Sentiment Analysis Before Presidential Election 2024 Using Naïve Bayes Classifier Based on Public Opinion in Twitter [Analisa Sentimen Jelang Pilpres 2024 Menggunakan Naïve Bayes Classifier Berdasarkan Opini Publik di Twitter]

Heri Prasetyo[1)], Arif Senja Fitrani[2)]
[1, 2)]Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
*Email Penulis Korespondensi: asfjim@umsida.ac.id[2]

Abstract. *This study aims to determine the performance of the Naïve Bayes Classifier algorithm and sentiment analysis tested on a dataset obtained from Twitter social media scrapping with the topic of 2024 presidential candidates. Three candidates frequently discussed in public spaces were used as keyword parameters in data mining: #anis, #ganjar, and #pilpres2024, resulting in 3021 tweets extracted from December 2022 to January 2023, which were successfully converted to ".xlsx" format documents. Public opinions extracted from the dataset were then pre-processed using the Python programming language, resulting in 2610 cleaned tweets. The data that passed the pre-processing stage was then labeled as positive or negative sentiment. Sentiment analysis was performed using the Naïve Bayes Classifier algorithm with three testing experiments using different training and testing data compositions in each experiment. The results of the study showed that the best Naïve Bayes model was obtained in the first experiment with a 10% testing data and 90% training data composition, resulting in 71% accuracy, 93% precision, 66% recall, and an f-measure score of 77%. The conclusion of the study is that the electability of the 2024 presidential candidates shapes public opinion and generates public sentiment in the form of positive and negative tweets. Positive tweets had a higher percentage of 71.5% (1543), while negative sentiment tweets accounted for 28.5% (614). Further research is expected to produce different information by using different classification algorithms and larger data sets*
Keywords *- Naïve Bayes Classifier, Public Opinion, Sentiment Analysis, Twitter*

Abstrak. *Penelitian ini bertujuan mengetahui performa algoritma naïve bayes classifier dan Analisa sentimen yang diuji pada dataset hasil scrapping pada sosial media twitter dengan topik kandidat pilpres 2024. 3 kandidat yang sering dibicarakan diruang publik menjadi parameter keyword pada penambangan data #anis, #ganjar dan #pilpres2024 sehingga didapatkan 3021 tweet dengan rentang waktu Desember 2022 sd. Januari 2023 yang berhasil di ekstrak menjadi format dokumen '.xlsx. Opini publik yang berhasil di ekstrak kemudian dilakukan proses pre-processing menggunakan Bahasa pemrogaman python yang kemudian menghasilkan data 2610 tweet yang telah dibersihkan, data yang berhasil melewati proses pre-processing kemudian di berikan label positif dan negatif sehingga mempunyai sifat sentimental. Analisa sentimen dilakukan menggunakan algoritma naïve bayes classifier dengan 3 kali percobaan pengujian data dengan komposisi pembagian data uji dan data latih berbeda pada tiap kali percobaan. Hasil penelitan menunjukkan bahwa model naïve bayes terbaik terdapat pada percobaan pertama dengan pembagian data 10% data uji dan 90% data latih, pada percobaan pertama didapatkan 71% keakurasian nya, dengan nilai precission 93%, recall 66% dan f - measure scored 77%. Kesimpulan penelitian elektabilitas pilpres 2024 membentuk opini publik dan menimbulkan sentiment publik berupa tweet yang bersifat positif, negatif, perbincangan/tweet pada laman sosial media twitter positif memiliki persentase lebih tinggi dengan 71,5% (1543) dan percakapan yang memiliki sentiment negatif 28,5% (614). Pada penelitian selanjutnya diharapkan menghasilkan sesuatu informasi yang berbeda dengan menggunakan algoritma klasifikasi yang lain dan jumlah data yang lebih banyak.*
Kata Kunci *- Analisa Sentimen, Naïve Bayes Classifier, Opini Publik, Twitter.*

I. PENDAHULUAN

Kontestasi pemilihan presiden (pilpres) selalu menjadi topik yang sangat menarik diulas [1], pergerakan berita dari berbagai macam unsur selalu muncuk ke permukaan, kandidat yang sedang diusung sudah mulai diapungkan oleh partai – partai politik dan media [2]. Pencarian mengenai topik pilpres 2024 mulai ramai di halaman pencarian, google maupun social media, salah satunya media social twitter.

Twitter merupakan media berbagi informasi berupa foto, video dan narasi apasaja yang tidak dilarang oleh kebijakan perusahaan(twitter) [3]. Pejabat tinggi pemerintah bahkan lembaga di Indonesia banyak menggunakan social media ini [4] begitu pula dengan tokoh kandidat calon presiden pada pilpres 2024, banyak narasi yang dilontaran untuk menarik aspirasi public yang bertujuan mendeklarasikan opininya jelang pilpres 2024, banyak masyarakat pengguna twitter merespon kejadian tersebut di kolom komentar(re-tweet) sampai mengungkapkan narasinya pada

laman pribadinya(tweet) untuk menanggapai narasi yang lontarkan kandidat pilpres 2024, dari yang pro kepada calon tersebut bahkan tanggapan negatif tentangnya [5].

Narasi yang dibangun selalu mempunyai sifat positif dan negative, itulah yang disebut dengan sentiment. Analisis sentimen adalah salah satu bentuk analisis data yang bertujuan untuk mengetahui dan mengevaluasi opini, sentimen, dan emosi yang ada dalam suatu narasi/topik tertentu, seperti kasus pejabat publik, polemik atau suatu produk terntentu [6]. Analisis sentimen sering digunakan untuk memahami pandangan dan persepsi masyarakat terhadap suatu topik tertentu, termasuk juga dalam konteks politik, seperti pemilihan presiden.

Pemilihan presiden sendiri merupakan ajang pesta demokrasi untuk memilih suatu tokoh terbaik pada suatu negara sehingga dapat dipilih oleh masyarakat dan dijadikan suatu pimpinan tertinggi di negara tersebut [7]. Proses politik ini yang seringkali menimbulkan banyak polemik, sehingga menarik dikupas dan dijadikan suatu pembelajaran dalam taraf intelktual akademis maupun teoritis bagi para pengamat, pelaku, pakar politik [8].

Data yang didapatkan pada penelitian ini bersumber dan memanfaatkan opini public yang dituangkan pada laman twitter, proses penambangan datanya sendiri disebut dengan teks mining dengan metode scapping menggunakan Bahasa pemrogaman python. Teknik scrapping merupakan salah satu cara untuk mengunduh data di ruang media public yang cukup efektif [9], pada kasus ini digunakan untuk mengunduh opini public yang bersinggungan dengan kandidat capres pada pilpres 2024.

Hasil scrapping merupakan data mentah yang belum dapat kita gunakan untuk menganalisa suatu kasus, harus melalui beberapa tahapan seperti teks preprocessing, eksplorasi data, pemodelan topik menggunakan algoritma Naïve Bayes Classifier (NBC), evaluasi dan sehingga mendapatkan tujuan penelitian [10].

Algoritma Naïve Bayes Classifier sendiri merupakan metode klasifikasi berdasarkan teorema “bayes” menggunakan pembelajaran mesin, dengan perhitungan probabilitas frekuensi nilai yang sering muncul [11].

Oleh karena itu dalam penelitian ini ingin melakukan pengujian kinerja algoritma NBC pada study topik pilpres 2024 dan diketahui elektabilitas potensi kandidat bakal calon presiden pada pemilihan presiden 2024 mendatang berdasarkan opini publik.

II. METODE

A. Tahapan penelitian

Penelitian ini dilakukan melalui beberapa tahapan proses sesuai kaidah pengolahan data yang sering digunakan pada penelitian – penelitan sebelumnya. Yakni meliputi penambangan data(scrapping), preprocessing, labeling, pemodelan topik, analisis dan evaluasi. Seperti pada gambar 1 dibawah ini.

Gambar 1. Alur kerja penelitian

B. Penambangan data

Scrapping (atau web scraping) adalah proses ekstraksi data dari sebuah website atau sumber informasi lainnya secara otomatis dengan menggunakan software atau bot tertentu. Proses ini dilakukan dengan cara mengekstrak data secara sistematis dari website, kemudian menyimpan data tersebut dalam format yang bisa diakses dan diolah dengan mudah. Tujuan dari scrapping bisa bermacam-macam, mulai dari pengambilan informasi produk pada situs e-commerce, penelitian pasar, hingga pengumpulan data untuk analisis dan penelitian ilmiah. Scrapping biasanya dilakukan dengan memanfaatkan tools atau bahasa pemrograman seperti Python, JavaScript, atau R untuk membaca dan mengambil data dari website [12].

C. Pre-processing

Preprocessing merupakan proses menstandarkan data yang didapatkan dari public melalui tahapan – tahapan agar data terstandar, sehingga mengurangi error-rate pada proses Analisa data [13]. Tahapan preprocessing seperti berikut :

1. Case Folding : adalah proses mengubah seluruh huruf menjadi huruf kecil. Pada proses ini karakter-karakter 'A'-'Z' yang terdapat pada data diubah kedalam karakter 'a'-'z'., seperti pada tabel 1.

Tabel 1. Case Folding

No	Sebelum	Sesudah
1	Nyopras Nyapres Aduhh	nyopras nyapres aduhh
2	Rambut PUTIH Pasti di bantu PRESIDEN	rambut putih pasti di bantu presiden

2. Cleansing : merupakan proses untuk membersihkan kata yang tidak diperlukan dengan beberapa teknik seperti memperkecil noise, membetulkan data yang tidak konsisten, mengisi data kosong, mengidentifikasi atau membuang outlier. Kata yang akan dihilangkan meliputi URL, Hashtag (#), Username (@). Selain itu juga akan menghilangkan tanda baca seperti koma (,), titik (.), Tanda seru (!), dan tanda baca lainnya, seperti pada tabel 2.

Tabel 2. Cleansing

No	Sebelum	Sesudah
1	@vinci, nyopras - nyapres aduhh !!!!	vinci nyopras nyapres aduhh
2	#rambut putih, pasti di bantu presiden…..	rambut putih pasti di bantu presiden

3. Tokenizing : merupakan proses pemisahan kalimat menjadi suatu token atau kata yang terpotong. Manfaat memisahkan menjadi perkata ini adalah untuk memudahkan pada proses selanjutnya, yakni proses stopword dan stemming, karena dua proses tersebut akan mencocokan kata per kata dengan library root nya, seperti tabel 3.

Tabel 3. Tokenazing

No	Sebelum	Sesudah
1	vinci nyopras nyapres aduhh	[vinci] [nyopras] [nyapres] [aduhh]
2	rambut putih pasti di bantu presiden	[rambut] [putih] [pasti] [di] [bantu] [presiden]

4. Stopward : merupakan proses membersihkan kata yang tidak diperlukan, apakah termasuk di dalam daftar kata tidak penting (stoplist) atau tidak, sehingga yang tersisa adalah kata penting saja atau disebut keywords, seperti pada tabel 4.

Tabel 4. Stopword-removal

No	Sebelum	Sesudah
1	vinci nyopras nyapres aduhh	vinci nyopras nyapres aduh
2	rambut putih pasti di bantu presiden	rambut putih pasti bantu presiden

Pada tabel diatas terlihat bahwa kata imbuhan/penghubung yang tidak diperlukan seperti “di” dapat dibuang secara otomatis menggunakan library stopword.

5. Stemming : berfungsi untuk mereduksi kata yang bukan stopword menjadi ke-root word yang sesuai, dengan dilakukan nya proses *stemm* sehingga menghilangkan awalan dan akiran atau kata imbuhan.

D. Labeling

Merupakan tahapan untuk memberikan label pada setiap opini / tweet yang datanya berhasil di standarisasikan pada tahapan pre-processing, proses labeling sendiri bertujuan memberikan paradigma yang bersifat positif dan negative pada masing – masing kalimat, sehinggi mempunyai sentimen [14], seperti pada tabel 5 dibawah.

Tabel 5. Labeling

Username	Tweet	Sentimen
@yudistir_	vinci nyopras nyapres aduhh, padahal ga bagus juagaa	Negatif
@miklok.kerenabies	rambut putih pasti di bantu presiden, insyaallah menang	Positif

Pada tabel 5 diatas terdapat 2 contoh untuk pelabelan positif dan negative sehingga tweet hasil scrapping dapat diterjemankan dalam Bahasa program untuk dilihat kecenderungan antara komentar positif dan negative nya.

E. Pemodelan topik

Pada penelitian ini yang digunakan adalah Algoritma Naive Bayes Classifier merupakan teknik klasifikasi berdasarkan Teorema Bayes dengan asumsi independensi di antara para prediktor. Naive Bayes Classifier memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Dalam istilah sederhana, penggolongan Naive Bayes menganggap bahwa kehadiran fitur tertentu di kelas tidak terkait dengan kehadiran fitur lainnya. Keuntungan penggunaan adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menentukan estimasi parameter yg diperlukan dalam proses pengklasifikasian. Karena yang diasumsikan sebagai variabel independen, maka hanya varians dari suatu variabel dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarians.

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

Dengan P merupakan probabilitas, X adalah data dengan class yang belum diketahuai, H adalah hipotesis data X yang merupakan suatu class spesifik, P(H|X) merupakan probabilitas hipotesis H berdasarkan kondisi X (posteriori prob.), P(H) adalah probabilitas hipotesis H (prior prob), P(X|H) adalah probabilitas X berdasarkan kondisi tersebut, P(X) merupakan probabilitas dari X [15].

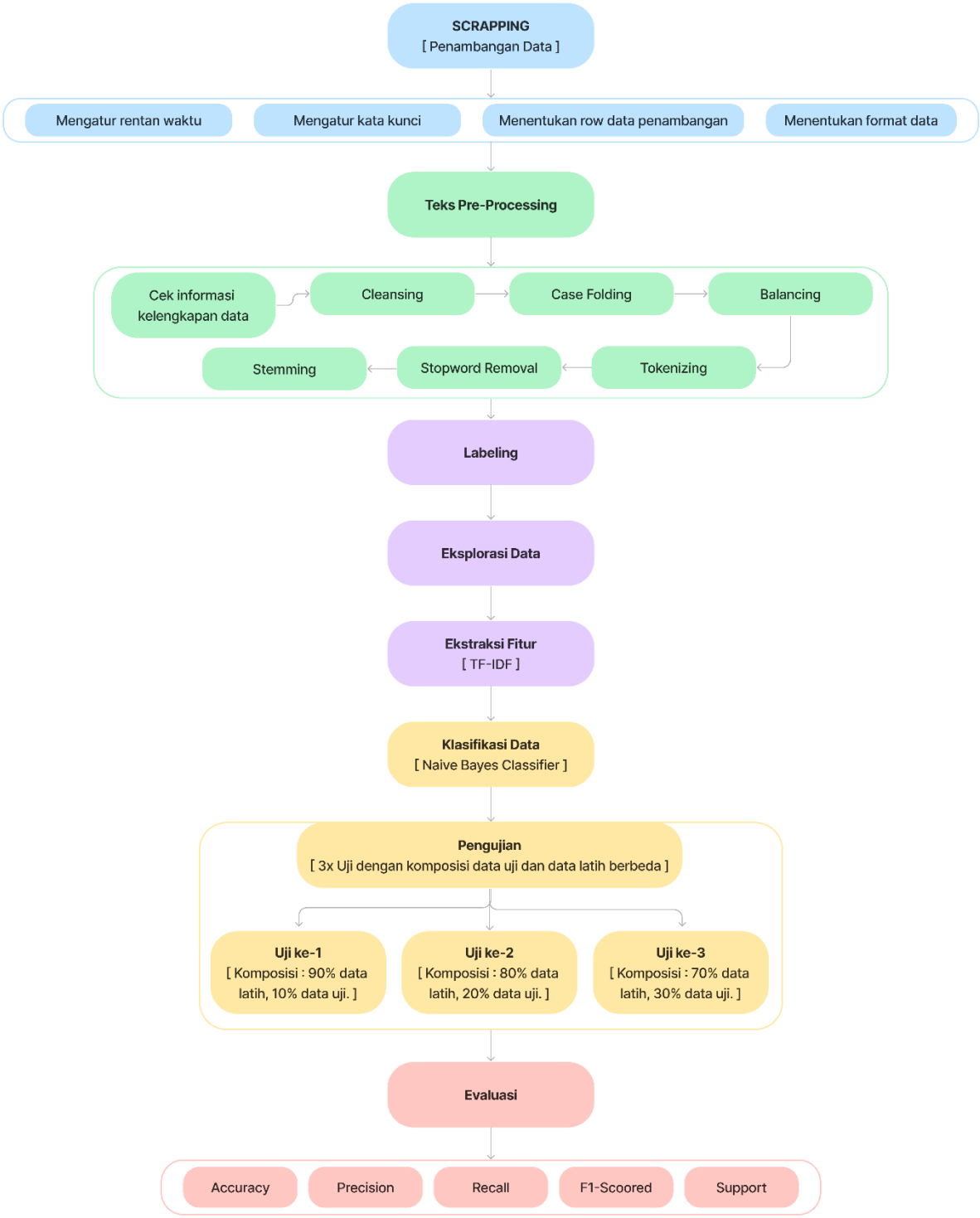
F. Analisa dan evaluasi

Setelah melakukan preprocesing dan labeling, kemudian data dibagi menjadi data training dan data testing untuk melakukan pengujian model data, untuk mengetahui keakurasian pemodelan data atau metode yang digunakan maka dapat mengetahui nilai precission, recall & accuration serta f-measure. Proses Analisa sendiri bisa meliputi eksplorasi data seperti mengetahui karakteristik dan keberhasilan preprocesing yang dapat dilihat dari wordcloud seperti gambar 2 dibawah ini.

Gambar 2. Eksplorasi data

III. HASIL DAN PEMBAHASAN

Penelitian dilakukan melalui beberapa tahapan sesuai teknik pengolahan data teks pada umumnya, seperti tergambar pada diagram dibawah ini.



Gambar 3. Kerangka Berfikir

A. Scrapping

Proses scrapping atau yang biasa dikenal dengan penambangan data dilakukan dengan rentang waktu desember 2022 sampai dengan januari 2023, dengan filter atau kata kunci “pilpres 2024”, “ganjar”, “anis”. Library pada bahasa pemrograman phyton yang digunakan adalah snscrap sehingga memungkinkan menambang data pada lama twitter tanpa menggunakan API, kode program yang digunakan seperti pada gambar 4 berikut.

```
import snscrape.modules.twitter as sntwitter
import csv

keyword = "ganjar"
since_date = "2023-05-01"
until_date = "2023-07-30"
query = f"{keyword} since:{since_date} until:{until_date}"

tweets = sntwitter.TwitterSearchScraper(query).get_items()

with open('pilpres-ganjar1.csv', 'w', newline='', encoding='utf8') as f:
    writer = csv.writer(f)
    writer.writerow(['id', 'date', 'content', 'retweets', 'favorites', 'user_id', 'username'])
    for tweet in tweets:
        writer.writerow([tweet.id, tweet.date, tweet.content, tweet.retweetCount, tweet.likeCount, tweet.user.id, tweet.user.username])
```

Gambar 4. Scrapping

Sehingga didapatkan data sejumlah 3021 dengan data row (id twitter, tanggal, konten/tweet, jumlah retweet, jumlah menyukai, id pengguna, nama pengguna) dengan format .xlsx/excel. Hasil penambangan data dapat dilihat pada gambar 5.

	id	date	tweet	retweets	favorites	user_id	username
0	1609479185160100096	2023-01-01 09:18:35+00:00	@DoankWarto Apa pun itu.hati warga Jakarta cin...	0.0	0.0	1.554809e+18	IdalestariLest2;;;
1	1609479855070129920	2023-01-01 09:21:14+00:00	@Muhammad_Saewad Lagi ngomongin si yohanes Ani...	0.0	0.0	1.321775e+18	Kucinganggora_8;;;
2	1609480619418939904	2023-01-01 09:24:17+00:00	@rahmansandre Anis Baswedan	1.0	1.0	1.270434e+18	NalijahSaidi;;;
3	1609480817704819968	2023-01-01 09:25:04+00:00	@M45Broo Ternyata sama ya kaya si Yohanes Anis...	0.0	1.0	1.321775e+18	Kucinganggora_8;;;
4	1609508523876179968	2023-01-01 11:15:09+00:00	@bachrum_achmadi Yohanes Anis Baswedan demi pr...	0.0	1.0	1.570257e+18	JakaYuda11;;;
5	1609513470697050112	2023-01-01 11:34:49+00:00	@sutanmangara Kayaknya Anis Baswedan sibuk pen...	0.0	0.0	1.442893e+18	AfhamPranoto;;;
6	1609515124083259904	2023-01-01 11:41:23+00:00	@caul_tanjung Kemanakah Anis Baswedan... Kok b...	0.0	0.0	1.442893e+18	AfhamPranoto;;;
7	1609519244722720000	2023-01-01 11:57:46+00:00	@bachrum_achmadi Bukannya buzzeRp nyalain anis...	0.0	0.0	1.471303e+18	MatKlowor2;;;
8	1609532058891840000	2023-01-01 12:48:41+00:00	@zola_papazola2 @tatakujiyati @aniesbaswedan @...	0.0	0.0	2.519471e+09	Bakhjatul;;;
9	1609534583216940032	2023-01-01 12:58:43+00:00	@Relawananies Sehat selalu pak Yohanes Anis Ba...	1.0	0.0	1.290610e+18	benuaandin;;;

Gambar 5. Hasil Scrapping

B. Teks preprocessing

Pada tahapan ini banyak proses yang dilakukan, preprocessing perlu dilakukan untuk menambah keakurasian evaluasi dan eksplorasi data kedepannya. Setelah melakukan scrapping data perlu dilakukan pengecekan kualitas, kuantitas dan komposisi data, sehingga diketahui informasi mengenai data tersebut. Setelah itu dilakukan proses :

1. Cleansing

Pembersihan data dilakukan dengan menghapus data yang error, null, duplicate. Kode program seperti pada gambar 6 berikut :

```
# DATA CLEANING
#CLEANING kolom yang tidak digunakan
import pandas as pd

data = data.drop(columns=['id', 'date', 'retweets', 'favorites', 'user_id'])
data.head(10)
```

Gambar 6. Cleansing

Pada tahap ini juga menghapus beberapa row yang tidak dibutuhkan untuk proses analisis sentiment seperti (id pengguna, tanggal, retweet, favorite, user id) sehingga dapat mempercepat kinerja machine learning dan meningkatkan kecepatan komputasi pengolahan data. Data yang berhasil di bersihkan seperti terdapat pada gambar 7.

	tweet	username
0	@DoankWarto Apa pun itu.hati warga Jakarta cin...	IdalestariLest2;;;
1	@Muhammad_Saewad Lagi ngomongin si yohanes Ani...	Kucinganggora_8;;;
2	@rahmansandre Anis Baswedan	NalijahSaidi;;;
3	@M45Broo Ternyata sama ya kaya si Yohanes Anis...	Kucinganggora_8;;;
4	@bachrum_achmadi Yohanes Anis Baswedan demi pr...	JakaYuda11;;;
5	@sutanmangara Kayaknya Anis Baswedan sibuk pen...	AfhamPranoto;;;

Gambar 7. Hasil Cleansing

Selain menghapus data tersebut pada proses cleansing juga menghapus huruf/symbol-simbol yang tidak diperlukan seperti tanda tanya (?), titik (.), koma (,), tanda seru (!) dan symbol lainnya. Kode program dan hasil dapat dilihat pada gambar 8 dan 9.

```
# proses cleansing remove regex (cleansing) seperti tanda baca dan angka angka
import re
import string

def cleansing(tweet):
    tweet = tweet.strip(" ")
    tweet = re.sub(r'[?|$|.|!_:@")(-+,]', '', tweet)
    tweet = re.sub(r'[?|$|.|!_:@")(-+,]', '', tweet)
    tweet = re.sub(r'\d+', '', tweet)
    tweet = re.sub(r"\b[a-zA-Z]\b", "", tweet)
    tweet = re.sub('\s+',' ', tweet)
    return tweet
data['tweet'] = data['tweet'].apply(cleansing)

data.head(10)
```

Gambar 8. Remove regex

	tweet
0	DoankWarto Apa pun ituhati warga Jakarta cinta...
1	MuhammadSaewad Lagi ngomongin si yohanes Anis ...
2	rahmansandre Anis Baswedan
3	MBroo Ternyata sama ya kaya si Yohanes Anis Ba...
4	bachrumachmadi Yohanes Anis Baswedan demi pres...
5	sutanmangara Kayaknya Anis Baswedan sibuk penc...

Gambar 9. Hasil Remove regex

2. Case Folding

Pada tahap ini berfungsi merubah teks/huruf yang terdapat pada suatu kalimat menjadi huruf kecil atau yang disebut *lowercase* semua mulai ‘A-Z’ menjadi ‘a-z’.

3. Balancing

Selanjutnya adalah balancing untuk mengetahui data kita seimbang atau tidak, fungsi pada tahapan ini adalah untuk menambah keakurasian saat evaluasi data. Tahap ini dilakukan dengan cara menecek komposisi masing-masing row setalah dilakukan cleansing. Proses dapat dilakukan seperti pada gambar diabawah berikut.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2641 entries, 0 to 2678
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   tweet     2641 non-null   object
 1   username  2610 non-null   object
dtypes: object(2)
memory usage: 61.9+ KB
```

Gambar 10. Info balancing

Pada gambar 10 terlihat bahwa komposisi data pada kolom tweet dan kolom username tidak seimbang, maka dari itu perlu dilakukan perbaikan pada masing – masing kolom, sehingga hasil yang diharapkan data dapat seimbang jumlahnya, seperti pada gambar 11 berikut.

```
data.dropna(inplace=True) # menghapus data kosong
data.fillna(data.mean(), inplace=True) # mengisi data kosong dengan rata-rata kolom
data.fillna('unknown', inplace=True) # mengisi data kosong dengan nilai 'unknown'

data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2610 entries, 0 to 2678
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   tweet     2610 non-null   object
 1   username  2610 non-null   object
dtypes: object(2)
memory usage: 61.2+ KB
```

Gambar 11. Hasil balancing

Setelah proses cleansing, semula data global berjumlah 3021 baris menjadi 2610 baris data yang siap di prose ke tahap selanjutnya.

4. Tokenizing

Tokenizing dilakukan untuk memisahkan kalimat menjadi penggalan kata, fungsi pada tahap ini adalah untuk memudahkan proses stopward removal. Proses ini dilakukan seperti terlihat pada gambar 12.

	tweet	username
0	[doankwarto, apa, pun, ituhati, warga, jakarta...	IdalestariLest2;;;
1	[muhammadsaewad, lagi, ngomongin, si, yohanes,...	Kucinganggora_8;;;
2	[rahmansandre, anis, baswedan]	NalijahSaidi;;;
3	[mbroo, ternyata, sama, ya, kaya, si, yohanes,...	Kucinganggora_8;;;

Gambar 12. Tokenizing

5. Stopword removal

Proses stopword berfungsi untuk menghapus kata yang tidak diperlukan sehingga dapat meningkatkan akurasi dan mengurangi error rate saat dilakukan visualisasi data maupun evaluasi data.

6. Stemming

Stemming adalah metode untuk mencari kata dasar dari sebuah kata. Proses stemming memiliki pengaruh dalam tingkat akurasi temu kembali informasi. Stemming dilakukan dengan cara menghilangkan imbuhan yang terdapat pada kata. Proses stemming ini cenderung memakan waktu yang cukup lama, berkisar 2-3 jam, tergantung masing – masing jumlah dataset yang digunakan.

Keseluruhan proses pada tahap teks pre-processing dapat dilihat pada tabel 6 dibawah ini :

Tabel 6. Proses Pre-processing

Proses	Sebelum	Sesudah
Case Folding	PILPRES 2024 Semoga mempunyai hasil yang baik, bagus dan berkualitas. Indonesia Maju !!!	pilpres 2024 semoga mempunyai hasil yang baik, bagus dan berkualitas. indonesia maju !!!
Cleansing	PILPRES 2024 Semoga mempunyai hasil yang baik, bagus dan berkualitas. Indonesia Maju !!!	pilpres 2024 semoga mempunyai hasil yang baik bagus dan berkualitas indonesia maju
Tokenizing	PILPRES 2024 Semoga mempunyai hasil yang baik, bagus dan berkualitas. Indonesia Maju !!!	‘pilpres’ ‘2024’ ‘semoga’ ‘mempunyai’ ‘hasil’ ‘yang’ ‘baik’ ‘bagus’ ‘dan’ ‘berkualitas’ ‘indonesia’ ‘maju’
Stopword	PILPRES 2024 Semoga mempunyai hasil yang baik, bagus dan berkualitas. Indonesia Maju !!!	‘pilpres’ ‘2024’ ‘semoga’ ‘mempunyai’ ‘hasil’ ‘baik’ ‘bagus’ ‘berkualitas’ ‘indonesia’ ‘maju’
Stemming	PILPRES 2024 Semoga mempunyai hasil yang baik, bagus dan berkualitas. Indonesia Maju !!!	‘pilpres’ ‘2024’ ‘semoga’ ‘punya’ ‘hasil’ ‘baik’ ‘bagus’ ‘dan’ ‘kualitas’ ‘indonesia’ ‘maju’

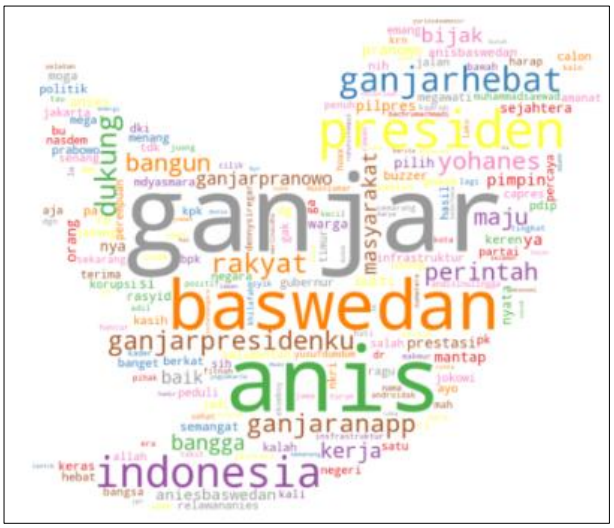
C. Labeling data

Tahapan ini sangat menentukan jumlah sentimen. Proses labeling dilakukan dengan bantuan library python machine learning, pada tahapan ini perlu juga bantuan translator dari google dengan cara memasukkan google translate kedalam machine learning, hal ini diperlukan karena algoritma yang saat ini ada (2023) belum mampu memproses labeling data yang berbahasa Indonesia, sehingga perlu merubah data menjadi bahasa inggris.

D. Eksplorasi data

Eksplorasi data dilakukan untuk menggali informasi yang terdapat di dalam dataset tersebut, sehinggal dapat memberikan informasi untuk dipahami, selain itu eksplorasi data juga bertujuan melihat semua komponen data sebelum dilakukan proses pemodelan data dan evaluasi. Pada penelitian ini eksplorasi data dilakukan untuk melihat komposisi sentiment, jumlah sentiment, wordcloud, seperti gambar dibawah ini.

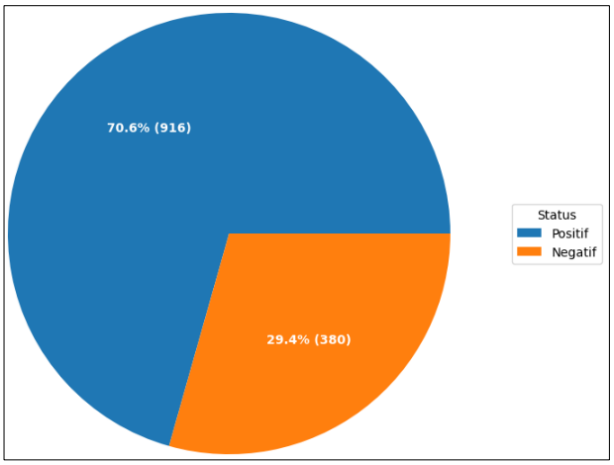
- Wordcloud



Gambar 13. Wordcloud

Wordcloud diatas jelas sekali menggambarkan kata yang sering muncul, kata tersebut didapat dari hasil ekstrasi proses – proses sebelumnya, teks yang paling besar menggambarkan bahwa kata tersebut sering muncul, sehingga mempunyai komposisi paling besar dalam gambar tersebut.

- Diagram Lingkaran

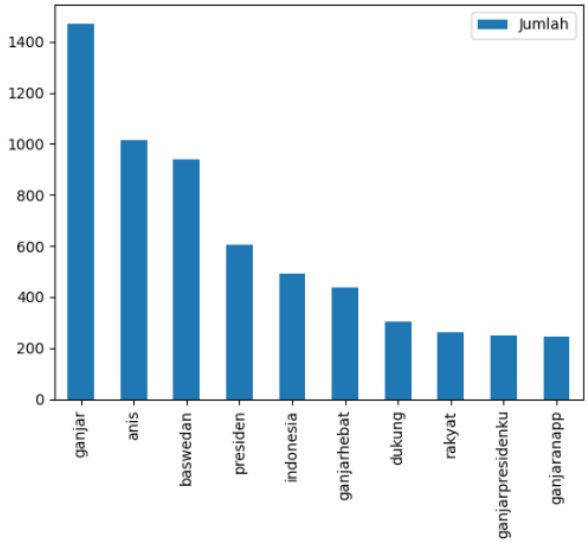


Gambar 14. Sentimen Diagram Lingkaran

Diagram lingkaran diatas dapat menginformasikan bahwasanya pada dataset dalam penelitian ini terdapat 70.6% komentar positif perihal pilpres 2024 dan 29.4% berkomentar negative.

- Diagram Batang

Pada gambar 15 berikut terdapat informasi jumlah kata yang sering muncul dari yang tertinggi hingga terendah yang terdapat dalam dataset penelitian ini. Kita dapat melihat kata “ganjar” menjadi yang paling tertinggi, sehingga tidak salah bahwasannya pada gambar 13 kata “ganjar” menjadi yang paling dominan atas gambar wordcloud tersebut.



Gambar 15. Diagram Batang

E. Ekstraksi fitur

Dokumen teks yang sebelumnya telah melalui banyak proses kemudian ditambahkan ke dalam proses pembobotan kata dengan menggunakan algoritma Term Frequency –Invers Document Frequency (TF-IDF). Dimana jika nilai yang didapatkan dari tahap pembobotan kata semakin tinggi bobot katanya maka mengindikasikan bahwa kata tersebut semakin layak digunakan sebagai keyword terhadap kalimat tersebut. Tahapan ekstraksi fitur menggunakan TF-IDF dapat dilihat pada gambar 16 berikut:

```
# PROSES TF IDF

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = CountVectorizer()
X = vectorizer.fit_transform(data_clean['tweet'].astype('U'))

tf = TfidfVectorizer()
text_tf = tf.fit_transform(data_clean['tweet'].astype('U'))
print(text_tf)
```

Gambar 16. TF-IDF

```
  (2, 1655)     0.10183407308691947
  (2, 2746)     0.19410233457798368
  (2, 1530)     0.24185191502433376
  (2, 3989)     0.3075142563623541
  (2, 581)      0.2592611742633071
  (2, 1969)     0.3075142563623541
  :     :
  (2155, 4528)  0.2744176397355608
  (2155, 2884)  0.255920205984347
  (2155, 2380)  0.24373625784928868
  (2155, 1484)  0.20227003542477864
  (2155, 2350)  0.2744176397355608
  (2155, 2549)  0.49863637222867
  (2155, 3706)  0.255920205984347
  (2155, 4009)  0.1747821872461326
```

Gambar 17. Hasil TF-IDF

Pada gambara 16 paragraf pertama kode tersebut berfungsi untuk memanggil fitur / library pada proses TF-IDF, kemudian pada paragraph kedua membuat variable vectorizer berisi jumlah perhitungan pembobotan kata. Kemudian pada paragraph ke tiga membuat variabel vectorizer berisi TfidfVectorizer(). Lalu pada baris paling bawah digunakan untuk menampilkan hasil pembobotan dari proses TF-IDF. Hasil pemrosesan seperti pada gambar 17 diatas.

F. Klasifikasi data

Pada tahapan ini akan dilakukan klasifikasi menggunakan algoritma naïve bayes classifier, dengan skema 3 kali uji coba untuk mencari model terbaik. Sebelum itu diperlukan proses pembagian komposisi data latih dan data uji dengan komposisi masing – masing yang berbeda.

Uji ke-1

Pada uji pertama dilakukan pembagian data menjadi 90% data latih dan 10% data uji, hasil dari pengujian ini dapat dilihat pada gambar 18.

	precision	recall	f1-score	support
negatif	0.45	0.85	0.59	53
positif	0.93	0.66	0.77	163
accuracy			0.71	216
macro avg	0.69	0.76	0.68	216
weighted avg	0.81	0.71	0.73	216

Gambar 18. Hasil percobaan ke – 1

Pada percobaan ke 1 menggunakan metode naïve bayes classifier dengan komposisi pembagian data latih dan data uji seperti dijelaskan diatas menghasilkan nilai accuracy 71%, precision negative 45%, precision positif 93%, recall negative 85%, recall positif 66%, f1-score negative 59% f1-score positif 77%.

Uji ke-2

Pada uji kedua dilakukan pembagian data menjadi 80% data latih dan 20% data uji, maka hasil percobaannya seperti pada gambar 19 berikut :

	precision	recall	f1-score	support
negatif	0.46	0.83	0.59	115
positif	0.91	0.65	0.76	317
accuracy			0.69	432
macro avg	0.69	0.74	0.67	432
weighted avg	0.79	0.69	0.71	432

Gambar 19. Hasil percobaan ke – 2

Pada percobaan kedua menggunakan metode naïve bayes classifier dengan komposisi pembagian data latih dan data uji seperti dijelaskan diatas menghasilkan nilai accuracy 69%, precision negative 46%, precision positif 91%, recall negative 83%, recall positif 65%, f1-score negative 59% f1-score positif 76%.

Uji ke-3

Pada uji terakhir dilakukan pembagian data menjadi 70% data latih dan 30% data uji, hasil percobaan seperti terdapat pada gambar 20 berikut.

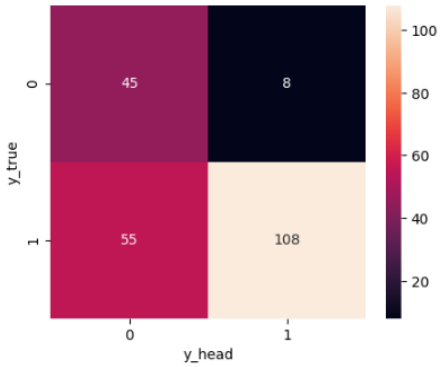
	precision	recall	f1-score	support
negatif	0.46	0.82	0.59	179
positif	0.90	0.63	0.74	469
accuracy			0.68	648
macro avg	0.68	0.72	0.67	648
weighted avg	0.78	0.68	0.70	648

Gambar 20. Hasil percobaan ke – 3

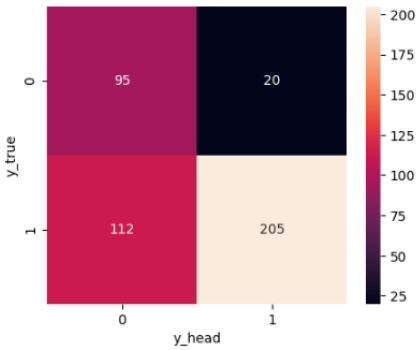
Pada percobaan ketiga menggunakan metode naïve bayes classifier dengan komposisi pembagian data latih dan data uji seperti dijelaskan diatas menghasilkan nilai accuracy 68%, precision negative 46%, precision positif 90%, recall negative 82%, recall positif 63%, f1-score negative 59% f1-score positif 74%.

G. Evaluasi

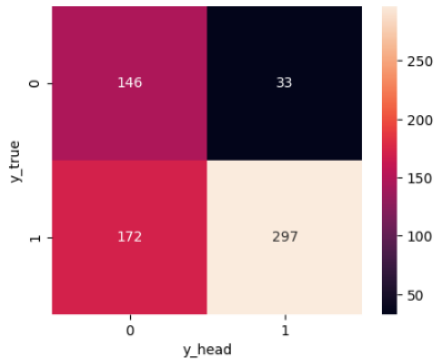
Data training dan data testing adalah dua sub-proses yang membentuk proses validasi. Model algoritma yang dipilih selama tahap pemodelan dilatih menggunakan sub-proses data training. Setelah pemodelan algoritma dilatih pada tahap sub-proses training, selanjutnya akan dilakukan testing. Adapun evaluasi atau pengujian hasil klasifikasi Naïve Bayes Classifier (NBC) menggunakan confusion matrix yang merupakan alat untuk melakukan analisis terhadap seberapa baik klasifikasi yang telah dihasilkan dan mengenali tuple dari kelas yang berbeda. Hasil confusion matrix metode NBC dapat dilihat pada Gambar 21 untuk percobaan pertama, gambar 22 untuk percobaan kedua dan confusion matrix pada gambar 23 untuk percobaan ketiga.



Gambar 21. Confussion Matrix Uji ke-1



Gambar 22. Confussion Matrix Uji ke-2



Gambar 23. Confussion Matrix Uji ke-3

Pada Gambar confusional matrix tersebut merupakan hasil dari matrix yang dihasilkan dan divisualisasikan dalam bentuk confusional matrix. Setelah proses yang sudah dilakukan maka dapat digambarkan dalam bentuk tabel matrix evaluasi dapat dilihat pada Tabel 7 sebagai berikut :

Tabel 7. Rekap percobaan

Percobaan	Kelas	Akurasi	*Precission*	*Recall*	*f-measure*
Ke – 1	Negatif	0.71	0.45	0.85	0.59
	Positif		0.93	0.66	0.77
Ke – 2	Negatif	0.68	0.46	0.82	0.59
	Positif		0.90	0.63	0.74
Ke – 3	Negatif	0.69	0.48	0.82	0.60
	Positif		0.90	0.64	0.75

IV. SIMPULAN

Berdasarkan percobaan yang telah dilakukan sehingga mendapatkan hasil yang ingin dicari, hasil percobaan pengujian metode Naïve Bayes Classifier terhadap data hasil scrapping opini publik pada sosial media twitter dengan topik pilpres 2024 dengan kata kunci pencarian "pilpres 2024", "anis", "ganjar" dan kata pendekatan atau yang berhubungan lainnya. Data hasil scrapping 3021 dilakukan pre-processing sehingga menjadi 2610 kata, kemudian diberikan label sentimen yang bermuatan positif berjumlah 1543 (71,5%) dan negatif berjumlah 614 (28,5%). Percobaan yang telah dilakukan berturut – turut sebanyak 3 kali percobaan dengan hasil terbaik diperoleh pada percobaan ke – 1 sehingga di dapatkan accuracy scored 71% dengan nilai precission 93% class positif 45% negatif, recall 66% class negatif 85% positif dan f - measure scored 59% pada class negative 77% pada class positif. Sehingga dapat disimpulkan bahwa percobaan terbaik yang dilakukan adalah pada percobaan ke – 1 dengan pembagian 10% data uji dan 90% data latih. Pada penelitian selanjutnya diharapkan menggunakan algoritma lain dengan data yang lebih banyak sehingga dapat mendapatkan hasil yang berbeda.

UCAPAN TERIMAKASIH

Terimakasih kepada Universitas Muhammadiyah Sidoarjo yang telah memberikan fasilitas laboratorium computer informatika sehingga penelitian ini dapat diselesaikan dengan baik, kemudian dapat memberikan pembelajaran dan pengetahuan secara umum pada publikasi artikel ini.

REFERENSI

[1] M. A. Firmansyah, D. Mulyana, S. Karlinah, and S. Sumartias, "Kontestasi Pesan Politik dalam Kampanye Pilpres 2014 di Twitter: Dari Kultwit Hingga Twitwar," *JIK*, vol. 16, no. 1, p. 42, Jan. 2018, doi: 10.31315/jik.v16i1.2681.

[2] A. Septiana, "Analisis Fungsi Partai Politik Pada Pilkada Musi Banyuasin 2017 (Studi Terhadap Partai Politik Pengusung Pasangan Dodi Reza Dan Beni Hernedi)," *jssp*, vol. 3, no. 1, pp. 28–41, Jun. 2019, doi: 10.19109/jssp.v3i1.4066.

[3] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, "Who says what to whom on twitter," in *Proceedings of the 20th international conference on World wide web*, Hyderabad India: ACM, Mar. 2011, pp. 705–714. doi: 10.1145/1963405.1963504.

[4] P. Patmawati and M. Yusuf, "Analisis Topik Modelling Terhadap Penggunaan Sosial Media Twitter oleh Pejabat Negara," *bits*, vol. 3, no. 3, pp. 122–129, Dec. 2021, doi: 10.47065/bits.v3i3.1012.

[5] I. W. D. Gafatia and N. Hadinata, "Analisis Pro Kontra Vaksin Covid 19 Menggunakan Sentiment Analysis Sumber Media Sosial Twitter," *JPSII*, vol. 2, no. 1, pp. 34–42, Nov. 2021, doi: 10.47747/jpsii.v2i1.544.

[6] M. D. Devika, C. Sunitha, and A. Ganesh, "Sentiment Analysis: A Comparative Study on Different Approaches," *Procedia Computer Science*, vol. 87, pp. 44–49, 2016, doi: 10.1016/j.procs.2016.05.124.

[7] I. Kurniawan and A. Susanto, "Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019," *Jurnal Eksplora Informatika*, vol. 9, no. 1, pp. 1–10, Sep. 2019, doi: 10.30864/eksplora.v9i1.237.

[8] E. I. Saptanti, "Analisis Manajemen Impresi Ma'ruf Amin dalam Debat Pilpres 2019," *ULTIMA Comm*, vol. 12, no. 2, pp. 262–284, Dec. 2020, doi: 10.31937/ultimacomm.v12i2.1573.

[9] Y. Sahria, "Implementasi Teknik Web Scraping pada Jurnal SINTA Untuk Analisis Topik Penelitian Kesehatan Indonesia," 2020.

[10] N. L. P. M. Putu, Ahmad Zuli Amrullah, and Ismarmiaty, "Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation," *RESTI*, vol. 5, no. 1, pp. 123–131, Feb. 2021, doi: 10.29207/resti.v5i1.2587.

[11] H. Annur, "Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes," *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 160–165, Aug. 2018, doi: 10.33096/ilkom.v10i2.303.160-165.
[12] D. T. Anggraeni, "FORECASTING HARGA SAHAM MENGGUNAKAN METODE SIMPLE MOVING AVERAGE DAN WEB SCRAPPING," *jurnalmatrik*, vol. 21, no. 3, pp. 234–241, Dec. 2019, doi: 10.33557/jurnalmatrik.v21i3.726.
[13] D. Gunawan, "Metode Klasifikasi pada Data Preprocessing Data," no. 1, 2016.
[14] D. Darwis, E. S. Pratiwi, and A. F. O. Pasaribu, "PENERAPAN ALGORITMA SVM UNTUK ANALISIS SENTIMEN PADA DATA TWITTER KOMISI PEMBERANTASAN KORUPSI REPUBLIK INDONESIA," *Edutic*, vol. 7, no. 1, Nov. 2020, doi: 10.21107/edutic.v7i1.8779.
[15] S. Raschka, "Naive Bayes and Text Classification I - Introduction and Theory." arXiv, Feb. 14, 2017. Accessed: May 25, 2023. [Online]. Available: http://arxiv.org/abs/1410.5329

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.