Aisyah Journal of Informatics and Electrical Engineering
Vol. 8, No. 1, Februari 2026
ISSN 2686-0139 (Print), ISSN 2685-9556 (Online)
https://jti.aisyahuniversity.ac.id/index.php/AJIEE

Perbandingan Metode Naive Bayes dan Support Vector Machine Untuk Analisis Sentimen Pengguna X Terhadap Penyakit Virus HMPV di Indonesia

Muhamad Hartawan [1], Panji Bintoro [2], Hafsah Mukaromah [3], Agus Wantoro [4]*
[1,4]Program Studi Teknik Informatika, Universitas Aisyah Pringsewu, Pringsewu, Indonesia
[2]Program Studi Rekayasa Perangkat Lunak, Universitas Aisyah Pringsewu, Pringsewu, Indonesia
[3]Program Studi Farmasi, Universitas Aisyah Pringsewu, Pringsewu, Indonesia

Info Artikel

Riwayat Artikel:

Received January 26, 2026
Revised February 4, 2026
Accepted February 10, 2026

Abstract *– Human Metapneumovirus (HMPV) is a respiratory virus from the Pneumoviridae family that can infect both the upper and lower respiratory tract. Although no official cases have been identified in Indonesia, the spread of HMPV in several countries has raised public concern. Social media, particularly platform X (formerly Twitter), has become a primary platform for the public to express opinions, information, and discussions related to health issues. Sentiment analysis is needed to understand public perceptions, thus providing a basis for the government and medical professionals to formulate more appropriate communication strategies. This study aims to classify public sentiment regarding HMPV into positive, negative, and neutral categories and to compare the performance of two text classification methods: Naïve Bayes and Support Vector Machine (SVM). The research methodology uses the CRISP-DM framework with the following stages: data collection of 1,476 tweets using keywords related to HMPV, text preprocessing (cleaning, case folding, tokenizing, filtering, stemming), automatic labeling using IndoBERT, data balancing through resampling, and feature extraction with TF-IDF. The data was then split into training-test ratios (90:10, 80:20, 70:30, 60:40), followed by modeling and evaluation using a confusion matrix, accuracy, precision, recall, and F1-score. The results showed that SVM outperformed Naïve Bayes, with higher accuracy and F1-score in almost all data split scenarios. These findings confirm that SVM is superior in analyzing the sentiment of Indonesian-language text on social media. This research is expected to support public opinion monitoring, strengthen health communication strategies, and serve as a reference for further research on other health issues.*

Keywords: *Sentiment Analysis, HMPV, Naïve Bayes, SVM, IndoBERT*

Corresponding Author:
Panji Bintoro
Email:
panjibintoro09@aisyahuniversity.ac.id

Abstrak *– Human Metapneumovirus (HMPV) merupakan virus pernapasan dari famili Pneumoviridae yang dapat menginfeksi saluran pernapasan atas dan bawah. Meskipun belum ada laporan kasus resmi di Indonesia, penyebaran HMPV di beberapa negara telah menimbulkan kekhawatiran publik. Media sosial, khususnya platform X (sebelumnya Twitter), menjadi sarana utama masyarakat untuk menyampaikan informasi, opini, dan diskusi terkait isu kesehatan. Oleh karena itu, analisis sentimen diperlukan untuk memahami persepsi publik sebagai dasar perumusan strategi komunikasi kesehatan yang lebih efektif. Penelitian ini bertujuan mengklasifikasikan sentimen publik terhadap HMPV ke dalam kategori positif, negatif, dan netral, serta membandingkan kinerja metode Naïve Bayes dan Support Vector Machine (SVM). Metodologi penelitian mengacu pada kerangka kerja CRISP-DM yang meliputi pengumpulan 1.476 tweet dengan kata kunci terkait HMPV, pra-pemrosesan teks (pembersihan, case folding, tokenisasi, filtering, dan stemming), pelabelan otomatis menggunakan IndoBERT, penyeimbangan data melalui resampling, serta ekstraksi fitur menggunakan TF-IDF. Data dibagi ke dalam beberapa rasio pelatihan dan pengujian, kemudian dievaluasi menggunakan matriks kebingungan, akurasi, presisi, recall, dan F1-score. Hasil penelitian menunjukkan bahwa SVM secara konsisten menghasilkan akurasi dan F1-score lebih tinggi dibandingkan Naïve Bayes. Temuan ini mengindikasikan bahwa SVM lebih efektif untuk analisis sentimen teks berbahasa Indonesia di media sosial dan dapat mendukung pemantauan opini publik serta strategi komunikasi kesehatan.*

Kata Kunci*: Analisis Sentimen, HMPV, Naïve Bayes, SVM, IndoBERT*

I. PENDAHULUAN

Karena penyebarannya di banyak negara, *Human Metapneumovirus* (HMPV) adalah salah satu virus pernapasan yang menjadi perhatian dunia [1]. Hingga saat ini, belum ada kasus resmi HMPV di Indonesia. Namun, penyebaran informasi tentang virus ini di media sosial telah menyebabkan berbagai tanggapan dan pendapat dari masyarakat [2]. Infeksi saluran pernapasan akut (ISPA) dan patogen pernapasan yang umum diawasi secara teratur di banyak negara. Tingkat penyakit mirip flu seperti penyakit (ILI) dan infeksi saluran pernapasan akut (ISPA) meningkat di beberapa negara di daerah beriklim sedang di belahan bumi bagian utara dalam beberapa minggu terakhir,

mengikuti tren musiman biasa. Virus HMPV adalah virus pernafasan yang beredar di banyak negara dari musim dingin hingga musim semi. Namun, tidak semua negara secara teratur menguji dan mempublikasikan tren HMPV. Sebagian besar orang yang terinfeksi memiliki gejala yang ringan. seperti gejala pernapasan seperti flu yang hilang dalam beberapa hari. Jumlah kasus infeksi saluran pernafasan akut telah meningkat dalam beberapa minggu terakhir di Tiongkok, terutama di provinsi utara negara itu. Data yang diterbitkan oleh Tiongkok menunjukkan peningkatan kasus influenza musiman, rhinovirus, RSV, dan HMPV [3]. Di Indonesia, belum ada kasus HMPV yang dilaporkan hingga saat ini. Namun, masyarakat disarankan untuk mempertahankan kesehatan dengan menjalani gaya hidup bersih dan sehat [4]. Untuk memperkuat daya tahan tubuh dan mencegah penularan berbagai virus yang dapat membahayakan kesehatan, hal ini sangat penting. Selain itu, pemerintah Indonesia terus memantau perkembangan wabah HMPV di China dan negara lain. Peningkatan kewaspadaan di pintu masuk negara, termasuk pengawasan kekarantinaan kesehatan bagi pelaku perjalanan internasional yang menunjukkan gejala influenza seperti penyakit (ILI), dilakukan untuk mencegah hal ini terjadi [4].

Infeksi saluran pernapasan atas atau bawah disebabkan oleh virus HMPV. Gejala yang ditimbulkannya termasuk batuk, pilek, hidung tersumbat, dan demam [5]. Dalam kasus yang parah, virus ini dapat menyebabkan komplikasi seperti bronkitis atau pneumonia. Meskipun virus ini biasanya tidak berbahaya bagi orang dewasa yang sehat, anak-anak, orang tua, dan orang dengan sistem kekebalan tubuh yang lemah, seperti diabetes, gangguan pernapasan, atau penyakit jantung, lebih rentan. Sampai saat ini, belum ada pengobatan khusus untuk HMPV. Menjaga pola hidup sehat, mencuci tangan secara teratur, dan menggunakan masker di tempat umum dapat membantu mengurangi kemungkinan tertular penyakit menular [4].

Meskipun berbagai penelitian telah mengkaji analisis sentimen menggunakan metode Naïve Bayes dan Support Vector Machine (SVM), sebagian besar penelitian tersebut masih berfokus pada topik umum seperti politik, layanan publik, dan e-commerce. Penelitian yang secara khusus membahas isu kesehatan, terutama terkait Human Metapneumovirus (HMPV) di Indonesia, masih sangat terbatas. Selain itu, belum banyak penelitian yang memanfaatkan pelabelan otomatis berbasis model bahasa seperti IndoBERT serta menangani permasalahan ketidakseimbangan data secara sistematis. Oleh karena itu, diperlukan penelitian yang secara khusus mengkaji perbandingan performa Naïve Bayes dan SVM pada analisis sentimen isu HMPV berbasis data media sosial berbahasa Indonesia dengan pendekatan preprocessing dan resampling yang terstruktur.

Berdasarkan data yang diperoleh, penelitian ini berfokus pada analisis pendapat masyarakat tentang penyebaran penyakit Human Metapneumovirus (HMPV) di Indonesia. Dengan meningkatnya diskusi tentang masalah ini di media sosial, kita perlu memahami lebih baik bagaimana masyarakat merespon virus ini, baik dalam bentuk kekhawatiran tentang penyebarannya maupun tanggapan mereka terhadap tindakan pencegahan yang diambil pemerintah. Dengan menggunakan metode Naive Bayes dan Support Vector Machine (SVM), penelitian ini bertujuan untuk mengkategorikan opini publik tentang penyakit HMPV menjadi sentimen positif, negative, dan netral. Metode ini dipilih karena memiliki kemampuan untuk mengolah teks dan mengklasifikasikan sentimen secara efektif. Selain itu, telah terbukti akurat dalam banyak penelitian sebelumnya. Salah satu hasil yang diharapkan dari penelitian ini adalah memperoleh pemahaman tentang bagaimana masyarakat menggunakan media sosial untuk menanggapi penyakit HMPV dan mengukur akurasi model klasifikasi yang digunakan. Penelitian ini diharapkan dapat membantu pemerintah, tenaga kesehatan, dan pihak terkait membuat rencana komunikasi yang lebih baik untuk meningkatkan kesadaran masyarakat tentang HMPV dan mengontrol konten media sosial yang beredar.

II. METODE

A. Objek Penelitian

Objek penelitian merupakan fokus utama yang menjadi pusat perhatian dalam suatu penelitian. Dalam penelitian ini, objek yang diangkat adalah respon atau opini masyarakat Indonesia di media sosial X terhadap penyakit virus *HMPV (Human Metapneumovirus).* Penyakit ini belakangan menjadi perbincangan hangat di berbagai platform digital karena penyebarannya yang cepat dan gejala yang mirip dengan infeksi saluran pernapasan lainnya [6]. Media sosial, khususnya X, sering digunakan oleh masyarakat untuk menyampaikan opini, keluhan, hingga informasi mengenai isu-isu kesehatan. Banyaknya cuitan yang muncul terkait virus *HMPV* ini memunculkan berbagai sentimen, baik positif, negatif ataupun netral.

B. Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan dengan metode *crawling* menggunakan bantuan alat *Tweet Harvest*, yang merupakan salah satu perangkat lunak untuk pengambilan data dari media sosial *X* secara otomatis [7]. Proses *crawling* dilaksanakan dengan menentukan rentang waktu sejak tanggal 27 Januari 2022 hingga 20 Juni 2025. Dalam proses ini, digunakan sejumlah kata kunci berbahasa Indonesia yang relevan dengan topik penelitian, yaitu "*Virus HMPV*", "*Infeksi HMPV*", "*Penyakit HMPV*", dan "*Gejala HMPV*". Berdasarkan proses *crawling* tersebut, diperoleh sebanyak 1.476 *tweet* yang sesuai dengan kriteria pencarian dan berbahasa Indonesia. Data hasil crawling

ini selanjutnya digunakan sebagai sumber utama untuk analisis dalam penelitian. Untuk proses *crawling* data dapat dilihat pada Gambar 1 dan hasil dari proses ini dapat dilihat pada Gambar 2.

```
[11] df = pd.read_csv('Dataset_HMPV.csv')
     df
```

	conversation_id_str	created_at	favorite_count	full_text	id_str
0	1891342457368891615	Mon Feb 17 04:22:35 +0000 2025	0	Merebaknya virus HMPV di kalangan anak-anak bi...	1891342457368891615
1	1891186977241797116	Sun Feb 16 18:04:46 +0000 2025	1	Virus hmpv lagi rame Mulai lagi pake masker ke...	1891186977241797116
2	1890027024049484031	Fri Feb 14 11:35:28 +0000 2025	0	@wannabekopisusu 3 hari kak? Kaya masa inkunas...	1890364232060821716
3	1890261392026161562	Fri Feb 14 04:46:50 +0000 2025	0	Dapatkan tips: Apa itu Virus HMPV? Gejala ...	1890261394534388143
4	1889244974212116530	Tue Feb 11 09:27:59 +0000 2025	0	Menariknya isu HMPV virus baru yg konon mirip ...	1889244986765389869
...	...	...	...	...	...
1471	1874079245426634827	Tue Dec 31 13:04:45	0	China mengalami lonjakan penyakit	1874079245426634827

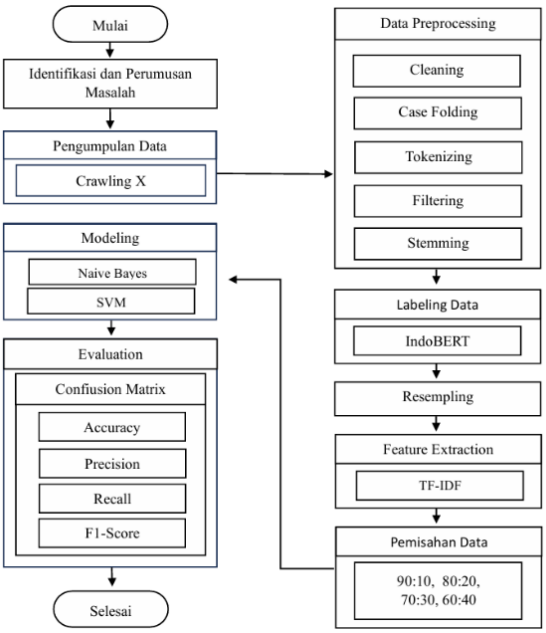
Gambar 1. Proses Crawling Data

```
# Craul Data

filename = 'Virus HMPV.csv'
search_keyword = 'Virus HMPV since:2020-01-01 until:2025-07-24 lang:id'
limits = 1000

!npx -y tweet-harvest@latest -o "{filename}" -s "{search_keyword}" --tab "LATEST" -l {limits} --token {twitter_auth_token}
opening twitter search page...

-- Scrolling... (1)
Filling in keywords: Virus HMPV since:2020-01-01 until:2025-07-24 lang:id

 (2) (3) (4)

Your tweets saved to: /content/tweets-data/Virus_HMPV.csv
Total tweets saved: 19

-- Scrolling... (1)

Your tweets saved to: /content/tweets-data/Virus_HMPV.csv
Total tweets saved: 98

-- Scrolling... (1)

Your tweets saved to: /content/tweets-data/Virus_HMPV.csv
Total tweets saved: 56
```

Gambar 2. Hasil Crawling Data

C. Pra-pemrosesan Data

Penelitian ini melakukan analisis sentimen terhadap opini masyarakat di X mengenai Virus *HMPV* di Indonesia dengan menerapkan metode *CRISP-DM.* Tahapan penelitian yang dilakukan dapat dilihat melalui diagram alir pada Gambar 3.



Gambar 3. Tahapan Penelitian

D. Data Preprocessing

Data yang telah diperoleh dari hasil *crawling* selanjutnya melalui tahapan *preprocessing* untuk memastikan kualitas dan konsistensi data sebelum dianalisis lebih lanjut. Proses *preprocessing* ini terdiri dari beberapa tahapan, yaitu *cleaning*, *case folding, tokenizing, filtering*, dan *stemming* [8].

- Cleaning

  Tahap *cleaning* dilakukan untuk menghapus karakter atau elemen yang tidak diperlukan, seperti tanda baca, angka, simbol, dan tautan (*URL*) [9]. Untuk hasil dari proses ini dapat dilihat pada Tabel 1.

TABEL 1
HASIL CLEANING DATA

Sebelum	Sesudah
@wannabekopisusu 3 hari kak? Kaya masa inkunasi virus HMPV	hari kak Kaya masa inkunasi virus HMPV
Menariknya isu HMPV virus baru yg konon mirip Covid19 berada pada perbatasan kuadran kiri bawah ke kuadran kanan atas.#evello #tiktok https://t.co/cec4ntRcZO	Menariknya isu HMPV virus baru yg konon mirip Covid berada pada perbatasan kuadran kiri bawah ke kuadran kanan atasevello tiktok

- Case Folding

  Selanjutnya, *case folding* bertujuan untuk menyeragamkan semua huruf menjadi huruf kecil (*lowercase*) guna menghindari perbedaan akibat kapitalisasi [10]. Untuk hasil dari proses ini dapat dilihat pada Table 2.

TABEL 2
HASIL CASE FOLDING

Sebelum	Sesudah
hari kak Kaya masa inkunasi virus HMPV	hari kak kaya masa inkunasi virus hmpv
Menariknya isu HMPV virus baru yg konon mirip Covid berada pada perbatasan kuadran kiri bawah ke kuadran kanan atasevello tiktok	menariknya isu hmpv virus baru yg konon mirip covid berada pada perbatasan kuadran kiri bawah ke kuadran kanan atasevello tiktok

- Tokenizing

  Proses *tokenizing* memecah teks menjadi satuan kata atau token [11]. Untuk hasil dari proses ini dapat dilihat pada Tabel 3.

TABEL 3
HASIL TOKENIZING

Sebelum	Sesudah
hari kak kaya masa inkunasi virus hmpv	['hari', 'kak', 'kaya', 'masa', 'inkunasi', 'virus', 'hmpv']
menariknya isu hmpv virus baru yg konon mirip covid berada pada perbatasan kuadran kiri bawah ke kuadran kanan atasevello tiktok	['menariknya', 'isu', 'hmpv', 'virus', 'baru', 'yg', 'konon', 'mirip', 'covid', 'berada', 'pada', 'perbatasan', 'kuadran', 'kiri', 'bawah', 'ke', 'kuadran', 'kanan', 'atasevello', 'tiktok']

- Filtering

  Pada tahap *filtering*, dilakukan penghapusan kata-kata yang termasuk dalam daftar *stopword* atau kata yang tidak memiliki makna penting dalam analisis [12]. Untuk hasil dari proses ini dapat dilihat pada Tabel 4.

TABEL 4
HASIL FILTERING

Sebelum	Sesudah
['hari', 'kak', 'kaya', 'masa', 'inkunasi', 'virus', 'hmpv']	['hari', 'kak', 'kaya', 'masa', 'inkunasi', 'virus', 'hmpv']
['menariknya', 'isu', 'hmpv', 'virus', 'baru', 'yg', 'konon', 'mirip', 'covid', 'berada', 'pada', 'perbatasan', 'kuadran', 'kiri', 'bawah', 'ke', 'kuadran', 'kanan', 'atasevello', 'tiktok']	['menariknya', 'isu', 'hmpv', 'virus', 'baru', 'konon', 'mirip', 'covid', 'berada', 'perbatasan', 'kuadran', 'kiri', 'bawah', 'kuadran', 'kanan', 'atasevello', 'tiktok']

- Stemming

  Terakhir, proses *stemming* digunakan untuk mengubah kata ke bentuk dasarnya (*root word*), sehingga variasi kata dengan makna yang sama dapat disatukan [13]. Untuk hasil dari proses ini dapat dilihat pada Tabel 5.

TABEL 5
HASIL STEMMING

Sebelum	Sesudah
['hari', 'kak', 'kaya', 'masa', 'inkunasi', 'virus', 'hmpv']	['hari', 'kak', 'kaya', 'masa', 'inkunasi', 'virus', 'hmpv']
['menariknya', 'isu', 'hmpv', 'virus', 'baru', 'konon', 'mirip', 'covid', 'berada', 'perbatasan', 'kuadran', 'kiri', 'bawah', 'kuadran', 'kanan', 'atasevello', 'tiktok']	['tarik', 'isu', 'hmpv', 'virus', 'baru', 'konon', 'mirip', 'covid', 'ada', 'batas', 'kuadran', 'kiri', 'bawah', 'kuadran', 'kanan', 'atasevello', 'tiktok']

Setelah semua proses telah dilaksanakan akan menghasilkan kalimat 1358 yang siap diolah ke proses berikutnya, untuk hasil final dari proses ini dapat dilihat pada Gambar 4.

```
--- SAMPLE 3 ---
Original     : @wannabekopisusu 3 hari kak? Kaya masa inkunasi virus HMPV
Cleaned      : hari kak Kaya masa inkunasi virus HMPV
Lowercase    : hari kak kaya masa inkunasi virus hmpv
Tokenized    : ['hari', 'kak', 'kaya', 'masa', 'inkunasi', 'virus', 'hmpv']
Filtered     : ['hari', 'kak', 'kaya', 'masa', 'inkunasi', 'virus', 'hmpv']
Stemmed      : ['hari', 'kak', 'kaya', 'masa', 'inkunasi', 'virus', 'hmpv']
Final Result : hari kak kaya masa inkunasi virus hmpv
```

Gambar 4. Hasil Run Program Akhir Data Preprocessing

E. Labeling

Proses *labeling* data pada penelitian ini dilakukan secara otomatis dengan memanfaatkan model bahasa berbasis *transformer*, yaitu *IndoBERT*. *IndoBERT* merupakan salah satu model pretrained yang telah dilatih secara khusus pada korpus berbahasa Indonesia, sehingga mampu memahami konteks dan struktur kalimat dalam bahasa Indonesia dengan baik [14]. Model ini digunakan untuk mengklasifikasikan sentimen dari setiap data teks yang telah melalui proses preprocessing, ke dalam kategori tertentu secara otomatis, tanpa perlu pelabelan manual. Penggunaan *IndoBERT* diharapkan dapat meningkatkan akurasi dan efisiensi dalam proses anotasi data, mengingat model ini telah terbukti memiliki performa yang unggul dalam berbagai tugas pemrosesan bahasa alami (*NLP*) berbahasa Indonesia [15]. Mengasilkan statistik labeling yang telah di jalankan dengan mengeluarkan distribusi label serta sampel hasil *labeling* berupa *text* dan *score* labelnya dapat dilihat pada Gambar 5.

```
=== STATISTIK LABELING ===
Total data berlabel: 1358

Distribusi label:
- netral: 1114 (82.0%)
- negatif: 222 (16.3%)
- positif: 22 (1.6%)

=== SAMPLE HASIL LABELING ===

Text 1: rebak virus hmpv kalang anakanak akibat fatal tahu cara cegah tahu dasar kena vi...
Label: negatif (confidence: 0.825)

Text 2: virus hmpv rame mulai pake masker mana ges kalau kelen betul baca tweet segera s...
Label: netral (confidence: 0.553)

Text 3: hari kak kaya masa inkunasi virus hmpv...
Label: negatif (confidence: 0.999)

Text 4: dapat tips apa virus hmpv gejala cara tular langkah cegah tangan jangan lewat ki...
Label: netral (confidence: 1.000)
```

Gambar 5. Hasil Run Program Labeling IndoBERT

F. Resampling

Dalam penelitian ini, dilakukan proses resampling terhadap dataset yang telah diberi label menggunakan *IndoBERT* karena distribusi kelas yang tidak seimbang secara signifikan. Terdapat 1.114 data berlabel netral, 222 negatif, dan hanya 22 positif. Ketimpangan ini dapat menyebabkan model pembelajaran mesin menjadi bias terhadap kelas mayoritas (netral), sehingga mengurangi akurasi dan performa dalam mendeteksi kelas minoritas (positif dan negatif). Untuk mengatasi permasalahan tersebut, dilakukan resampling menggunakan pendekatan *Random Oversampling* agar jumlah data di setiap kelas menjadi seimbang [16]. Proses ini dilakukan melalui kode Python yang telah dirancang untuk membaca dataset yang telah dilabeling (*step3_labeled_text.csv*) dengan tiga kolom utama: text, label, dan score. Dengan teknik ini, model pembelajaran seperti *Naive Bayes* dan *SVM* dapat dilatih secara lebih adil terhadap seluruh kelas, sehingga diharapkan mampu memberikan performa klasifikasi yang lebih representatif dan tidak bias terhadap kelas mayoritas. Untuk hasil dari proses ini dapat dilihat pada Tabel 6.

TABEL 6
HASIL RESAMLING

Label	Sebelum	Sesudah
Netral	1114	1114
Negatif	222	1114
Positif	22	1114

G. Feature Extraction

Setelah data diberi label dan melalui tahap preprocessing, langkah selanjutnya adalah proses ekstraksi fitur (*feature extraction*) yang bertujuan untuk mengubah data teks menjadi representasi numerik yang dapat diolah oleh model klasifikasi. Pada penelitian ini, metode yang digunakan untuk ekstraksi fitur adalah *Term Frequency–Inverse*

*Document Frequen*cy (*TF-IDF*). *TF-IDF* merupakan salah satu teknik pembobotan kata yang mengukur seberapa penting suatu kata dalam sebuah dokumen relatif terhadap keseluruhan korpus [17]. Nilai *TF-IDF* yang dihasilkan mencerminkan bobot atau relevansi kata dalam konteks dokumen tertentu, sehingga dapat digunakan sebagai masukan (input features) dalam proses klasifikasi data [18]. Untuk hasil dari proses ini menampilkan *TF-IDF vectorization*, analisis perkelas dan statistik dari TF-IDF dapat dilihat pada Gambar 6.

```
=== TF-IDF VECTORIZATION ===
Total data: 3342
Melakukan TF-IDF vectorization...
Shape TF-IDF matrix: (3342, 5000)
Jumlah fitur: 5000
Sparsity: 99.43%

=== STATISTIK TF-IDF ===
Nilai TF-IDF minimum : 0.023583
Nilai TF-IDF maksimum: 1.000000
Nilai TF-IDF rata-rata: 0.169989
Nilai TF-IDF median   : 0.153212

=== HASIL DISIMPAN ===
- step4_tfidf_matrix.csv
- step4_feature_importance.csv
- step4_tfidf_info.csv
```

Gambar 6. Hasil Run Program Feature Extraction TF-IDF

H. Pemisahana Data

Dalam penelitian ini, data yang telah melalui proses preprocessing, *labeling*, dan *feature extraction* kemudian dibagi ke dalam beberapa skenario proporsi data latih (*training* data) untuk mengevaluasi performa model secara lebih menyeluruh. Adapun pemisahan data dilakukan ke dalam empat skenario, yaitu sebanyak 10%, 20%, 30%, dan 40% dari total keseluruhan data. Tujuan dari variasi proporsi ini adalah untuk mengamati sejauh mana peningkatan jumlah data latih dapat mempengaruhi akurasi dan kinerja model klasifikasi [19]. Sisa data dari masing-masing skenario digunakan sebagai data uji (*testing* data), sehingga evaluasi model dilakukan secara konsisten berdasarkan variasi proporsi data latih yang digunakan. Untuk hasil dari proses ini dibagi menjadi 4 presentase pemisahan data yang ditunjukkan pada Tabel 7, distribusi ditunjukkan pada Tabel 8, distribusi label berdasarkan jumlah pemisahan data dapat dilihat pada Tabel 9.

TABEL 7
HASIL PEMISAHAN DATA

Pemisahan	Latih	Uji	Total
90/10%	3007	335	3342
80/20%	2673	669	3342
70/30%	2339	1003	3342
60/40%	2005	1337	3342

TABEL 8
DISTRIBUSI LABEL

Distribusi Label	Total
Netral (Net)	1114
Negatif (Neg)	1114
Positif (Pos)	1114

TABEL 9
DISTRIBUSI LABEL BERDASARKAN PEMISAHAN DATA

Pemisahan	Latih Net	Uji Net	Latih Neg	Uji Neg	Latih Pos	Uji Pos
90/10%	1003	1002	1002	112	112	111
80/20%	891	891	891	223	223	223
70/30%	780	780	779	335	334	334
60/40%	669	668	668	446	446	445

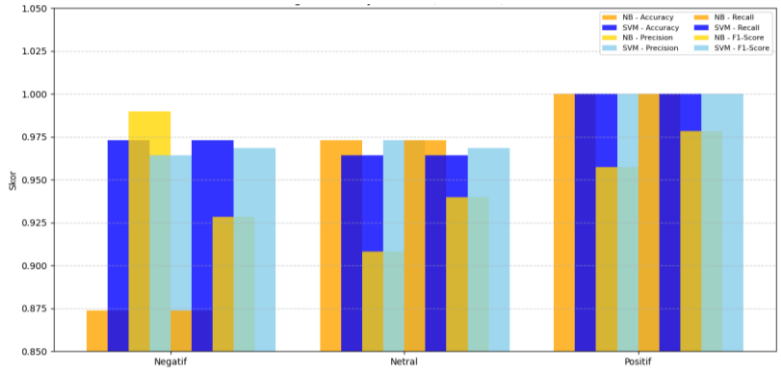
III. HASIL DAN PEMBAHASAN

A. Modeling dan Evaluasi

Pada tahap *modeling*, penelitian ini menerapkan dua algoritma klasifikasi yang umum digunakan dalam analisis teks, yaitu *Support Vector Machine (SVM)* dan *Naive Bayes*. Kedua algoritma ini dipilih karena dikenal memiliki performa yang baik serta efisien dalam mengolah data teks, khususnya dalam konteks klasifikasi berbasis sentimen. Setelah model dibangun, evaluasi dilakukan untuk menilai kinerja masing-masing algoritma menggunakan empat metrik utama, yaitu akurasi, presisi, recall, dan *F1-score* [20]. Akurasi digunakan untuk mengukur proporsi prediksi yang benar terhadap seluruh data, sedangkan presisi menunjukkan ketepatan model dalam mengklasifikasikan data yang relevan. Recall digunakan untuk menilai sejauh mana model mampu menangkap seluruh data yang tergolong positif, dan *F1-score* digunakan untuk memberikan gambaran keseimbangan antara presisi dan recall. Seluruh evaluasi dilakukan terhadap masing-masing model pada berbagai skenario proporsi data latih, yaitu 10%, 20%, 30%, dan 40%, guna memperoleh perbandingan performa yang komprehensif. Hasil evaluasi perkelas naïve bayes dan SVM pada pemisahan 90:10 ditunjukkan pada Tabel 10 dan Gambar 7.

TABEL 10
HASIL EVALUASI PERKELAS NAÏVE BAYES DAN SVM 90:10

Metric	NB(Neg)	NB(Net)	NB(Pos)	SVM(Neg)	SVM(Net)	SVM(Pos)
Akurasi	0.8739	0.9732	1.0000	0.9730	0.9643	1.0000
Presisi	0.9898	0.9083	0.9573	0.9643	0.9730	1.0000
Recall	0.8739	0.9732	1.0000	0.9730	0.9643	1.0000



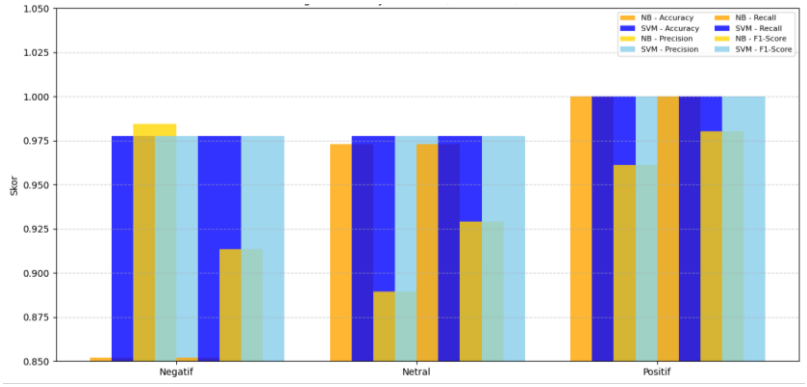
Gambar 7. Perbandingan *Naïve Bayes* vs *SVM* (Rasio 90:10)

Berdasarkan hasil evaluasi performa per kelas yang ditampilkan pada Tabel 10 dan Gambar 7, terlihat bahwa algoritma *Naïve Bayes* dan *SVM* menunjukkan performa yang sangat baik pada pemisahan data latih dan data uji sebesar 90:10. Untuk kelas negatif, akurasi *Naïve Bayes* sebesar 0,8739 dengan presisi tinggi mencapai 0,9898, recall 0,8739, dan F1-score sebesar 0,9282. Hal ini menunjukkan bahwa meskipun sistem sangat tepat dalam mengidentifikasi kelas negatif (presisi tinggi), terdapat beberapa data kelas negatif yang gagal dikenali (recall relatif lebih rendah). Pada kelas netral, performa *Naïve Bayes* juga cukup baik dengan akurasi 0,9732, presisi 0,9083, recall 0,9732, dan F1-score 0,9397, yang menunjukkan keseimbangan antara kemampuan model dalam mengenali dan memprediksi kelas tersebut. Untuk kelas positif, seluruh metrik evaluasi pada *Naïve Bayes* menunjukkan nilai sempurna atau mendekati sempurna, dengan akurasi dan recall sebesar 1.0000, serta presisi 0,9573 dan F1-score 0,9782.

Sementara itu, algoritma *SVM* secara umum menunjukkan performa yang lebih stabil di semua kelas. Untuk kelas negatif, *SVM* mencatatkan akurasi 0,9730, presisi 0,9643, recall 0,9730, dan F1-score 0,9686. Kinerja ini menandakan bahwa model mampu mengenali dan memprediksi kelas negatif dengan baik dan konsisten. Pada kelas netral, *SVM* juga memberikan hasil yang tinggi dengan akurasi 0,9643, presisi 0,9730, recall 0,9643, dan F1-score 0,9686. Adapun untuk kelas positif, seluruh metrik evaluasi menunjukkan nilai sempurna (1.0000), yang menandakan bahwa *SVM* mampu mengklasifikasikan seluruh data kelas positif dengan tepat tanpa kesalahan. Secara keseluruhan, baik *Naïve Bayes* maupun *SVM* menunjukkan kinerja yang sangat baik, namun *SVM* cenderung lebih stabil dan unggul secara konsisten di ketiga kelas. Hasil evaluasi perkelas naïve bayes dan SVM pada pemisahan 80:20 ditunjukkan pada Tabel 11 dan Gambar 8.

TABEL 11
HASIL EVALUASI PERKELAS NAÏVE BAYES DAN SVM 80:20

Metric	NB(Neg)	NB(Net)	NB(Pos)	SVM(Neg)	SVM(Net)	SVM(Pos)
Akurasi	0.8520	0.9731	1.0000	0.9776	0.9776	1.0000
Presisi	0.9845	0.8893	0.9612	0.9776	0.9776	1.0000
Recall	0.8520	0.9731	1.0000	0.9776	0.9776	1.0000

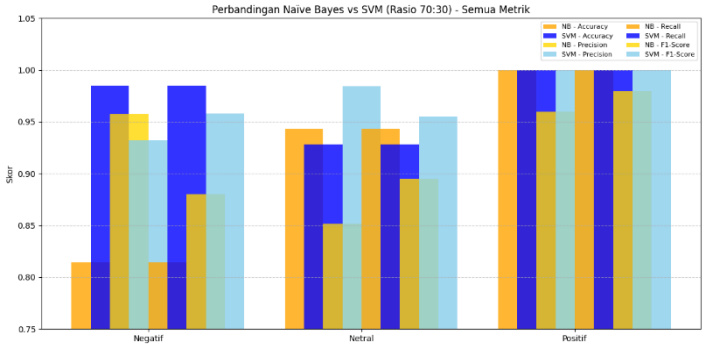


Gambar 8. Perbandingan *Naïve Bayes* vs *SVM* (Rasio 80:20)

Tabel 11 dan Gambar 8 menyajikan hasil evaluasi performa model klasifikasi *Naïve Bayes* dan *Support Vector Machine (SVM)* terhadap data latih dan data uji dengan rasio pemisahan 80:20, berdasarkan tiga label sentimen yaitu negatif, netral, dan positif. Berdasarkan hasil evaluasi, model *Naïve Bayes* menunjukkan akurasi sebesar 85,20% untuk kelas negatif, 97,31% untuk kelas netral, dan 100% untuk kelas positif. Presisi tertinggi dicapai oleh kelas negatif sebesar 98,45%, sementara recall tertinggi sebesar 100% diperoleh pada kelas positif. F1-score tertinggi juga diperoleh pada kelas positif, yakni sebesar 98,02%. Sementara itu, model *SVM* menunjukkan performa yang lebih stabil dan tinggi secara keseluruhan, dengan akurasi, presisi, recall, dan F1-score sebesar 97,76% untuk kelas negatif dan netral, serta mencapai nilai sempurna (100%) untuk kelas positif pada keempat metrik evaluasi. Hasil ini menunjukkan bahwa *SVM* memiliki keunggulan dalam mengklasifikasikan data sentimen secara konsisten dibandingkan *Naïve Bayes*. Hasil evaluasi perkelas naïve bayes dan SVM pada pemisahan 70:30 ditunjukkan pada Tabel 12 dan Gambar 9.

TABEL 12
HASIL EVALUASI PERKELAS NAÏVE BAYES DAN SVM 70:30

Metric	NB(Neg)	NB(Net)	NB(Pos)	SVM(Neg)	SVM(Net)	SVM(Pos)
Akurasi	0.8144	0.9431	1.0000	0.9850	0.9281	1.0000
Presisi	0.9577	0.8514	0.9599	0.9320	0.9841	1.0000
Recall	0.8144	0.9431	1.0000	0.9850	0.9281	1.0000

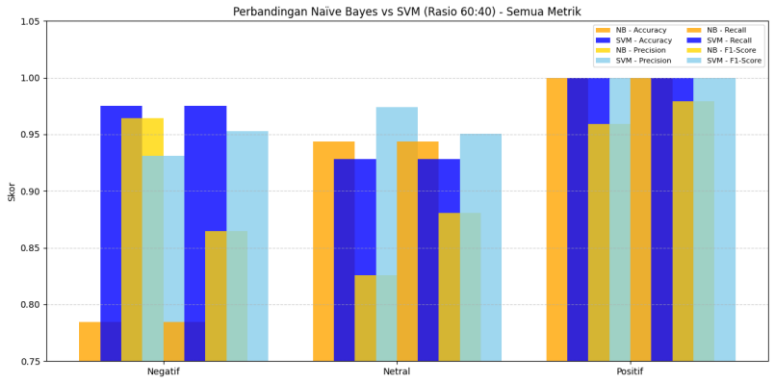


Gambar 9. Perbandingan *Naïve Bayes* vs *SVM* (Rasio 70:30)

Tabel 12 dan Gambar 9 menunjukkan hasil evaluasi kinerja model *klasifikasi Naïve Bayes* dan *Support Vector Machine (SVM)* berdasarkan metrik evaluasi per kelas, yaitu negatif, netral, dan positif dengan skenario pemisahan data latih dan data uji sebesar 70:30. Berdasarkan hasil yang ditampilkan, model *Naïve Bayes* menghasilkan akurasi sebesar 81,44% pada kelas negatif, 94,31% pada kelas netral, dan mencapai 100% pada kelas positif. Presisi untuk masing-masing kelas yaitu 95,77% (negatif), 85,14% (netral), dan 95,99% (positif), dengan recall yang sama besar dengan akurasi untuk tiap kelas. Sementara itu, model *SVM* menunjukkan performa yang lebih baik, terutama pada kelas negatif dengan akurasi dan recall sebesar 98,50%, serta presisi sebesar 93,20%. Pada kelas netral, *SVM* mencapai presisi tertinggi sebesar 98,41% dan recall sebesar 92,81%. Untuk kelas positif, baik *Naïve Bayes* maupun *SVM* sama-sama mencapai nilai sempurna (100%) pada semua metrik evaluasi. Secara umum, *SVM* menunjukkan performa yang lebih unggul dibandingkan *Naïve Bayes* pada kelas negatif dan netral, sedangkan keduanya sama-sama optimal pada kelas positif. Hasil evaluasi perkelas naïve bayes dan SVM pada pemisahan 60:40 ditunjukkan pada Tabel 13 dan Gambar 10.

TABEL 13
HASIL EVALUASI PERKELAS NAÏVE BAYES DAN SVM 60:40

Metric	NB(Neg)	NB(Net)	NB(Pos)	SVM(Neg)	SVM(Net)	SVM(Pos)
Akurasi	0.7843	0.9439	1.0000	0.9753	0.9283	1.0000
Presisi	0.9641	0.8255	0.9591	0.9313	0.9741	1.0000
Recall	0.7843	0.9439	1.0000	0.9753	0.9283	1.0000

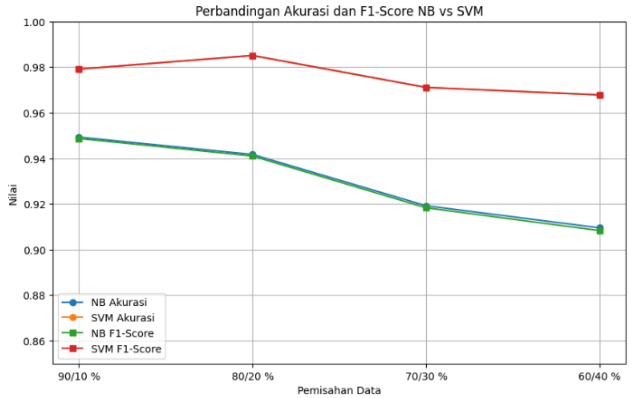


Gambar 10. Perbandingan *Naïve Bayes* vs *SVM* (Rasio 60:40)

Tabel 13 dan gambar 10 menunjukkan hasil evaluasi model *Naïve Bayes* dan *Support Vector Machine (SVM)* dengan pemisahan data latih dan uji sebesar 60:40 berdasarkan masing-masing kelas sentimen, yaitu negatif (Neg), netral (Net), dan positif (Pos). Pada model *Naïve Bayes*, akurasi tertinggi diperoleh pada kelas positif sebesar 1.0000, diikuti oleh kelas netral sebesar 0.9439, dan kelas negatif sebesar 0.7843. Nilai presisi juga menunjukkan hasil yang tinggi untuk semua kelas, dengan nilai tertinggi pada kelas negatif sebesar 0.9641, sedangkan nilai recall tertinggi juga diperoleh pada kelas positif sebesar 1.0000. Sementara itu, model *SVM* menunjukkan performa yang lebih konsisten dengan akurasi di atas 0.92 untuk semua kelas, dan akurasi tertinggi sebesar 1.0000 pada kelas positif. Model SVM juga mencatat nilai presisi tertinggi pada kelas netral sebesar 0.9741, dan nilai F1-Score tertinggi, yaitu 1.0000, pada kelas positif. Secara keseluruhan, baik *Naïve Bayes* maupun *SVM* menunjukkan performa yang sangat baik pada kelas positif, namun *SVM* cenderung memberikan hasil yang lebih stabil di ketiga kelas jika dibandingkan dengan *Naïve Bayes.* Rata-rata hasil akurasi dan F1-score naïve bayes dan SVM ditunjukkan pada Tabel 14 dan Gambar 11.

TABEL 14
RATA-RATA DARI HASIL AKURASI DAN F1-SCORE NAÏVE BAYES DAN SVM

Pemisahan	NB Akurasi	SVM Akurasi	NB F1-Score	SVM F1-Score
90/10 %	0.9493	0.9791	0.9487	0.9791
80/20 %	0.9417	0.9851	0.9410	0.9851
70/30 %	0.9192	0.9711	0.9183	0.9711
60/40 %	0.9095	0.9678	0.9083	0.9678



Gambar 11. Perbandingan Akurasi dan F1-Score Naïve Bayes vs SVM

Tabel 14 dan gambar 11 menyajikan rata-rata hasil akurasi dan *F1-score* dari metode klasifikasi *Naïve Bayes* dan *Support Vector Machine (SVM)* berdasarkan empat skema pemisahan data latih dan data uji, yaitu 90:10, 80:20, 70:30, dan 60:40. Berdasarkan hasil tersebut, terlihat bahwa kinerja metode *SVM* secara konsisten lebih unggul dibandingkan Naïve Bayes pada setiap skema pemisahan data. Pada pemisahan 90:10, *SVM* mencatatkan nilai akurasi dan *F1-score* tertinggi sebesar 0.9791, sedangkan *Naïve Bayes* berada pada angka 0.9493 untuk akurasi dan 0.9487 untuk *F1-score*. Pola yang sama juga terlihat pada pemisahan 80:20, di mana akurasi dan *F1-score SVM* mencapai 0.9851, mengungguli *Naïve Bayes* yang hanya meraih 0.9417 untuk akurasi dan 0.9410 untuk F1-score. Penurunan performa secara bertahap terjadi seiring dengan semakin kecilnya proporsi data latih, namun SVM tetap menunjukkan performa yang stabil dan tinggi dibandingkan *Naïve Bayes.* Dari data ini dapat disimpulkan bahwa metode *SVM* memberikan hasil klasifikasi yang lebih andal dan akurat dalam konteks analisis sentimen pada dataset yang digunakan. Hal ini menunjukkan bahwa *SVM* memiliki kemampuan generalisasi yang lebih baik dalam mengenali pola sentimen pada data teks, khususnya ketika digunakan dalam skenario dengan jumlah data latih yang beragam.

B. Word Cloud

Berikut ini adalah kata kata yang sering muncul di masing masing label positif, negatif dan netral yang sudah dibentuk ke dalam word cloud guna memperlihatkan ke dalam bentuk visual [21].

Gambar 12. Word Cloud Positif

Gambar 13. Word Cloud Negatif

Gambar 14. Word Cloud Netral

IV. KESIMPULAN

Berdasarkan hasil evaluasi model Naïve Bayes (NB) dan Support Vector Machine (SVM) pada analisis sentimen dengan tiga kelas (negatif, netral, dan positif) serta berbagai skema pemisahan data (90:10, 80:20, 70:30, dan 60:40), dapat disimpulkan bahwa data yang digunakan mengalami ketidakseimbangan dengan dominasi sentimen positif, sehingga diperlukan teknik resampling. Hasil pengujian menunjukkan bahwa SVM secara konsisten memiliki performa lebih unggul dibandingkan NB, dengan akurasi dan F1-score berkisar antara 96,78% hingga 98,51%, sedangkan NB berada pada kisaran 90,83% hingga 94,93%. Skema pemisahan data 80:20 memberikan hasil terbaik pada kedua model, khususnya pada SVM dengan nilai akurasi dan F1-score tertinggi sebesar 98,51%, meskipun skema 90:10 juga menunjukkan performa yang relatif optimal. Oleh karena itu, penelitian selanjutnya disarankan untuk menerapkan teknik resampling seperti SMOTE, ADASYN, atau undersampling, menguji model lain seperti Random Forest, Gradient Boosting, serta deep learning (LSTM atau BERT), dan menggunakan dataset yang lebih besar serta beragam guna meningkatkan akurasi dan kemampuan generalisasi model.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada Universitas Aisyah Pringsewu atas dukungan dan fasilitas yang diberikan selama penelitian ini berlangsung. Terima kasih juga kepada Bapak Panji Bintoro, S.Kom., M.Cs selaku dosen pembimbing. Tidak lupa apresiasi diberikan kepada keluarga dan sahabat yang senantiasa memberikan dukungan moril dan materil.

DAFTAR PUSTAKA

[1] N. M. S. Gálvez *et al.*, "Host components that modulate the disease caused by hmpv," *Viruses*, vol. 13, no. 3, pp. 1–21, 2021, doi: 10.3390/v13030519.
[2] W. Ji *et al.*, "Clinical and epidemiological characteristics of 96 pediatric human metapneumovirus infections in Henan, China after COVID-19 pandemic: a retrospective analysis," *Virol. J.*, vol. 21, no. 1, pp. 1–14, 2024, doi: 10.1186/s12985-024-02376-0.
[3] WHO, "Trends of acute respiratory infection, including human metapneumovirus, in the Northern Hemisphere," WHO. Accessed: Jan. 09, 2025. [Online]. Available: https://www.who.int/emergencies/disease-outbreak-news/item/2025-DON550
[4] Kemenkes RI, "Wabah Virus HMPV Merebak di China, Kemenkes Imbau Publik untuk Waspada," kemenkes. Accessed: Jan. 09, 2025. [Online]. Available: https://kemkes.go.id/id/Wabah-Virus-HMPV-Merebak-di-China, Kemenkes-Imbau-Publik-untuk-Waspada
[5] ALODOKTER, "HMPV, Ketahui Gejala, Penyebab, dan Pencegahannya." Accessed: Mar. 12, 2025. [Online]. Available: https://www.alodokter.com/hmpv-ketahui-gejala-penyebab-dan-pencegahannya
[6] X. Xia, X. Chen, S. Wang, X. Zhang, H. Cai, and J. Yu, "Clinical epidemiological features of hMPV infection in children with acute respiratory tract infection," *Rev. Psiquiatr. Clin.*, vol. 50, no. 6, pp. 78–84, 2023, doi: 10.15761/0101-60830000000711.
[7] A. Wafda, "Aspect-Based Sentiment Analysis terhadap Cuitan Platform X tentang Kurikulum Merdeka Menggunakan IndoBERT," 2025, [Online]. Available: https://dspace.uii.ac.id/handle/123456789/55157%0Ahttps://dspace.uii.ac.id/bitstream/handle/123456789/55157/22917022.pdf?sequence=1
[8] D. Duei Putri, G. F. Nama, and W. E. Sulistiono, "Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier," *J. Inform. dan Tek. Elektro Terap.*, vol. 10, no. 1, pp. 34–40, 2022, doi: 10.23960/jitet.v10i1.2262.
[9] F. Syadid, "Analisis Sentimen Komentar Netizen Terhadap Calon Presiden Indonesia 2019 Dari Twitter Menggunakan Algoritma Term Frequency-Invers Document Frequency (Tf- Idf) Dan Metode Multi Layer Perceptron (Mlp) Neural Network," *Skripsi Univ. Islam Negeri Syarif Hidayatullah Jakarta*, p. 72, 2019.
[10] F. N. Hidayat and S. Sugiyono, "Analisis Sentimen Masyarakat Terhadap Perekrutan Pppk Pada Twitter Dengan Metode Naive Bayes Dan Support Vector Machine," *J. Sains dan Teknol.*, vol. 5, no. 2, pp. 665–672, 2023, doi: 10.55338/saintek.v5i2.1359.
[11] S. Wulandari and F. N. Hasan, "Analisis Sentimen Masyarakat Indonesia Terhadap Pengalaman Belanja Thrifting Pada Media Sosial Twitter Menggunakan Algoritma Naïve Bayes," *J. Media Inform. Budidarma*, vol. 8, no. 2, p. 768, 2024, doi: 10.30865/mib.v8i2.7520.
[12] F. Amaliah and I. K. Dwi Nuryana, "Perbandingan Akurasi Metode Lexicon Based Dan Naive Bayes Classifier Pada Analisis Sentimen Pendapat Masyarakat Terhadap Aplikasi Investasi Pada Media Twitter," *J. Informatics Comput. Sci.*, vol. 3, no. 03, pp. 384–393, 2022, doi: 10.26740/jinacs.v3n03.p384-393.
[13] M. W. A. Putra, Susanti, Erlin, and Herwin, "Analisis Sentimen Dompet Elektronik Pada Twitter Menggunakan Metode Naïve Bayes Classifier," *IT J. Res. Dev.*, vol. 5, no. 1, pp. 72–86, 2020, doi: 10.25299/itjrd.2020.vol5(1).5159.
[14] U. Khairani, V. Mutiawani, and H. Ahmadian, "Pengaruh Tahapan Preprocessing Terhadap Model Indobert Dan Indobertweet Untuk Mendeteksi Emosi Pada Komentar Akun Berita Instagram," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 4, pp. 887–894, 2024, doi: 10.25126/jtiik.1148315.
[15] FAHRENDRA KHOIRUL IHTADA, "STUDI PERBANDINGAN METODE EKSTRAKSI FITUR UNTUK TOPIC MODELING BERBASIS ASPEK DAN SENTIMEN ANALISIS PADA ULASAN PRODUK E-COMMERCE," 2025.
[16] A. Syukron and A. Subekti, "Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit," *J. Inform.*, vol. 5, no. 2, pp. 175–185, 2018, doi: 10.31311/ji.v5i2.4158.
[17] E. A. Lisangan, A. Gormantara, and R. Y. Carolus, "Implementasi Naive Bayes pada Analisis Sentimen Opini Masyarakat di Twitter Terhadap Kondisi New Normal di Indonesia," *KONSTELASI Konvergensi Teknol. dan Sist. Inf.*, vol. 2, no. 1, pp. 23–32, 2022, doi: 10.24002/konstelasi.v2i1.5609.
[18] N. Agustina, D. H. Citra, W. Purnama, C. Nisa, and A. R. Kurnia, "Implementasi Algoritma Naive Bayes untuk Analisis Sentimen Ulasan Shopee pada Google Play Store," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 2, no. 1, pp. 47–54, 2022, doi: 10.57152/malcom.v2i1.195.
[19] M. Farhan, "ANALISIS PERBANDINGAN PENGARUH VARIASI DATA AUGMENTASI TERHADAP KINERJA MOBILENETV2 DALAM KLASIFIKASI PENYAKIT DAUN TEH," no. February, pp. 4–6, 2024.
[20] S. Puad, G. Garno, and A. Susilo Yuda Irawan, "Analisis Sentimen Masyarakat Pada Twitter Terhadap Pemilihan Umum 2024 Menggunakan Algoritma Naïve Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 3, pp. 1560–1566, 2023, doi: 10.36040/jati.v7i3.6920.
[21] H. P. Doloksaribu and Yusran Timur Samuel, "Komparasi Algoritma Data Mining Untuk Analisis Sentimen Aplikasi Pedulilindungi," *J. Teknol. Inf. J. Keilmuan dan Apl. Bid. Tek. Inform.*, vol. 16, no. 1, pp. 1–11, 2022, doi: 10.47111/jti.v16i1.3747.