Jurnal Teknik Informatika Vol. 14 No. 2, Oktober 2021
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v14i2.24024

JURNAL TEKNIK INFORMATIKA

Homepage : http://journal.uinjkt.ac.id/index.php/ti

ANALISIS SENTIMEN MASYARAKAT TERHADAP VAKSINASI COVID-19 BERDASARKAN OPINI PADA TWITTER MENGGUNAKAN ALGORITMA NAIVE BAYES

Zulfikar Firmansyah[1], Nila Feby Puspitasari[2]

[1] Program Studi Informatika, Fakultas Ilmu Komputer
[2] Program Studi Teknik Informatika, Fakultas Ilmu Komputer
[1,2]Universitas AMIKOM Yogyakarta
[1,2]Jl Ringroad Utara, Condongcatur, Depok, Sleman, Yogyakarta Indonesia 55283
E-mail: [1]zulfikar.firmansyah@students.amikom.ac.id, [2]nilafeby@amikom.ac.id

Artikel:
Diterima: 07 Januari, 2022
Direvisi: 17 Januari, 2022
Diterbitkan: 17 Januari, 2022

***Alamat Korespondensi:**
zulfikar.firmansyah@students.amikom.ac.id

ABSTRACT

Corona virus is a group of viruses that infect the respiratory tract. This virus is known as Covid-19 which is known to have originated from China, which appeared in December 2019. In early March 2020, the first time the Covid-19 virus was reported to have entered Indonesia and spread to all provinces in Indonesia. The steps taken by the government to prevent the spread of the virus include creating a Covid-19 vaccination program where this information can be obtained through social media, including Twitter, which is a popular social media in Indonesia and is currently a trending topic. Its users are free to have an opinion or opinion through posts or comments. There are various kinds of opinions from the public, there are positive, neutral, and negative opinions about the Covid-19 vaccination program.
Therefore, this study can be formulated that how to respond to Indonesian public opinion on the Covid-19 vaccination program using data taken from Twitter social media and conducting sentiment analysis using the Naive Bayes algorithm by classifying positive, neutral, and negative sentiments from Twitter using keywords. namely "Vaccine" and "Covid".
The results of the research that have been carried out show that the level of system accuracy in the application of the Naïve Bayes algorithm gets an accuracy value of 78% and testing using the k-fold cross validation method gets an accuracy value of 80%.

Keywords: *Sentiment Analysis, Naive Bayes, Vaccines, Twitter*

ABSTRAK

Virus corona merupakan sekumpulan virus yang menginfeksi saluran pernafasan. Virus ini dikenal dengan nama Covid-19 yang diketahui berasal dari negara China yang muncul pada Desember 2019. Pada awal bulan Maret tahun 2020, pertama kalinya virus Covid-19 dilaporkan masuk ke Indonesia dan menyebar ke seluruh provinsi di

ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v14i2.24024

indonesia. Adapun langkah yang dilakukan oleh pemerintah untuk mencegah penyebaran virus antara lain membuat program vaksinasi Covid-19 dimana informasi tersebut dapat diperoleh melalui sosial media antara lain twitter yang merupakan sosial media populer di Indonesia dan saat ini sedang menjadi trending topik. Para penggunanya bebas berpendapat atau beropini melalui postingan atau komentar. Terdapat berbagai macam opini dari masyarakat, ada yang positif, netral, dan opini negatif terhadap program vaksinasi Covid-19. Oleh karena itu, penelitian ini dapat dirumuskan bahwa bagaimana merespon opini masyarakat Indonesia terhadap program vaksinasi Covid-19 menggunakan data yang diambil dari sosial media twitter dan melakukan analisis sentimen menggunakan algoritma Naive Bayes dengan mengklasifikasikan sentimen positif, netral, dan negatif dari twitter menggunakan kata kunci yaitu "Vaksin" dan "Covid".
Hasil penelitian yang telah dilakukan, menunjukkan tingkat akurasi sistem dalam penerapan algoritma *Naïve Bayes* mendapatkan nilai akurasi sebesar 78% dan pengujian menggunakan metode k-fold cross validation mendapatkan nilai akurasi sebesar 80%.

Kata Kunci: *Analisis Sentimen, Naive Bayes, Vaksin, Twitter*

I. PENDAHULUAN

Virus corona merupakan penyakit menular yang menyerang pernafasan manusia. Virus ini bermula dari Negara China, lebih tepatnya di Kota Wuhan, Provinsi Hubei. Pada awal bulan maret tahun 2020 virus ini masuk ke Indonesia [1]. Seiring berjalannya waktu penyebaran virus corona semakin cepat dan korban mulai berjatuhan. Pemerintah berupaya mengatasi virus corona ini dengan mengeluarkan peraturan dan kebijakan seperti pembatasan kegiatan masyarakat, protokol kesehatan, dan saat ini memberikan vaksinasi kepada masyarakat secara bertahap untuk mencegah penyebaran virus.

Twitter merupakan salah satu media sosial yang sedang populer digunakan pada saat ini [2]. Pengguna twitter bebas mengunggah dan mengekspresikan apapun termasuk pendapatnya. Unggahan pada twitter dapat berupa fakta, saran, informasi, dan kritik terhadap sesuatu [3]. Banyak sekali informasi penting yang bisa didapatkan dari postingan twitter dan dapat digunakan sebagai sumber data penelitian terutama data mining [4].

Salah satu informasi yang bisa didapatkan pada twitter adalah tanggapan masyarakat terhadap vaksinasi Covid-19. Saat ini Covid-19 sedang menjadi trending topik, terdapat berbagai macam opini, rumor, informasi yang belum jelas kebenarannya, sehingga menimbulkan pro dan kontra dalam masyarakat terkait dengan vaksinasi dari pemerintah. Pada media sosial twitter belum terdapat fitur untuk klasifikasi tweet atau komentar tersebut termasuk sentimen positif, negatif, atau netral. Jadi perlu dilakukan analisis sentimen untuk mengklasifikasikan tweet terkait dengan vaksinasi Covid-19 dari Pemerintah.

Hasil Penelitian [5] tentang Sentimen Analisis Publik Terhadap Joko Widodo Terhadap Wabah Covid-19 Menggunakan Metode Machine Learning. Proses analisa menggunakan Data Mining dengan kata kunci "Jokowi" dan "Covid yang dijadikan sebagai dataset. Dalam data mining untuk melakukan sentimen analisis, dilakukan teknik seperti transformation, tokenizing, stemming, classification, dan lain-lain yang sangat berpengaruh terhadap tingkat akurasinya. Adapun Gata Framework digunakan untuk preprocessing, dan Rapidminer juga digunakan untuk melakukan Analisa dan membandingkan 3 (tiga) algoritma klasifikasi yaitu Naive Bayes, Support Vector Machine, dan k-NN. Dari ketiga algoritma klasifikasi tersebut, didapatkan akurasi terbaik yaitu Support Vector Machine dengan accuracy 84.58%, precision 82.14% dan recall 85.82%.

Hasil Penelitian [6] tentang Analisis Sentimen Pro dan Kontra Masyarakat Indonesia terhadap Vaksin Covid-19 pada Media Sosial Twitter. Tujuan penelitian ini adalah melakukan

ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v14i2.24024

Analisa terhadap respon masyarakat tentang wacana vaksinasi dengan cara melakukan klasifikasi respon tersebut ke dalam respon positif dan negatif dengan mengambil data dari twitter. Langkah berikutnya akan dilakukan pengelompokkan opini masyarakat dengan menggunakan model Latent Dirichlet Allocation (LDA) untuk menangkap berbagai macam topik pembicaraan masyarakat terkait dengan wacana vaksinasi tersebut pada media sosial twitter. Hasil analisa menunjukkan bahwa masyarakat lebih banyak memberikan respon positif terhadap wacana tersebut (30%) dibandingkan dengan respon negatifnya (26%). Kata-kata bersentimen yang paling sering muncul juga mengindikasikan lebih banyak kata yang bersentimen positif dibandingkan dengan kata yang bersentimen negatif.

Hasil Penelitian [7] tentang pengembangan analisis sentimen pada dokumen twitter mengenai dampak Virus Corona Menggunakan Metode Naive Bayes Classifier. Penelitian ini bertujuan untuk memperoleh analisis teks yang diambil dari twitter dengan mengklasifikasikan sentimen positif atau negatif masyarakat. Data pengujian pada penelitian ini sejumlah 796 dokumen tweet dengan 290 sentimen positif dan 506 sentimen negative. Hasil klasifikasi dievaluasi menggunakan accuracy dan error rate untuk mengetahui tingkat keakuratan dokumen. Hasil penelitian menunjukkan bahwa metode Naive Bayes mampu mengklasifikasi dokumen tweet dengan tingkat akurasi sebesar 67% dan error rate sebesar 33%. Percobaan dengan menggunakan 3 (tiga) jumlah varian data (100, 200, dan 500) dan menghasilkan selisih nilai akurasi yang tidak jauh berbeda yaitu 0,02. Hal ini menunjukkan metode Naive Bayes untuk klasifikasi data tweet terkait dampak virus Corona menghasilkan performa yang stabil, dikarenakan nilai accuracy yang diperoleh cukup baik.

Oleh karena itu, pada penelitian ini akan dirumuskan bahwa bagaimana respon dan opini masyarakat Indonesia terhadap program vaksinasi Covud-19 dengan pengambilan data dari media sosial twitter. Peneliti mengambil data dari twitter karena data mudah didapatkan secara langsung, tidak harus melakukan survei secara tradisional, dan tidak memerlukan biaya yang banyak. Algoritma yang digunakan pada penelitian ini adalah Naive Bayes, karena mempunyai kelebihan yaitu sederhana dan memiliki tingkat akurasi yang tinggi [5]. Tujuan penelitian ini adalah melakukan klasifikasi terhadap sentimen positif, netral, atau negatif tentang opini masyarakat pada twitter dengan kata kunci yaitu "vaksin" dan "covid". Hasil penelitian ini diharapkan dapat memberikan informasi kepada masyarakat mengenai penerapan vaksinasi COVID-19 apakah cenderung termasuk dalam sentimen positif, netral, atau negatif. Serta data hasil penelitian ini dapat digunakan sebagai bahan pertimbangan untuk menentukan model kebijakan pada institusi terkait.

II. METODOLOGI

2.1 Tahapan Pengumpulan Data

Adapun teknik pengumpulan data yang digunakan pada penelitian ini adalah sebagai sebagai berikut :

2.1.1 Studi Literatur

Studi literatur dilakukan dengan mengumpulkan teori-teori yang berkaitan dengan penelitian ini. Sumber teori berasal dari buku referensi, hasil penelitian yang telah dipublikasi dan terakreditasi serta artikel-artikel terkait. Selain itu peneliti juga mengunjungi situs-situs yang terkait dengan sentimen analisis, text mining, dan Algoritma Naive Bayes Classifier. Adapun literatur yang dijadikan acuan penelitian dapat dilihat di dalam daftar pustaka.

2.1.2 Observasi

Peneliti melakukan observasi dengan cara melakukan pengambilan data dari twitter API tentang opini masyarakat terhadap vaksinasi Covid-19 pada bulan Juli tahun 2021 ini. Proses pengambilan data dilakukan secara manual dengan memanfaatkan fitur dari sosial media twitter yang diperuntukkan untuk developer yang terdapat pada website https://developer.twitter.com dengan cara *crawling* data menggunakan bahasa pemrograman Pyhton.

2.2 Pengolahan Data

Pada tahapan ini akan dilakukan proses *preprocessing* yang merupakan tahapan penting, yaitu mengurangi sejumlah atribut yang tidak perlu untuk digunakan kedalam proses klasifikasi. Data yang dimasukkan pada tahap ini adalah data mentah yang masih kotor dari proses crawling data. Hasil dari proses

Jurnal Teknik Informatika Vol. 14 No. 2, Oktober 2021
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v14i2.24024

preprocessing tersebut berupa data yang berkualitas yang dapat mempermudah dalam proses klasifikasi [7]. Adapun proses preprocessing terdiri dari beberapa tahapan yaitu:

2.2.1 *Cleaning*

Pada tahapan *cleaning*, merupakan tahap pembersihan kata yang tidak berpengaruh terhadap hasil klasifikasi sentimen. Komponen dokumen tweet memiliki berbagai atribut yang tidak berpengaruh terhadap sentimen, karena setiap tweet hampir semua memiliki atribut tersebut. Contoh nya adalah mention yang diawali dengan atribut '@', hastag yang diawali dengan atribut '#', link yang diawali dengan atribut 'http','bit.ly' dan karakter simbol ~!@#$%^&*()_+?<>,.?:{}[]|. Atribut yang tidak berpengaruh tersebut akan dihilangkan dari dokumen kemudian akan digantikan dengan karakter spasi. Contoh *cleaning* ditunjukkan pada Tabel 1.

Tabel 1. Contoh cleaning

Sebelum *cleaning*	Sesudah *cleaning*
sebagai bentuk komitmen kemanusiaan melawan pandemi covid-19. mari bersama mensukseskan program vaksinasi #covid19 #vaksinasi	sebagai bentuk komitmen kemanusiaan melawan pandemi covid-19 mari bersama mensukseskan program vaksinasi

2.2.2 *Case Folding*

Tahapan *case folding* merupakan tahap untuk mengkonversi huruf kapital pada teks ulasan ke dalam huruf kecil. Tahapan ini bertujuan supaya dokumen teks ulasan memiliki bentuk standar. Berikut ini contoh *case folding* pada Tabel 2.

Tabel 2. Contoh case folding

Sebelum *case folding*	Sesudah *case folding*
Satgas Ingatkan Masyarakat untuk Tidak Unggah Sertifikat Vaksinasi COVID-19 ke Medsos	satgas ingatkan masyarakat untuk tidak unggah sertifikat vaksinasi covid-19 ke medsos

2.2.3 *Stopword Removal*

Pada tahapan stopword removal akan dilakukan proses menghilangkan kata yang tidak sesuai dengan topik dokumen, jika ada kata yang tidak berpengaruh terhadap akurasi dalam proses klasifikasi sentimen dokumen. Kata yang akan dihilangkan, akan dihimpun kedalam database kata stopword. Jika dalam dokumen tweet, terdapat kata yang sesuai dengan kata dalam stopword, maka kata tersebut akan dihilangkan dan diganti dengan karakter spasi. Berikut ini contoh *stopword removal* pada Tabel 3.

Tabel 3. Stopword removal

Sebelum	Sesudah
semoga program vaksinasi covid adalah langkah yang efektif untuk menekan penyebaran virus	semoga program vaksinasi covid langkah efektif menekan penyebaran virus

2.2.4 *Tokenizing*

Tahapan ini adalah proses tokenisasi yaitu membagi teks yang terdapat pada kalimat atau paragraph dengan cara memotong kata berdasarkan tiap kata yang menyusunnya menjadi bentuk potongan tunggal. Kata dalam dokumen yang dimaksud adalah kata yang dipisah oleh spasi. Sehingga hasil dari proses tokenisasi ini merupakan kata tunggal yang dimasukkan ke dalam database untuk keperluan pembobotan. Contoh *tokenizing* ditunjukkan pada Tabel 4.

Tabel 4. Contoh tokenisasi

Sebelum *tokenisasi*	Sesudah *tokenisasi*
semoga program vaksinasi covid adalah langkah yang efektif menekan penyebaran virus	'semoga', 'program', 'vaksinasi','covid','adalah', 'langkah','yang', 'efektif', 'untuk', 'menekan', 'penyebaran, 'virus'

2.2.5 *Stemming*

Pada tahapan ini adalah stemming, yaitu suatu proses yang dilakukan untuk mengubah kata yang terdapat dalam suatu dokumen ke dalam kata dasar dengan menggunakan aturan tertentu. Proses stemming bahasa Indonesia dilakukan dengan cara menghilangkan sufiks, prefix, dan konfiks pada dokumen. Contoh *stemming* ditunjukkan pada Tabel 5.

ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v14i2.24024

Tabel 5. Contoh stemming

Sebelum *stemming*	Sesudah *stemming*
semoga tuhan melindungi orang orang untuk mampu melewati pandemi covid	semoga tuhan lindung orang orang untuk mampu lewat pandemi covid

2.3 Metode Analisis Data

Algoritma Naive Bayes adalah algoritma klasifikasi yang menggunakan metode probabilitas dan statistik yang diperkenalkan oleh ilmuwan Inggris bernama Thomas Bayes. Algoritma Naive Bayes ini akan memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya, sehingga dikenal sebagai Teorema Bayes [8].

Menurut Taheri (2014) Naive Bayes adalah teknik klasifikasi populer yang mudah digunakan, serta memiliki tingkat efektivitas yang tinggi untuk data mining dan machine learning. Algoritma klasifikasi Naive Bayes mempertimbangkan bobot untuk probabilitas bersyarat. Fungsi tujuan adalah dimodelkan dan diperhitungkan, yang didasarkan pada struktur pengklasifikasi Naive Bayes dan atributnya adalah bobot. Bobot optimal ditentukan oleh lokal metode optimasi menggunakan metode *quasecant*. Pada pendekatan yang diusulkan, pengklasifikasi Naive Bayes diambil sebagai titik pangkal. Hasil eksperimen numerik pada beberapa kumpulan data dunia nyata dalam klasifikasi biner, yang menunjukkan efisiensi metode yang diusulkan [9]. Persamaan (1) Teorema Bayes:

$$P(H|X) = \frac{P(X|H)}{P(X)}.P(H) \quad (1)$$

Keterangan:

X	:	Data class yang belum diketahui
H	:	Hipotesis data merupakan suatu class spesifik
P(H|X)	:	Probabilitas hipotesis H berdasar kondisi X
P(H)	:	Probabilitas hipotesis H
P(X|H)	:	Probabilitas X berdasarkan kondisi pada hipotesis H
P(X)		Probabilitas X

2.4 Metode Pengujian

2.4.1 Confusion Matrix

Confusion Matrix digunakan sebagai metode untuk melakukan evaluasi dengan mengukur dan menguji peforma atau kinerja dari model klasifikasi yang telah dibuat. Variabel pengujian yang digunakan pada saat proses evaluasi adalah akurasi, yang proses perhitungannya diperoleh dari tabel matriks, dimana tabel yang ditampilkan dalam confusion matrix terdiri dari kelas predikasi dan kelas aktual [10]. Pada saat proses pengujian model, maka akan mendapatkan hasil nilai akurasi dan confusion matrix. Adapun *Confussion matrix* ditunjukkan pada Tabel 6.

Tabel 6 Confusion matrix

Confusion Matrix		Kelas Aktual	
		Positif	Negatif
Kelas Prediksi	Positif	TP	FP
	Negatif	FN	TN

Keterangan dari nilai matriks adalah:

a. *True Positive* (TP), adalah sejumlah data yang diprediksi positif dan kenyataannya positif.
b. *True Negative* (TN), adalah sejumlah data yang diprediksi negatif dan kenyataannya negatif.
c. *False Positive* (FP), adalah sejumlah data yang diprediksi positif dan kenyataannya negatif.
d. *False Negative* (FN), adalah sejumlah data yang diprediksi negatif dan kenyataannya positif.

2.4.2 K-Fold Cross Validation

Pada tahapan ini akan dilakukan pengujian menggunakan *K-Fold Cross Validation* untuk menentukan hasil uji dan evaluasi kinerja algoritma maksimal. Dengan metode ini pengujian yang digunakan dengan cara melipat data sebanyak k dan melakukan perulangan sebanyak k.

III. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Pada penelitian ini data diambil dari twitter dengan cara crawling data menggunakan bahasa pemrograman python. Pada saat proses pengambilan data twitter, peneliti menggunakan kata kunci ”vaksin” dan “covid”.

Jurnal Teknik Informatika Vol. 14 No. 2, Oktober 2021
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v14i2.24024

Pengambilan dimulai dari tanggal 06 Juli 2021 sampai dengan 11 Juli 2021. Jumlah data yang diambil sebanyak 1000 data tweet.

3.2 *Preprocessing*

Tahap selanjutnya adalah melakukan *cleaning* atau pembersihan data sebelum digunakan untuk proses klasifikasi. Tujuannya adalah mengubah data yang masih kotor menjadi data yang bersih dan terstruktur sehingga memudahkan untuk proses klasifikasi data. Proses *preprocessing* ditunjukkan pada Gambar 1.

```
1. Konversi lowercase
dataClean['text'] = dataClean['text'].apply(lambda x: " ".join(x.lower() for x in x.split()))
#2. Cleaning
#hapus b'
dataClean['text'] = dataClean['text'].str.replace("b'|b\"",'')
# Hapus "RT"
dataClean['text'] = dataClean['text'].str.replace(r'^RT[\s]+','')
dataClean['text'] = dataClean['text'].str.replace(r'^rt[\s]+','')
# Hapus Link
dataClean['text'] = dataClean['text'].str.replace('https?://[^\s]+', '')
dataClean['text'] = dataClean['text'].str.replace('http?://[^\s]+', '')
dataClean['text'] = dataClean['text'].str.replace('www\.[^\s]+', '')
# Hapus hashtag
dataClean['text'] = dataClean['text'].str.replace(r'#', '')
#Hapus Tanda baca >> eg: @, http etc
dataClean['text'] = dataClean['text'].str.replace('[^\w\s]','')
#Hapus angka, karakter tunggal
dataClean['text'] = dataClean['text'].str.replace(r'\d+','')
dataClean['text'] = dataClean['text'].str.replace(r"\s+"," ")
dataClean['text'] = dataClean['text'].str.replace(r"\b[a-zA-Z]\b","")
#3. Hapus Stopword
dataClean['text'] = dataClean['text'].apply(lambda x: " ".join(x for x in x.split() if x not in stopWordList))
#4. Tokenizing
# Import Plugin Tokenisasi
from spacy.lang.id import Indonesian
import spacy
nlp = Indonesian() # use directly
nlp = spacy.blank('id') # blank instance'
def tokenisasiKalimat(kalimat):
    Teks = nlp(kalimat)
    Token_kata = [token.text for token in Teks]
    return Token_kata
#5. Stemming
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
factory = StemmerFactory()
stemmer = factory.create_stemmer()
dataClean['text'] = dataClean['text'].apply(lambda x: " ".join([stemmer.stem(word) for word in x.split()]))
```

Gambar 1. Tahap *preprocessing*

3.3 Analisis Sentimen

3.3.1 Pembobotan TF-IDF

Pada tahap ini, akan dilakukan perhitungan pada setiap kata yang berdasarkan *term frequency* yang muncul pada data. *Term Frequency* berguna untuk menghitung frekuensi kemunculan kata dari suatu term dalam dokumen. *Inverse Document Frequency* adalah perhitungan bagaimana term disebarkan secara luas dalam dokumen. Contoh pembobotan ditunjukkan pada Tabel 7.

Tabel 7. Contoh pembobotan tf-idf

Kata	Term Frequence	IDF	TF-IDF
Covid	436	0.547055	238.515789
Vaksin	121	2.930492	354.589502
Mungkin	97	3.390024	328.832336
Gak	128	3.071570	577.455226

3.3.2 Pelabelan Data

Selanjutnya adalah tahap pelabelan data tweet, dalam proses ini setiap dokumen teks akan dilabeli dengan sentimen positif, negatif, atau netral. Pelabelan ini dilakukan secara manual dengan menggunakan library dari python, yaitu SentiStrength dengan kosa kata bahasa Indonesia. Proses pelabelan data ditunjukkan pada Gambar 2.

```
#Analisis Sentimen dengan fungsi SentiStrength
from Assets.SentiStrength.SentiStrength import *
teksSampel = 'zulfikar pintar dan ganteng sekali tetapi lintah darat :)'
senti.doSentiStrength(teksSampel)
senti.doSentiStrength('Sistemnya eror')
```

```
{'classified_text': 'zulfikar pintar [4] dan ganteng [5] sekali tetapi lintah darat [-4] :) [3]', 'tweet_text': ' zulfikar pintar dan ganteng sekali tetapi lintah darat :)', 'sentence_score': ['zulfikar pintar [4] dan ganteng [5] sekali tetapi lintah darat [-4] :) [3]'], 'max_positive': 5, 'max_negative': -4, 'kelas': 'positive'}

{'classified_text': 'Sistemnya eror [-4]',
 'tweet_text': ' Sistemnya eror',
 'sentence_score': ['Sistemnya eror [-4]'],
 'max_positive': 1,
 'max_negative': -4,
 'kelas': 'negative'}
```

```
# Fungsi ambil kelas sentimen
def getKelasSentimen(kalimat):
    resTmp = senti.doSentiStrength(kalimat)
    return resTmp['kelas']

teksSampel = dataClean['text'][0]
print(teksSampel,' (',getKelasSentimen(teksSampel),')')
```

```
vaksin covid di mula capai di juta dosis ppkmdaruratsolusijitu  ( neutral )
```

```
#Hasil Sentimen
dataClean['Sentimen'] = dataClean['text'].apply(lambda x: getKelasSentimen(x))
dataClean = checkInfo(dataClean)
dataClean.head()
```

Gambar 2. Pelabelan sentimen

Sedangkan hasil pelabelan data ditunjukkan pada Tabel 8.

Tabel 8. Hasil pelabelan data

Tweet	Label
vaksin covid capai juta dosis	neutral
Sinovac efeknya buruk	negative
vaksin efeknya bagus untuk kesehatan	Positive
sukseskan vaksin covid	Positive

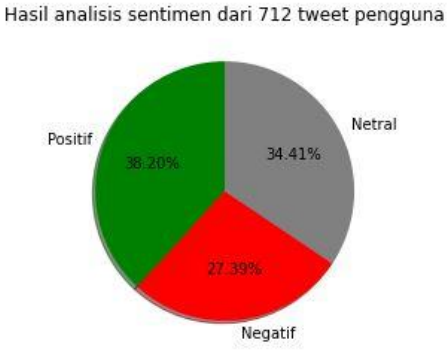
Langkah Selanjutnya adalah visualisasi dari hasil pelabelan sentimen dengan menggunakan *library matplotlib* dari *python* untuk gambar grafiknya, berikut ini adalah prosesnya pada Gambar 3.

Jurnal Teknik Informatika Vol. 14 No. 2, Oktober 2021
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v14i2.24024

```
#Visualisasi Sentimen
import matplotlib.pyplot as plt
from collections import Counter

kolomSentimen = dataClean['Sentimen']

#Persiapan variabel - hitung
counter = Counter(kolomSentimen)
positive = counter['positive']
negative = counter['negative']
neutral = counter['neutral']

#Inisialisasi propertis grafis
labels = 'Positif', 'Negatif', 'Netral'
sizes = [positive, negative, neutral]
colors = ['green', 'red', 'grey']
judul = 'Hasil analisis sentimen dari '+str(kolomSentimen.count())+' tweet pengguna'

#Cetak grafis
plt.pie(sizes, labels = labels, colors = colors, shadow = True, startangle = 90, autopct='%1.2f%%')
plt.title(judul)
plt.show();
```

Gambar 3. Visualisasi sentiment

Adapun visualisasi hasil pelabelan yaitu berupa grafik ditunjukkan pada Gambar 4.



Gambar 4. Hasil visualisasi pelabelan

Berdasarkan pada hasil visualisasi pelabelan menunjukkan jumlah sentimen positif 38,20%, sentimen negatif 27.39%, dan sentimen netral 34.41%.

3.3.3 Klasifikasi *Naive Bayes*

Pada tahapan ini, akan dilakukan proses perhitungan klasifikasi menggunakan algoritma naive bayes dengan bahasa pemrograman *python.* Pada proses ini, data dibagi menjadi 2 (dua) yaitu data *training* (data latih) dan data testing (data uji). Jumlah dari data diacak oleh sistem dengan data training sejumlah 75% atau 0,75 dan data testing sejumlah 25% atau 0,25. Berikut ini proses klasifikasi menggunakan algoritma *naive bayes* ditunjukkan pada Gambar 5.

```
#Implementasi Naive Bayes menggunakan python
import pandas as pd
import numpy as np
import csv
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer, TfidfVectorizer
from sklearn.model_selection import train_test_split, KFold, cross_val_score
from sklearn import metrics
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score, f1_score, precision_score, recall_score
from numpy import mean
data = pd.read_csv('Hasil/14.Hasil Analisis Sentimen.csv', encoding='latin-1')
vectorizer = TfidfVectorizer(use_idf=True, lowercase=True, strip_accents='ascii')
txt=vectorizer.fit_transform(data['text'].values.astype('U'))
X_train, X_test, y_train, y_test = train_test_split(txt, data['Sentimen'], test_size=0.25, random_state=0)
MNB = MultinomialNB()
MNB.fit(X_train, y_train)
predicted = MNB.predict(X_test)
conmatrix = np.array(confusion_matrix(y_test, predicted))
print("Accuracy Score={0:.2f}%".format(accuracy_score(y_test, predicted)))
print("F1 Score= {0:.2f}%".format(f1_score(y_test, predicted, average='macro')))
print("Precision Score= {0:.2f}%".format(precision_score(y_test, predicted, average='macro')))
print("Recall Score= {0:.2f}%".format(recall_score(y_test, predicted, average='macro')))
print("\n")
print("Confusion Matrix=")
print(format(conmatrix))
print("\n")
print(classification_report(y_test, predicted))
```

Gambar 5. Perhitungan naive bayes

3.4 Evaluasi

3.4.1 Confusion Matrix

Untuk mengetahui akurasi performa dari Algoritma Naïve Bayes, maka perlu dilakukan pengujian terhadap model klasifikasi yang telah dibuat. Hasil evaluasi model menggunakan metode confusion matrix mendapatkan nilai yang ditunjukkan pada Tabel 9.

Tabel 9. Hasil pengujian

Confusion Matrix	Nilai
Accuracy	78%
Precison Score	81%
F1 Score	79%
Recall	79%

3.4.2 K-fold Cross Validation

Pengujian ulang digunakan untuk menentukan hasil uji dan evaluasi kinerja algoritma maksimal dengan menggunakan metode k-fold cross validation. Penelitian ini menggunakan nilai k berjumlah 5 (lima), yang berarti dataset dibagi menjadi 5 (lima) bagian yang sama dan melakukan perulangan sebanyak 5 (lima) kali. Hasil pengujian ditunjukkan pada Tabel 10.

Tabel 10. Hasil pengujian k-fold

K-Fold	Nilai Akurasi (*Accuracy Score*)
1	0.80373832
2	0.74766355
3	0.78504673
4	0.8411215
5	0.85849057
Rata-Rata	0.8072

ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v14i2.24024

IV. PENUTUP

Berdasarkan hasil penelitian yang telah dilakukan menunjukkan tingkat akurasi sistem dalam penerapan algoritma *Naïve Bayes* untuk klasifikasi data twitter menggunakan skenario pengujian metode confussion matrix mendapatkan nilai akurasi sebesar 78% dan pengujian menggunakan metode k-fold cross validation dengan nilai k sebanyak 5 (lima) perulangan mendapatkan nilai akurasi sebesar 80%. Hasil akurasi dalam penelitian dapat dipengaruhi oleh beberapa hal yaitu semakin bersih dataset yang digunakan maka hasil akurasi semakin bagus, banyaknya data latih dan data uji dalam pengujian akurasi, selain itu dalam pelabelan sentimen pada data seharusnya dilakukan oleh pakar bahasa agar tingkat akurasinya tinggi, namun pada penelitian ini dilakukan secara manual dengan library bahasa pemrograman python, konsekuensi dari pelabelan data pada penelitian ini adalah terdapat beberapa data tweet yang tidak sesuai dengan analisis semtimen sebenarnya.

DAFTAR PUSTAKA

[1] N. M. A. J. Astari, Dewa Gede Hendra Divayana, and Gede Indrawan, "Analisis Sentimen Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode Naive Bayes Classifier," *J. Sist. dan Inform.*, vol. 15, no. 1, pp. 27–29, 2020.

[2] N. Hardi, Y. Alkahfi, P. Handayani, W. Gata, and M. R. Firdaus, "Analisis Sentimen Physical Distancing pada Twitter Menggunakan Text Mining dengan Algoritma Naive Bayes Classifier," *Sistemasi*, vol. 10, no. 1, p. 131, 2021.

[3] M. D. Mulyawan and I. Slamet, "ANALISIS SENTIMEN TERKAIT VAKSIN COVID-19 PADA DATA TWITTER MENGGUNAKAN SUPPORT VECTOR MACHINE," pp. 133–139, 2021.

[4] D. A. Muthia and H. Rachmi, "Implementation of Text Mining in Predicting Consumer Interest on Digital Camera Products," *2018 6th Int. Conf. Cyber IT Serv. Manag.*, no. Citsm, pp. 1–7, 2018.

[5] S. Hikmawan, A. Pardamean, and S. N. Khasanah, "Sentimen Analisis Publik Terhadap Joko Widodo terhadap wabah Covid-19 menggunakan Metode Machine Learning," *J. Kaji. Ilm.*, vol. 20, no. 2, pp. 167–176, 2020.

[6] F. F. Rachman and S. Pramana, "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter," *Heal. Inf. Manag. J.*, vol. 8, no. 2, pp. 100–109, 2020.

[7] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, Computers. Amsterdam: Elsevier, 2011.

[8] R. L. Hale, "Cluster analysis in school psychology: An example," *J. Sch. Psychol.*, vol. 19, no. 1, pp. 51–56, 1981.

[9] S. Taheri, J. Yearwood, M. Mammadov, and S. Seifollahi, "Attribute weighted Naive Bayes classifier using a local optimization," *Neural Comput. Appl.*, vol. 24, no. 5, pp. 995–1002, 2014.

[10] A. Novantirani, M. K. Sabariah, and V. Effendy, "Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine," *e-Proceeeding Eng.*, vol. 2, no. 1, pp. 1–7, 2015.