JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i3.6612

Perbandingan Kinerja *Naive Bayes* dan *KNN* dalam Analisis Sentimen Komentar X dengan dan tanpa Text *Preprocessing* (Studi Kasus: Danantara)

Master Edison Siregar[1*], Steven Dermawan[2], Ahmad Abdillah Hisyam[3]

[1]Program Studi S1 Informatika, Universitas Pradita
[2,3]Program Studi S1 Sistem Informasi, Universitas Pradita

Keywords:
Naive Bayes;
K-Nearest Neighbor;
Analisis Sentimen;
Preprocessing Teks;
Danantara;

Corespondent Email:
steven.dermawan@student.pradita.ac.id

Abstrak. Penelitian ini bertujuan untuk membandingkan kinerja dua algoritma klasifikasi, yaitu *Naive Bayes* (*NB*) dan *K-Nearest Neighbor* (*KNN*), dalam menganalisis sentimen komentar masyarakat terhadap topik Danantara di X. Data diperoleh melalui proses *crawling* menggunakan Google Collab, kemudian diolah dengan pendekatan kuantitatif eksperimental. Penelitian ini menguji empat skenario *preprocessing* teks untuk mengevaluasi pengaruhnya terhadap akurasi klasifikasi sentimen, dengan implementasi dilakukan melalui RapidMiner. Hasil analisis menunjukkan bahwa kombinasi *cleaning*, *case folding*, *tokenizing*, dan *stemming* merupakan skema *preprocessing* paling optimal, menghasilkan akurasi tertinggi sebesar 66,59% untuk NB dan 68,89% untuk *KNN*. Sebaliknya, tanpa *preprocessing*, akurasi menurun drastis menjadi 51,38% (*NB*) dan 51,15% (*KNN*). Skema parsial lainnya memberikan hasil akurasi yang mendekati, seperti kombinasi *cleaning*, *case folding*, *tokenizing*, dan *stopword removal* yang menghasilkan 65,21% (*NB*) dan 66,13% (*KNN*). Secara umum, algoritma *KNN* menunjukkan kinerja yang lebih unggul dibandingkan *NB* dalam setiap skenario. Analisis visual menggunakan *word cloud* mengungkapkan bahwa kata "Indonesia" dan "semangat" mendominasi sentimen positif, "korupsi" dan "koruptor" mendominasi sentimen negatif, serta "presiden" dan "BUMN" muncul dominan pada sentimen netral. Temuan ini diharapkan dapat menjadi acuan dalam pengembangan sistem klasifikasi sentimen, sekaligus memberikan gambaran umum tentang pandangan publik terhadap isu Danantara.

Abstract. *This study aims to compare the performance of two classification algorithms, namely Naive Bayes (NB) and K-Nearest Neighbor (KNN), in analyzing public comment sentiment on the topic of Danantara in X. Data were obtained through a crawling process using Google Collab, then processed with an experimental quantitative approach. This study tested four text preprocessing scenarios to evaluate their effect on sentiment classification accuracy, with implementation carried out through RapidMiner. The results of the analysis showed that the combination of cleaning, case folding, tokenizing, and stemming was the most optimal preprocessing scheme, producing the highest accuracy of 66.59% for NB and 68.89% for KNN. Conversely, without preprocessing, the accuracy decreased drastically to 51.38% (NB) and 51.15% (KNN). Other partial schemes provided close accuracy results, such as the combination of cleaning, case folding, tokenizing, and stopword removal which produced*

65.21% (NB) and 66.13% (KNN). In general, the KNN algorithm showed superior performance compared to NB in every scenario. Visual analysis using word cloud revealed that the words "Indonesia" and "spirit" dominate positive sentiment, "corruption" and "corruptors" dominate negative sentiment, and "president" and "BUMN" appear dominant in neutral sentiment. These findings are expected to be a reference in developing a sentiment classification system, as well as providing an overview of public views on the Danantara issue.

1. PENDAHULUAN

Perkembangan media sosial di era digital telah menciptakan ruang baru bagi masyarakat dalam menyampaikan opini, kritik, dan aspirasi terhadap isu-isu sosial maupun kebijakan pemerintah [1]. Di Indonesia, X merupakan salah satu platform yang paling aktif digunakan, dengan lebih dari 24 juta masyarakat pada tahun 2024, menjadikannya negara dengan jumlah masyarakat X terbanyak keempat di dunia [2]. Tingginya aktivitas masyarakat dalam membicarakan isu-isu aktual menjadikan X sebagai sumber data yang kaya untuk memahami persepsi publik secara real-time dan berbasis teks alami. Berbagai topik kebijakan, program pemerintah, dan institusi publik kerap menjadi bahan diskusi warganet, yang kemudian terekam dalam bentuk komentar dan cuitan.

Topik yang akhir-akhir ini menarik perhatian publik di media sosial adalah Danantara, yakni lembaga investasi strategis yang dibentuk oleh pemerintah Indonesia untuk mengelola aset negara secara lebih produktif [3]. Pembentukan Danantara memunculkan ragam reaksi di kalangan masyarakat. Beberapa pihak menanggapinya secara positif sebagai inovasi dalam pengelolaan kekayaan negara, namun terdapat pula komentar yang bernada skeptis, kritis, atau bahkan negatif terhadap konsep dan urgensi lembaga ini [4]. Besarnya volume opini yang tersebar di media sosial menunjukkan pentingnya analisis lebih lanjut guna memahami sentimen publik terhadap kebijakan ini secara sistematis.

Analisis sentimen merupakan pendekatan komputasional yang digunakan untuk mengidentifikasi dan mengklasifikasikan sikap emosional dalam teks, baik positif, negatif, maupun netral. Teknik ini sering digunakan dalam kajian opini publik, pemantauan media, dan pemrosesan bahasa alami (*natural language processing*) [5]. Di antara berbagai algoritma yang umum digunakan dalam klasifikasi sentimen, *Naive Bayes* dan *K-Nearest Neighbor* (*KNN*) menonjol karena kesederhanaannya dan kemampuannya dalam menangani data teks berbahasa alami. *Naive Bayes* bekerja berdasarkan prinsip probabilistik dengan asumsi independensi antar fitur, sedangkan *KNN* mengandalkan kedekatan antar data dalam ruang vektor fitur [6].

Keberhasilan klasifikasi sentimen tidak hanya ditentukan oleh algoritma yang digunakan, tetapi juga sangat dipengaruhi oleh tahap awal pengolahan data teks atau yang dikenal dengan text *preprocessing*. Proses ini melibatkan serangkaian tahapan seperti *cleaning, case folding, tokenizing, stopword removal*, dan *stemming* yang bertujuan untuk membersihkan dan menyederhanakan data teks mentah menjadi bentuk yang lebih terstruktur dan mudah diproses [7]. Kombinasi tahapan *preprocessing* yang digunakan dapat memberikan dampak signifikan terhadap performa model klasifikasi, baik dari sisi akurasi, presisi, maupun recall.

Sayangnya, penelitian yang membandingkan secara eksplisit performa algoritma klasifikasi sentimen berdasarkan variasi skenario *preprocessing* pada data berbahasa Indonesia masih relatif terbatas. Oleh karena itu, penelitian ini bertujuan untuk mengevaluasi pengaruh empat skenario *preprocessing* terhadap performa algoritma *Naive Bayes* dan *KNN* dalam mengklasifikasikan sentimen komentar X terkait Danantara. Keempat skenario tersebut meliputi: tanpa *preprocessing* sama sekali, *preprocessing* lengkap (*cleaning*, *case folding*, *tokenizing*, *stopword removal*, dan *stemming*), *preprocessing* tanpa *stopword removal*, dan *preprocessing* tanpa *stemming*.

Selain membandingkan tingkat akurasi dari masing-masing algoritma dalam setiap skenario, penelitian ini juga menyajikan visualisasi *word cloud* berdasarkan klasifikasi sentimen positif, negatif, dan netral. Visualisasi ini berfungsi untuk menggambarkan kata-kata yang paling sering muncul dalam tiap kategori sentimen, sehingga memberikan pemahaman tematik yang lebih mendalam terhadap kecenderungan opini publik. Dengan pendekatan ini, penelitian ini diharapkan mampu memberikan kontribusi terhadap pengembangan analisis sentimen berbasis bahasa Indonesia serta menawarkan wawasan berbasis data terhadap persepsi masyarakat terhadap kebijakan publik seperti Danantara.

2. TINJAUAN PUSTAKA

2.1. Analisis Sentimen

Analisis sentimen merupakan proses pengolahan data teks untuk mengidentifikasi dan mengklasifikasikan opini atau emosi yang terkandung dalam sebuah pernyataan, biasanya dalam bentuk positif, negatif, atau netral [8]. Teknik ini banyak digunakan dalam studi media sosial, termasuk dalam menganalisis komentar masyarakat di platform seperti X. Analisis sentimen bergantung pada representasi fitur teks dan pemilihan algoritma klasifikasi yang tepat [9].

2.2. *Naive Bayes*

Naive Bayes merupakan algoritma klasifikasi berbasis probabilistik yang mengasumsikan independensi antar fitur. Algoritma ini sering digunakan dalam pengolahan bahasa alami karena efisiensinya dalam menangani data teks berdimensi tinggi. Dalam konteks analisis sentimen, *Naive Bayes* efektif karena dapat mengkalkulasi kemungkinan suatu dokumen termasuk ke dalam kelas tertentu berdasarkan distribusi kata-kata dalam dokumen tersebut [10].

2.3. *K-Nearest Neighbor* (*KNN*)

K-Nearest Neighbor adalah algoritma klasifikasi berbasis *instance-based learning* yang menentukan kelas suatu data berdasarkan mayoritas kelas dari k tetangga terdekatnya. Dalam analisis teks, performa *KNN* sangat dipengaruhi oleh representasi fitur dan *preprocessing*, karena algoritma ini mengandalkan pengukuran jarak antar dokumen dalam ruang vektor [11].

2.4. *Preprocessing* Teks

Preprocessing teks adalah tahap penting dalam analisis data teks yang bertujuan membersihkan dan menstandarisasi data sebelum dianalisis lebih lanjut. Tahapan *preprocessing* dapat dijelaskan sebagai berikut [12]:

a) *Cleaning* Data: Pada tahap ini, karakter dan tanda baca yang tidak diperlukan dihapus untuk mengurangi *noise* dalam data. Hal ini termasuk menghapus *mention*, *hashtag*, *URL*, serta karakter selain huruf dan angka.
b) *Case Folding*: Proses ini mengubah semua huruf dalam teks menjadi huruf kecil untuk menghindari inkonsistensi dalam huruf besar dan kecil, yang sering terjadi dalam postingan di platform seperti X.
c) Penghapusan *Stopword*: Tujuan dari tahapan ini adalah untuk menghapus kata-kata yang sering muncul namun kurang berperan dalam analisis teks, seperti "yang", "tapi", dan kata sejenisnya.
d) *Stemming*: Proses ini berfungsi untuk mengubah kata-kata menjadi bentuk dasar atau akarnya.

Penerapan *preprocessing* yang tepat memiliki peran penting dalam meningkatkan performa model analisis, khususnya dalam konteks klasifikasi teks [13].

2.5. Penelitian Terdahulu

Penelitian ini merujuk kepada beberapa penelitian terdahulu yang membahas dua topik berbeda dalam analisis sentimen. Pertama, penelitian terkait pengaruh teknik *preprocessing* terhadap performa analisis sentimen. Sebuah studi mengenai sentimen masyarakat terhadap konflik Israel–Palestina menggunakan algoritma *Support Vector Machine* (SVM) menunjukkan bahwa skema *preprocessing* dengan *case folding* dan *stemming* menghasilkan F1-Score tertinggi—98% untuk ulasan positif dan 93% untuk negatif. Namun, penambahan *stopword removal* justru menurunkan performa menjadi 97% dan 85% untuk ulasan positif dan negatif [14]. Penelitian lain yang membandingkan tiga

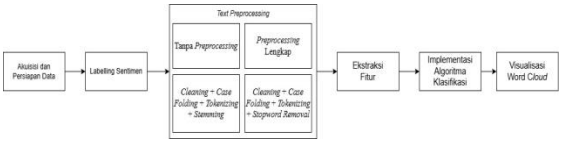
perlakuan—tanpa *preprocessing*, *stopword removal*, dan *stemming*—menunjukkan hasil yang mengejutkan, yakni data tanpa *preprocessing* menghasilkan akurasi tertinggi, yang menunjukkan bahwa efektivitas teknik *preprocessing* sangat bergantung pada karakteristik data dan model yang digunakan [15].

Kedua, ada penelitian yang membahas perbandingan antara algoritma dalam analisis sentimen. Sebuah studi terkait sentimen masyarakat selama pandemi COVID-19 di Twitter menemukan bahwa kombinasi teknik *cleaning*, normalisasi, dan *stemming* menghasilkan akurasi 77,77% dengan menggunakan SVM serta metode Mutual Information untuk seleksi fitur [16]. Sementara itu, penelitian lain yang menganalisis opini publik mengenai vaksinasi AstraZeneca menunjukkan bahwa *Naive Bayes* lebih unggul dibandingkan *K-Nearest Neighbor* (*KNN*), dengan akurasi 90,71% dibandingkan 74,78%, menunjukkan bahwa *Naive Bayes* lebih cocok untuk menganalisis opini publik di Twitter dalam konteks tertentu [17].

Namun, penelitian tentang opini masyarakat terkait pemindahan Ibu Kota Nusantara menunjukkan bahwa *KNN* memberikan performa yang lebih baik dengan akurasi 88,12% dan *precision* 93,98%, mengungguli *Naive Bayes* yang hanya mencapai 82,27% [18]. Meskipun sejumlah penelitian telah membahas pengaruh teknik *preprocessing* maupun melakukan perbandingan terhadap berbagai algoritma, belum banyak studi yang secara komprehensif menggabungkan kedua aspek tersebut dalam satu kerangka evaluasi yang terintegrasi, yakni mengevaluasi berbagai teknik *preprocessing* sekaligus membandingkan performa algoritma dalam kondisi yang seragam.

Oleh karena itu, penelitian ini menawarkan kebaruan dengan menguji secara komprehensif pengaruh kombinasi teknik *preprocessing* seperti *cleaning*, *case folding*, *tokenizing*, *stopword removal*, dan *stemming* terhadap performa dua algoritma, *Naive Bayes* dan *KNN*, untuk menemukan pendekatan terbaik dalam analisis sentimen teks berbahasa Indonesia.

3. METODE PENELITIAN



Gambar 1. Metodologi Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan membandingkan performa dua algoritma klasifikasi, yaitu *Naive Bayes* dan *K-Nearest Neighbor* (KNN), dalam menganalisis sentimen komentar masyarakat di platform X terhadap topik Danantara. Seluruh proses pengolahan data dan evaluasi dilakukan menggunakan perangkat lunak RapidMiner, yang memungkinkan pelaksanaan tahapan *preprocessing*, penerapan algoritma klasifikasi, serta perhitungan akurasi secara terintegrasi. Evaluasi dilakukan melalui empat skenario *preprocessing* teks yang berbeda, dengan tujuan untuk mengidentifikasi pengaruh masing-masing tahapan *preprocessing* terhadap tingkat akurasi klasifikasi. Adapun tahapan metodologi dalam penelitian ini digambarkan dalam Gambar 1 yang dijabarkan sebagai berikut:

3.1 Akuisisi dan Persiapan Data

Tahapan awal dalam penelitian ini adalah pengumpulan data yang dilakukan dengan teknik *crawling data* melalui platform X (sebelumnya Twitter) menggunakan Google Collab. Data dikumpulkan berdasarkan pencarian dengan kata kunci "Danantara" pada rentang waktu Februari hingga April 2025. Proses pengumpulan dilakukan secara bertahap untuk menyesuaikan dengan batasan rate limit dari X. Setelah seluruh proses *crawling* selesai, diperoleh sebanyak 1.453 data tweet yang kemudian dikompilasi menjadi satu dataset mentah berformat CSV untuk keperluan analisis lebih lanjut.

3.2 Labeling Sentimen

Proses pelabelan sentimen dalam penelitian ini dilakukan secara manual oleh peneliti tanpa bantuan leksikon atau alat otomatis lainnya. Setiap tweet dianalisis berdasarkan konteks dan diklasifikasikan ke dalam tiga kategori sentimen, yaitu positif, negatif, dan netral. Sentimen positif

mencerminkan dukungan terhadap peran Danantara sebagai pengelola aset negara, apresiasi terhadap kinerja atau strategi yang dilakukan, serta optimisme terhadap masa depan institusi ini. Sentimen negatif ditandai dengan adanya kritik, keraguan, atau kekhawatiran terhadap transparansi, efektivitas, maupun potensi penyalahgunaan dalam pengelolaan Danantara. Sementara itu, sentimen netral terdiri dari pernyataan yang bersifat informatif dan tidak memuat opini emosional, seperti penjelasan teknis atau deskripsi struktural.

3.3 Text *Preprocessing*

Penelitian ini membandingkan performa model pada empat kombinasi *preprocessing* untuk mengkaji pengaruhnya terhadap klasifikasi sentimen:

- Kasus 1 – Tanpa *Preprocessing*: Data dianalisis dalam bentuk mentah tanpa pembersihan.
- Kasus 2 – *Preprocessing* Lengkap: Meliputi tahapan *cleaning*, *case folding*, *tokenizing*, *stopword removal*, dan *stemming*.
- Kasus 3 – *Cleaning* + *Case Folding* + *Tokenizing* + *Stemming*: Tidak melibatkan *stopword removal*.
- Kasus 4 – *Cleaning* + *Case Folding* + *Tokenizing* + *Stopword Removal*: Tidak melibatkan *stemming*.

Setiap kombinasi digunakan untuk mengevaluasi bagaimana tahapan *preprocessing* memengaruhi performa klasifikasi model.

3.4 Ekstraksi Fitur

Setelah *preprocessing*, data teks dikonversi menjadi vektor numerik menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). Teknik ini mengukur pentingnya sebuah kata dalam suatu dokumen relatif terhadap seluruh korpus, memungkinkan model untuk memahami bobot dan relevansi kata terhadap konteks sentimen.

3.5 Implementasi Algoritma Klasifikasi

Dua algoritma klasifikasi diterapkan untuk mengevaluasi efektivitas model dalam berbagai skenario *preprocessing*:

- *Naive Bayes*: Algoritma probabilistik berbasis teorema Bayes dengan asumsi independensi antar fitur.
- *K-Nearest Neighbor (KNN):* Algoritma berbasis kedekatan jarak (*distance-based*) yang mengklasifikasikan data berdasarkan kemiripan dengan tetangga terdekatnya.

Pembagian data dilakukan menggunakan metode *Holdout Validation*, dengan proporsi 70% untuk data latih (*training set*) dan 30% untuk data uji (*test set*).

3.6 Evaluasi Kinerja

Kinerja algoritma *Naive Bayes* dan *KNN* dievaluasi berdasarkan nilai akurasi yang dihasilkan secara langsung melalui RapidMiner pada berbagai skenario *preprocessing* teks. Setiap skenario *preprocessing* diterapkan pada kedua algoritma, dan nilai akurasi yang diperoleh dicatat untuk masing-masing kombinasi. Selanjutnya, dilakukan analisis perbandingan akurasi antara kedua algoritma untuk setiap skenario guna mengidentifikasi metode *preprocessing* yang paling efektif dalam meningkatkan performa klasifikasi sentimen. Hasil evaluasi disajikan dalam bentuk tabel yang menampilkan nilai akurasi dari masing-masing kombinasi serta perbedaan performa antara kedua algoritma dalam setiap skenario.

3.7 Visualisasi *Word Cloud*

Untuk melengkapi analisis, dilakukan visualisasi *word cloud* pada hasil *preprocessing* lengkap. *Word cloud* ini dibuat terpisah untuk setiap kategori sentimen (positif, negatif, dan netral), dengan tujuan untuk mengidentifikasi kata-kata dominan yang muncul dalam masing-masing kelompok sentimen. Visualisasi ini memberikan gambaran yang lebih jelas mengenai kata-kata yang sering digunakan dalam setiap kategori, sehingga dapat membantu memahami pola sentimen yang terkandung dalam data.

4. HASIL DAN PEMBAHASAN

4.1. Analisis Tanpa *Text Preprocessing*

accuracy: 51.38%

	true Positif	true Negatif	true Netral	class precision
pred. Positif	0	0	0	0.00%
pred. Negatif	135	222	76	51.27%
pred. Netral	0	0	1	100.00%
class recall	0.00%	100.00%	1.30%	

Gambar 2. Akurasi Performa *Naive Bayes* Tanpa *Preprocessing*

accuracy: 51.15%

	true Positif	true Negatif	true Netral	class precision
pred. Positif	0	0	0	0.00%
pred. Negatif	135	222	77	51.15%
pred. Netral	0	0	0	0.00%
class recall	0.00%	100.00%	0.00%	

Gambar 3. Akurasi Performa *KNN* Tanpa *Preprocessing*

Pada tahap awal, analisis dilakukan terhadap data komentar tanpa melalui proses *text preprocessing*. Dengan demikian, teks masih mengandung berbagai elemen yang tidak standar seperti simbol, emotikon, huruf kapital, hingga bentuk kata yang beragam. Kondisi ini menyebabkan data menjadi *noisy* dan tidak homogen, sehingga menyulitkan model dalam mengenali pola-pola sentimen secara efektif.

Hasil pengujian menunjukkan bahwa algoritma *Naive Bayes* menghasilkan akurasi sebesar 51,38%, sedangkan K-Nearest Neighbor (*KNN*) memperoleh akurasi 51,15%. Kedua nilai ini menunjukkan performa yang rendah karena data mentah menyebabkan representasi vektor menjadi sangat bervariasi dan *sparse*. Hal ini berdampak langsung pada kemampuan klasifikasi model yang menjadi kurang optimal. Gambar 2 dan Gambar 3 menggambarkan akurasi rendah yang diperoleh kedua algoritma dalam kondisi tanpa *preprocessing*.

4.2. Analisis dengan *Preprocessing* Lengkap

accuracy: 65.21%

	true Positif	true Negatif	true Netral	class precision
pred. Positif	82	16	6	78.85%
pred. Negatif	37	170	40	68.83%
pred. Netral	16	36	31	37.35%
class recall	60.74%	76.58%	40.26%	

Gambar 4. Akurasi Performa *Naive Bayes* dengan *Preprocessing* Lengkap

accuracy: 66.13%

	true Positif	true Negatif	true Netral	class precision
pred. Positif	100	33	13	68.49%
pred. Negatif	27	164	41	70.69%
pred. Netral	8	25	23	41.07%
class recall	74.07%	73.87%	29.87%	

Gambar 5. Akurasi Performa *KNN* dengan *Preprocessing* Lengkap

Pada skenario ini, dilakukan *preprocessing* lengkap yang meliputi *cleaning*, *case folding*, *tokenizing*, *stopword removal*, dan *stemming*. Dengan tahapan ini, data menjadi lebih bersih, seragam, dan bermakna bagi model.

Merujuk pada Gambar 4 dan Gambar 5, hasil akurasi *Naive Bayes* hanya mencapai 65,21%, dan *KNN* sebesar 66,13%. Walaupun terjadi peningkatan dibanding skenario tanpa *preprocessing*, akurasi ini masih lebih rendah dibanding skenario dengan *preprocessing* parsial, yang akan dibahas pada bagian selanjutnya. Hal ini menunjukkan bahwa *preprocessing* berlebihan dapat menghapus konteks penting dalam teks. Kata-kata ekspresif yang mengandung sentimen emosional bisa saja hilang akibat proses *stemming* atau penghapusan *stopword*.

4.3. Analisis dengan *Cleaning*, *Case Folding*, *Tokenizing*, dan *Stemming*

accuracy: 66.59%

	true Positif	true Negatif	true Netral	class precision
pred. Positif	77	11	8	80.21%
pred. Negatif	45	183	40	68.28%
pred. Netral	13	28	29	41.43%
class recall	57.04%	82.43%	37.66%	

Gambar 6. Akurasi Performa *Naive Bayes* dengan *Cleaning, Case Folding, Tokenizing,* dan *Stemming*

accuracy: 68.89%

	true Positif	true Negatif	true Netral	class precision
pred. Positif	98	25	13	72.06%
pred. Negatif	30	174	37	72.20%
pred. Netral	7	23	27	47.37%
class recall	72.59%	78.38%	35.06%	

Gambar 7. Akurasi Performa *KNN* dengan *Cleaning, Case Folding, Tokenizing,* dan *Stemming*

Pada skenario ini, tahapan *preprocessing* mencakup *cleaning*, *case folding*, *tokenizing*, dan *stemming*, namun tidak disertai dengan *stopword removal*. Hasil pengujian

menunjukkan akurasi sebesar 66,59% untuk algoritma *Naive Bayes* dan 68,89% untuk *KNN*.

Berdasarkan Gambar 6 dan Gambar 7, terlihat bahwa penambahan proses *stemming* tidak memberikan peningkatan akurasi yang signifikan. Hal ini mengindikasikan bahwa penyederhanaan kata melalui *stemming* kurang memberikan kontribusi berarti dalam konteks data media sosial yang kaya akan ekspresi. Kemungkinan besar, bentuk asli dari kata justru lebih informatif dalam menangkap nuansa sentimen yang terkandung dalam teks.

4.4. Analisis dengan *Cleaning, Case Folding, Tokenizing*, dan *Stopword Removal*

accuracy: 65.21%

	true Positif	true Negatif	true Netral	class precision
pred. Positif	82	16	6	78.85%
pred. Negatif	37	170	40	68.83%
pred. Netral	16	36	31	37.35%
class recall	60.74%	76.58%	40.26%	

Gambar 8. Akurasi Performa *Naive Bayes* dengan *Cleaning, Case Folding, Tokenizing*, dan *Stopword Removal*

accuracy: 66.13%

	true Positif	true Negatif	true Netral	class precision
pred. Positif	100	33	13	68.49%
pred. Negatif	27	164	41	70.69%
pred. Netral	8	25	23	41.07%
class recall	74.07%	73.87%	29.87%	

Gambar 9. Akurasi Performa *KNN* dengan *Cleaning, Case Folding, Tokenizing*, dan *Stopword Removal*

Dalam skenario ini, *preprocessing* dilakukan dengan *cleaning, case folding, tokenizing*, dan *stopword removal*, tanpa menggunakan *stemming*. Hasilnya adalah 65,21% untuk *Naive Bayes* dan 66,13% untuk *KNN*—sama persis dengan *preprocessing* lengkap.

Merujuk pada Gambar 8 dan Gambar 9, terlihat bahwa proses *stopword removal* cenderung mengurangi efektivitas klasifikasi sentimen. Hal ini disebabkan karena beberapa *stopword* justru membawa makna penting, terutama dalam konteks negasi dan penekanan. Oleh karena itu, *stopword removal* perlu dipertimbangkan ulang dalam konteks data sentimen publik.

4.6. *Word cloud* Kategori Sentimen

Gambar 10. *Word Cloud* Sentimen Positif

Gambar 11. *Word Cloud* Sentimen Negatif

Gambar 12. *Word Cloud* Sentimen Netral

Untuk memahami lebih dalam karakteristik data dalam tiap kategori sentimen, dilakukan visualisasi dalam bentuk *word cloud* terhadap tiga kategori utama: sentimen positif, negatif, dan netral. *Word cloud* ini dihasilkan setelah seluruh tahapan *preprocessing* data selesai dilakukan, termasuk *cleaning, case folding, tokenizing, stopword removal*, dan *stemming*.

- **Gambar 10** menampilkan *word cloud* untuk kategori sentimen positif. Terlihat bahwa kata-kata seperti *Indonesia*, *semangat*, dan *positif* mendominasi representasi visual, menunjukkan konteks

ujaran yang bernada optimis dan mendukung.

- **Gambar 11** menampilkan *word cloud* untuk kategori sentimen negatif. Kata-kata seperti *negara*, *korupsi*, dan *koruptor* muncul dengan frekuensi tinggi, mencerminkan ekspresi kekecewaan atau kritik terhadap Danantara.
- **Gambar 12** menampilkan *word cloud* untuk kategori sentimen netral. Kata-kata yang muncul seperti *presiden*, *BUMN*, dan *kelola* cenderung bersifat informatif dan deskriptif, tanpa muatan emosi yang kuat.

Visualisasi ini sangat membantu dalam memahami konteks dan kecenderungan kata yang sering muncul pada tiap kategori sentimen. Selain itu, *word cloud* juga memberikan gambaran awal tentang bagaimana fitur-fitur tekstual dapat berkontribusi terhadap performa model dalam membedakan polaritas sentimen. Melalui frekuensi dan ukuran kata yang ditampilkan, kita dapat mengidentifikasi kata-kata kunci yang berperan penting dalam klasifikasi sentimen.

4.7. Perbandingan Kinerja *Naive Bayes* dan *KNN* pada Empat Kasus

Tabel 1. Perbandingan Kinerja *Naive Bayes* dan *KNN* pada Empat Kasus

No	Skenario *Preprocessing*	Akurasi *Naive Bayes* (%)	Akurasi *KNN* (%)
1	Tanpa *Preprocessing*	51,38	51,15
2	*Cleaning* + *Case Folding* + *Tokenizing* + *Stemming*	66,59	68,89
3	*Cleaning* + *Case Folding* + *Tokenizing* + *Stopword removal*	65,21	66,13
4	*Preprocessing* Lengkap (*Cleaning*, *Case Folding*, *Tokenizing*, *Stopword removal*, *Stemming*)	65,21	66,13

Tabel 1 merangkum performa algoritma *Naive Bayes* dan *K-Nearest Neighbor* (*KNN*) dalam empat skenario *preprocessing* yang berbeda. Berdasarkan data tersebut, terlihat bahwa *KNN* cenderung lebih unggul pada berbagai skenario, sementara *Naïve Bayes* cenderung lebih rendah.

Menariknya, skenario dengan kombinasi *cleaning*, *case folding*, *tokenizing*, dan *stemming* menghasilkan akurasi tertinggi pada kedua algoritma, yakni 66,59% untuk *Naive Bayes* dan 68,89% untuk *KNN*. Hal ini menunjukkan bahwa tahap *stemming* memiliki dampak yang sangat besar terhadap efektivitas klasifikasi, meskipun tidak disertai *stopword removal*.

Sebaliknya, *preprocessing* lengkap yang mencakup semua tahap (*cleaning* + *case folding*, *tokenizing*, *stopword removal*, dan *stemming*) justru menghasilkan akurasi yang lebih rendah, meskipun tetap lebih baik dibandingkan skenario tanpa *preprocessing*. Temuan ini mengindikasikan bahwa penambahan *stopword removal* tidak selalu memberikan kontribusi signifikan, bahkan berpotensi mengurangi konteks yang penting dalam data.

Dengan demikian, dapat disimpulkan bahwa kombinasi *preprocessing* yang paling optimal dalam penelitian ini adalah *cleaning*, *case folding*, *tokenizing*, dan *stemming*, bukan *preprocessing* lengkap seperti yang lazim diasumsikan. Temuan ini memperkuat pentingnya pendekatan eksperimen yang bertahap dalam menentukan strategi *preprocessing* yang paling efektif, khususnya dalam klasifikasi sentimen terhadap komentar X yang berkaitan dengan topik Danantara.

5. KESIMPULAN

Penelitian ini membandingkan performa algoritma *Naive Bayes* dan *K-Nearest Neighbor* (*KNN*) dalam analisis sentimen komentar X terhadap topik Danantara melalui

empat skenario *preprocessing* teks: tanpa *preprocessing*, *cleaning + case folding + tokenizing + stemming, cleaning + case folding + tokenizing + stopword removal*, *preprocessing* lengkap. Hasil menunjukkan bahwa skenario *preprocessing* memiliki dampak signifikan terhadap akurasi klasifikasi. Skema *preprocessing* terbaik berdasarkan hasil akurasi adalah *cleaning + case folding + tokenizing + stemming*, dengan akurasi *KNN* mencapai 68,89% dan *Naive Bayes* 66,59%. Sebaliknya, skema *preprocessing* lengkap justru menunjukkan penurunan akurasi menjadi 66,13% (*KNN*) dan 65,21% (*Naive Bayes*), yang mengindikasikan bahwa penambahan proses tidak selalu menjamin peningkatan performa.

Secara keseluruhan, algoritma *KNN* menunjukkan performa unggul dalam skenario terbaik, meskipun cenderung fluktuatif di skenario lainnya. Sebaliknya, *Naive Bayes* tampil lebih stabil dan konsisten berkat pendekatan probabilistiknya yang lebih toleran terhadap data berdimensi tinggi. Analisis *word cloud* dari hasil klasifikasi memperlihatkan bahwa kata-kata seperti "Indonesia", "semangat", dan "positif" mendominasi sentimen positif; "negara", "korupsi", dan "koruptor" muncul dalam sentimen negatif; sedangkan "presiden", "BUMN", dan "kelola" lebih banyak ditemukan dalam sentimen netral. Berdasarkan temuan ini, kombinasi algoritma KNN dengan tahapan *preprocessing* berupa *cleaning*, *case folding*, *tokenizing*, dan *stemming* direkomendasikan untuk analisis sentimen komentar X. Namun demikian, *Naive Bayes* tetap menjadi pilihan yang andal untuk memperoleh hasil yang lebih konsisten di berbagai kondisi.

DAFTAR PUSTAKA

[1] J. Hafizd, F. S. Nurfalah, M. A. P. Ramadhan, P. Kaerudin, and K. Elok, "Peran Media Sosial dalam Penyampaian Aspirasi Masyarakat untuk Perubahan yang Lebih Baik," *Strat. Soc. Humanit. Stud.*, vol. 1, no. 2, pp. 147–155, 2023, doi: 10.59631/sshs.v1i2.108.

[2] "Dikabarkan Akan Diblokir, Ternyata Masyarakat X Indonesia Keempat Terbanyak di Dunia - GoodStats Data." Accessed: May 05, 2025. [Online]. Available: https://data.goodstats.id/statistic/dikabarkan-akan-diblokir-ternyata-masyarakat-x-indonesia-keempat-terbanyak-di-dunia-hCYcS

[3] "Danantara Resmi Diluncurkan, Rosan Roeslani Tegaskan Komitmen Tata Kelola yang Transparan dan Berintegritas | Sekretariat Negara." Accessed: May 04, 2025. [Online]. Available: https://www.setneg.go.id/baca/index/danantara_resmi_diluncurkan_rosan_roeslani_tegaskan_komitmen_tata_kelola_yang_transparan_dan_berintegritas

[4] "Danantara Tuai Pro dan Kontra Warganet, Diresmikan Prabowo 24 Februari." Accessed: May 04, 2025. [Online]. Available: https://inet.detik.com/cyberlife/d-7784882/danantara-tuai-pro-dan-kontra-warganet-diresmikan-prabowo-24-februari

[5] F. Aftab *et al.*, "A Comprehensive Survey on Sentiment Analysis Techniques," *Int. J. Technol.*, vol. 14, no. 6, pp. 1288–1298, 2023, doi: 10.14716/ijtech.v14i6.6632.

[6] A. A. Nisa, N. H. Safitri, L. Stianingsih, and F. H. Saputri, "Analysis of Customer Sentiment towards Wardah Beauty Products on Instagram Using K-Nearest Neighbors (KNN) and Naïve Bayes Algorithms," *G-Tech J. Teknol. Terap.*, vol. 9, no. 2, pp. 919–928, 2025.

[7] M. U. Albab, Y. Karuniawati P, and M. N. Fawaiq, "Optimization of the Stemming Technique on Text preprocessing President 3 Periods Topic," *J. Transform.*, vol. 20, no. 2, pp. 1–10, 2023, doi: 10.26623/transformatika.v20i2.5374.

[8] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, pp. 1–19, 2021, doi: 10.1007/s13278-021-00776-6.

[9] R. Ayatulloh, K. Noor, N. T. Romadloni, F. Ramadhani, U. M. Karanganyar, and K. Karanganyar, "Perbandingan kinerja algoritma klasifikasi pada review masyarakat aplikasi netflix," *JITET (Jurnal Inform. dan Tek. Elektro Ter.*, vol. 13, no. 2, 2025.

[10] D. M. Efendi, S. Mintoro, . S., S. H. Lubis, and S. Lestari, "KLASIFIKASI KINERJA PEMBAYARAN ANGSURAN DENGAN ALGORITMA NAIVE BAYES (Studi Kasus : Data Nasabah Koperasi Simpan Pinjam Pembiayaan Syariah Bina Bersama)," *J. Inf. dan Komput.*, vol. 10, no. 1, pp. 57–61, 2022, doi: 10.35959/jik.v10i1.305.

[11] M. Rahmadiah and P. Suparman, "Penerapan Metode K-Nearest Neighbour Untuk Sistem Penentuan Peminjaman Modal Nasabah Bank Syariah Indonesia Cabang Cikarang Berbasis

Website," *J. Inf. dan Komput.*, vol. 10, no. 2, pp. 189–197, 2022.

[12] L. G. Irham, A. Adiwijaya, and U. N. Wisesty, "Klasifikasi Berita Bahasa Indonesia Menggunakan Mutual Information dan Support Vector Machine," *J. Media Inform. Budidarma*, vol. 3, no. 4, p. 284, 2019, doi: 10.30865/mib.v3i4.1410.

[13] E. Alshdaifat, D. Alshdaifat, A. Alsarhan, F. Hussein, and S. Moh, "The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance," *Data*, vol. 6, no. 11, 2021.

[14] A. A. Syam, G. H. M, A. Salim, D. F. Surianto, and M. F. B, "Analisis teknik preprocessing pada sentimen masyarakat terkait konflik israel-palestina menggunakan support vector machine," vol. 9, no. 3, pp. 1464–1472, 2024.

[15] B. Hakim, "Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning," *JBASE - J. Bus. Audit Inf. Syst.*, vol. 4, no. 2, pp. 16–22, 2021, doi: 10.30813/jbase.v4i2.3000.

[16] 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.

[17] S. H. Ramadhani and M. I. Wahyudin, "Analisis Sentimen Terhadap Vaksinasi Astra Zeneca pada Twitter Menggunakan Metode Naïve Bayes dan K-NN," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 6, no. 4, pp. 526–534, 2022, doi: 10.35870/jtik.v6i4.530.

[18] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," *J. KomtekInfo*, vol. 10, pp. 1–7, 2023, doi: 10.35134/komtekinfo.v10i1.330.