JURNAL JTIK
(JURNAL TEKNOLOGI INFORMASI DAN KOMUNIKASI)
KITA Komunitas Informasi Teknologi Aceh
OPEN ACCESS
E-ISSN: 2580-1643
Volume 11, Nomor 1, Januari 2027 | Pages: 191-201 | DOI: https://doi.org/10.35870/jtik.v11i1.7526

Analisis Perbandingan Algoritma *Naïve Bayes* dan *Support Vector Machine* dalam Klasifikasi Opini dan Fakta pada Berita Banjir Sumatera

Dessy Natalia Reba [1*], Lilis Indrayani [2], Christian Dwi Suhendra [3]

[1*,2,3] Program Studi Teknik Informatika, Fakultas Teknik, Universitas Papua, Kabupaten Manokwari, Provinsi Papua Barat, Indonesia.

Corresponding Email: rebadavina@gmail.com [1*]

Article History: Received: 12 June 2026, Revision: 8 July 2026, Accepted: 14 July 2026, Available Online: 1 January 2027.

Abstract

The rapid growth of online media has accelerated the dissemination of information related to flood disasters, creating the need for an automatic method to classify news into factual and opinion categories. This study aims to compare the performance of the Naïve Bayes and Support Vector Machine (SVM) algorithms in classifying factual and opinion-based news on flood events in Sumatra using Term Frequency–Inverse Document Frequency (TF-IDF) weighting. The research employed several text preprocessing stages, including cleaning, case folding, tokenization, stopword removal, and stemming, followed by TF-IDF weighting, classification, and model evaluation using accuracy, precision, recall, and F1-score. The experimental results showed that the Naïve Bayes algorithm achieved an accuracy of 92.34%, outperforming the Support Vector Machine algorithm, which achieved an accuracy of 91.88%. In addition, Naïve Bayes obtained higher precision, recall, and F1-score values for the opinion class. These findings indicate that Naïve Bayes is more effective than Support Vector Machine in classifying factual and opinion-based news related to flood events in Sumatra.

Keyword: Text Mining; Naïve Bayes; Support Vector Machine; TF-IDF; News Classification.

Abstrak

Perkembangan media daring menyebabkan penyebaran informasi mengenai bencana banjir berlangsung sangat cepat sehingga diperlukan metode untuk mengklasifikasikan berita ke dalam kategori fakta dan opini secara otomatis. Penelitian ini bertujuan membandingkan performa algoritma Naïve Bayes dan Support Vector Machine (SVM) dalam klasifikasi berita fakta dan opini pada pemberitaan banjir Sumatera menggunakan pembobotan Term Frequency-Inverse Document Frequency (TF-IDF). Penelitian dilakukan melalui tahapan preprocessing yang meliputi cleaning, case folding, tokenizing, stopword removal, dan stemming, dilanjutkan dengan pembobotan TF-IDF, proses klasifikasi, serta evaluasi menggunakan accuracy, precision, recall, dan f1-score. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes memperoleh accuracy sebesar 92,34%, lebih tinggi dibandingkan Support Vector Machine yang memperoleh 91,88%. Selain itu, Naïve Bayes juga menunjukkan nilai precision, recall, dan f1-score yang lebih baik pada kelas opini. Berdasarkan hasil tersebut, Naïve Bayes lebih efektif dibandingkan Support Vector Machine dalam mengklasifikasikan berita fakta dan opini pada pemberitaan banjir Sumatera.

Kata Kunci: Text Mining; Naïve Bayes; Support Vector Machine; TF-IDF; Klasifikasi Berita.

E-ISSN: 2580-1643

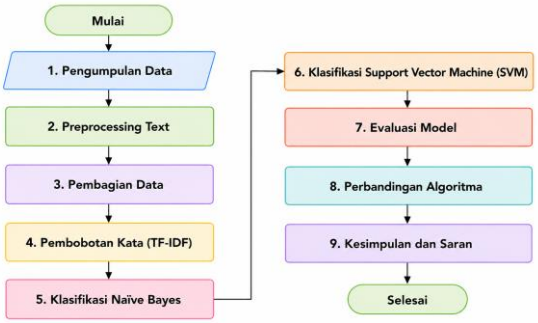
1. Pendahuluan

Pada tahun 2025, banjir besar terjadi di berbagai wilayah Pulau Sumatera, terutama di Provinsi Aceh, Sumatera Utara, dan Sumatera Barat. Bencana tersebut mengakibatkan kerugian yang besar bagi masyarakat serta menjadi perhatian utama berbagai media daring. Banyaknya pemberitaan mengenai bencana banjir menyebabkan beredarnya informasi yang mengandung fakta maupun opini. Keberadaan berita opini yang bercampur dengan berita fakta berpotensi memengaruhi persepsi masyarakat terhadap suatu peristiwa. Rahmatika dan Prisanto (2022) menyatakan bahwa karakteristik penyajian berita berpengaruh terhadap tingkat kepercayaan masyarakat terhadap media sebesar 28,4%. Oleh karena itu, diperlukan suatu metode yang mampu mengelompokkan berita secara otomatis berdasarkan kategori fakta dan opini sehingga informasi yang diterima masyarakat dapat dianalisis secara lebih sistematis. Salah satu pendekatan yang banyak digunakan untuk mengklasifikasikan dokumen teks adalah *text mining*. Dalam proses klasifikasi teks, pemilihan algoritma sangat memengaruhi performa model yang dihasilkan. Algoritma Naïve Bayes dikenal memiliki proses komputasi yang sederhana, cepat, dan efisien sehingga banyak digunakan dalam berbagai penelitian klasifikasi teks (Habib *et al.*, 2022). Sementara itu, *Support Vector Machine* (SVM) merupakan algoritma yang memiliki kemampuan klasifikasi yang baik pada data berdimensi tinggi dan sering digunakan dalam berbagai penelitian *text mining*.

Penelitian Nanda *et al.* (2022) menunjukkan bahwa algoritma SVM memperoleh akurasi sebesar 88% pada klasifikasi berita menggunakan kernel linear. Selain itu, Kowsari *et al.* (2019) menjelaskan bahwa *Naïve Bayes* maupun SVM merupakan dua algoritma yang memiliki performa yang baik pada berbagai permasalahan klasifikasi teks, terutama ketika dipadukan dengan representasi fitur TF-IDF. Meskipun berbagai penelitian sebelumnya telah membandingkan algoritma Naïve Bayes dan SVM, sebagian besar penelitian dilakukan pada analisis sentimen, klasifikasi berita secara umum, maupun data media sosial. Penelitian yang secara khusus membandingkan kedua algoritma dalam mengklasifikasikan berita fakta dan opini pada domain pemberitaan bencana banjir masih sangat terbatas. Padahal, karakteristik bahasa pada berita bencana memiliki perbedaan dibandingkan berita umum karena memuat informasi mengenai lokasi kejadian, dampak bencana, data korban, serta berbagai interpretasi mengenai penyebab maupun penanganan bencana. Berdasarkan kondisi tersebut, penelitian ini memiliki kebaruan dengan membandingkan performa algoritma *Naïve Bayes* dan *Support Vector Machine* pada klasifikasi berita fakta dan opini menggunakan dataset berita banjir yang dikumpulkan dari CNN Indonesia, Detik.com, dan Mongabay Indonesia. Representasi teks dilakukan menggunakan pembobotan TF-IDF, sedangkan evaluasi model menggunakan *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*. Selain mengevaluasi performa kedua algoritma, model terbaik juga diterapkan untuk mengklasifikasikan berita yang belum memiliki label. Hasil penelitian ini diharapkan dapat memberikan gambaran mengenai algoritma yang lebih sesuai untuk klasifikasi berita fakta dan opini pada domain pemberitaan bencana banjir serta menjadi referensi bagi pengembangan sistem klasifikasi informasi kebencanaan secara otomatis.

2. Metode Penelitian

Riset ini ialah telaah dalam ranah *text mining yang* berorientasi guna memproses informasi kontekstual non-struktural berwujud informasi yang dapat dianalisis menggunakan metode *machine learning*. Penelitian dilakukan untuk mengklasifikasikan berita banjir Sumatera ke dalam kategori fakta dan opini menggunakan pembobotan kata *TF-IDF*. Data berita yang telah dikumpulkan kemudian melewati jenjang *preprocessing* yang mengompilasi *cleaning*, *case folding*, *tokenizing*, *stopword removal*, *stemming*, serta *word cloud* mendahului pengerjaan metode kalkulasi bobot dan pengelompokan. Segenap rangkaian investigasi dieksekusi memanfaatkan sintaks instruksi *Python* pada platform *Google Colaboratory*, sementara pengujian model diselenggarakan memakai parameter *accuracy, precision, recall, f1-score,* serta *confusion matrix*.



Gambar 1. Tahapan penelitian

Berikut penjelasan tahapan penelitian sesuai gambar diatas:

1) Pengumpulan Data
   Pada tahap ini dilakukan pengumpulan data berita menggunakan teknik *web scraping* dengan bahasa pemrograman *Python*. Data berita diperoleh dari portal berita CNN Indonesia, Detik.com, dan Mongabay Indonesia. Data

E-ISSN: 2580-1643

yang dikumpulkan merupakan berita terkait banjir Sumatera pada periode November 2025 hingga Februari 2026. Jumlah data awal yang berhasil dikumpulkan sebanyak 3.500 data berita.

2) *PreprocessingText*
Pasca ketersediaan informasi, berikutnya himpunan tersebut melewati operasi *text preprocessing yang* diselenggarakan mendahului integrasi menuju mekanisme pengelompokan informasi. Di fase *text preprocessing* bakal mensterilkan berita lepas dari elemen *yang* tak dibutuhkan. Berikut urutan proses *preprocessing text* dalam penelitian ini (Habib *et al*., 2022).

Tabel 1. *Preprocessing*

No	*Preprocessing*	Penjelasan
1	*Cleaning*	Proses membersihkan teks dari tanda baca, simbol, angka, URL, dan karakter yang tidak diperlukan.
2	*CaseFolding*	Proses mengubah seluruh huruf pada teks menjadi huruf kecil.
3	*Tokenizing*	Proses memisahkan kalimat menjadi potongan kata atau token.
4	*Stopword Removal*	Proses menghapus kata umum yang tidak memiliki makna penting seperti "dan", "yang", dan "di"
5	*Stemming*	Proses mengubah kata berimbuhan menjadi kata dasar.
6	*Wordcloud*	Frekuensi kemunculan yang tinggi pada keseluruhan dataset berita

3) Pembagian Data
Setelah melalui proses *preprocessing,* Kuantitas data *yang* diaplikasikan pada telaah ilmiah berakumulasi menjadi 3196 data berita fakta dan opini yang dijelaskan oleh Lestari *et al* (2019).

4) Pembobotan Kata (*TF-IDF*)
Pembobotan kata dilakukan menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF) untuk mengubah data teks menjadi representasi numerik yang dapat diproses oleh algoritma klasifikasi. Dataset kemudian dibagi menggunakan metode *train-test split* dengan rasio 80:20, yaitu 80% (2.557 data) sebagai data latih dan 20% (639 data) sebagai data uji. Rasio ini dipilih karena umum digunakan dalam penelitian *machine learning* sehingga mampu menghasilkan data latih yang memadai sekaligus data uji yang representatif. Selanjutnya, model mempelajari hubungan antara bobot TF-IDF dan label kelas untuk melakukan klasifikasi pada data uji maupun data baru yang belum memiliki label (Habib *et al*., 2022). Berikut rumus *Term Frequency (TF):*

$$tf_{ij} = \frac{f_d}{max\ f_d(j)}$$

Keterangan:
tf_{ij} : Jumlah kemunculan kata (*term*) dalam setiap dokumen
f_d : Frekuensi terdapat kata (*term*) i dalam dokumen j
$max\ f_d(j)$: Total kata (*term*) pada dokumen j

Seterusnya guna mengalkulasi leksikon spesifik yang kerap mengemuka di dalam suatu berkas rumusnya sebagai berikut:

$$idf_{ij} = \ln(\frac{D}{df_i}) + 1$$

Keterangan:
idf_{ij} : Total munculnya kata (term) dalam sebuah dokumen
D : Total semua dokumen
df_i : Total dokumen yang terdapat kata (term) didalamnya

TF-IDF digunakan sebagai tolak ukur perhitungan statistik untuk menganalisa pentingnya keberada dan suatu kata pada sebuah naskah yang diformulasikan selaku representasi di bawah ini:

$$w_{ij} = tf_{ij}\ x\ idf_{ij}$$

Keterangan:
w_{ij} : Kuantitas munculnya kata (term) pada suatu berkas
tf_{ij} : Kuantitas semua dokumen
idf_{ij} : Kuantitas dokumen yang ada kata (term) didalamnya

5) Klasifikasi *Naïve Bayes Classifier*
Naïve Bayes merupakan algoritma klasifikasi berbasis probabilitas yang menerapkan Teorema Bayes untuk menentukan peluang suatu dokumen termasuk ke dalam kelas tertentu. Algoritma ini banyak digunakan pada klasifikasi teks karena sederhana, cepat, dan memiliki performa yang baik pada data berdimensi tinggi (Nanda *et al*., 2022). Pada penelitian ini, Naïve Bayes digunakan untuk mengklasifikasikan berita ke dalam dua

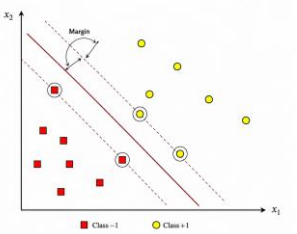
E-ISSN: 2580-1643

kategori, yaitu fakta dan opini. Persamaan dasar Teorema Bayes ditunjukkan pada Persamaan (4):

$$P(Y|Z) = \frac{P(Y)P(Z|Y)}{P(Z)}$$

Keterangan:
$P(Y)$: Probabilitas Kejadian Y
$P(Z)$: Probabilitas kejadian Y berdasarkan kejadian Z
$P(Z|Y)$: Probabilitas kejadian Y berdasarkan kejadian Z

6) Klasifikasi *Support Vector Machine (SVM)*
Support Vector Machine (SVM) merupakan algoritma klasifikasi yang bekerja dengan mencari hyperplane terbaik untuk memisahkan dua kelas data sehingga memiliki margin maksimum. Algoritma ini dikenal memiliki performa yang baik pada data berdimensi tinggi dan banyak digunakan dalam klasifikasi teks (Nanda *et al.*, 2022). Ilustrasi hyperplane pada SVM ditunjukkan pada Gambar 2, sedangkan persamaan dasarnya ditunjukkan pada Persamaan (5).



Gambar 2. *Hyper plane* pada *SVM Linear*

$$\boldsymbol{x_i w} + \mathbf{b} \geq +1, y_i = +1$$
$$\boldsymbol{x_i w} + \mathbf{b} \leq -1, y_i = -1$$

Keterangan:
w = vektor normal terhadap *hyperplane*
x = data masukan (*feature vector*)
b = konstanta (*bias*)

7) Evaluasi Model
Tahapan evaluasi model dilakukan untuk mengetahui performa algoritma *Naïve Bayes* dan *Support Vector Machine* dalam klasifikasi berita fakta dan opini. Evaluasi model dilakukan menggunakan beberapa metrik pengujian, yaitu *accuracy, precision, recall, f1-score, dan confusion matrix*.

8) Perbandingan Algoritma
Tahap terakhir yaitu membandingkan hasil evaluasi algoritma *Naïve Bayes* dan *Support Vector Machine* berlandaskan skor *accuracy*, *precision*, *recall*, serta *confusion matrix*. Temuan perbandingan difungsikan guna mendeteksi algoritma dengan unjuk kerja terunggul dalam pengelompokan berita opini dan fakta pada berita banjir Sumatera. Pemrosesan informasi di dalam telaah ilmiah ini dieksekusi memakai bahasa pemrograman *Python* pada platform *Google Colaboratory (Google Colab*). *Python* digunakan untuk proses *preprocessing*, pembobotan *TF-IDF*, klasifikasi, dan evaluasi model, sedangkan *Google Colab* digunakan sebagai lingkungan komputasi berbasis *cloud* yang mendukung pengembangan dan analisis data.

3. Hasil dan Pembahasan

3.1 Hasil

Dibawah ini merupakan berita yang dikumpulkan dari 3 portal berita, berikut hasil berita yang didapatkan dari setiap portal:

Tabel 2. Jumlah Data

No	Portal	Jumlah
1	CNN Indonesia	1566
2	Detik.com	1330
3	Mongabay Indonesia	300

Tabel 3. Jumlah Fakta dan Opini

No	Portal	Jumlah Fakta	Jumlah Opini
1	CNN Indonesia	1513	56
2	Detik.com	1246	83
3	Mongabay Indonesia	102	196

Tabel 4. Klasifikasi

No	Portal	Berita	Kategori
1	Detik.com	Akhir 2025 menjadi salah satu periode paling kelam dalam sejarah kebencanaan Indonesia. Banjir bandang dan tanah longsor melanda Aceh, Sumatera Utara, dan Sumatera Barat secara hampir simultan sejak penghujung November dan terus berlanjut hingga	Opini

E-ISSN: 2580-1643

		Desember.Hampir mendekati angka jutaan orang mengungsi, ratusan rumah. ...	
2	Detik.com	Dirangkum dari detik News, Pemerintah Provinsi Aceh mengirim surat permintaan bantuan kepada United Nations Development Program (UNDP) dan United Nations International Children's Emergency Fund (UNICEF). Seperti disebutkan sebelumnya, dua lembaga itu memiliki pengalaman saat masa pemulihan dan rehabilitasi pascabencana tsunami di Aceh pada 2004. "Benar (sudah melayangkan surat), (karena) mempertimbangkan mereka lembaga resmi PBB yang ada di Indonesia, maka meminta keterlibatan mereka dalam pemulihan. Kami rasa sangat dibutuhkan," kata Muhammad MTA. "Secara khusus pemerintah Aceh secara resmi juga telah menyampaikan permintaan keterlibatan beberapa lembaga internasional atas pertimbangan pengalaman bencana tsunami 2004 seperti UNDP dan UNICEF," kata Muhammad MTA dilansirdetikSumut, Senin (15/12). h...	Fakta
3	CNN Indonesia	Akhir pekan lalu sejumlah wilayah di Kalimantan Selatan (Kalsel) diterjang air bah, termasuk banjir bandang di Kabupaten Balangan. Banjir di kabupaten itu bahkan dilaporkan mencapai ketinggian dua meter. Kapolres Balangan AKBP Yulianor Abdi mengatakan bencana terparah terjadi di Kecamatan Tebing Tinggi, tepatnya di Desa Juuh, Sungsum, dan Gunung Batu. BPBD Kabupaten Balangan melaporkan sebanyak 10.949 jiwa terdampak dan 3.511 rumah terendam akibat banjir di 27 desa yang tersebar di tujuh kecamatan	Fakta
4	CNN Indonesia	Peneliti Hidrologi Hutan dan Konservasi DAS Universitas Gadjah Mada (UGM) Hatma Suryatmojo menduga terdapat dosa ekologis atau deforestasi masif di balik banjir bandang hingga longsor yang melanda Aceh, Sumatra Utara, dan Sumatra Barat. Menurutnya, fenomena tersebut bukan peristiwa yang berdiri sendiri, melainkan bagian dari pola berulang bencana hidrometeorologi akibat kombinasi faktor alam dan ulah manusia.	Opini
5	Mongabay Indonesia	Penelitian selanjutnya disarankan menggunakan dataset yang lebih besar dan beragam, melibatkan lebih banyak sumber berita, serta menerapkan optimasi parameter pada algoritma klasifikasi. Selain itu, penelitian dapat membandingkan algoritma lain seperti Random Forest, Logistic Regression, maupun model berbasis Deep Learning, serta menggunakan representasi teks yang lebih modern seperti Word2Vec, FastText, atau BERT untuk meningkatkan performa klasifikasi.	Fakta
6	Mongabay	Sejak awal berbagai kalangan mengkritik RUU Cipta Kerja karena dinilai memberikan karpet merah kepada investor dengan mengabaikan lingkungan. Reynaldo G. Sembiring menyatakan bahwa pengelolaan lingkungan seharusnya menjamin akses informasi, partisipasi, dan keadilan, sedangkan ketiga aspek tersebut justru dikurangi dalam UU Cipta Kerja.	Opini

Berdasarkan proses pelabelan yang telah dilakukan, setiap berita diklasifikasikan ke dalam dua kategori, yaitu fakta dan opini. Perlu ditegaskan bahwa pelabelan tersebut tidak dimaksudkan untuk menentukan apakah suatu berita benar atau salah, melainkan untuk mengidentifikasi karakteristik penyajian informasi dalam berita. Dilakukan pelabelan (*labeling*) secara manual terhadap data berita untuk mengelompokkan setiap berita ke dalam kategori fakta atau opini. Proses pelabelan dilakukan oleh penulis dengan mengacu pada karakteristik fakta dan opini yang dijelaskan oleh Lestari *et al.* (2019). Berita dikategorikan sebagai fakta apabila isi berita didominasi oleh informasi yang dapat diverifikasi, seperti data, peristiwa, maupun pernyataan narasumber. Sebaliknya, berita dikategorikan sebagai opini apabila lebih banyak memuat penilaian, interpretasi, argumentasi, atau pendapat mengenai suatu peristiwa. Seluruh berita dibaca secara menyeluruh sebelum diberikan label untuk menjaga konsistensi proses pelabelan. Hasil pelabelan kemudian digunakan sebagai *ground truth* dalam proses pelatihan (*training*) dan pengujian (*testing*) model klasifikasi. Menurut KBBI, fakta merupakan hal atau peristiwa yang benar-benar

E-ISSN: 2580-1643

terjadi dan dapat dibuktikan kebenarannya, sedangkan opini adalah pendapat, pikiran, atau pendirian seseorang. Oleh karena itu, berita berlabel fakta merupakan berita yang dominan menyajikan informasi berdasarkan peristiwa, data, maupun keterangan narasumber yang dapat diverifikasi. Sebaliknya, berita berlabel opini merupakan berita yang mengandung penilaian, interpretasi, argumentasi, atau pendapat yang disampaikan oleh penulis maupun narasumber. Pemberian label pada penelitian ini bukan merupakan penilaian terhadap validitas atau kebenaran suatu berita, melainkan sebagai proses pengelompokan jenis informasi yang terkandung dalam berita untuk keperluan klasifikasi menggunakan algoritma *Naïve Bayes* dan *Support Vector Machine (SVM)*.

Text Preprocessing

Cleaning

Cleaning dilakukan untuk membersihkan teks dari karakter yang tak dibutuhkan layaknya tanda ortografis, bilangan, lambang, *url*, serta sela marjinal berlebih. Di bagian bawah ini ialah perolehan fase *cleaning*

Isi Berita	cleaning
Menteri Perdagangan (Mendag) Budi Santoso memastikan konflik geopolitik di Timur Tengah belum berdampak pada harga pangan di dalam negeri. Menurutnya, pasokan dan harga kebutuhan pokok masih terkendali. "Kalau kita ngomongin dampak perang sebenarnya nggak mengganggu kita di dalam negeri, apalagi kan kalau kita kontribusi terhadap PDB itu kan sebenarnya, belanja dalam negeri yang paling besar," ujar Budi saat ditemui di kantornya, Jakarta Pusat, Kamis (5/3/2026). Hal ini terpantau langsung	Menteri Perdagangan Mendag Budi Santoso memastikan konflik geopolitik di Timur Tengah belum berdampak pada harga pangan di dalam negeri Menurutnya pasokan dan harga kebutuhan pokok masih terkendali Kalau kita ngomongin dampak perang sebenarnya nggak mengganggu kita di dalam negeri apalagi kan kalau kita kontribusi terhadap PDB itu kan sebenarnya belanja dalam negeri yang paling besar ujar Budi saat ditemui di kantornya Jakarta Pusat Kamis Hal ini terpantau langsung dari pengecekan

Gambar 3. *CleaningData*

Case Folding

Pada tahap *case folding* huruf diubah dari huruf besar pada teks menjadi huruf kecil supaya penulisan kata kian homogen serta tak dianggap berbeda oleh sistem. Di bawah ini merupakan hasil tahapan *case folding*.

cleaning	casefold
Menteri Perdagangan Mendag Budi Santoso memastikan konflik geopolitik di Timur Tengah belum berdampak pada harga pangan di dalam negeri Menurutnya pasokan dan harga kebutuhan pokok masih terkendali Kalau kita ngomongin dampak perang sebenarnya nggak mengganggu kita di dalam negeri apalagi kan kalau kita kontribusi terhadap PDB itu kan sebenarnya belanja dalam negeri yang paling besar ujar Budi saat ditemui di kantornya Jakarta Pusat Kamis Hal ini terpantau langsung dari pengecekan	menteri perdagangan mendag budi santoso memastikan konflik geopolitik di timur tengah belum berdampak pada harga pangan di dalam negeri menurutnya pasokan dan harga kebutuhan pokok masih terkendali kalau kita ngomongin dampak perang sebenarnya nggak mengganggu kita di dalam negeri apalagi kan kalau kita kontribusi terhadap pdb itu kan sebenarnya belanja dalam negeri yang paling besar ujar budi saat ditemui di kantornya jakarta pusat kamis hal ini terpantau langsung dari pengecekan

Gambar 4. *Case folding*

Tokeninzing

Tokenizing fungsinya adalah untuk memecah kalimat menjadi potongan kata yang kemudian disebut token. Tahapan ini bertujuan agar selanjutnya dijuluki *token*. Fase ini berorientasi supaya tiap-tiap kosakata berkemampuan ditransformasikan secara mandiri. Di bawah ini merupakan hasil tahapan *tokenizing*.

casefold	tokenizing
menteri perdagangan mendag budi santoso memastikan konflik geopolitik di timur tengah belum berdampak pada harga pangan di dalam negeri menurutnya pasokan dan harga kebutuhan pokok masih terkendali kalau kita ngomongin dampak perang sebenarnya nggak mengganggu kita di dalam negeri apalagi kan kalau kita kontribusi terhadap pdb itu kan sebenarnya belanja dalam negeri yang paling besar ujar budi saat ditemui di kantornya jakarta pusat kamis hal ini terpantau langsung dari pengecekan lapangan di berbagai daerah termasuk	['menteri', 'perdagangan', 'mendag', 'budi', 'santoso', 'memastikan', 'konflik', 'geopolitik', 'di', 'timur', 'tengah', 'belum', 'berdampak', 'pada', 'harga', 'pangan', 'di', 'dalam', 'negeri', 'menurutnya', 'pasokan', 'dan', 'harga', 'kebutuhan', 'pokok', 'masih', 'terkendali', 'kalau', 'kita', 'ngomongin', 'dampak', 'perang', 'sebenarnya', 'nggak', 'mengganggu', 'kita', 'di', 'dalam', 'negeri', 'apalagi', 'kan', 'kalau', 'kontribusi', 'terhadap', 'pdb', 'itu', 'kan', 'sebenarnya', 'belanja', 'dalam', 'negeri', 'yang', 'paling', 'besar', 'ujar', 'budi', 'saat', 'ditemui', 'di', 'kantornya',

Gambar 5. *Tokenizing*

Stopword Removal

Fase ini berorientasi demi mengeliminasi leksikon-leksikon tak mengonsep makna ataupun kosakata-kosakata yang kerap direpetisi, seperti "dan", "yang", "di", "ke", dan "dari". Di bawah ini merupakan hasil tahapan *stopword removal*.

tokenizing	stopword
['menteri', 'perdagangan', 'mendag', 'budi', 'santoso', 'memastikan', 'konflik', 'geopolitik', 'di', 'timur', 'tengah', 'belum', 'berdampak', 'pada', 'harga', 'pangan', 'di', 'dalam', 'negeri', 'menurutnya', 'pasokan', 'dan', 'harga', 'kebutuhan', 'pokok', 'masih', 'terkendali', 'kalau', 'kita', 'ngomongin', 'dampak', 'perang', 'sebenarnya', 'nggak', 'mengganggu', 'kita', 'di', 'dalam', 'negeri', 'apalagi', 'kan', 'kalau', 'kita', 'kontribusi', 'terhadap', 'pdb', 'itu',	['menteri', 'perdagangan', 'mendag', 'budi', 'santoso', 'memastikan', 'konflik', 'geopolitik', 'timur', 'tengah', 'berdampak', 'harga', 'pangan', 'negeri', 'menurutnya', 'pasokan', 'harga', 'kebutuhan', 'pokok', 'terkendali', 'kalau', 'ngomongin', 'dampak', 'perang', 'sebenarnya', 'mengganggu', 'negeri', 'kan', 'kalau', 'kontribusi', 'pdb', 'kan', 'sebenarnya', 'belanja', 'dalam', 'negeri', 'paling', 'besar', 'ujar', 'budi', 'ditemui', 'kantornya', 'jakarta', 'pusat', 'kamis',

Gambar 6. *Stopword removal*

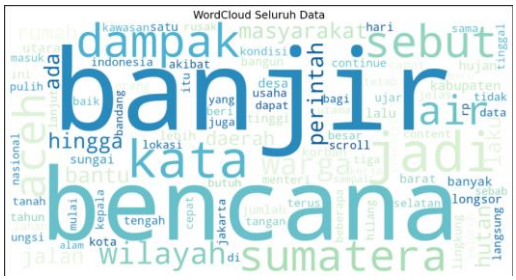
Stemming

Tahap *stemming* merupakan proses mengubah kata yang memiliki imbuhan menjadi kata dasar. Imbuhan yang dihilangkan dapat berupa awalan, akhiran, sisipan, maupun gabungan dari beberapa imbuhan. Tujuan dari proses ini adalah menyamakan bentuk kata sehingga kata-kata yang memiliki makna sama dapat dikenali sebagai satu kata dasar. Hasil dari proses *stemming* ditampilkan pada tabel berikut.

stopword	stemming
['menteri', 'perdagangan', 'mendag', 'budi', 'santoso', 'memastikan', 'konflik', 'geopolitik', 'timur', 'tengah', 'berdampak', 'harga', 'pangan', 'negeri', 'menurutnya', 'pasokan', 'harga', 'kebutuhan', 'pokok', 'terkendali', 'kalau', 'ngomongin', 'dampak', 'perang', 'sebenarnya', 'mengganggu', 'negeri', 'kan', 'kalau', 'kontribusi', 'pdb', 'kan', 'sebenarnya', 'belanja', 'dalam', 'negeri', 'paling', 'besar', 'ujar', 'budi', 'ditemui', 'kantornya', 'jakarta', 'pusat', 'kamis', 'terpantau', 'langsung', 'pengecekan',	['menteri', 'dagang', 'mendag', 'budi', 'santoso', 'pasti', 'konflik', 'geopolitik', 'timur', 'tengah', 'dampak', 'harga', 'pangan', 'negeri', 'turut', 'pasok', 'harga', 'butuh', 'pokok', 'kendali', 'kalau', 'ngomongin', 'dampak', 'perang', 'benar', 'ganggu', 'negeri', 'kan', 'kalau', 'kontribusi', 'pdb', 'kan', 'benar', 'belanja', 'dalam', 'negeri', 'paling', 'besar', 'ujar', 'budi', 'temu', 'kantor', 'jakarta', 'pusat', 'kamis', 'pantau', 'langsung', 'kece', 'lapang', 'bagai', 'daerah', 'masuk', 'daerah', 'dampak',

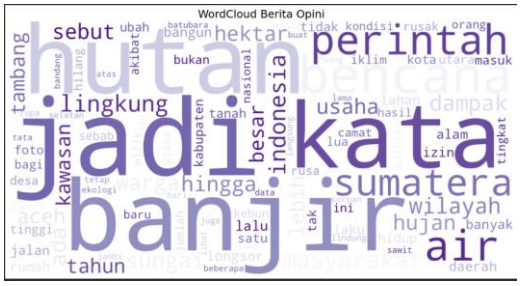
Gambar 7. *Stemming*

E-ISSN: 2580-1643



Gambar 8. *Word Cloud Umum*

Gambar 9. *Word Cloud Fakta*



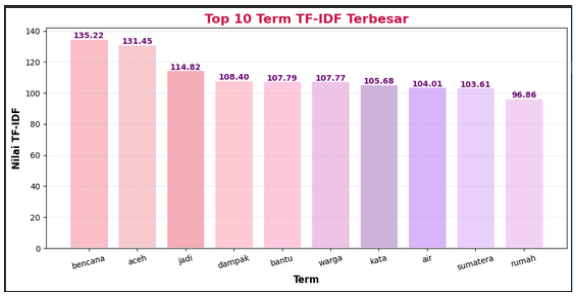
Gambar 10. *Word cloud Opini*

Berdasarkan visualisasi *Word Cloud,* terlihat bahwa pada seluruh dataset kata yang paling dominan adalah "banjir", diikuti oleh "kata", "bencana", "dampak", "sebut", "sumatera", dan "perintah". Dominasi kata-kata tersebut menunjukkan bahwa keseluruhan berita yang digunakan dalam penelitian berfokus pada peristiwa banjir di wilayah Sumatera, kondisi yang terjadi, serta dampak yang ditimbulkan. Pada *Word Cloud* berita fakta, kata yang paling sering muncul adalah "bencana", "banjir", "kata", "aceh", "dampak", "jadi", dan "wilayah", yang menggambarkan bahwa berita fakta lebih banyak menyampaikan informasi mengenai lokasi kejadian, kondisi lapangan, dampak bencana, serta fakta-fakta yang dapat diverifikasi. Sementara itu, pada *Word Cloud* berita opini, kata yang paling dominan adalah "hutan", "banjir", "kata", "perintah", "indonesia", "sumatera", "air", dan "lingkung", yang menunjukkan bahwa berita opini lebih banyak membahas isu lingkungan, peran pemerintah, serta pandangan atau interpretasi mengenai penyebab dan dampak banjir. Perbedaan pola kemunculan kata-kata dominan tersebut menunjukkan bahwa berita fakta lebih menitikberatkan pada penyampaian informasi mengenai peristiwa yang terjadi, sedangkan berita opini lebih banyak memuat pembahasan, argumentasi, dan interpretasi terhadap fenomena banjir.

Kata	Nilai TF-IDF
bencana	135.224508
aceh	131.452513
jadi	114.818644
dampak	108.403058
bantu	107.786564
warga	107.774457
kata	105.679790
air	104.008865
sumatera	103.610503
rumah	96.862502

Gambar 11. *Nilai TF-IDF*

Berdasarkan Gambar 11, nilai *TF-IDF* menunjukkan tingkat kepentingan suatu kata terhadap keseluruhan kumpulan dokumen. Semakin besar nilai *TF-IDF* suatu kata, maka kata tersebut dianggap semakin representatif dalam menggambarkan isi dokumen dan memiliki kontribusi yang lebih besar dalam proses klasifikasi. Sebaliknya, kata dengan nilai *TF-IDF* yang rendah memiliki kemampuan yang lebih kecil dalam membedakan karakteristik antar dokumen karena kemunculannya cenderung lebih umum atau kurang spesifik.



Gambar 12. *Top 10 Term TF-IDF Terbesar*

Naïve Bayes Classifier

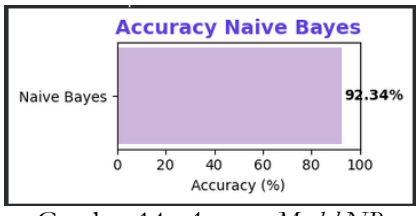
Proses klasifikasi diawali dengan melakukan pelatihan model menggunakan data *training* yang telah dibagi sebelumnya menggunakan metode *train-test split*. Data *training* yang digunakan berupa hasil pembobotan *TF-IDF* dari teks berita yang telah memiliki label fakta dan opini. Pada tahap ini algoritma *Naïve Bayes* dan *SVM* mempelajari pola pemisahan terbaik antar kategori berita berdasarkan nilai bobot kata yang diperoleh dari proses *TF-IDF*. Berdasarkan hasil evaluasi model, algoritma *Naïve Bayes* memperoleh nilai *accuracy* sebesar 92% dengan rata-rata *precision* sebesar 0,91, *recall* sebesar 0,92, dan *f1-score* sebesar 0,91. Pada kategori Fakta, model memperoleh nilai *precision* sebesar 0,94, *recall* sebesar 0,98, dan *f1-score* sebesar 0,96, yang menunjukkan bahwa sebagian besar berita fakta berhasil diklasifikasikan dengan benar. Sementara itu, pada kategori Opini, model memperoleh *precision* sebesar 0,74, *recall* sebesar

E-ISSN: 2580-1643

0,42, dan sebesar 0,53. Nilai recall yang lebih rendah menunjukkan bahwa masih terdapat berita opini yang diprediksi sebagai berita fakta.

	precision	recall	f1-score	support
Fakta	0.940000	0.980000	0.960000	573.000000
Opini	0.740000	0.420000	0.530000	67.000000
accuracy	0.920000	0.920000	0.920000	0.920000
macro avg	0.840000	0.700000	0.750000	640.000000
weighted avg	0.910000	0.920000	0.910000	640.000000

Gambar 13. *Evaluasi model NB*

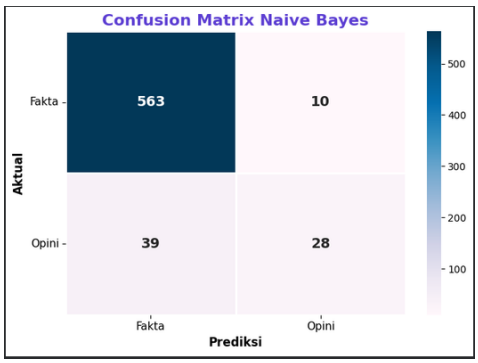


Gambar 14. *Accuracy Model NB*

Hasil pengujian menunjukkan bahwa algoritma *Naïve Bayes* memperoleh nilai *accuracy* sebesar 92,34%. Nilai tersebut menunjukkan bahwa model memiliki kemampuan yang baik dalam mengklasifikasikan berita fakta dan opini pada dataset penelitian. Dari temuan *confusion matrix* diperoleh:

1) Sebanyak 563 data fakta terbukti diproyeksikan secara tepat sfakta
2) Sebanyak 28 data opini terbukti diproyeksikan secara tepat sebagai opini
3) Tterdapat 39 berita opini yang salah diprediksi sebagai fakta
4) 10 berita fakta yang salah diprediksi sebagai opini.

Hasil *confusion matrix* menunjukkan bahwa model lebih baik dalam mengenali kategori fakta dibandingkan kategori opini. Hal ini disebabkan jumlah data fakta yang kian melimpah disandingkan informasi opini akibatnya sistem kian dominan mendalami regulasi leksikal pada berita fakta.



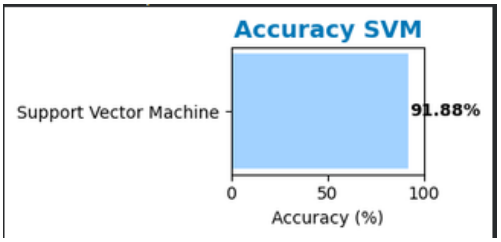
Gambar 15. *Confusion matrix NB*

Klasifikasi (SVM)

Berdasarkan hasil evaluasi model, algoritma *Support Vector Machine (SVM)* memperoleh nilai *accuracy* sebesar 92% dengan rata-rata *precision* sebesar 0,91, *recall* sebesar 0,92, dan *f1-score* sebesar 0,91. Pada kategori Fakta, model memperoleh *precision* sebesar 0,93, *recall* sebesar 0,98, dan *f1-score* sebesar 0,96. Sedangkan pada kategori Opini, model memperoleh *precision* sebesar 0,71, *recall* sebesar 0,37, dan *f1-score* sebesar 0,49. Nilai tersebut menunjukkan bahwa model masih mengalami kesulitan dalam mengidentifikasi sebagian berita opini.

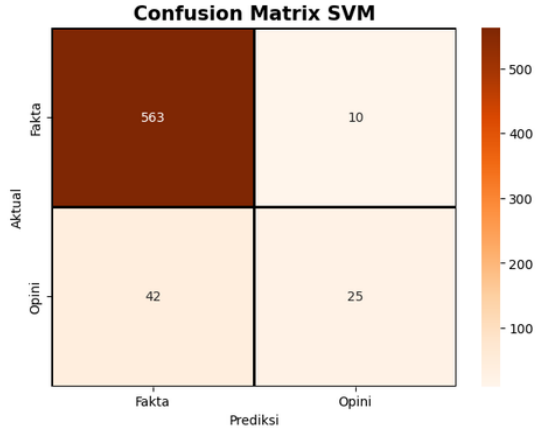
	precision	recall	f1-score	support
Fakta	0.930000	0.980000	0.960000	573.000000
Opini	0.710000	0.370000	0.490000	67.000000
accuracy	0.920000	0.920000	0.920000	0.920000
macro avg	0.820000	0.680000	0.720000	640.000000
weighted avg	0.910000	0.920000	0.910000	640.000000

Gambar 16. *Evaluasi model SVM*



Gambar 17. *Accuracy Model SVM*

Hasil pengujian menunjukkan bahwa algoritma *Support Vector Machine* memperoleh nilai *accuracy* sebesar 91,88%. Nilai tersebut menunjukkan bahwa model mampu mengklasifikasikan berita fakta dan opini dengan tingkat ketepatan yang baik meskipun sedikit lebih rendah dibandingkan *Naïve Bayes*.



Gambar 18. *Confusion matrix SVM*

Berdasarkan hasil *confusion matrix* diperoleh:

1) Sebanyak 563 data fakta berhasil diprediksi dengan benar sebagai fakta.

E-ISSN: 2580-1643

2) Sebanyak 25 data opini berhasil diprediksi dengan benar sebagai opini.
3) Terdapat 42 data opini yang salah diprediksi sebagai fakta.
4) Terdapat 10 data fakta yang salah diprediksi sebagai opini.

Hasil tersebut menunjukkan bahwa model *SVM* memiliki performa yang baik dalam mengenali berita fakta, tetapi masih mengalami kesalahan dalam mengklasifikasikan sebagian berita opini.

Tabel 5. *Accuracy*

No	Algoritma	*Accuracy*
1	*Naïve Bayes*	92.34%
2	*SVM*	91.88%

Tabel 6. Hasil Evaluasi

No	Metrik	*NB* Fakta	*SVM* Fakta	*NB* Opini	*SVM* Opini
1	Precision	0.94	0.93	0.74	0.71
2	Recall	0.98	0.98	0.42	0.37
3	F1-Score	0.96	0.96	0.53	0.49

Berdasarkan hasil evaluasi pada Tabel , algoritma *Naïve Bayes* dan *Support Vector Machine* memiliki nilai *recall* yang sama pada kategori fakta, yaitu sebesar 0,98. Hal ini menunjukkan bahwa kedua algoritma mampu mengenali sebagian besar data berita fakta dengan baik. Pada kategori opini, algoritma *Naïve Bayes* memperoleh nilai *precision* sebesar 0,74, *recall* sebesar 0,42, dan *f1-score* sebesar 0,53, sedangkan algoritma *Support Vector Machine* memperoleh nilai *precision* sebesar 0,71, *recall* sebesar 0,37, dan *f1-score* sebesar 0,49. Hasil tersebut menunjukkan bahwa *Naïve Bayes* memiliki performa yang sedikit lebih baik dalam mengklasifikasikan berita opini dibandingkan *Support Vector Machine.*

Berdasarkan hasil perbandingan pada Tabel 6, algoritma *Naïve Bayes* memperoleh nilai *accuracy* sebesar 92,34%, sedangkan algoritma *Support Vector Machine* memperoleh nilai accuracy sebesar 91,88%. Meskipun selisih akurasi kedua algoritma relatif kecil, *Naïve Bayes* menunjukkan performa yang lebih baik berdasarkan nilai *accuracy* maupun metrik evaluasi pada kategori opini. Berdasarkan hasil pengujian pada dataset yang digunakan dalam penelitian ini, algoritma *Naïve Bayes* menunjukkan performa yang sedikit lebih baik dibandingkan *Support Vector Machine* dalam mengklasifikasikan berita fakta dan opini. Perbedaan performa antara *Naïve Bayes* dan *Support Vector Machine* diduga dipengaruhi oleh karakteristik dataset yang digunakan. Dataset penelitian didominasi oleh kategori fakta sehingga distribusi kata pada masing-masing kelas lebih sesuai dengan pendekatan probabilistik yang digunakan oleh *Naïve Bayes.* Kondisi tersebut memungkinkan *Naïve Bayes* memperoleh performa yang sedikit lebih baik dibandingkan *Support Vector Machine* pada penelitian ini.

3.2 Pembahasan

Hasil penelitian ini menunjukkan bahwa algoritma Naïve Bayes memiliki performa yang lebih baik dibandingkan Support Vector Machine (SVM) dalam mengklasifikasikan berita fakta dan opini terkait bencana banjir di Sumatera. Naïve Bayes mencapai akurasi sebesar 92,34%, sementara SVM hanya memperoleh akurasi 91,88%. Temuan ini sejalan dengan penelitian yang dilakukan oleh Nanda *et al.* (2022), yang juga menunjukkan bahwa Naïve Bayes memiliki keunggulan dalam klasifikasi berita. Penelitian tersebut mengindikasikan bahwa algoritma Naïve Bayes lebih efektif dalam menangani data yang tidak seimbang, di mana jumlah berita fakta lebih banyak dibandingkan berita opini. Hal ini diperkuat oleh analisis Kowsari *et al.* (2019), yang menyatakan bahwa performa algoritma klasifikasi sangat dipengaruhi oleh karakteristik dataset dan representasi fitur yang digunakan, seperti TF-IDF. Dalam konteks penelitian ini, pemilihan fitur TF-IDF sebagai representasi teks terbukti memberikan kontribusi signifikan terhadap kinerja Naïve Bayes, sehingga memudahkan model dalam mengenali pola yang ada dalam berita. Selain itu, perbedaan karakteristik bahasa dalam berita bencana, yang sering kali mengandung informasi faktual yang dapat diverifikasi, memungkinkan Naïve Bayes untuk lebih unggul dalam klasifikasi berita fakta dibandingkan dengan SVM. Dengan demikian, hasil penelitian ini tidak hanya memberikan wawasan baru tentang efektivitas algoritma dalam klasifikasi berita bencana, tetapi juga menekankan pentingnya pemilihan algoritma yang tepat sesuai dengan karakteristik data yang tersedia.

E-ISSN: 2580-1643

4. Kesimpulan

Berdasarkan hasil pengujian, algoritma *Naïve Bayes* memperoleh *accuracy* sebesar 92,34%, sedangkan *Support Vector Machine* memperoleh 91,88%. Selain memiliki nilai *accuracy* yang lebih tinggi, *Naïve Bayes* juga menghasilkan nilai p*recision, recall,* dan *F1-score* yang lebih baik pada kelas opini. Hasil tersebut menunjukkan bahwa *Naïve Bayes* lebih efektif dalam mengenali karakteristik berita opini pada dataset yang digunakan. Perbedaan performa kedua algoritma diduga dipengaruhi oleh distribusi dataset yang tidak seimbang, di mana jumlah berita fakta lebih banyak dibandingkan berita opini. Kondisi tersebut menyebabkan kedua model memiliki kemampuan yang sangat baik dalam mengidentifikasi berita fakta, tetapi masih mengalami kesulitan dalam mengenali berita opini. Hasil penelitian ini berbeda dengan penelitian Nanda *et al.* (2022) yang menunjukkan bahwa *SVM* memiliki performa lebih baik dibandingkan *Naïve Bayes* pada klasifikasi berita. Perbedaan tersebut diduga dipengaruhi oleh karakteristik dataset, distribusi kelas, serta domain berita yang digunakan. Temuan ini sejalan dengan Kowsari *et al.* (2019) yang menyatakan bahwa performa algoritma klasifikasi sangat dipengaruhi oleh karakteristik data dan representasi fitur *TF-IDF*. Penelitian ini memiliki beberapa keterbatasan. Pertama, dataset hanya berasal dari tiga portal berita, yaitu CNN Indonesia, Detik.com, dan Mongabay Indonesia, sehingga belum sepenuhnya mewakili seluruh pemberitaan banjir di Indonesia. Kedua, penelitian hanya membandingkan dua algoritma klasifikasi tanpa melakukan optimasi parameter maupun membandingkannya dengan algoritma lain.

Selain itu, proses pelabelan dilakukan secara manual oleh satu peneliti berdasarkan pedoman Lestari *et al.* (2019) dan belum menerapkan pengukuran inter-rater reliability seperti *Cohen's Kappa*. Oleh karena itu, konsistensi pelabelan masih bergantung pada penilaian peneliti. Penelitian selanjutnya disarankan menggunakan dataset yang lebih besar dan beragam, melibatkan lebih banyak sumber berita, serta menerapkan optimasi parameter pada algoritma klasifikasi. Selain itu, penelitian dapat membandingkan algoritma lain seperti *Random Forest, Logistic Regression,* maupun model berbasis *Deep Learning,* serta menggunakan representasi teks yang lebih modern seperti *Word2Vec, FastText,* atau *BERT* untuk meningkatkan performa klasifikasi.

5. Daftar Pustaka

Agustin, Y. H., Mulyani, N. C., & Prasetya, W. S. (2025). Analisis sentimen opini publik menggunakan algoritma Naïve Bayes dan TF-IDF. *Jurnal Algoritma, 22*(2), 1373–1384.

Bangun, E. P., Koagouw, F. V. I. A., & Kalangi, J. S. (2019). Analisis isi unsur kelengkapan berita pada media online ManadoPostOnline.com. *Jurnal Ilmiah, 1*(1), 1–13.

Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan algoritma Naïve Bayes untuk analisis sentimen review data Twitter BMKG Nasional. *Jurnal Tekno Kompak, 15*(1), 131–145.

Deolika, A., Kusrini, & Luthfi, E. T. (2019). Analisis pembobotan kata pada klasifikasi text mining. *Jurnal Teknologi Informasi, 3*(2), 179–184.

Habib, S. M., Haerani, E., Gusti, S. K., & Ramadhani, S. (2022). Klasifikasi berita menggunakan metode Naïve Bayes classifier. *Jurnal Nasional Komputasi dan Teknologi Informasi, 5*(2), 248–258.

Hermawan, L., & Ismiati, M. B. (2020). Pembelajaran text preprocessing berbasis simulator untuk mata kuliah information retrieval. *TRANSFORMATIKA, 17*(2), 188–199.

Khairunnisa, S., Adiwijaya, A., & Al Faraby, S. (2021). Pengaruh text preprocessing terhadap analisis sentimen komentar masyarakat pada media sosial Twitter (studi kasus pandemi COVID-19). *Jurnal Media Informatika Budidarma, 5*(2), 406–414.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. E. (2019). Text classification algorithms: A survey. *Information, 10*(4), Article 150. https://doi.org/10.3390/info10040150.

Krisnandi, D., Ambarwati, R. N., Asih, A. Y., Ardiansyah, A., & Pardede, H. F. (2023). Analisis komentar cyberbullying terhadap kata yang mengandung toksisitas dan agresi menggunakan bag of words dan TF-IDF dengan klasifikasi SVM. *Jurnal Linguistik Komputasional, 6*(2), 36–41.

Lestari, R., Sudiyana, B., & Wahyuni, T. (2019). Fakta dan opini dalam teks tajuk rencana pada surat kabar Kompas. *KLITIKA: Jurnal*

E-ISSN: 2580-1643

Ilmiah Pendidikan Bahasa dan Sastra Indonesia, 1(1), 1–10.

Mettler, M., & Mondak, J. J. (2024). Fact-opinion differentiation. *Harvard Kennedy School Misinformation Review, 5*(2). https://doi.org/10.37016/mr-2020-136.

Nanda, R., Haerani, E., Gusti, S. K., & Ramadhani, S. (2022). Klasifikasi berita menggunakan metode support vector machine. *Jurnal Nasional Komputasi dan Teknologi Informasi, 5*(2), 269–278.

Putri, A. K., & Nur, D. I. (2023). Penggunaan bahasa Python untuk analisis dan visualisasi data penduduk di Desa Sumberjo, Nganjuk. *KARYA: Jurnal Pengabdian kepada Masyarakat, 3*(3), 206–217.

Rahmatika, N., & Prisanto, G. F. (2022). Pengaruh berita clickbait terhadap kepercayaan pada media di era attention economy. *Avant Garde, 10*(2), 190–200.

Shidiq, M. F. A., & Alita, D. (2025). Analisis sentimen masyarakat terhadap kasus judi online menggunakan data dari media sosial X pendekatan Naïve Bayes dan SVM. *Jurnal Sistem Informasi dan Informatika (SIMIKA), 8*(1), 24–35.