cache, up to 3.40 GHz), RAM 8,00 GB, sedangkan perangkat lunak yang digunakan Microsoft Excel, Python versi 3 dan Google Colab untuk pengolahan data. Sedangkan bahan yang digunakan diambil dari hasil *crawling* data Twitter dengan tag Indonesia.