



# Klasifikasi Sentimen Masyarakat Di Twitter Terhadap Prabowo Subianto Sebagai Bakal Calon Presiden 2024 Menggunakan M-KNN

Abdul Halim\*, Yusra, Muhammad Fikry, Muhammad Irsyad, Elvia Budianita

Fakultas Sains dan Teknologi, Program Studi Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru

Panam, Jl. HR. Soebrantas No.Km. 15, RW.15, Simpang Baru, Kota Pekanbaru, Riau, Indonesia

Email: <sup>1,\*</sup>11950113404@students.uin-suska.ac.id, <sup>2</sup>yusra@uin-suska.ac.id, <sup>3,\*</sup>muhammad.fikry@uin-suska.ac.id, irsyadtech@uin-suska.ac.id, elvia.budianita@uin-suska.ac.id

Email Penulis Korespondensi: 11950113404@students.uin-suska.ac.id

Submitted: 09/08/2023; Accepted: 23/10/2023; Published: 26/10/2023

**Abstrak**—Pemilihan presiden diadakan setiap lima tahun sekali dan setiap calon presiden akan mendapatkan dukungan dari beberapa partai politik untuk maju dalam pencalonan dalam pemilihan tersebut. Dalam sistem multi partai, jumlah partai peserta pemilu sangatlah banyak, sehingga perspektif pemilih terhadap para aktor politik, termasuk calon presiden yang akan maju dalam pemilu tahun 2024, menjadi beragam. Hasil survei dari Polling Indonesia (SPIN) yang dilakukan pada tanggal 7 hingga 16 Oktober 2022 menunjukkan bahwa Prabowo Subianto memiliki elektabilitas tertinggi dengan skor 31,6%, berdasarkan survei kepemimpinan nasional. Dalam penelitian ini, dilakukan pengujian dengan mengklasifikasikan data tweet dari masyarakat yang dikumpulkan di aplikasi Twitter dari bulan Januari hingga Desember 2022 menggunakan metode Modified k-Nearest Neighbor untuk menganalisis sentimen masyarakat terkait pemilihan yang akan datang. Data yang dikumpulkan sebanyak 2.100 data dengan kategori positif dan negatif yang terkait dengan "Calon Presiden" dan "Prabowo Subianto" serta pelaksanaan klasifikasi Modified k-Nearest Neighbor dilakukan menggunakan Google Colab. Berdasarkan hasil pengujian confusion matrix dari klasifikasi Modified k-Nearest Neighbor dengan tiga perbandingan yang dilakukan (yaitu perbandingan 70%:30%, 80%:20% dan 90%:10%) dan menggunakan k = 3, 5, 7, 9, 11, saat uji coba perbandingan 90:10 pada K=3 didapatkan hasil akurasi tertinggi dengan nilai 93,3%.

**Kata Kunci:** Klasifikasi; Modified k-Nearest Neighbor; Prabowo Subianto; Sentimen; Twitter; Calon Presiden

**Abstract**—Presidential elections are held every five years and each presidential candidate will get support from several political parties to run for candidacy in the election. In a multi-party system, the number of parties participating in the election is very large, so that the perspectives of voters on political actors, including presidential candidates who will advance in the 2024 elections, are varied. The survey results from Polling Indonesia (SPIN) conducted from 7 to 16 October 2022 show that Prabowo Subianto has the highest electability with a score of 31.6%, based on a national leadership survey. In this study, a test was carried out by classifying tweet data from the public collected on the Twitter application from January to December 2022 using the Modified k-Nearest Neighbor method to analyze public sentiment regarding the upcoming election. Data collected as many as 2,100 data with positive and negative categories related to "Presidential Candidate" and "Prabowo Subianto" and the implementation of the Modified k-Nearest Neighbor classification was carried out using Google Colab. Based on the results of the confusion matrix test from the Modified k-Nearest Neighbor classification with three comparisons made (ie comparisons 70%:30%, 80%:20% dan 90%:10%) and using K=3, 5, 7, 9, 11 when testing a comparison of 90:10 at K=3 the highest accuracy results were obtained with a value of 93,3%.

**Keywords:** Classification; Modified k-Nearest Neighbor; Prabowo Subianto; Sentiment; Twitter; Presidential Candidate

## 1. PENDAHULUAN

Menurut hasil penelitian yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), diperkirakan jumlah pengguna internet di Indonesia akan mencapai 210,03 juta orang selama periode 2021 hingga 2022[1]. Tingginya jumlah pengguna internet di Indonesia dari tahun ke tahun yang membuat dunia internet menjadi salah satu komoditas informasi yang sangat efektif. Berdasarkan pengamatan We Are Social di bulan Januari 2022, jumlah pengguna aktif media sosial di Indonesia mencapai 191 juta orang, merupakan peningkatan sebesar 12,35% daripada tahun sebelumnya yang berjumlah 170 juta orang[2]. Dengan melihat tren yang ada, dapat diketahui bahwa jumlah pengguna media sosial di Indonesia terus mengalami peningkatan setiap tahunnya, cenderung meningkat dari tahun 2014 hingga 2022 dengan pertumbuhan tidak selalu stabil. Banyak media sosial yang sering digunakan oleh masyarakat, salah satunya Twitter. Pada tahun 2022, diperkirakan pengguna layanan Twitter di Indonesia mengalami peningkatan 18,45 juta pengguna, jumlah ini sebanding dengan 4,23% pengguna Twitter di seluruh dunia yang diperkirakan akan mencapai 436 juta pengguna[3].

Pemilihan presiden diadakan setiap 5 tahun sekali. Setiap calon presiden akan didukung oleh beberapa partai untuk maju dalam pemilihan. Penggunaan sistem multi partai dalam pemilu mengakibatkan adanya banyak partai yang berpartisipasi, sehingga perspektif pemilih terhadap para aktor politik menjadi sangat beragam, termasuk para calon presiden yang akan maju dalam pilpres 2024. Semua calon presiden akan berusaha untuk memikat hati para pemilih dengan cara yang berbeda-beda, karena tentu saja mereka memiliki pendekatan dan visi yang berbeda antara satu calon presiden dengan calon presiden lainnya. Berdasarkan hasil Survei Opini Publik Indonesia pada 7 sampai 16 Oktober 2022, Prabowo Subianto menempati posisi elektoral tertinggi dengan skor 31,6%, merupakan Survei Kepemimpinan Nasional yang melibatkan 1.230 responden dari 34 provinsi di Indonesia[4]. Prabowo subianto juga banyak diperbincangkan di Twitter karna bakal calon pilpres 2024. Menurut

data media Liputan Enam tanggal 7 juli 2022, akun Twitter Prabowo Subianto diikuti sebanyak 4,5 juta pengikut[5].

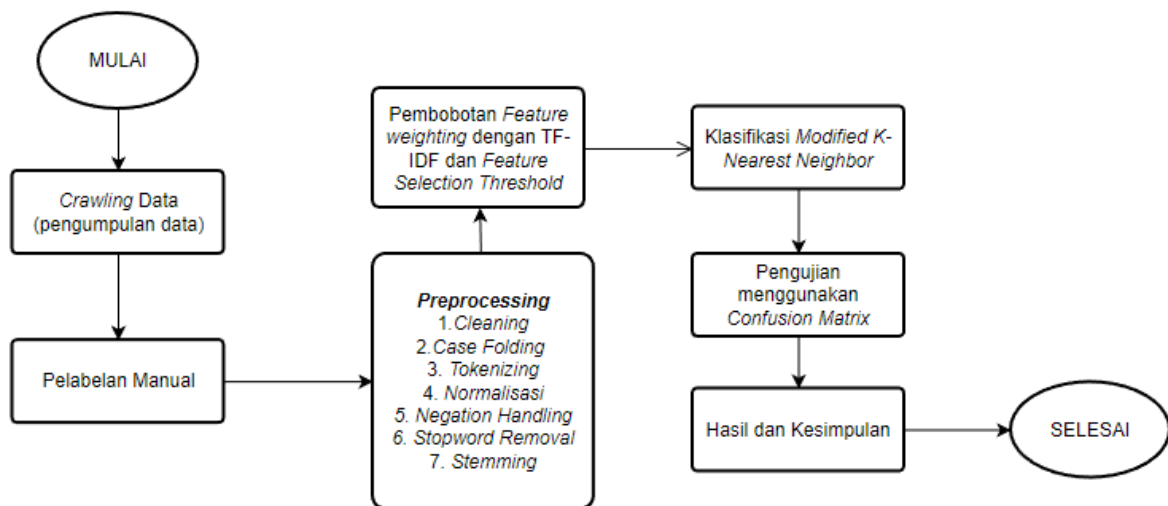
Klasifikasi sentimen merupakan salah satu teknik dalam analisis teks atau pemrosesan bahasa alami yang digunakan untuk mengekstrak informasi tentang sikap atau pandangan seseorang terhadap suatu isu atau kejadian[6]. Tujuan utama dari klasifikasi sentimen adalah untuk mengelompokkan polaritas atau sentimen dari teks menjadi beberapa kategori, seperti positif atau negatif. Pada penelitian ini akan dilakukan dengan algoritma Modified k-Nearest Neighbor (Mk-NN). Modified k-Nearest Neighbor adalah salah satu algoritma pada text mining hasil dari modifikasi lebih baik dari algoritma awalnya yaitu k-Nearest Neighbour (k-NN). Metode algoritma Mk-NN menggabungkan 2 proses klasifikasi baru untuk mengitung sebuah nilai validitas antar data training dan menghitung bobot masing-masing tetangga yaitu perhitungan weight voting[7]. Kombinasi dari kedua proses klasifikasi baru ini membantu meningkatkan akurasi dan kualitas klasifikasi dalam metode Mk-NN yang mempertimbangkan validitas antar data training dan pembobotan tetangga terdekat untuk memberikan hasil analisis yang lebih baik.

Penelitian terkait oleh Imaning Dyah Larasati, Ahmad Afif S, dan M Tanzil Furqon, Tahun 2019 tentang Prediksi Kelulusan Mahasiswa Berdasarkan Prestasi Akademik dengan algoritma Modified k-Nearest Neighbour (Mk-NN), menggunakan data alumni akademik FILKOM sejumlah 200. Data tersebut kemudian ditransformasikan pada IPK semester, sks dan kelulusan 1-5. Penelitian ini membandingkan akurasi metode Mk-NN dan k-NN pada saat k=5 akurasi pengujian nilai k dengan pengujian jumlah data latih didapatkan tingkat akurasi tertinggi masing-masing mencapai 85,25% untuk Mk-NN dan 84% untuk k-NN[8]. Pada penelitian lainnya, Muhammad Anas Ziaulhaq juga melakukan penelitian pada tahun 2022, melakukan analisis terhadap sentimen masyarakat di Twitter tentang pembelajaran online akibat covid-19 menggunakan metode Modified k-Nearest Neighbor yang dimodifikasi dengan diperoleh data 1000, 700, 500, 300 dan 100, hasil pengujian k=7 data hingga 300 dengan akurasi maksimal 87%, presisi dengan maksimal k=7 data hingga 300 pada 100%, dan k=3 data hingga 100 dengan recall 100%[9]. Penelitian tersebut menunjukkan bahwa kinerja metode Modified K-Nearest Neighbor bekerja sangat baik.

Berdasarkan temuan tersebut, menggunakan metode Modified k-Nearest Neighbor (Mk-NN) hasil pengembangan dari k-NN yang digunakan dalam klasifikasi sentimen publik di media sosial Twitter. Penelitian ini fokus pada data sebanyak 2100 yang terkait dengan Prabowo Subianto sebagai calon presiden 2024. Penelitian ini dimulai dari mengumpulkan data mentah dari Twitter, pelabelan data, pemrosesan data teks (preprocessing), klasifikasi sentimen, pembobotan kata serta menseleksi fitur dan pengujian menggunakan confusion matrix. Metode ini mengadopsi pendekatan tetangga terdekat (k) dalam Klasifikasi Opini Masyarakat di Twitter untuk Pencalonan Presiden Prabowo Subianto pada tahun 2024.

## 2. METODOLOGI PENELITIAN

Penelitian akan dilakukan menurut proses yang ditampilkan gambar 1.



**Gambar 1.** Proses awal dan akhir penelitian Mk-NN

### 2.1 Data

Pengumpulan informasi dilakukan bertujuan untuk mendapatkan sebuah data yang digunakan untuk tercapainya tujuan dari riset. Informasi dari sosial media ataupun social media mining jadi salah satu metode untuk mengenali tingkatan partisipasi dari pengguna[10]. Data diperoleh melalui Twitter API yang terkait dengan "Calon Presiden" dan "Prabowo Subianto". Peneliti mengambil data tweet dari pengguna yang terjadi antara tanggal 29 November

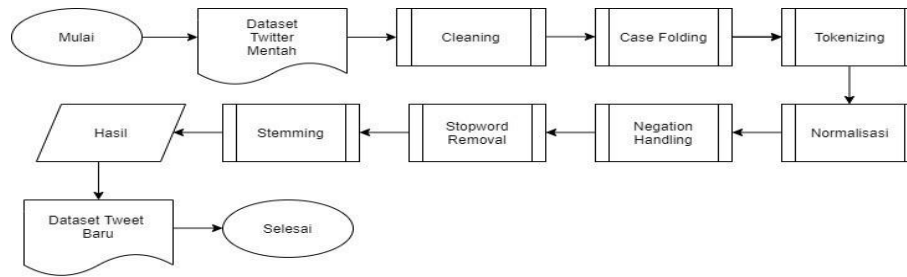
2022 hingga 03 Januari 2023. Total data yang diambil sebanyak 2100 tweet yang akan diimbangi antara kelas Positif dan kelas Negatif untuk mencapai akurasi klasifikasi yang optimal.

## 2.2 Pelabelan Data

Data dilabelkan secara manual menggunakan bantuan Dosen Bahasa Indonesia UIN SUSKA RIAU. Pelabelan yang dilakukan adalah label kelas pada data dengan jumlah data 1050 untuk setiap kelas yaitu positif dan negatif.

## 2.3 Tahapan Text Processing

Pre-processing atau pemrosesan data mentah adalah proses yang digunakan untuk memproses data mentah menjadi data yang terstruktur, sehingga data yang diubah dapat digunakan untuk melakukan analisis informasi[11]. Proses pre-processing ini membantu dalam mengoptimalkan data agar lebih siap digunakan untuk analisis informasi selanjutnya, gambar 2 menampilkan alur proses Text Processing.



**Gambar 2.** Text Processing

## 2.4 Pembobotan

Pembobotan adalah sebuah proses untuk membobotkan sebuah kata (term) pada suatu dokumen berdasarkan frekuensi kemunculan kata tersebut. Bobot ini mencerminkan tingkat pentingnya suatu kata dalam dokumen tersebut, sehingga kata-kata yang muncul lebih sering akan diberi bobot yang lebih tinggi. Pembobotan juga memperhitungkan seberapa umum suatu term dalam seluruh dokumen, dengan menggunakan konsep Inverse Document Frequency (IDF). Pada tahap awal pembobotan, dilakukan perhitungan nilai Term Frequency (TF), yang mengacu pada jumlah kemunculan suatu kata dalam dokumen tersebut. Pada langkah berikutnya, dihitung nilai Inverse Document Frequency (IDF) yang berperan dalam mengurangi bobot kata-kata yang muncul secara umum di semua dokumen[12]. Dengan menggabungkan nilai TF dan IDF, dapat diperoleh bobot akhir suatu term dalam dokumen. Bobot ini membantu dalam menggambarkan seberapa pentingnya suatu term dalam konteks dokumen tersebut, serta membantu dalam analisis informasi dan klasifikasi dokumen, perhitungan TF-IDF pada persamaan berikut:

$$W_{i,j} = \frac{n_{i,j}}{\sum_{j=1}^P n_{j,i}} \log_2 \frac{D}{d_j} \quad (1)$$

Keterangan:

$W_{i,j}$  = pembobotan tf-idf kata (term) ke-j pada dokumen ke-i

$n_{i,j}$  = jumlah kemunculan kata (term) ke-j pada dokumen ke-i

P = jumlah kata (term) yang terbentuk

$\sum_{j=1}^P n_{j,i}$  = jumlah total kemunculan semua kata (term) dalam dokumen ke-i

D = jumlah seluruh dokumen

$d_j$  = jumlah dokumen yang mengarah kata (term) ke-j

## 2.5 Seleksi Fitur Threshold

Setiap fitur setelah pembobotan akan diseleksi dengan proses pemilihan fitur yang relevan dan penghapusan fitur tersebut yang redundan atau tidak relevan dalam analisis data. Dalam metode ini, digunakan sebuah threshold (batas) yang menentukan bobot minimum yang harus dimiliki oleh suatu fitur agar dianggap relevan dan dipertahankan[13]. Menerapkan threshold, fitur-fitur yang memiliki bobot atau nilai di bawah batas tersebut akan dihapus dari data, sehingga hanya fitur-fitur yang dianggap signifikan yang tetap dipertahankan.

## 2.6 Klasifikasi

Klasifikasi dilakukan dengan algoritma Modified k-Nearest Neighbor (Mk-NN) adalah hasil perbaikan dari metode k-Nearest Neighbor (k-NN). Proses algoritma yang membedakan dua algoritma tersebut terletak pada dua proses tambahan di perhitungan algoritma Mk-NN yang menggunakan nilai validitas dan weight voting[14]. Dengan adanya perhitungan nilai validitas, Mk-NN dapat menentukan sejauh mana tetangga-tetangga terdekat relevan atau sesuai dengan data yang sedang diprediksi. Dengan mempertimbangkan validitas data latih, Mk-NN

dapat memberikan bobot yang lebih tinggi kepada tetangga-tetangga yang lebih relevan dalam proses pengambilan keputusan dengan persamaan sebagai berikut:

a. Perhitungan jarak Euclidean

$$d(x_i, y_i) = \sqrt{\sum_{i=0}^n (x_i - y_i)^2} \tag{2}$$

Keterangan:

d = jarak antar data x dan y

x = titik data latih

y = titik data uji

b. Perhitungan nilai validitas

$$\text{Validity}(x) = \frac{1}{h} \sum_{i=0}^n S(\text{lbl}(x), \text{lbl}(\text{Ni}(x))) \tag{3}$$

Keterangan:

H = jumlah (k) atau tetangga

lbl(x) = kelas (x)

lbl(Ni(x)) = kelas tetangga (x)

S = persamaan dalam perhitungan antara titik x dan data ke I dari (k)

c. Perhitungan weight voting

$$w(i) = \text{Validity}(i) \times \frac{1}{d(i)+0,5} \tag{4}$$

Keterangan:

(i) = Pencarian Weight Voting

Validit(i) = Nilai Validitas

d(i) = Jarak Euclidean

## 2.7 Pengujian

Setelah pengelompokan data, hasil akan di proses dengan prediksi Confusion matrix untuk mengevaluasi efektifitas suatu model klasifikasi dalam data pelatihan dan data pengujian[15]. Tabel 1 menggambarkan perbandingan antara prediksi model dengan label sebenarnya dari data.

**Tabel 1.** Rancangan Analisis Komputasi

Label (kelas)	Hasil	
	POSITIF	NEGATIF
POSITIF	True Positive	True Negative
NEGATIF	False Positive	False Negative

Keterangan:

True Positive (TP) = Jumlah sampel yang diklasifikasikan dengan benar sebagai kelas positif.

True Negative (TN) = Jumlah sampel yang diklasifikasikan dengan benar sebagai kelas negatif.

False Positive (FP) = Jumlah sampel yang salah diklasifikasikan sebagai kelas positif.

False Negative (FN) = Jumlah sampel yang salah diklasifikasikan sebagai kelas negatif.

Menghitung nilai dari akurasi (accuracy), presisi (precision) dan sensitivitas (recall) seperti persamaan:

$$\text{accuracy} = \frac{TP + TN}{\text{Total}} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{7}$$

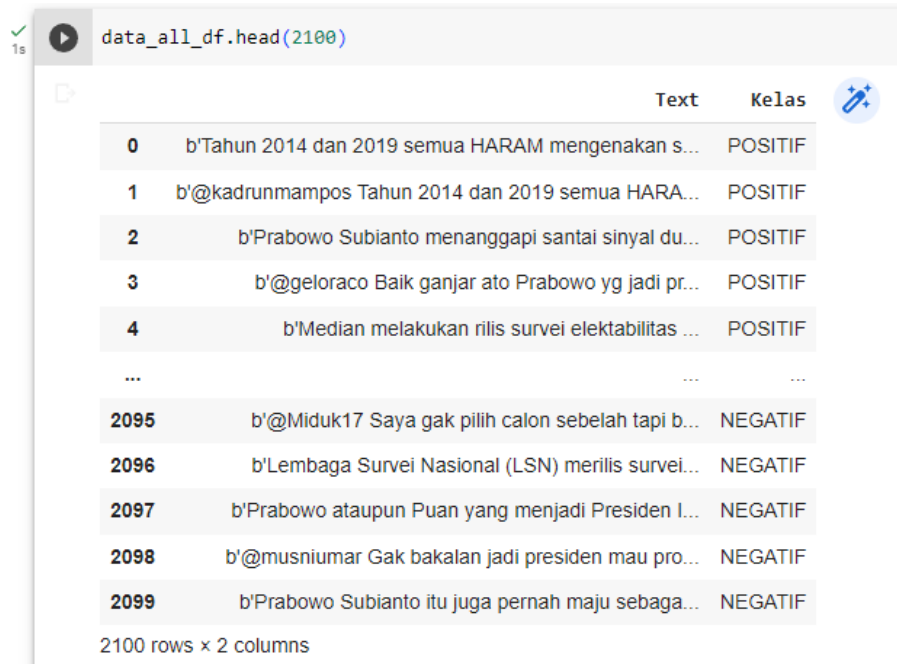
$$\text{F1 Score} = \frac{2 \times P \times R}{P + R} \tag{8}$$

## 3. HASIL DAN PEMBAHASAN

### 3.1 Data

Keseluruhan data penelitian dikumpulkan untuk analisis sentimen terhadap tweet Prabowo Subianto di aplikasi Twitter sebagai calon presiden 2024. Proses pendataan ini dibatasi sampai bakal Calon Presiden Prabowo Subianto, data dikumpulkan melalui teknik crawling dengan menggunakan fungsi Application Interface (API)

pada aplikasi tersebut. Jumlah data yang diekstraksi dari Twitter sebanyak 2100 data, dengan pembagian seimbang antara 1050 data positif dan 1050 data negatif. Hal ini dilakukan untuk mencapai hasil klasifikasi yang optimal dalam analisis. Proses pengambilan data tweet dari API Twitter berbentuk sebuah teks. Hasil seperti gambar 3 yang diproses menggunakan Google Colab.



**Gambar 3.** Tampilan Data Mentah

### 3.2 Pelabelan Data

Setelah pengumpulan data selesai, data dilabeli secara manual, berdasarkan bantuan Roza Afifah, S.Pd., dosen Bahasa Indonesia di Universitas Islam Negeri Sultan Syarif Kasim Riau, data telah dilabeli menjadi dua kelas, yaitu positif dan negatif. Tabel 2 menampilkan kedua kelas tersebut, yaitu kelas positif dan negatif, yang telah ditentukan berdasarkan proses pelabelan tersebut.

**Tabel 2.** Tweet Positif dan Negatif

Tweet Positif	Tweet Negatif
b'Tahun 2014 dan 2019 semua HARAM mengenakan simbol kafir dan masuk ke Gereja KECUALI Prabowo\n\nTahun 2021 - 2024 semua HARAM mengenakan simbol kafir dan masuk ke Gereja KECUALI anoes\n\nTahun ""XXXX"" semua HARAM mengenakan simbol kafir dan masuk ke Gereja KECUALI calon presiden kadrun'	b'Si Anies belum menjabat aja dikritik langsung ngamuk FPI dan pendukung khilafahnya apa lagi sudah menjabt langsung meninggal ditempat mungkin si kakek ini\n\n#PakaiMolis #BersamaMandiri #jeketiselcaday #HPSMinisterID #KirimTakGentar Arsenal Maneh PKS Prabowo Cak Imin Ganjar Erick
b'@kadrunmampos Tahun 2014 dan 2019 semua HARAM mengenakan simbol kafir dan masuk ke Gereja KECUALI Prabowo\n\nTahun 2021 - 2024 semua HARAM mengenakan simbol kafir dan masuk ke Gereja KECUALI anoes\n\nTahun ""XXXX"" semua HARAM mengenakan simbol kafir dan masuk ke Gereja KECUALI calon presiden kadrun'	b'Cak Imin Ancam Koalisi Bubar, bila ada Duet Prabowo-Ganjar - Tribun Bekasi <a href="https://t.co/zpSBPkmczk">https://t.co/zpSBPkmczk</a> #PrabowoUnggul #BangkitBersama",
b'Prabowo Subianto menanggapi santai sinyal dukungan Presiden Joko Widodo (Jokowi) terhadap bakal calon presiden yang nantinya berkontestasi di pemilihan presiden 2024.	b'Cak Imin Still In The Game!\nPKB Tak Tertarik Usung Ganjar\nJadi Pendamping Prabowo...\nPKB Maunya Prabowo-Cak Imin\nGaspol Sudah! Rem Blong!\nJangan Kasih Rambut Putih...'
b'@geloraco Baik ganjar ato Prabowo yg jadi presiden 2024 pemenangnya adalah oligarki, krn skenario hanya dua pasang calon itu yg akan di mainkan, kecuali anies yg menang, baru jokowi nasib nya akan seperti PM razak mantan penguasa malaysia'	b'@LekMarteg Prabowo 2 X di Jatim aja gk laku apalagi amies yg rekam jejaknya lebih buruk',
b'Median melakukan rilis survei elektabilitas tokoh calon presiden dan wakil presiden 2024. Hasilnya	,b'@HelmiFelis_ Jadi kalau 2024 Ganjar atau Prabowo menang ..ada kecurangan ..gitu ya



Tweet Positif	Tweet Negatif
elektabilitas Prabowo berada di urutan pertama dengan 24,2%. <a href="https://t.co/aNNoV1B2fn">https://t.co/aNNoV1B2fn</a>	cyy\nJadi kalau 2024 wan Anies kalah..tuh kan prediksi sy benar ..nih ada video saya..kita dicurangin ..lol'

### 3.3 Tahapan Text Processing

#### a. Cleaning Data

Data yang diperoleh merupakan data mentah dari tweet pengguna aplikasi Twitter, dilakukan cleaning pada data seperti yang ditampilkan di tabel 3. Pembersihan dilakukan untuk menghilangkan duplikasi data[16].

**Tabel 3.** Cleaning Data

Cleaning Data	Case Folding
@Cak Imin Sebut PKB Tak Tertarik Usung Prabowo Ganjar di Pilpres 2024	Cak Imin Sebut PKB Tak Tertarik Usung Prabowo Ganjar di Pilpres 2024

#### b. Case Folding

Case folding merupakan teknik pembersihan yang digunakan sebagai tahap mengubah seluruh huruf dalam suatu teks menjadi huruf kecil[17]. Hal ini dilakukan untuk mencapai konsistensi dalam representasi teks dan mempermudah proses analisis atau pemrosesan yang ditampilkan pada tabel 4.

**Tabel 4.** Case Folding

Cleaning Data	Case Folding
Cak Imin Sebut PKB Tak Tertarik Usung Prabowo Ganjar di Pilpres 2024	cak imin sebut pkb tak tertarik usung prabowo ganjar di pilpres 2024

#### c. Tokenizing

Tokenisasi adalah proses untuk membagi atau memecah teks menjadi unit-unit yang lebih kecil menjadi token-token atau kata-kata[18]. Umumnya, setiap kata dipisahkan dari kata lain dengan karakter dan spasi yang dicontohkan di tabel 5.

**Tabel 5.** Tokenizing

Case Folding	Tokenizing
cak imin sebut pkb tak tertarik usung prabowo ganjar di pilpres 2024	['cak', 'imin', 'sebut', 'pkb', 'tak', 'tertarik', 'usung', 'prabowo', 'ganjar', 'di', 'pilpres', '2024']

#### d. Normalisasi

Normalisasi kata akan mengubah kata yang tidak baku jadi baku serta mengganti singkatan jadi kata asalnya[19]. Contoh normalisasi di proses pada tabel 6.

**Tabel 6.** Normalisasi

Tokenizing	Normalisasi
Cak	Cak
imin	imin
sebut	sebut
pkb	pkb
<b>tak</b>	<b>tidak</b>
tertarik	tertarik
usung	usung
prabowo	prabowo
ganjar	ganjar
di	di
pilpres	pilpres
2024	2024

#### e. Negation Handling

Negasi bertujuan untuk menegasi, membantah, penegasan dalam kalimat interogatif, imperaktif dan prediktif[20]. Contoh kata negasi adalah “tidak”, “tanpa”, “bukan”, “jangan” seperti yang ditampilkan tabel 7.

**Tabel 7.** Negation Handling

Normalisasi	Negation Handling
['cak', 'imin', 'sebut', 'pkb', ' <b>tidak</b> ', 'tertarik', 'usung', 'prabowo', 'ganjar', 'di', 'pilpres', '2024']	['cak', 'imin', 'sebut', 'pkb', ' <b>diabaikan</b> ', 'usung', 'prabowo', 'ganjar', 'di', 'pilpres', '2024']

f. Stopword Removal

Proses stopwords removal dilakukan untuk menghapus sebuah kata umum atau stopwords dari kata dalam teks yang terlalu sering timbul namun tidak memiliki makna yang signifikan terhadap analisis teks[21]. kata-kata seperti "dan", "atau", "di", "dari", "ke", "yang", stopwords dilakukan pada tabel 8.

**Tabel 8.** Stopword Removal

Negation Handling	Stopword Removal
['cak', 'imin', 'sebut', 'pkb', 'diabaikan', 'usung', 'prabowo', 'ganjar', 'di', 'pilpres', '2024']	['cak', 'imin', 'pkb', 'diabaikan', 'usung', 'prabowo', 'ganjar', 'pilpres', '2024']

g. Stemming

Proses stemming yang digunakan untuk melakukan teknik pencarian kata dasar pada term (kata) dalam sebuah dokumen[22]. Stemming dilakukan untuk menghilangkan atau menghapus imbuhan awalan, sisipan dan akhiran seperti yang dicontohkan tabel 9.

**Tabel 9.** Stemming

Stopword Removal	Stemming
['cak', 'imin', 'pkb', 'diabaikan', 'usung', 'prabowo', 'ganjar', 'pilpres', '2024']	['cak', 'imin', 'pkb', 'abai', 'usung', 'prabowo', 'ganjar', 'pilpres', '2024']

**3.4 Pembobotan**

Setelah proses pengolahan teks selesai, kata-kata akan diberi bobot menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). Pemberian bobot setiap kata bertujuan memberi nilai pada kata yang berulang sama. Term Frequency merupakan proses menentukan jumlah sebuah kata yang timbul dalam sebuah dokumen. Setiap dokumen memiliki perbedaan frekuensi sebuah term yang sama, kemungkinan sebuah term yang sama lebih sering muncul pada dokumen yang lebih besar daripada dokumen kecil. Sedangkan, Inverse Document Frequency (IDF) merujuk pada jumlah dokumen dimana sebuah term yang sama muncul. Jika sebuah term jarang muncul pada tiap dokumen, maka semakin rendah nilai bobotnya, seperti yang ditampilkan gambar 4.

	00	000t	002	02	07	08	10	100	110	115	...	yusril	ywh	zalm	zhone	zhonk	zon	zonauang	zonky	zul kifli	Kelas	
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	POSITIF
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	POSITIF
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	POSITIF
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	POSITIF
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	POSITIF

5 rows x 3533 columns

**Gambar 4.** Pembobotan

**3.5 Seleksi Fitur**

Dalam seleksi fitur, digunakan threshold (ambang batas) nilai untuk menentukan apakah suatu fitur harus dipertahankan atau dieliminasi. Threshold ini biasanya berada dalam rentang 0 hingga 1, dan dapat diatur sesuai dengan kebutuhan dan tujuan analisis[23]. Seleksi diaplikasikan pada Google Colab yang ditampilkan oleh gambar 5.

Hasil Feature Selection																						
	08	20102	2024	2029	agama	ahmad	akrab	aku	alonely	ancam	...	unggul	ungkap	usung	wacana	wakil	warga	widodo	yea	yess	Kelas	
0	0.0	0.0	0.030228	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	POSITIF	
1	0.0	0.0	0.030228	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	POSITIF	
2	0.0	0.0	0.089897	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.260121	0.0	0.0	POSITIF	
3	0.0	0.0	0.061823	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	POSITIF	
4	0.0	0.0	0.089287	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.192877	0.0	0.000000	0.0	0.0	POSITIF	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2095	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.207152	0.0	0.000000	0.0	0.0	NEGATIF	
2096	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	NEGATIF	
2097	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	NEGATIF	
2098	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	NEGATIF	
2099	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.164686	0.0	0.000000	0.0	0.0	NEGATIF	

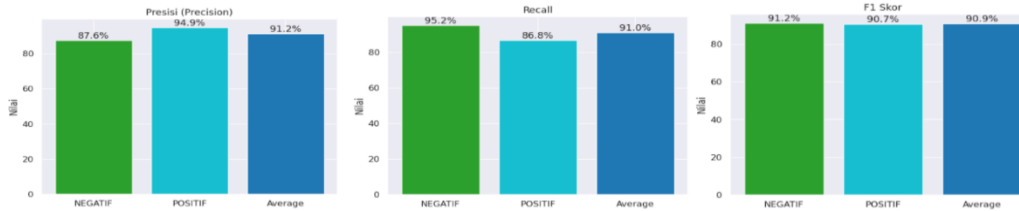
2100 rows x 176 columns

**Gambar 5.** Hasil Threshold

**3.6 Word Cloud**

Barakhirnya proses yang di klasifikasi, dilanjutkan dengan memvisualisasikan kata-kata yang paling banyak digunakan oleh masyarakat di Twitter yang berkaitan dengan "Calon Presiden" dan "Prabowo Subianto" yang

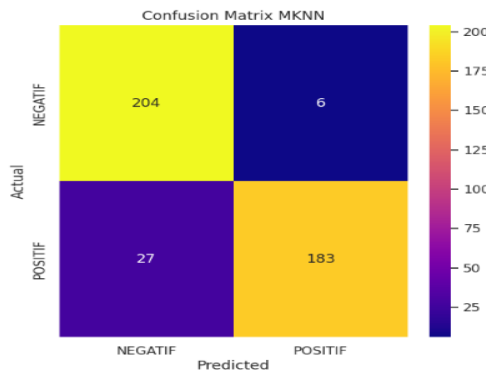




**Gambar 9.** Nilai Precision, Recall dan f1score pengujian 70:30

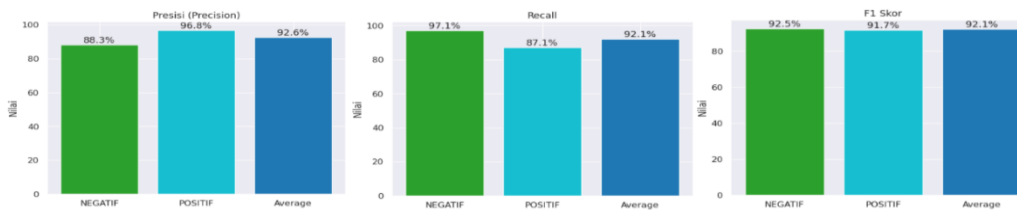
b. Pengujian pada pembagian 80% dan 20%

Dalam pengujian dengan pembagian data latih dan uji sebesar 80% dan 20% mendapatkan akurasi prediksi tertinggi saat  $K = 3$  dengan persentase sebesar 92,1%. Hasil implementasi pengujian confusion matrix tersebut ditampilkan pada Gambar 10 di Google Colab.



**Gambar 10.** Confusion Matrix Perbandingan 80:20

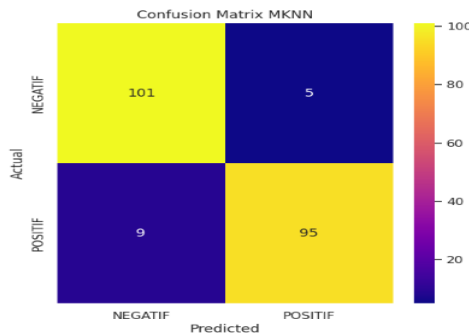
Pada Gambar 11, ditampilkan nilai presisi sebesar 88,3% negatif dan 96,8% positif dengan rata-rata nilai presisi adalah 92,6%. Selanjutnya, untuk nilai recall 97,1% negatif dan 87,1% positif, sehingga rata-rata nilai recall adalah 92,1%. Kemudian, untuk f1-score 92,5% negatif dan 91,7% positif, dengan rata-rata f1-score adalah 92,1%.



**Gambar 11.** Nilai Precision, Recall dan f1score Pengujian 80:20

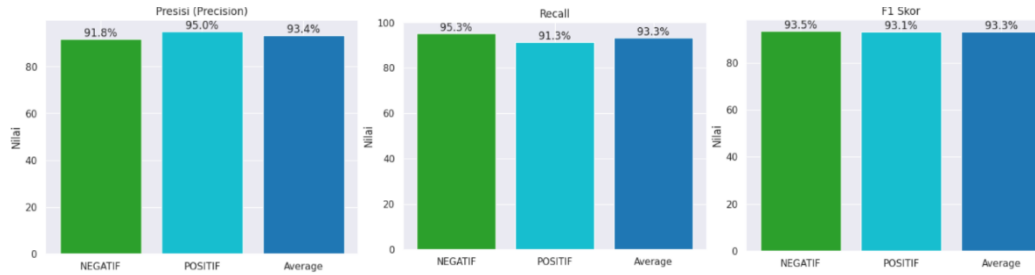
c. Pengujian pada pembagian 90% dan 10%

Dalam pengujian dengan pembagian data latih dan uji sebesar 90% dan 10% mendapatkan akurasi prediksi tertinggi saat  $K = 3$  dengan persentase sebesar 93,3%. Hasil implementasi pengujian confusion matrix tersebut ditampilkan pada Gambar 12 di Google Colab.



**Gambar 12.** Confusion Matrix Perbandingan 90:10

Pada Gambar 13, ditampilkan nilai presisi 91,8% negatif dan 95% positif dengan rata-rata nilai presisi adalah 91,4%. Selanjutnya, untuk nilai recall 95,3% negatif dan 91,3% positif dengan rata-rata nilai recall adalah 93,3%. Kemudian, untuk f1-score 93,5% negatif dan 93,1% positif dengan rata-rata f1-score adalah 93,3%.



**Gambar 13.** Nilai Precision, Recall dan f1 score pengujian 90:10

## 4. KESIMPULAN

Berdasarkan penelitian klasifikasi sentimen ini, disimpulkan bahwa metode Modified k-Nearest Neighbor (Mk-NN) dapat berhasil digunakan untuk melakukan proses klasifikasi dari sentimen masyarakat dengan kata yang terkait "calon presiden" dan "Prabowo Subianto". Penelitian dimulai dengan mengumpulkan data mentah dari aplikasi Twitter, kemudian dilakukan preprocessing data, pembobotan kata dan seleksi fitur (threshold). Proses klasifikasi menggunakan metode Modified k-Nearest Neighbor dan hasilnya dievaluasi menggunakan confusion matrix. Pengujian dilakukan dengan menggunakan tiga perbandingan (70%:30%, 80%:20% dan 90%:10%) menggunakan parameter  $k = 3, 5, 7, 9, 11$ . Dari hasil pengujian, ditemukan bahwa perbandingan data latih dan data uji sebesar 90%:10% dengan menggunakan nilai  $K = 3$  menghasilkan akurasi tertinggi, yaitu mencapai 93,3%. Selain itu, hasil precision, recall, dan f1-score mencapai 93,4%, 93,3% dan 93,3%. Penelitian ini mendapatkan hasil yang maksimal dalam klasifikasi sentimen menggunakan Modified k-Nearest Neighbor sebagai metode yang digunakan.

## REFERENCES

- [1] B. Dimas, "APJII: Pengguna Internet Indonesia Tembus 210 Juta pada 2022," 2022. <https://dataindonesia.id/digital/detail/apjii-pengguna-internet-indonesia-tembus-210-juta-pada-2022> (accessed Dec. 04, 2022).
- [2] M. M. Ivan, "Pengguna Media Sosial di Indonesia Capai 191 Juta pada 2022," 2022. <https://dataindonesia.id/digital/detail/pengguna-media-sosial-di-indonesia-capai-191-juta-pada-2022> (accessed Dec. 04, 2022).
- [3] A. R. Monavia, "Pengguna Twitter di Indonesia Capai 18,45 Juta pada 2022," 2022. <https://dataindonesia.id/digital/detail/pengguna-twitter-di-indonesia-capai-1845-juta-pada-2022> (accessed Dec. 05, 2022).
- [4] S. K. Melalusa, "Survei SPIN: Prabowo tempati elektabilitas tertinggi capres 2024 - ANTARA News," 2022. <https://www.antaraneews.com/berita/3200145/survei-spin-prabowo-tempati-elektabilitas-tertinggi-capres-2024> (accessed Dec. 04, 2022).
- [5] M. D. Agustinus, "Cek Jumlah Follower Twitter Sejumlah Tokoh Publik yang Berpotensi Jadi Capres 2024 - Tekno Liputan6.com," 2022. <https://www.liputan6.com/tekno/read/5007928/cek-jumlah-follower-twitter-sejumlah-tokoh-publik-yang-berpotensi-jadi-capres-2024> (accessed Dec. 05, 2022).
- [6] M. I. Aditama, R. Irfan Pratama, K. Hafizzana, U. Wiwaha, and N. A. Rakhmawati, "Analisis Klasifikasi Sentimen Pengguna Media Sosial Twitter Terhadap Pengadaan Vaksin COVID-19," JIEET (Journal Inf. Eng. Educ. Technol., vol. 4, no. 2, pp. 90–92, Dec. 2020, doi: 10.26740/JIEET.V4N2.P90-92.
- [7] A. Fairuzatul. Jannah; Ginanjar. Abdurrahman; Qurrota, "Jurnal Smart Teknologi Analisa Klasifikasi Data Kualitas Kadar Karat Emas Menggunakan Metode Modified Analysis Of Classification Of Gold Carat Quality Data Using The Modified K-Nearest Neighbor (MK-NN) Method Jurnal Smart Teknologi," vol. 3, no. 5, pp. 511–519, 2022.
- [8] M. T. Larasati, Imaning D; Supianto, Ahmad A; Furqon, "Prediksi Kelulusan Mahasiswa Berdasarkan Kinerja Akademik Menggunakan Metode Modified K-Nearest Neighbor (MK-NN) | Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer," 2019. <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/5284> (accessed Dec. 05, 2022).
- [9] M. Anas Ziaulhaq, "Analisis Sentimen Data Twitter Tentang Pembelajaran Online di Indonesia Akibat Covid-19 Dengan Menggunakan Metode Modified K-Nearest Neighbor - Repository Tugas Akhir," 2022. <http://repota.jti.polinema.ac.id/805/> (accessed Dec. 05, 2022).
- [10] L. F. Narulita and D. H. Sulistyawati, "Pengumpulan Data Twitter Tentang Covid-19 di Indonesia untuk Menghitung Tingkat Engagement Pengguna," J. Teknol. Inf. dan Ilmu Komput., vol. 8, no. 3, p. 565, 2021, doi: 10.25126/jtiik.2021834626.
- [11] G. F. Grandis, Y. Arumsari, and Indriati, "Seleksi Fitur Gain Ratio pada Analisis Sentimen Kebijakan Pemerintah Mengenai Pembelajaran Jarak Jauh dengan K-Nearest Neighbor," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 5, no. 8, pp. 3507–3514, 2021.
- [12] R. Z. Supono and S. Muhammad Azis, "Perbandingan Metode TF-ABS dan TF-IDF Pada Klasifikasi Teks Helpdesk Menggunakan K-Nearest Neighbor," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 5, no. 5, pp. 911–918, 2021, doi: 10.29207/resti.v5i5.3403.
- [13] Y. H. Kristian, K. R. Prilianti, and P. L. Tirma, "Implementasi Text Mining Untuk Analisis Tempat Wisata Di Indonesia," J. SimanteC, vol. 7, no. 2, pp. 73–82, 2019.



- [14] W. A. Istiqhfarani, I. Cholissodin, and F. A. Bachtiar, “Klasifikasi Penyakit Dental caries menggunakan Algoritme Modified K- Nearest Neighbor,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 5, pp. 1499–1506, 2020, [Online]. Available: <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/7265/3498>
- [15] A. I. Saputra, H. Oktavianto, H. Azizah, and A. Faruq, “Penerapan Algoritma Modified K-Nearest Neighbour (MKNN) pada Klasifikasi Masa Studi Mahasiswa Teknik Informatika,” *J. Smart Teknol.*, vol. 3, no. 1, pp. 2774–1702, 2021.
- [16] A. P. Giovani, A. Ardiansyah, T. Haryanti, L. Kurniawati, and W. Gata, “Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritma Klasifikasi,” *J. Teknoinfo*, vol. 14, no. 2, p. 115, 2020, doi: 10.33365/jti.v14i2.679.
- [17] Y. A. V. Gunawan, S. N. Agus, M. M. Ida Bagus, W. I. Made, N. A. C. P. I Gusti, and A. G. A. K. I Gusti, “Analisis Sentimen Ulasan Aplikasi Transportasi Online Menggunakan Multinomial Naïve Bayes dan Query Expansion Ranking,” *JELIKU (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 11, no. 1, p. 121, 2022, doi: 10.24843/jlk.2022.v11.i01.p13.
- [18] I. Najiyah and I. Hariyanti, “Sentimen Analisis Covid-19 Dengan Metode Probabilistic Neural Network Dan Tf-Idf,” *J. Responsif Ris. Sains dan Inform.*, vol. 3, no. 1, pp. 100–111, 2021, doi: 10.51977/jti.v3i1.488.
- [19] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, “Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19),” *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 406, 2021, doi: 10.30865/mib.v5i2.2835.
- [20] K. Saputri, H. Faizah, and C. Charlina, “Negasi dalam Tuturan Peserta Diskusi Indonesia Lawyers Club,” *SASTRANESIA J. Progr. Stud. Pendidik. Bhs. dan Sastra Indones.*, vol. 9, no. 1, p. 29, 2021, doi: 10.32682/sastronesia.v9i1.1788.
- [21] A. Santosa, I. Purnamasari, and R. Mayasari, “Pengaruh Stopword Removal dan Stemming Terhadap Performa Klasifikasi Teks Komentar Kebijakan New Normal Menggunakan Algoritma,” *J. Sains Komput. Inform.*, vol. 6, no. 1, pp. 81–93, 2022.
- [22] I. Z. Simanjuntak, “Analisa Kombinasi Algoritma Stemming Dan Algoritma Soundex Dalam Pencarian Kata Bahasa Indonesia,” *Inf. dan Teknol. Ilm.*, vol. 10, no. 1, pp. 24–30, 2022, [Online]. Available: <http://ejournal.stmik-budidarma.ac.id/index.php/inti/article/view/5040>
- [23] R. H. Faturrahman, W. Astuti, and M. D. Purbolaksono, “Klasifikasi Sentimen Ulasan Film Menggunakan Support Vector Machine , Information Gain , dan N-Grams,” *e-Proceeding Eng.*, vol. 9, no. 3, pp. 1928–1933, 2022.
- [24] H. B. Tambunan and T. W. D. Hapsari, “Analisis Opini Pengguna Aplikasi New PLN Mobile Menggunakan Text Mining,” *Petir*, vol. 15, no. 1, pp. 121–134, 2021, doi: 10.33322/petir.v15i1.1352.