

## **ANALYSIS AND IMPLEMENTATION OF SENTIMENT SYSTEM ON THE ELECTABILITY OF INDONESIAN PRESIDENTIAL CANDIDATES 2024 USING SUPPORT VECTOR MACHINE METHOD**

Jasmine Avrile Kaniyasari Harahap<sup>\*1</sup>, Wahyu Syaifullah JS<sup>2</sup>, Mohammad Idhom<sup>3</sup>

<sup>1,2,3</sup>Data Science, Faculty of Computer Science, Universitas Pembangunan Nasional “Veteran” Jawa Timur, Indonesia  
Email: <sup>1</sup>[20083010012@student.upnjatim.ac.id](mailto:20083010012@student.upnjatim.ac.id), <sup>2</sup>[wahyu.s.j.saputra.if@upnjatim.ac.id](mailto:wahyu.s.j.saputra.if@upnjatim.ac.id), <sup>3</sup>[idhom@upnjatim.ac.id](mailto:idhom@upnjatim.ac.id)

(Article received: May 01, 2024; Revision: May 30, 2024; published: July 29, 2024)

### **Abstract**

Indonesia is a country that implements democracy in choosing presidential candidates through the election process. People have their own views on the presidential candidates they support, and in this digital era, social media is the main platform for people to express their opinions. Public opinion can be positive or negative, public opinion, hate speech, and various other comments that can cause hostility, insults, debates, and disputes. In this study, data modeling using the Support Vector Machine (SVM) method will be evaluated using a confusion matrix. The data used for Anies data is 1607 tweets, Prabowo data is 1761 tweets, and Ganjar data is 1607 tweets with the keywords “anies baswedan”, “prabowo subianto”, and “ganjar pranowo” with the data collection period from November - December 2023. The results of this study show that the sentiment classification model has good performance. For Anies Baswedan data, the SVM model achieved accuracy of 86.64%, precision of 86.69%, recall of 86.64%, and f1-score of 86.62%. For Prabowo Subianto data, the model achieved an accuracy of 90.65%, precision of 90.81%, recall of 90.65%, and f1-score of 90.61%. Meanwhile, for Ganjar Pranowo data, the model achieved an accuracy of 93.78%, precision of 93.67%, recall of 93.78%, and f1-score of 93.72%. These results show that the system is able to classify people's sentiment.

**Keywords:** Election, Presidential Candidates, Sentiment Analysis, Support Vector Machine.

## **ANALISIS DAN IMPLEMENTASI SISTEM SENTIMEN TERHADAP ELEKTABILITAS CALON PRESIDEN INDONESIA 2024 MENGGUNAKAN METODE SUPPORT VECTOR MACHINE**

### **Abstrak**

Indonesia adalah negara yang menerapkan demokrasi dalam memilih calon presiden melalui proses pemilu. Masyarakat memiliki pandangan tersendiri mengenai calon presiden yang didukung, dan pada era digital ini, media sosial menjadi wadah utama bagi masyarakat untuk menyampaikan opini. Opini masyarakat dapat berupa positif atau negatif, penggiringan opini publik, ujaran kebencian, serta beragam komentar lainnya yang dapat menimbulkan permusuhan, penghinaan, perdebatan, dan perselisihan. Dalam penelitian ini, pemodelan data menggunakan metode Support Vector Machine (SVM) yang akan dievaluasi menggunakan confusion matrix. Data yang digunakan untuk data Anies sebanyak 1607 tweet, data Prabowo 1761 tweet, dan data Ganjar 1607 tweet dengan kata kunci “anies baswedan”, “prabowo subianto”, dan “ganjar pranowo” dengan periode pengumpulan data dari bulan November - Desember 2023. Hasil dari penelitian ini menunjukkan bahwa model klasifikasi sentimen memiliki performa yang baik. Untuk data Anies Baswedan, model SVM mencapai akurasi sebesar 86.64%, presisi sebesar 86.69%, recall sebesar 86.64%, dan f1-score sebesar 86.62%. Untuk data Prabowo Subianto, model mencapai akurasi sebesar 90.65%, presisi sebesar 90.81%, recall sebesar 90.65%, dan f1-score sebesar 90.61%. Sedangkan untuk data Ganjar Pranowo, model mencapai akurasi sebesar 93.78%, presisi sebesar 93.67%, recall sebesar 93.78%, dan f1-score sebesar 93.72%. Hasil ini menunjukkan bahwa sistem mampu mengklasifikasikan sentimen masyarakat dengan tingkat akurasi yang tinggi dan dapat diimplementasikan dalam sebuah aplikasi website untuk memantau opini publik terkait pemilu 2024.

**Kata kunci:** Analisis Sentimen, Calon Presiden, Pemilu, Support Vector Machine.

## 1. PENDAHULUAN

Indonesia merupakan salah satu negara yang menganut sistem demokrasi. Tahun politik menjadi sebuah tahun yang seringkali ditunggu oleh sebagian masyarakat Indonesia karena ketika memasuki tahun politik akan banyak fenomena atau peristiwa politik yang unik terjadi, salah satunya adalah komentar atau penyampaian opini yang beragam dari masyarakat di media sosial. Seorang politikus yang mencalonkan diri sebagai calon presiden Indonesia di pemilu, tentunya mempertimbangkan tingkat popularitas melalui opini atau komentar masyarakat. Popularitas tersebut didukung oleh paparan media, baik yang bersifat publik maupun swasta. Mereka memanfaatkan berbagai teknik untuk meningkatkan popularitas dan mencapai tingkat keberhasilan yang tinggi di masyarakat [1].

Berdasarkan hal tersebut, muncul beberapa survei elektabilitas terkait dengan tokoh masyarakat dengan kualifikasi yang baik, sehingga memungkinkan akan menjadi calon presiden Indonesia dalam pemilihan presiden tahun 2024. Beberapa lembaga survei seperti Lembaga Survei Indonesia (LSI), Saiful Mujani Research & Consulting (SMRC), Indonesia Political Opinion (IPO), dan Lembaga Survei Nasional (LSN) menunjukkan bahwa terdapat tiga tokoh publik dengan nilai elektabilitas tertinggi, yaitu Prabowo Subianto, Ganjar Pranowo, dan Anies Baswedan [2]–[4]. Menurut hasil survei LSN yang melibatkan sebanyak 1.420 responden, Prabowo Subianto memperoleh elektabilitas sebesar 38,5%, diikuti oleh Ganjar Pranowo dengan 32,8%, dan Anies Baswedan dengan 21,9% dalam tingkat elektabilitas nasional [5]. Hal tersebut memungkinkan masyarakat untuk menyampaikan komentar atau opini publik terkait dengan hasil survei pada pemilu 2024 dan baris berita yang menerbitkan berita di media sosial [6].

Berdasarkan fenomena pemilu pada tahun-tahun sebelumnya, tahun politik akan identik dengan adanya acara debat politik, dimana pada acara tersebut mengundang bakal calon presiden Indonesia tahun 2024, yakni Ganjar Pranowo, Prabowo Subianto, dan Anies baswedan untuk melakukan adu gagasan dari ketiga bakal calon presiden tersebut. Hal ini pun mendapatkan sambutan publik yang sangat meriah serta ramai diperbincangkan oleh masyarakat di media sosial salah satunya adalah aplikasi Twitter. Pada platform media sosial tersebut terdapat fitur trending topics, fitur ini berguna untuk melihat fenomena yang sedang ramai diperbincangkan oleh netizen.

Berdasarkan pada peristiwa tersebut, timbulah komentar-komentar dari masyarakat mengenai ketiga bakal calon presiden Indonesia, terdapat beragam komentar di media sosial khususnya Twitter, dalam komentar tersebut mengandung pro kontra terhadap ketiga bakal calon presiden tersebut, penggiringan opini publik baik itu positif ataupun negatif, ujaran

kebencian terhadap ketiga calon presiden tersebut, serta komentar-komentar beragam lainnya yang dapat menimbulkan permusuhan, penghinaan, perdebatan, dan perselisihan di Twitter. Sejumlah komentar yang berada di media sosial Twitter tersebut, terdapat banyak sekali informasi didalamnya untuk dapat dijadikan sebagai penelitian dengan tema analisis sentimen. Analisis sentimen dari komentar-komentar masyarakat di Twitter terhadap elektabilitas calon presiden Indonesia tersebut dinilai penting karena dapat digunakan untuk melihat mana calon presiden yang memiliki image positif dan negatif di mata masyarakat, dapat mempengaruhi pilihan masyarakat dan menentukan hasil akhir dari pemilu 2024.

Berdasarkan pada latar belakang di atas, perlu dilakukan sebuah analisis terhadap teks tersebut, yakni dengan melakukan analisis sentimen. Analisis sentimen atau dapat disebut dengan opinion mining merupakan bidang studi yang menganalisis pendapat, sentimen, evaluasi, sikap, dan emosi seseorang terhadap entitas sebuah produk, layanan, individu, organisasi, peristiwa, masalah, topik dan atribut lainnya [7]. Pada penelitian ini, penulis akan menggunakan salah satu pendekatan dalam analisis sentimen dengan menggunakan Support Vector Machine yang dapat memprediksi sentimen teks dengan mengklasifikasikan teks kedalam sentimen positif, dan negatif.

Penelitian yang serupa dilakukan oleh Darwis mengenai Analisis Sentimen Twitter Terhadap Wacana Penundaan Pemilu Dengan Metode Support Vector Machine. Data sampel yang digunakan 100 data tweet dengan pengelompokan sentimen positif sebanyak 40 data tweets dan sentimen negatif sebanyak 60 data tweets. Hasil pengujian klasifikasi pada data tweets dengan tingkat akurasi yang baik sebesar 92% [8]. Penelitian Rian Tinenger mengenai Analisis Sentimen Layanan pada Indihome dari Twitter menggunakan algoritma SVM dengan 280 data tweet, menghasilkan 87% akurasi, 86% presisi, 95% *recall*, 90% *f1-score* [9]. Penelitian Amar mengenai Analisis Sentimen keberlanjutan belajar daring dengan 200 data tweet memakai metode SVM dan pemilihan Fitur TF-IDF evaluasi tertinggi dihasilkan pada 8-Fold Cross Validation dengan hasil accuracy 86,00%, precision,38%, dan recall 85,02% [10].

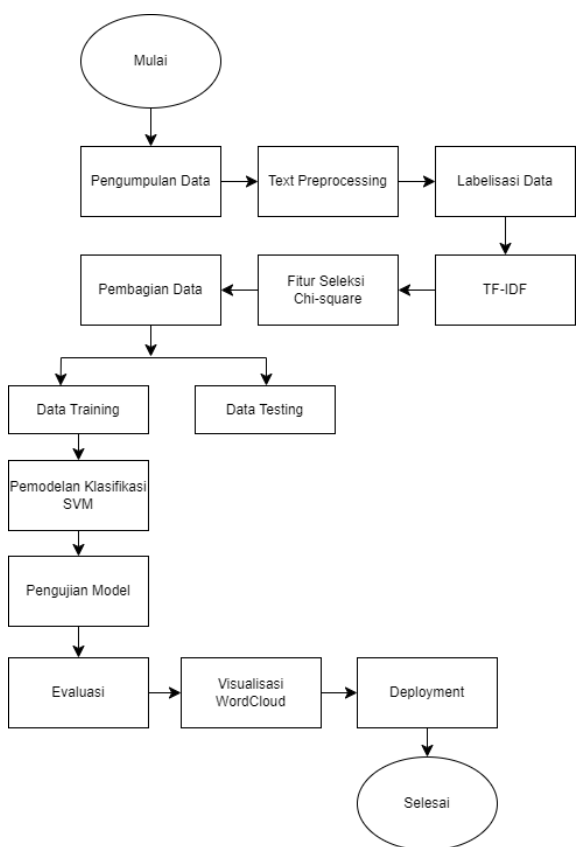
Dalam hal ini, penerapan algoritma SVM dalam melakukan analisis sentimen menggunakan data yang diperoleh dari media sosial Twitter, bertujuan untuk melakukan klasifikasi post atau tweet dari pengguna media sosial Twitter kedalam sentimen positif, dan negatif. SVM merupakan algoritma machine learning yang sering kali digunakan untuk melakukan klasifikasi dan memberikan hasil yang cukup baik dalam riset analisis [11]. Dengan melakukan pemilihan fungsi kernel yang berbeda pada algoritma SVM diharapkan dapat meningkatkan akurasi pada karakteristik data tertentu [12]. Pemilihan fungsi kernel yang tepat menjadi faktor penting dalam

membangun model klasifikasi untuk mengidentifikasi sentimen positif dan negatif dari pengguna media sosial Twitter [13].

Berdasarkan penjelasan di atas, penelitian ini akan mengungkap bagaimana analisis sentimen menggunakan SVM untuk mengklasifikasikan pendapat masyarakat terhadap calon presiden Indonesia 2024. Tujuan dari penelitian ini adalah untuk mengklasifikasikan komentar masyarakat dari twitter ke dalam sentimen positif dan negatif serta mengimplementasikannya ke dalam sebuah sistem berbasis website.

## 2. METODE PENELITIAN

Pada penelitian ini, data yang akan digunakan adalah data primer (data yang diambil secara langsung) yang didapatkan dari scrapping media sosial Twitter.



Gambar 1. Flowchart penelitian

Berikut merupakan tahapan yang dilakukan pada penelitian ini yang terdapat pada Gambar 1, yakni:

### 1. Pengumpulan data

Pada tahap ini dilakukan pengambilan data di media sosial Twitter dengan menggunakan library *Tweet Harvest*. Data yang diambil adalah data tweet atau komentar masyarakat terkait dengan bakal calon presiden Indonesia 2024. Hasil dari pengambilan data tersebut nantinya akan berupa file dengan ekstensi CSV. Terdapat tiga data yang akan dikumpulkan,

yaitu data calon presiden Anies Baswedan, Prabowo Subianto, dan Ganjar Pranowo.

### 2. Text Preprocessing

*Text preprocessing* merupakan proses yang dilakukan agar dapat meningkatkan kualitas pada teks atau dalam melakukan penyeleksian data *text* agar menghilangkan *noise* pada data [14].

#### a. Cleaning Text

*Cleaning text* bertujuan untuk membersihkan atau menghapus karakter, tanda baca ataupun komponen-komponen yang tidak berpengaruh terhadap proses klasifikasi sentimen.

#### b. Case folding

Proses *case folding* adalah mengubah kalimat dalam dataset yang memiliki huruf kapital menjadi huruf kecil agar dapat terdeteksi dengan setara atau arti yang sama.

#### c. Tokenization

Proses *tokenization* adalah proses yang dilakukan dengan mengubah kalimat menjadi potongan-potongan kata untuk diproses lebih lanjut.

#### d. Normalisasi

Tahapan normalisasi dilakukan dengan mengubah kata tidak baku atau *slang word* menjadi kata baku. Selain itu, kata dengan salah penulisan, seperti penyingkatan kata juga akan diperbaiki dalam proses normalisasi.

#### e. Stopword Removal (Filtering)

Pada tahapan *stopword removal* atau *filtering* dilakukan penghilangan kata-kata yang tidak memiliki makna. Hal ini dilakukan agar sistem dapat terfokus pada kata-kata yang lebih memiliki makna dalam proses klasifikasi. Hasil dari tahapan ini akan mengubah jumlah kata pada kalimat menjadi lebih sedikit.

#### f. Stemming

Tahapan *stemming* akan dilakukan perubahan dari kata yang memiliki imbuhan menjadi bentuk kata dasar.

### 3. Labelisasi Data

Labelisasi adalah proses pemberian terhadap kelas berdasarkan dari karakteristik yang terdapat dalam dokumen atau kalimat [15]. Tahap labelisasi pada data dengan menggunakan kamus Lexicon. Kamus Lexicon yang akan digunakan adalah kamus Lexicon InSet (Indonesia Sentiment Lexicon). InSet berasal dari sekumpulan kata-kata berbahasa Indonesia yang memiliki bobot nilai -5 hingga +5, InSet berisikan 3.609 kata positif dan 6.609 kata negatif. Hasil dari tahap ini adalah sebuah data yang sudah memiliki label positif dan negatif.

### 4. Fitur Ekstraksi TF-IDF

Term Frequency (TF) adalah salah satu metode pembobotan term yang paling sederhana, juga dikenal sebagai bag of words. TF mengukur frekuensi kemunculan suatu term dalam sebuah dokumen. Untuk mengurangi skala angka yang besar, frekuensi ini kemudian dilogaritmakan, sehingga menghasilkan nilai yang lebih kecil dan lebih mudah diolah.



### 3.2. Text Preprocessing

#### a. Cleaning Text

Pada proses cleaning text dilakukan pembersihan atau penghapusan function pada teks yang meliputi tanda baca, simbol-simbol yang terdapat dalam teks, seperti "\$%&'()\*+,-./:;<=>?@[\\]^\_`{|}~", penghapusan karakter HTML (link), hashtag (“#”), dan URL (Uniform Resource Locator). Berikut merupakan *syntax* untuk melakukan proses cleaning text, dapat dilihat pada Gambar 4.

```

# Preprocessing Cleaning
df['tweet_clean'] = df['tweet'].apply(lambda x: clean_text(x))

def clean_text(text):
    # Remove HTML tags
    text = re.sub('<.*>', '', text)
    # Remove URLs
    text = re.sub('https?://.*', '', text)
    # Remove hashtags
    text = re.sub('#.*', '', text)
    # Remove punctuation
    text = re.sub('[!@#$%^&*()_+~\.\:;<=>?@[\]^_`{|}~]', '', text)
    # Lowercase
    text = text.lower()
    return text

df['tweet_clean'] = df['tweet'].apply(lambda x: clean_text(x))
    
```

Gambar 4. Syntax cleaning text

#### b. Case folding

Tahapan ini menggunakan fungsi `str.lower()` yang terdapat dalam library Pandas, bertujuan untuk mengubah huruf kapital pada atribut ‘tweet\_clean’ ke dalam bentuk huruf kecil agar dapat dideteksi dengan arti yang sama atau setara. Berikut merupakan *syntax* untuk melakukan proses case folding, dapat dilihat pada Gambar 5.

```

# Case Folding
df['case_folding'] = df['tweet_clean'].str.lower()

df.loc[0:1607, ['tweet_clean', 'case_folding']]
    
```

	tweet_clean	case_folding
0	Genjar sudah mulai memantau IKN dimana Genjar ...	genjar sudah mulai memantau ikn dimana genjar ...
1	Ini dia nih jagoan gw yg akan menempati Istana.	ini dia nih jagoan gw yg akan menempati istana.
2	Genjar tuh bebas ngak kayak yang pencari cel...	genjar tuh bebas ngak kayak yang pencari cel...
3	Genjar mahud bukan hanya jang kosong telap...	genjar mahud bukan hanya jang kosong telap...
4	Warga Jatim mendukung Genjar!Mahud karena mere...	warga jatin mendukung genjarmahud karena mere...
...	...	...
1602	Keberhasilan di Jatim adalah bukti bahwa Pak G...	keberhasilan di jatin adalah bukti bahwa pak g...
1603	Proses yang dijalani oleh Pak Genjar adalah bu...	proses yang dijalani oleh pak genjar adalah bu...
1604	Genjar Pranowo adalah pemimpin yang mampu meng...	genjar pranowo adalah pemimpin yang mampu meng...

Gambar 5. Syntax case folding

#### c. Tokenization

Proses tokenization bertujuan untuk untuk mengubah kalimat pada atribut ‘case\_folding’, menjadi potongan-potongan kata. Hal ini dilakukan agar dapat mempermudah proses selanjutnya. Berikut merupakan *syntax* untuk melakukan proses tokenization, dapat dilihat pada Gambar 6.

```

# Import word_tokenize & FreqDist from NLTK
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist

# ----- Tokenizing -----

# NLTK word tokenize
def word_tokenize_wrapper(text):
    return word_tokenize(text)

df['tokenize'] = df['case_folding'].apply(word_tokenize_wrapper)
df.loc[0:1608, ['case_folding', 'tokenize']]
    
```

Gambar 6. Syntax tokenization

#### d. Normalisasi

Dalam penelitian ini penulis melakukan tahapan normalisasi menggunakan kamus normalisasi yang bersumber dari github repository kuncachyosetyonugroho dengan nama ‘key\_norm.csv’, dimana dalam kamus normalisasi tersebut terdapat sebanyak 3723 kata, kemudian penulis melakukan penambahan kata pada kamus tersebut, sehingga menjadi 3779 kata dan disimpan dengan nama ‘key\_normalisasi.xlsx’. Berikut merupakan *syntax* untuk melakukan proses normalisasi, dapat dilihat pada Gambar 7 dan Gambar 8.

```

[10] import urllib.request

url = "https://raw.githubusercontent.com/knugroho/klasifikasi-spam-sm/master/data/key_norm.csv"
filename = "key_norm.csv"

urllib.request.urlretrieve(url, filename)

('key_norm.csv', <http.client.HTTPMessage at 0x7e9746ea25c0>)

key_norm = pd.read_csv('key_norm.csv')
print(key_norm.head())

key_norm.shape

   _id  singkat  hasil
0    1    abis    habis
1    2  accent  tekanan
2    3  accept  terima
3    4  accident kecelakaan
4    5  achievement prestasi
(3728, 3)
    
```

Gambar 7. Download kamus key\_norm pada Github

Gambar 8 merupakan *syntax* untuk melakukan tahap normalisasi, file yang sudah di download pada Github dapat dimasukkan ke dalam *syntax* ini.

```

normalized_word = pd.read_csv("key_norm.csv")
normalized_word_dict = {}

for index, row in normalized_word.iterrows():
    if row[0] not in normalized_word_dict:
        normalized_word_dict[row[0]] = row[1]

def normalized_term(document):
    return [normalized_word_dict[term] if term in normalized_word_dict else term for term in document]

df['normalisasi'] = df['tokenize'].apply(normalized_term)

df.loc[0:1607, ['tokenize', 'normalisasi']]
    
```

Gambar 8. Syntax normalisasi

#### e. Stopword Removal (Filtering)

*Syntax* code untuk proses *filtering* atau *stopword removal* terdapat pada gambar 9 dimana pada gambar tersebut merupakan *syntax code* untuk melakukan import library menggunakan library NLTK. Berikut merupakan *syntax* untuk melakukan proses *stopword removal (filtering)*, dapat dilihat pada Gambar 9.

```

[15] from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.corpus import stopwords

list_stopwords = stopwords.words("Indonesian")

len(list_stopwords)

758

list_stopwords[:10]

['ada',
 'adalah',
 'adanya',
 'adapun',
 'agak',
 'agaknya',
 'agar',
 'akan',
 'akankah',
 'akhir']

[21] from nltk.corpus import stopwords
    
```

Gambar 9. Import library NLTK

Gambar 10 merupakan *syntax code* untuk melakukan proses *stopword removal* dengan menambahkan beberapa kata tambahan yang akan

digunakan dalam proses tersebut. Hasil dari proses stopword removal terdapat pada Gambar 10.

```

from nltk.corpus import stopwords

# ----- get stopwords from NLTK stopwords -----
# get stopwords Indonesia
list_stopwords = stopwords.words('Indonesian')

# ----- manually add stopwords -----
# append additional stopwords
list_stopwords.extend(['yg', 'dg', 'nt', 'dgn', 'ny', 'd', 'klo', 'boleh',
                      'kalo', 'amp', 'bian', 'bikin', 'bilang', 'sekali',
                      'gak', 'ga', 'kmn', 'nya', 'nh', 'sini', 'saja',
                      'si', 'tau', 'tdk', 'tuh', 'utk', 'ya', 'begitu',
                      'jd', 'jgn', 'sdh', 'aja', 'n', 't', 'nggak',
                      'nyg', 'hehe', 'pen', 'u', 'nan', 'loh', 'rt',
                      '&amp;', 'yah', 'gw', 'lu', 'lo', 'gtw', 'bukan',
                      'lyaa', 'si', 'ruamm', 'itu', 'gue', 'dan', 'juga',
                      'ca', 'can', 'emg', 'mau', 'kall', 'jadi', 'begini',
                      'billa', 'nanti'])

# ----- add stopwords from txt file -----
# read txt stopwords using pandas
txt_stopword = pd.read_csv('stopwords.txt', names=['stopwords'], header = None)

# convert stopwords string to list & append additional stopwords
list_stopwords.extend(txt_stopword['stopwords'][0].split(' '))
    
```

Gambar 10. Syntax stopwords removal

f. *Stemming*

Dalam tahap ini, digunakan library Sastrawi yang memproses bahasa alami (Natural Language Processing) untuk bahasa Indonesia. Berikut merupakan *syntax* untuk melakukan proses stemming, dapat dilihat pada Gambar 11.

```

▼ Stemming

[ ] from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
import pandas as pd

# Create stemmer
factory = StemmerFactory()
stemmer = factory.create_stemmer()

# Stemmed wrapper function
def stemmed_wrapper(term):
    return stemmer.stem(term)

# Create a dictionary for unique terms
term_dict = {}
for document in df['stopwords']:
    for term in document:
        if term not in term_dict:
            term_dict[term] = ''

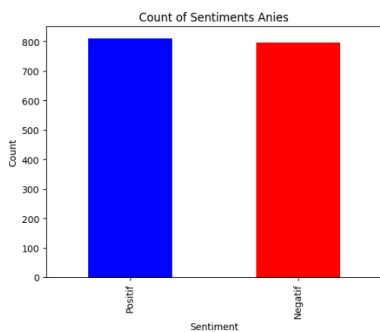
# Apply stemming to unique terms
for term in term_dict:
    term_dict[term] = stemmed_wrapper(term)

# Annotate stemmed term to the DataFrame
    
```

Gambar 11. Syntax stemming

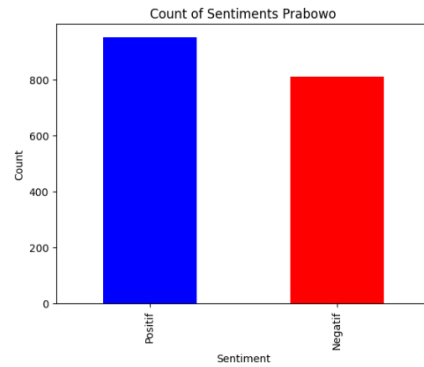
3.3. Labelisasi Data

Metode pelabelan dengan menggunakan Lexicon InSet ini dengan cara apabila kata dalam dokumen tersebut memiliki polarity score > 0 maka akan masuk ke dalam label positif, namun bila kata dalam dokumen tersebut memiliki polarity score < 0 maka akan masuk ke dalam label negatif. Gambar 12 menunjukkan visualisasi untuk data Anies yang sudah diberi label terdapat sebanyak 810 data dengan label positif dan sebanyak 797 data dengan label negatif.



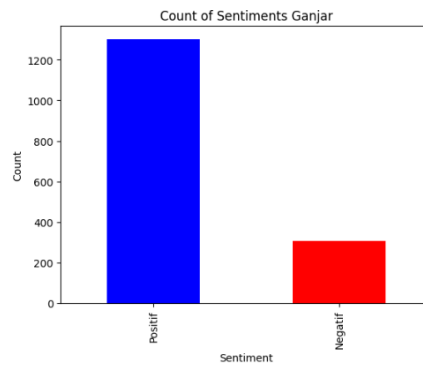
Gambar 12. Hasil sentimen dataset Anies

Gambar 13 menunjukkan visualisasi untuk data Prabowo yang sudah diberi label terdapat sebanyak 951 data dengan label positif dan 810 data dengan label negatif.



Gambar 13. Hasil sentimen dataset Prabowo

Gambar 14 menunjukkan hasil untuk data Ganjar yang sudah diberi label terdapat sebanyak 1301 data dengan label positif dan 306 data dengan label negatif.



Gambar 14. Hasil sentimen dataset Ganjar.

3.4. Fitur Ekstraksi TF-IDF

Fitur ekstraksi atau dapat dikatakan sebagai pembobotan kata dilakukan untuk membobot kata yang terdapat dalam dataset agar dapat melihat seberapa penting kata/*term* tersebut. Tahapan fitur ekstraksi dilakukan dengan menggunakan metode TF-IDF (*Term Frequency Inverse Document Frequency*) yang dapat membantu dalam proses evaluasi pentingnya kata atau term terkait dengan sentimen yang terkandung di dalam dokumen. Sehingga, dapat melihat seberapa banyak fitur yang akan digunakan. Hasil perhitungan dari pembobotan kata didapatkan dari kedua nilai yaitu nilai TF (*Term Frequency*) dan IDF (*Inverse Document Frequency*). Perhitungan bobot TF-IDF penelitian menggunakan *library sklearn* untuk menjalankan perhitungan pada proses TF-IDF dengan mengimport *TfidfVectorizer*, perhitungan ini dilakukan terhadap dataset dari kolom 'sebelum' yang terdapat dalam Tabel 1.

Tabel 1. Data sampel perhitungan TF IDF

Doc	Sebelum	Sesudah
D1	@kumparan Pak anies bisa jadi inspirasi buat semula	['anies', 'inspirasi']
D2	@aniesbaswedan Apresiasi bulat Anies yang bikin pasar tradisional lebih nyaman dan kekinian!	['apresiasi', 'anies', 'pasar', 'tradisional', 'nyaman', 'kini']
D3	@Meltro_TV Yang mau perubahan udah pasti milih Anies 😊	['ubah', 'milih', 'anies']
D4	@BangPino_ Tetap celalu dukung pak Anies	['dukung', 'anies']

Berikut merupakan hasil perhitungan TF-IDF untuk menghitung bobot setiap kata dari kelima sampel data yang ditunjukkan pada Tabel 2. Pada tabel 2 dapat dilihat bahwa hasil dari perhitungan TF-

IDF kelima sampel data yang ada pada tabel 1 yaitu terdapat angka 0.23299 merupakan nilai dari pembobotan terhadap kata dengan TF-IDF.

Tabel 2. Hasil pembobotan TF-IDF

Kata	TF-IDF				
	D1	D2	D3	D4	D5
anies	0	0	0	0	0
apresiasi	0.23299	0	0	0	0
bangsa	0	0	0	0	0.1165
cak	0	0	0	0	0.1165
dukung	0	0	0	0.34949	0
harap	0	0	0	0	0.1165
inspirasi	0.23299	0	0	0	0
imin	0	0	0	0	0.1165
kini	0	0.13979	0	0	0
nyaman	0	0.13979	0	0	0
milih	0	0	0.23299	0	0
pasar	0	0.13979	0	0	0
tradisional	0	0.13979	0	0	0
ubah	0	0	0.13265	0	0.06632

### 3.5. Fitur Seleksi Chi-Square

Penelitian ini menggunakan nilai taraf nyata/alpha ( $\alpha$ ) yakni 0.01, 0.025, 0.05, 0.1, 0.2, 0.4, dan 0.5. Kemudian peneliti mencoba menggunakan ketujuh nilai taraf nyata/alpha ( $\alpha$ ) tersebut ke dalam penelitian ini dan digunakan untuk melihat tingkat hasil performa dari ketujuh nilai taraf nyata/alpha ( $\alpha$ ) tersebut. Berdasarkan hasil evaluasi terhadap beberapa eksperimen, yaitu perbandingan performa pada hasil seleksi fitur, peneliti memutuskan untuk memilih hasil evaluasi pada eksperimen dengan nilai taraf nyata/alpha ( $\alpha$ ) 0.4 karena terdapat beberapa faktor, yakni dilihat dari nilai akurasi sebesar 86.64%, dan nilai *recall* sebesar 86.64%. Hasil evaluasi menunjukkan bahwa hasil seleksi fitur pada eksperimen dengan nilai taraf nyata/alpha ( $\alpha$ ) 0.4, pemilihan nilai taraf nyata/alpha ( $\alpha$ ) yakni dilihat dari nilai akurasi sebesar 90.65% dan nilai *recall* sebesar 90.65%. Hasil evaluasi selanjutnya menunjukkan bahwa hasil seleksi fitur pada eksperimen dengan nilai taraf nyata/alpha ( $\alpha$ ) 0.4, pemilihan nilai taraf nyata/alpha ( $\alpha$ ) yakni dilihat dari nilai akurasi sebesar 93.78% dan nilai *recall* sebesar 93.78%.

### 3.6. Klasifikasi Model SVM

Sebelum proses klasifikasi dilakukan pembagian data terlebih dahulu. Pembagian data dilakukan dengan cara membagi data menjadi dua kategori yaitu data latih dan data uji. Skenario pertama adalah pembagian data 70:30, skenario kedua adalah pembagian data 80:20, skenario ketiga adalah pembagian data 90:10. Ketiga skenario ini

dilakukan pada dataset Anies, dataset Prabowo, dan dataset Ganjar.

Tabel 3 merupakan pembagian data dari dataset anies dengan proporsi data 70:30.

Tabel 3. Pembagian data latih 70:30 data anies

	Positif	Negatif	Jumlah Data
Data Latih	559	565	1124
Data Uji	251	232	483
Jumlah Data Anies	810	797	1607

Tabel 4 merupakan pembagian data dari dataset anies dengan proporsi data 80:20.

Tabel 4. Pembagian data 80:20 data anies

	Positif	Negatif	Jumlah Data
Data Latih	643	642	1285
Data Uji	167	155	322
Jumlah Data Anies	810	797	1607

Tabel 5 merupakan pembagian data dari dataset anies dengan proporsi data latih 90:10.

Tabel 5. Pembagian data 90:10 data anies

	Positif	Negatif	Jumlah Data
Data Latih	734	712	1446
Data Uji	76	85	161
Jumlah Data Anies	810	797	1607

Tabel 6 merupakan pembagian data dari dataset prabowo dengan proporsi data 70:30.

Tabel 6. Pembagian data 70:30 data prabowo

	Positif	Negatif	Jumlah Data
Data Latih	666	566	1232
Data Uji	285	244	529
Jumlah Data Prabowo	951	810	1761

Tabel 7 merupakan pembagian data dari dataset prabowo dengan proporsi data 80:20.

Tabel 7. Pembagian data 80:20 data prabowo

	Positif	Negatif	Jumlah Data
Data Latih	760	648	1408
Data Uji	191	162	353
Jumlah Data Prabowo	951	810	1761

Tabel 8 merupakan pembagian data dari dataset prabowo dengan proporsi data 90:10.

Tabel 8. Pembagian data 90:10 data prabowo

	Positif	Negatif	Jumlah Data
Data Latih	854	730	1584
Data Uji	97	80	171
Jumlah Data Prabowo	951	810	1761

Tabel 9 merupakan pembagian data dari dataset ganjar dengan proporsi data 70:30.

Tabel 9. Pembagian data 70:30 data ganjar

	Positif	Negatif	Jumlah Data
Data Latih	895	229	1124
Data Uji	406	77	483
Jumlah Data Prabowo	1303	306	1607

Tabel 10 merupakan pembagian data dari dataset ganjar dengan proporsi data 80:20.

Tabel 10. Pembagian data 80:20 data ganjar

	Positif	Negatif	Jumlah Data
Data Latih	760	648	1408
Data Uji	191	162	353
Jumlah Data Prabowo	1303	306	1607

Tabel 11 merupakan pembagian data dari dataset ganjar dengan proporsi data 90:10.

Tabel 11. Pembagian data latih 90:10 data ganjar

	Positif	Negatif	Jumlah Data
Data Latih	854	730	1584
Data Uji	97	80	177
Jumlah Data Prabowo	1303	306	1607

Pada penelitian ini akan membandingkan tiga kernel, yakni kernel *linear*, *polynomial* dan *rbf*. Skenario untuk uji parameter yang akan digunakan adalah parameter *cost* (C) = [0.1, 1, 10, 100], dan parameter *gamma* ( $\gamma$ ) = [0.001, 0.01, 0.1, 1]. Nilai dari data anies yang peneliti dapatkan menunjukkan bahwa nilai taraf nyata/alpha terbaik adalah 0.4 dengan jumlah fitur sebanyak 617 dengan model kernel rbf dan nilai parameter *cost* (C) = 100 parameter *gamma* ( $\gamma$ ) = 0.01 dengan nilai akurasi sebesar 86.62%.

Nilai dari data prabowo menunjukkan bahwa nilai taraf nyata/alpha terbaik adalah 0.4 dengan jumlah fitur sebanyak 635 dengan model kernel rbf dan nilai parameter *cost* (C) = 100 dan parameter *gamma* ( $\gamma$ ) = 1 dengan nilai akurasi sebesar 90.61%. Nilai dari data ganjar menunjukkan bahwa nilai taraf nyata/alpha terbaik adalah 0.4 dengan jumlah fitur sebanyak 366 dengan model kernel rbf dan nilai parameter *cost* (C) = 10 dan parameter *gamma* ( $\gamma$ ) = 1 dengan nilai akurasi sebesar 93.72%.

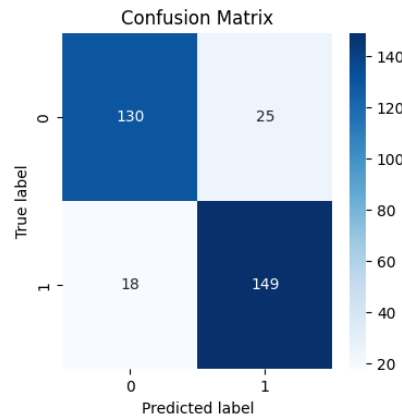
### 3.7. Evaluasi

Pengujian data Anies Pada tahapan ini didapatkan hasil dari classification report sebagai berikut. Berdasarkan pada Gambar 17 hasil dari classification report data anies didapatkan hasil nilai *accuracy* sebesar 87%, nilai *precision* sebesar 87%, nilai *recall* sebesar 87%, dan nilai F1-score sebesar 87%.

Classification report:

	precision	recall	f1-score	support
Negatif	0.88	0.84	0.86	155
Positif	0.86	0.89	0.87	167
accuracy			0.87	322
macro avg	0.87	0.87	0.87	322
weighted avg	0.87	0.87	0.87	322

Gambar 17. Classification report data Anies



Gambar 18. Confusion matrix data Anies

Berdasarkan pada Gambar 18 hasil dari confusion matrix data anies menampilkan hasil keseluruhan confusion matrix seperti di bawah ini.

- Label aktual negatif yang diprediksi negatif terdiri dari 130 data
- Label aktual positif yang diprediksi negatif terdiri dari 18 data
- Label aktual negatif yang diprediksi positif terdiri dari 25 data
- Label aktual positif yang diprediksi positif terdiri dari 149 data

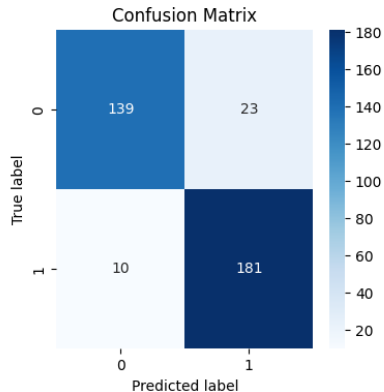
Berdasarkan pada Gambar 19 hasil dari classification report data prabowo didapatkan hasil nilai *accuracy* sebesar 91%, nilai *precision* sebesar 91%, nilai *recall* sebesar 91%, dan nilai F1-score sebesar 91%.

Classification report:

	precision	recall	f1-score	support
Negatif	0.93	0.86	0.89	162
Positif	0.89	0.95	0.92	191
accuracy			0.91	353
macro avg	0.91	0.90	0.91	353
weighted avg	0.91	0.91	0.91	353

Gambar 19. Classification report data Prabowo





Gambar 20. Confusion matrix data Prabowo

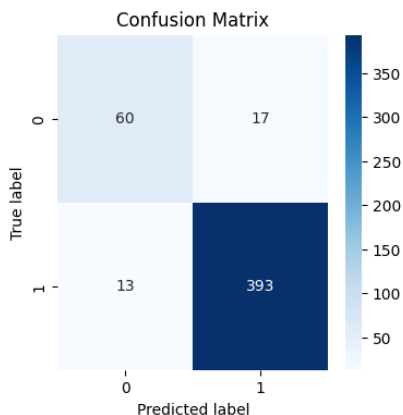
Berdasarkan pada Gambar 20 di atas hasil dari confusion matrix data prabowo menampilkan hasil keseluruhan confusion matrix seperti di bawah ini.

- Label aktual negatif yang diprediksi negatif terdiri dari 139 data
- Label aktual positif yang diprediksi negatif terdiri dari 10 data
- Label aktual negatif yang diprediksi positif terdiri dari 23 data
- Label aktual positif yang diprediksi positif terdiri dari 181 data

Berdasarkan pada Gambar 21 hasil dari classification report data ganjar didapatkan hasil nilai accuracy selbelsar 94%, nilai precision sebesar 94%, nilai recall sebesar 94%, dan nilai F1-score sebesar 94%.

Classification report:				
	precision	recall	f1-score	support
Negatif	0.82	0.78	0.80	77
Positif	0.96	0.97	0.96	406
accuracy			0.94	483
macro avg	0.89	0.87	0.88	483
weighted avg	0.94	0.94	0.94	483

Gambar 21. Classification report data Ganjar



Gambar 22. Confusion matrix data Ganjar

Berdasarkan pada Gambar 22 di atas hasil dari confusion matrix data anies menampilkan hasil keseluruhan confusion matrix seperti di bawah ini.

- Label aktual negatif yang diprediksi negatif terdiri dari 60 data

- Label aktual positif yang diprediksi negatif terdiri dari 13 data
- Label aktual negatif yang diprediksi positif terdiri dari 17 data
- Label aktual positif yang diprediksi positif terdiri dari 393 data.

### 3.8. Visualisasi WordCloud

Visualiasasi dilakukan untuk melihat gambaran terhadap hasil dari penelitian yang telah dilakukan, selain itu visualisasi juga dapat memudahkan pembaca dalam memahami hasil dari penelitian ini. Hasil penelitian akan menampilkan wordcloud dari sentimen terhadap ketiga calon presiden Indonesia 2024 serta dapat menampilkan informasi yang telah didapatkan dari penelitian ini. Berikut merupakan visualisasi wordcloud ketiga calon presiden Indonesia 2024 setelah preprocessing data.



Gambar 23. WordCloud data Anies

Pada Gambar 23 merupakan visulalisasi wordcloud dari data Aniels seltellah melalui tahapan preprocessing, tf-idf dan chi square. Belrdasarkan visualiasasi yang terdapat dalam Gambar 23 terdapat banyak kata yang sering muncul. Kata tersebut diantaranya adalah “dukung”, “pesona”, “pilih”, “baik”, “basi”, dan lainnya. Berdasarkan visualisasi tersebut dapat dilihat bahwa kata yang sering muncul dalam visualiasasi tersebut merupakan kata-kata umum yang memiliki arti atau makna yang dapat menggambarkan dengan jelas sentimen masyarakat di twitter terhadap salah satu calon presiden Indonesia.



Gambar 24. WordCloud data Prabowo

Gambar 24 menunjukkan visualisasi wordcloud dari data Prabowo setelah melalui tahap preprocessing, tf-idf, dan chi square. Dalam visualisasi ini, banyak kata sering muncul, seperti “panik”, “drama”, “tulus”, “positif”, “klaim”, dan lain-lain. Kata-kata ini memiliki arti atau makna yang jelas dan dapat menggambarkan sentimen masyarakat di Twitter terhadap salah satu calon presiden Indonesia.



Gambar 25. WordCloud data Ganjar

Pada gambar 25 menampilkan visualisasi wordcloud dari data Ganjar setelah melalui tahap preprocessing, tf-idf, dan chi square. Visualisasi tersebut menunjukkan bahwa semua kata yang muncul memiliki arti atau makna. Kata-kata yang muncul antara lain “pilih”, “bukti”, “jilat”, “tuntas”, “serius”, “inspirasi”, dan lain-lain. Dari visualisasi ini, dapat dilihat bahwa kata-kata yang sering muncul adalah kata-kata umum yang bermakna dan dapat menggambarkan dengan jelas sentimen masyarakat di Twitter terhadap salah satu calon presiden Indonesia.

### 3.9. Deployment

Deployment aplikasi menggunakan library Streamlit yang terdapat dalam Python. Implementasi model SVM menggunakan model SVM dengan melihat hasil dari confusion matrix, yakni nilai dari akurasi dan recall, kemudian model tersebut akan disimpan untuk digunakan pada tahap Deployment. Berikut adalah kode atau syntax untuk menyimpan model, dapat dilihat pada Gambar 26.

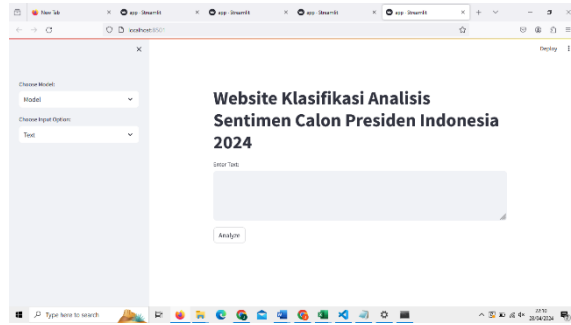
```

Export Model
[45] joblib.dump(model, 'svm_prabowo_2.pk1')
['svm_prabowo_2.pk1']
[46] joblib.dump(tf_idf, 'svm_prabowo_2.pk1')
['svm_prabowo_2.pk1']
    
```

Gambar 26. Syntax export model

Setelah model berhasil disimpan, maka disusun code script yang dapat membuat aplikasi berbasis website untuk melakukan prediksi sentimen komentar

masyarakat terhadap calon presiden Indonesia 2024. Setelah code script tersebut dibuat, maka Gambar 27 merupakan tampilan dari halaman deployment aplikasi.



Gambar 27. Tampilan halaman website

### 4. DISKUSI

Dari ketiga [8]–[10] penelitian terkait yang sudah dipaparkan pada bagian pendahuluan peneliti akan melakukan penelitian atau pengamatan yang sama, namun dengan kasus atau objek penelitian yang berbeda dan pengujian dengan banyak data yang berbeda pada algoritma yang dipilih, dengan tujuan sebagai perbandingan. Para peneliti melakukan penelitian analisis sentimen untuk mengetahui opini atau tanggapan seseorang terhadap topik tertentu, dengan melakukan klasifikasi algoritma untuk mengetahui nilai akurasi mengenai data topik yang dipilih, agar dapat mengetahui opini atau tanggapan masyarakat Indonesia terutama pada topik pemilu 2024 dari tweet media sosial twitter, peneliti melakukan penelitian analisis sentimen menggunakan algoritma Support Vector Machine dengan uji coba kernel dan uji coba pembagian proporsi data.

Analisis sentimen membantu untuk mengetahui tanggapan atau opini masyarakat Indonesia mengenai pemilu 2024 dapat dilihat pada gambar 15, 16, dan 17 hasil dari persentase kelas sentimen ketiga calon presiden Indonesia 2024. Persentase kelas sentimen positif lebih unggul. Hal ini menunjukkan bahwa masyarakat Indonesia menyambut baik adanya pemilu 2024 dengan memberikan tanggapan positif dan negatif terhadap calon presiden Indonesia 2024. Dan penelitian analisis sentimen ini, menghasilkan nilai akurasi yang cukup baik dari pengujian algoritma SVM kernel rbf, dengan evaluasi Confusion Matrix yaitu tahap pengujian untuk mendapatkan nilai precision, recall dan F1-score, untuk lebih jelasnya dapat dilihat pada hasil classification report.

### 5. KESIMPULAN

Hasil analisis sentimen masyarakat di media sosial Twitter terhadap elektabilitas calon presiden Indonesia 2024 menunjukkan data sebagai berikut: untuk Anies, terdapat 1607 tweet dengan 810 tweet positif dan 797 tweet negatif. Visualisasi wordcloud

setelah preprocessing menampilkan kata-kata yang menggambarkan sentimen masyarakat terhadap Anies, seperti “dukung”, “baik”, dan lainnya. Untuk Prabowo, terdapat 1761 tweet dengan 951 tweet positif dan 810 tweet negatif. Visualisasi wordcloud untuk Prabowo menampilkan kata-kata seperti “drama”, “panik”, dan lainnya. Sedangkan untuk Ganjar, terdapat 1607 tweet dengan 1301 tweet positif dan 306 tweet negatif. Visualisasi wordcloud untuk Ganjar menampilkan kata-kata seperti “jilat”, “inspirasi”, dan lainnya. Pengujian model klasifikasi sentimen menggunakan metode Support Vector Machine (SVM) menunjukkan hasil sebagai berikut: untuk data Anies, model mencapai akurasi 86.64%, presisi 86.69%, *recall* 86.64%, dan *f1-score* 86.62%, menggunakan kernel RBF dengan parameter *cost* (C) = 100 dan *gamma* ( $\gamma$ ) = 0.01, serta proporsi data 70:30. Untuk data Prabowo, model mencapai akurasi 90.65%, presisi 90.81%, *recall* 90.65%, dan *f1-score* 90.61%, menggunakan kernel RBF dengan parameter *cost* (C) = 100 dan *gamma* ( $\gamma$ ) = 1, serta proporsi data 80:20. Untuk data Ganjar, model mencapai akurasi 93.78%, presisi 93.67%, *recall* 93.78%, dan *f1-score* 93.72%, menggunakan kernel RBF dengan parameter *cost* (C) = 10 dan *gamma* ( $\gamma$ ) = 1, serta proporsi data 80:20.

#### DAFTAR PUSTAKA

- [1] T. R. Putra and A. Nurcholis, “Pengaruh Media Sosial terhadap Partisipasi Pemilih Pemula pada Pemilihan Presiden 2019: Studi pada Mahasiswa FISIPOL UGM,” *J. PolGov*, vol. 2, no. 1, 2020, doi: 10.22146/polgov.v2i1.1372.
- [2] LSI, “Rilis Survei Nasional Isu-Isu Nasional Dan Peta Kompetensi Pilpres & Pileg,” *Lemb. Surv. Indones.*, no. 40, pp. 3–9, 2023.
- [3] Saiful Mujani Research & Consulting (SMRC), *Trend Dukungan Kepada Bakal Calon Presiden*. 2023.
- [4] P. Opinion, “INDONESIA Peta Elektabilitas Koalisi , Capres,” 2024.
- [5] A. Setiawan and S. A. Girsang, “Elektabilitas Calon Presiden: Realitas Atau Penggiringan Opini?,” *J. Kaji. Budaya dan Hum.*, vol. 5, no. 2, pp. 221–227, 2023, doi: 10.61296/jkbh.v5i2.148.
- [6] M. R. Fais Sya’ bani, U. Enri, and T. N. Padilah, “Analisis Sentimen Terhadap Bakal Calon Presiden 2024 Dengan Algoritme Naïve Bayes,” *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 2, p. 265, 2022, doi: 10.30865/jurikom.v9i2.3989.
- [7] M. Rizki, M. F. Hidayattullah, and D. Intan, “Klasifikasi Opini Publik di Twitter Terhadap Bakal Calon Presiden Indonesia Tahun 2024 Menggunakan LSTM Secara Realtime Berbasis Website,” vol. 14, no. 02, pp. 286–294, 2024, doi: 10.35970/infotekmesin.v14i2.1908.
- [8] D. R. Manalu, M. C. L. Tobing, and M. Yohanna, “Analisis Sentimen Twitter Terhadap Wacana Penundaan Pemilu Dengan Metode Support Vector Machine,” *METHOMIKA J. Manaj. Inform. dan Komputerisasi Akunt.*, vol. 6, no. 6, pp. 149–156, 2022, doi: 10.46880/jmika.vol6no2.pp149-156.
- [9] R. Tinages, A. Triayudi, and I. D. Sholihati, “Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (SVM),” *J. Media Inform. Budidarma*, vol. 4, no. 3, p. 650, 2020, doi: 10.30865/mib.v4i3.2181.
- [10] A. P. Natasuwarna, “Seleksi Fitur Support Vector Machine pada Analisis Sentimen Keberlanjutan Pembelajaran Daring,” *Techno.Com*, vol. 19, no. 4, pp. 437–448, 2020, doi: 10.33633/tc.v19i4.4044.
- [11] H. P. P. Zuriel and A. Fahrurrozi, “Implementasi Algoritma Klasifikasi Support Vector Machine Untuk Analisa Sentimen Pengguna Twitter Terhadap Kebijakan Psbb,” *J. Ilm. Inform. Komput.*, vol. 26, no. 2, pp. 149–162, 2021, doi: 10.35760/ik.2021.v26i2.4289.
- [12] C. Shofiya and S. Abidi, “Sentiment analysis on covid-19-related social distancing in Canada using twitter data,” *Int. J. Environ. Res. Public Health*, vol. 18, no. 11, 2021, doi: 10.3390/ijerph18115993.
- [13] Z. Alhaq, A. Mustopa, S. Mulyatun, and J. D. Santoso, “Penerapan Metode Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter,” *J. Inf. Syst. Manag.*, vol. 3, no. 2, pp. 44–49, 2021, doi: 10.24076/joism.2021v3i2.558.
- [14] S. Sumayah, F. Sembiring, and W. Jatmiko, “Analysis of Sentiment of Indonesian Community on Metaverse Using Support Vector Machine Algorithm,” *J. Tek. Inform.*, vol. 4, no. 1, pp. 143–150, 2023, doi: 10.52436/1.jutif.2023.4.1.417.
- [15] M. R. Arjunanto and S. Waluyo, “Analisis Sentimen Pada Media Sosial Twitter Terhadap Infrastruktur Jalan Di Lampung Menggunakan Algoritma K-Nearest Neighbor,” *Pros. Semin. Nas. ...*, vol. 2, no. September, pp. 810–817, 2023.
- [16] D. Septiani and I. Isabela, “Analisis Term Frequency Inverse Document Frequency (Tf-Idf) Dalam Temu Kembali Informasi Pada Dokumen Teks,” *SINTESIA J. Sist. dan Teknol. Inf. Indones.*, vol. 1, no. 2, pp. 81–88, 2022.

- [17] R. F. Ananda, A. Syahri, and F. N. Hasan, "Sentiment Analysis of Customer Satisfaction in Gojek and Grab Application Reviews Using the Naive Bayes Algorithm," *JUTIF J. Tek. Inform.*, vol. 5, no. 1, pp. 233–241, 2024.