DSI DATA SCIENCES INDONESIA

# Deteksi Berita Hoax pada Platform X Menggunakan Pendekatan Text Mining dan Algoritma Machine Learning

**Aulia Kartika Dewi[1)], Noni Fauzia Rahmadani[2)], Rifdah Syahputri[3)], Luftia Rahma Nasution[4)] Mhd. Furqan[5)]**
[1,2,3,4,5)] Universitas Islam Negeri Sumatera Utara

Received: 3 June 2025 Accepted: 18 June 2025 Published: 1 July 2025

OPEN ACCESS

*auliakartikadewi949@gmail.com

**Kata Kunci:** Hoax, Penambangan Teks, Mesin Vektor pendukung, Deteksi Otomatis, Klasifikasi Teks

**DSI: Jurnal Data Science Indonesia** is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

**Abstrak :** Penyebaran berita hoax di media sosial seperti Platform X meningkat pesat dan berdampak negatif pada masyarakat. Penelitian ini bertujuan untuk mengembangkan sistem deteksi hoax berbasis teks menggunakan pendekatan text mining dan algoritma machine learning, khususnya Support Vector Machine (SVM). Dataset yang digunakan terdiri dari 3.000 entri teks berbahasa Indonesia yang dikumpulkan dari Platform X dan telah divalidasi melalui sumber resmi seperti TurnBackHoax dan Kominfo. Proses preprocessing mencakup tokenisasi, stopword removal, dan stemming, dengan ekstraksi fitur berbasis TF-IDF menghasilkan 1.200 fitur utama. Model SVM dengan kernel Radial Basis Function (RBF) dipilih melalui grid search untuk memperoleh performa optimal. Evaluasi dilakukan menggunakan lima kali uji validasi silang, dengan hasil akurasi mencapai 83,2%, precision 82,5%, recall 84,1%, dan F1-score 83,3%. Kontribusi utama penelitian ini adalah penerapan kombinasi preprocessing TF-IDF dan SVM kernel RBF secara efektif untuk deteksi hoax dalam Bahasa Indonesia pada Platform X. Temuan ini menunjukkan potensi pendekatan ini dalam mengurangi penyebaran hoax secara otomatis.

## PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi yang pesat telah mengubah secara drastis cara masyarakat mengakses dan menyebarkan informasi. Media sosial kini menjadi salah satu platform utama dalam penyebaran informasi karena kemudahan dan kecepatannya. Namun, di balik manfaat tersebut, media sosial juga membuka ruang luas bagi penyebaran informasi yang tidak akurat, termasuk berita hoax. Berita hoax yang sengaja disebarluaskan untuk menyesatkan publik dapat menimbulkan keresahan sosial, kerugian ekonomi, serta merusak reputasi individu atau kelompok. Oleh karena itu, deteksi otomatis terhadap berita hoax menjadi krusial untuk meminimalisasi dampak negatif yang ditimbulkannya [1].

Mendeteksi berita hoax secara manual tidak lagi efektif karena volume data yang beredar di platform digital sangat besar dan terus meningkat. Untuk itu, dibutuhkan sistem deteksi otomatis yang mampu mengidentifikasi dan mengklasifikasikan konten hoax secara cepat, akurat, dan efisien. Salah satu pendekatan yang banyak digunakan adalah kombinasi metode *text mining* dan algoritma *machine learning*. Text mining memungkinkan ekstraksi informasi penting dari data teks tidak terstruktur, sementara machine learning memungkinkan sistem belajar dari data dan melakukan prediksi secara otomatis [2].

Dalam penelitian ini, algoritma Support Vector Machine (SVM) dipilih sebagai metode klasifikasi karena kemampuannya dalam menangani data berdimensi tinggi seperti teks. SVM telah menunjukkan performa yang baik dalam berbagai studi deteksi hoax. Misalnya, Fani [3] menggunakan SVM untuk mendeteksi hoax dari situs TurnBackHoax dan memperoleh akurasi sebesar 83% dengan recall pada kelas hoax mencapai 99%. Penelitian lain oleh Joergensen dan Munthe [4] menggabungkan algoritma K-Nearest Neighbors (KNN) dan SVM, menghasilkan akurasi hingga 93,31%.

Meskipun SVM telah banyak digunakan, masih terdapat keterbatasan dalam penelitian terdahulu, terutama dalam konteks deteksi hoax berbahasa Indonesia yang bersumber dari Platform X. Selain itu,

belum banyak studi yang secara spesifik mengeksplorasi integrasi preprocessing berbasis TF-IDF dan pemilihan kernel secara eksplisit dalam penerapan SVM pada platform tersebut. Berdasarkan celah tersebut, penelitian ini bertujuan untuk mengembangkan model klasifikasi hoax menggunakan pendekatan text mining dan algoritma SVM dengan kernel RBF, yang diharapkan dapat meningkatkan efektivitas deteksi hoax dalam Bahasa Indonesia pada Platform X. Penelitian ini juga diharapkan dapat berkontribusi dalam pengembangan sistem pendeteksi hoax otomatis yang adaptif dan aplikatif, serta mendorong peningkatan literasi informasi digital di masyarakat.

**TINJAUAN LITERATUR**

*Hoax* atau informasi palsu merupakan suatu fenomena yang marak terjadi akhir-akhir ini di media sosial, dan merupakan akibat yang ditimbulkan dari kemudahan akses teknologi yang terjadi Informasi dapat dikatakan palsu dikarenakantidak jelas sumber serta kebenarannya atau (fake news), kemudian disebarkanmelalui media sosial atau pun media chatting tanpa klarifikasi dan tanpa bisa melakukan klarifikasi. Mereka seakan tidak mau tahu apakah berita yang mereka terima terbukti kebenarannya. Banyak masyarakat yang langsung mengeluarkan doktrin atau penilaiannya pada sebuah berita tanpa hadir di sekitar tempat kejadian. *Hoax* berasal dari kata "hocus pocus"yang berasal dari bahasa latin "hoc est corpus", yang mempunyai arti"ini adalah tubuh". Kata ini biasanya digunakan oleh penyihir untuk mengklaim bahwa sesuatu adalah benar, padahal belum tentu benar, lebih lanjut dijelaskan bahwa kata hoaxbersumberdari bahasa Inggris yang memiliki artitipuan, menipu, berita bohong, berita palsu, dan kabar burung Jadi, hoaxbisa diartikan dengan ketidakbenaran dari suatu informasi [5].

Menurut definisi dari KBBI, kata "*hoax*" berarti informasi bohong. Kehadiran berita digital telah membawa kembali dan meningkatkan frekuensi pembuatan atau kemunculan hoax, atau "jurnalisme kuning". Berita palsu sering tersebar sebagai informasi yang salah di media sosial, namun kadang-kadang juga muncul di media umum [6]. Hoaks biasanya mengandung ujaran kebencian yang dapat menghasut, menyudutkan, dan bahkan memprovokasi agama, tokoh ideologi, dan lainlain. Hal ini terjadi ketika interaksi non-tatap muka antar orang-orang di media sosial mendorong keberanian untuk mengungkapkan pendapat atau pernyataan yang melibatkan penghinaan, ujaran kebencian, dan perundungan [7].

*Hoax* memiliki sisianegatif yang merugikan banyak pihak yang terlibat didalamnya. Bahkan, hoax dapat mengganggu stabilitas kehidupan itu sendiri. Oleh karena itu, hoax perlu disikapi secara bijak. Berita hoax sendiri biasanya memiliki efek yang mencengangkan atau bombastis bagi para pembacanya. Pada dasarnya perilaku pembaca lebih cenderung percaya informasi *hoax*, jika informasinya sesuai dengan opini atau sikap yang dimiliki Dan dengan efek tersebut para pembaca merasa harus membagikannya. Tidak ada kata repot dalam membagikan sebuah informasi diera internet sekarang ini, hanya beberapa tombol maka dalam hitungan detik, informasi tersebut dapat tersebar dengan cepat. Pada prosesnya itupun, tidak ada biaya apapun. Konsekuensi dari hal ini, para pelaku penyebaran tidak akan befikir dua kali untuk menyebarkannya walaupun berita tersebut salah [8].

X adalah platform digital yang memfasilitasi berbagi pemikiran dan kritik melalui konten tertulis. Sejumlah individu dan organisasi atau perusahaan memanfaatkan media sosial seperti X untuk mengetahui sentimen masyarakat umum saat akan mengambil kebijakan. Namun, karena keterbatasan menulis, yang tidak mampu mentransmisikan isyarat nonverbal seperti gerakan tubuh, ekspresi wajah, dan intonasi, seringkali sulit untuk mengidentifikasi tanda-tanda implisit seperti sarkas [9].

Di era digital ini sudah membawa perubahan yang sangat signifikan dalam cara manusia berinteraksi dan komunikasi, aplikasi X menjadi tempat utama untuk berbagi informasi. Aplikasi X sebelumnya dikenal sebagai twitter. Aplikasi X merupakan salah satu jaringan sosial terbesar didunia situs web yang paling banyak dikunjungi ke-5 didunia. Aplikasi X memungkinkan pengguna untuk berbagi teks, gambar, dan video. Banyak tokoh masyarakat, influencer dan bahkan masyarakat umum menggunakan aplikasi X untuk berkomunikasi dengan pengikutnya dan mengungkapkan pendapat mereka. Aplikasi X digunakan oleh individu, organisasi, dan perusahaan untuk berkomunikasi, membagikan informasi, dan mempromosikan produk atau layanan mereka [10].

Platform seperti Twitter dapat digunakan pada penyebaran berita informasi dengan berbagai cara.

Platform ini meningkatkan kemungkinan berita yang berpusat pada individu dengan memberikan posisi yang lebih luas, membuat dan berbagi konten [11]. Platform ini memberikan akses yang lebih luas dan langsung bagi warga negara untuk terlibat dalam berita luar, membuat suatu opini mereka di dengar dan mempengaruhi orang lain [12].

*Text Mining* adalah suatu proses penambangan intisari dari suatu dokumen data berupa teks yang bentuknya lebih tidak teratur yang dapat dibuat pola untuk menghasilkan sebuah informasi yang berguna. Proses penambangan data dapat bersumber melalui website, yang datanya dapat berupa data dokumen berbentuk teks dari website yang di dalamnya berisi ulasan, opini, kritik dan *feedback* [13]. Tujuan *text mining* adalah untuk pengkategorian data teks, baik itu untuk menemukan kategori yang sesuai dengan kelas yang ditentukan (supervised learning) maupun berdasarkan kesamaan karakteristik dan memberikan label pada kelas yang belum diketahui (unsupervised learning) [14].

*Text Mining* adalah bidang penelitian ilmu komputer yang baru dan menarik yang mencoba memecahkan krisis informasi yang berlebihan dengan menggabungkan teknik dari data mining, pembelajaran mesin, pemrosesan bahasa alami, pencarian informasi, dan pengetahuan pengelolaan. Text mining dapat didefinisikan secara luas sebagai proses intensif pengetahuan di mana pengguna berinteraksi dengan dokumen dari waktu ke waktu menggunakan seperangkat alat analisis [15]. Secara umum, *text mining* mengacu pada proses ekstrak informasi dari dokumen tidak terstruktur [10]. Pendefinisian *text mining* dapat dijelaskan dengan adanya penemuan informasi baru atau tidak diketahui sebelumnya oleh komputer dan informasi dari berbagai sumber yang berhasil diekstrak merupakan kunci dari proses text mining [16].

*Machine learning* adalah bagian dari *artificial intelligence* yang didefinisikan sebagai metode yang belajar dari pengalaman dan dapat menyesuaikan dengan input baru untuk melakukan suatu tindakan otomatis dan membuat keputusan yang tepat dan sesuai. *Machine learning* membantu peran manusia dalam pengambilan keputusan dengan mengambil keputusan secara otomatis, belajar dari data, dan menentukan pola [17]. Algoritma *machine learning* adalah metodeatau langkah yang digunakan untuk membuat model dari data. Algoritma bertujuan untuk mengidentifikasi pola dalam data atau hubungan dalam data dan memanfaatkannya untuk menyelesaikan berbagai masalah. Setiap algoritma memiliki kelebihan dan kelemahan tergantung pada jenis data dan masalah yang dihadapi [18].

*Machine Learning* bekerja dengan diawali tahapan input dataset terkait permasalahan yang ingin diproses untuk kemudian dipelajari melalui algoritma yang telah ditetapkan. Algoritma Machine Learning terdiri dari beberapa jenis yang dapat diterapkan dalam pengujian, salah satunya adalah algoritma jenis Supervised Learning yang diterapkan dalam model prediksi ini. *Supervised Learning* merupakan salah satu bagian dari pengelompokkan yang ada pada metode pembelajaran machine learning. *Supervised Learning* diartikan sebagai konsep ajar yang dalam pengawasan atau dalam satu jalan yang terarah. *Supervised Learning* mengenakan label atas data train untuk kemudian dijadikan acuan dalam pengamatan label data yang baru. Model dalam Supervised Learning diciptakan guna sebagai bahan menghasilkan ouput yang terperinci dengan dasar masukan sebelumnya [19].

Algoritma Support Vector Machine (SVM) digunakan sebagai salah satu pendekatan utama dalam klasifikasi berita palsu (fake news) berdasarkan judul dan isi artikel. SVM dipilih karena kemampuannya dalam mengenali pola dan membedakan dua kelas data dengan margin maksimum, yang sangat berguna dalam tugas klasifikasi biner seperti deteksi hoaks. Hasil eksperimen menunjukkan bahwa SVM menghasilkan akurasi tertinggi di antara seluruh algoritma yang diuji ketika proses pembelajaran hanya menggunakan judul berita, dengan nilai akurasi mencapai 94,19%. Meski demikian, waktu pelatihan model SVM jauh lebih tinggi dibandingkan metode lain, yaitu sekitar 186 detik, menjadikannya metode paling lambat dalam konteks ini. Hal ini menyoroti trade-off antara performa klasifikasi dan efisiensi komputasi. Oleh karena itu, meskipun SVM unggul dari segi akurasi, penggunaannya dalam sistem real-time harus mempertimbangkan waktu proses yang relatif lambat dibandingkan metode ensemble seperti Random Forest atau AdaBoost yang memberikan hasil sedikit lebih rendah tetapi dengan waktu pelatihan yang jauh lebih singkat [20].

Dalam penelitian M. Sudhakar & K.P. Kaliyamurthie, algoritma Support Vector Machine (SVM) digunakan untuk mendeteksi berita hoaks yang bersumber dari media sosial, khususnya Twitter, dengan

pendekatan berbasis pembelajaran mesin. Penulis membandingkan SVM dengan berbagai algoritma lain seperti Logistic Regression, Naïve Bayes, dan LSTM, dan hasilnya menunjukkan bahwa SVM memberikan performa terbaik kedua dengan akurasi mencapai 98%, hanya sedikit di bawah Logistic Regression yang mencapai 95,2%. SVM dipilih karena kemampuannya dalam membentuk *hyperplane* terbaik yang memisahkan dua kelas secara optimal, bahkan pada data non-linier melalui penerapan kernel trick. Penelitian ini juga menggunakan teknik ekstraksi fitur TF-IDF, serta menerapkan preprocessing seperti penghapusan karakter spesial, normalisasi teks, dan tokenisasi. Meski menghasilkan akurasi tinggi, model SVM juga memerlukan sumber daya komputasi yang cukup besar dan pelatihan yang relatif lama, namun tetap diakui sebagai algoritma yang sangat efektif dan stabil dalam mendeteksi hoaks dalam format teks pendek seperti tweet [21].

Berikut adalah beberapa penelitian terdahulu yang relevan dengan topik deteksi hoax berbasis teks. Analisis ini menunjukkan posisi dan kontribusi penelitian yang sedang dilakukan.

Tabel 1. Anakisis Kritir Penelitian

| **Peneliti** | **Metode** | **Dataset** | **Preprocessing** | **Hasil** | **Kelebihan** | **Kelemahan** |
|---|---|---|---|---|---|---|
| Fani (2022) | SVM | Situs TurnBackHoax (2000 data) | TF-IDF | Akurasi 83%, Recall 99% | Fokus data lokal, hasil tinggi | Tidak dijelaskan kernel SVM yang digunakan |
| Joergensen & Munthe (2021) | KNN + SVM | Media online nasional | Tidak dijelaskan | Akurasi 93,31% | Gabungan algoritma | Dataset belum divalidasi resmi |
| Sari et al. (2023) | NB, SVM | Twitter Indonesia | TF dan TF-IDF | SVM lebih baik dari NB | Ada perbandingan algoritma | Tidak fokus pada kernel atau parameter tuning |
| Penelitian ini | SVM (RBF kernel) | Platform X (3000 tweet divalidasi) | TF-IDF | Akurasi 83,2%, F1 83,3% | Fokus Bahasa Indonesia, kernel eksplisit | Belum dibandingkan dengan model lain |

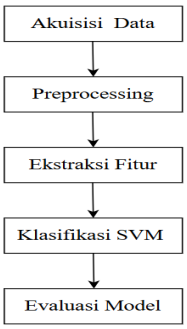
Dari hasil analisis, dapat disimpulkan bahwa:

a. Sebagian besar studi sebelumnya belum berfokus secara khusus pada Platform X sebagai sumber data hoax dalam Bahasa Indonesia.
b. Belum banyak penelitian yang secara eksplisit menyebut dan mengevaluasi pemilihan kernel dalam algoritma SVM.
c. Dataset yang digunakan dalam banyak studi belum tervalidasi melalui sumber resmi seperti TurnBackHoax atau Kominfo.
d. Belum ada integrasi menyeluruh antara preprocessing TF-IDF, pemilihan kernel, dan evaluasi klasifikasi secara end-to-end dalam konteks Indonesia.

Oleh karena itu, penelitian ini menempati posisi yang strategis dalam mengisi celah-celah tersebut dengan pendekatan sistematis berbasis TF-IDF dan SVM kernel RBF terhadap data hoax dari Platform X yang telah divalidasi.

## METODE PENELITIAN

Penelitian ini menggunakan algoritma Support Vector Machine (SVM) untuk klasifikasi berita hoax. Metode Knowledge Discovery in Database (KDD) digunakan untuk menemukan pola atau teknik dari data yang tersedia. Terdapat lima tahapan utama dalam proses *text mining* dalam kerangka KDD, yaitu: (1) pemilihan data, (2) preprocessing, (3) transformasi, (4) data mining, dan (5) evaluasi.




Gambar 1. Tahapan utama

Berbeda dengan strategi *neural network* yang mencoba menemukan *hyperplane* pemisah kelas secara iteratif, SVM berusaha menemukan *hyperplane* optimal dalam ruang input. Prinsip dasar SVM adalah pengklasifikasi linier, namun dapat dikembangkan untuk masalah non-linier dengan penerapan konsep *kernel trick*, yang memetakan data ke dalam ruang berdimensi tinggi agar dapat dipisahkan secara linier. Perkembangan teknik ini telah meningkatkan minat penelitian di bidang pengenalan pola, khususnya dalam klasifikasi berbasis teks.

Support Vector Machine merupakan algoritma klasifikasi yang termasuk dalam kategori *supervised learning*. Konsep dasarnya adalah mencari *hyperplane* terbaik yang mampu memisahkan dua kelas secara optimal. Namun, dalam kondisi tertentu, kernel linier tidak cukup kuat untuk memisahkan data dengan baik. Untuk mengatasi hal ini, digunakan kernel non-linier seperti *Radial Basis Function* (RBF) yang memungkinkan pembentukan *hyperplane* optimal dalam ruang fitur berdimensi tinggi [22]. Dalam penelitian ini, penulis menggunakan metode eksperimen yang bersifat kuantitatif. Penelitian ini dilakukan secara sistematis untuk menghubungkan sebab dan akibat, melalui manipulasi variabel dan observasi terhadap hasil. Salah satu tahapan krusial adalah preprocessing data, karena data yang diperoleh dari media sosial sering kali mengandung *noise*, *missing value*, *outlier*, serta struktur data yang tidak konsisten. Oleh karena itu, preprocessing diperlukan untuk memastikan kualitas data sebelum dilakukan klasifikasi. Beberapa tahapan preprocessing yang dilakukan antara lain: *cleansing* (menghapus tanda baca dan karakter khusus), *case folding* (mengubah huruf menjadi huruf kecil), *tokenizing* (memecah kalimat menjadi kata), *filtering* (menghapus *stopword* atau kata-kata umum yang tidak bermakna), dan *lemmatizing/stemming* (mengubah kata ke bentuk dasar) [23].

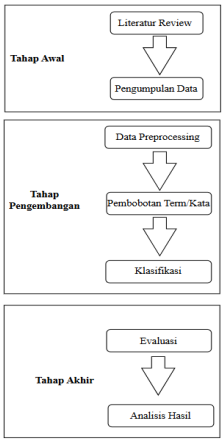
Definisi SVM adalah sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi linier dalam fitur berdimensi tinggi, dan dilatih menggunakan algoritma berbasis optimasi matematis. SVM pertama kali diperkenalkan oleh Vapnik pada tahun 1992, dan hingga saat ini menjadi metode andalan dalam pengenalan pola. Akurasi model SVM sangat bergantung pada pemilihan kernel dan parameter yang digunakan. Berdasarkan karakteristik datanya, SVM terbagi menjadi linier dan non-linier. SVM linier digunakan jika data dapat dipisahkan secara linier, sedangkan SVM non-linier menggunakan fungsi kernel untuk memisahkan data dalam ruang yang lebih kompleks [24].

Dalam penelitian ini digunakan 3.000 tweet berbahasa Indonesia yang dikumpulkan melalui Twitter API (Platform X). Dataset ini telah divalidasi kebenarannya menggunakan sumber terpercaya seperti TurnBackHoax.id dan Kominfo, serta terbagi secara seimbang antara dua kelas, yaitu 1.500 data hoax dan 1.500 data non-hoax. Data bersifat balanced, sehingga tidak memerlukan penyesuaian tambahan terkait ketidakseimbangan kelas. Proses ekstraksi fitur dilakukan menggunakan metode TF-IDF, yang menghasilkan 1.200 fitur utama sebagai representasi vektor dari data teks. Untuk klasifikasi, digunakan algoritma Support Vector Machine (SVM) dengan jenis kernel Radial Basis Function (RBF), yang dipilih melalui proses *grid search* guna memperoleh performa optimal. Parameter tuning yang digunakan meliputi C = 1.0, gamma = 'scale', dan kernel = 'rbf'.

## 1. Pengembangan Metode

Dalam melakukan penelitian ini, setidaknya dibagi menjadi 3 tahap besar, yaitu tahap awal, tahap

pengembangan, dan tahap akhir. Konsep SVM dapat dijelaskan secara sederhana sebagai upaya untuk menemukan hyperplane2 terbaik yang berfungsi sebagai pemisah dua kelas dalam ruang input. Dokumen pada umumnya memiliki struktur yang berubah-ubah atau tidak terstruktur. Oleh karena itu, diperlukan suatu proses yang dapat mengubah bentuk data yang sebelumnya tidak terstruktur menjadi bentuk data yang terstruktur. Proses konversi ini dikenal dengan istilah text preprocessing.

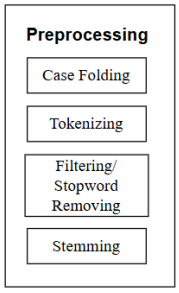



Gambar 2. Tahapan Metodologi

Pada proses klasifikasi, algoritma Support Vector Machine (SVM) digunakan untuk mengenali pola pada data training yang telah diberi label dan kemudian menghasilkan model machine learning. Model yang dihasilkan kemudian digunakan untuk mengenali data tweet baru yang belum dilabeli sehingga menghasilkan prediksi kelas sentimennya apakah positif, negatif, atau netral.

**2. Data Proses**

Proses pengambilan data dilakukan dengan mengumpulkan data-data pendukung yang diambil selama proses penelitian berlangsung. Beberapa cara yang dapat digunakan adalah melalui studi literatur, dan salah satu cara yang dapat digunakan adalah dengan mengumpulkan dan mempelajari beberapa referensi dari berbagai sumber yang berkaitan dengan judul penelitian yang dilakukan. Dataset diambil melalui Twitter API dengan proses crawling. Dataset dalam bentuk file berformat csv. Gambar 3. Ilustrasi pengambilan data melalui proses crawling.




Gambar 3. *Preprocessing Stages*

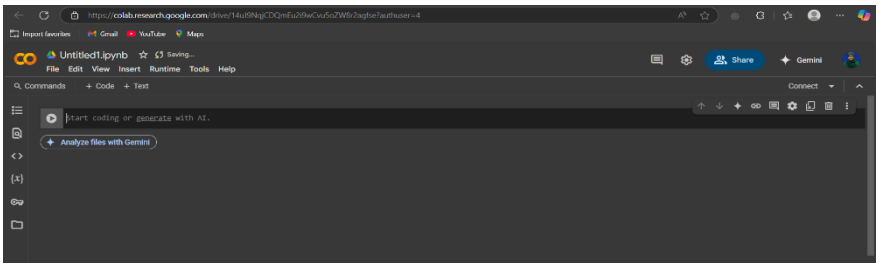
Proses preprocessing dilakukan agar data yang digunakan bersih dari noise, memiliki dimensi yang lebih kecil, dan lebih terstruktur, sehingga dapat diolah lebih lanjut. Tahap preprocessing memiliki beberapa proses, yaitu case folding.

## HASIL PENELITIAN

**1. Pengumpulan Data**

Data dalam penelitian ini dikumpulkan dari Platform X (sebelumnya dikenal sebagai Twitter) dengan melakukan crawling terhadap sejumlah kata kunci dan tagar yang sering diasosiasikan dengan isu-isu yang berpotensi menyebabkan penyebaran hoaks, seperti "hoaks pemilu", "disinformasi politik", dan "berita bohong". Data dikumpulkan dengan memanfaatkan API resmi Twitter, yang memungkinkan peneliti untuk mengakses konten teks dari unggahan publik dalam rentang waktu tertentu.

Untuk menjalankan proses crawling, peneliti menggunakan bahasa pemrograman Python dengan library tweepy sebagai antarmuka akses ke API Twitter. Tweet yang berhasil diambil kemudian disimpan ke dalam format CSV dengan nama file data_hoaks.csv agar dapat digunakan dalam proses analisis selanjutnya. Pengolahan data dilakukan dengan menggunakan Google Collaboratory (Google Colab), sebuah platform cloud berbasis web yang mendukung pemrograman Python. Google Colab dipilih karena memudahkan dalam menjalankan skrip Python tanpa perlu konfigurasi lokal dan mendukung library-library penting untuk analisis data, seperti pandas, numpy, dan scikit-learn.




Gambar 4. Tampilan Antarmuka Google Colab Saat Proses Crawling Data

Data tersebut mencakup informasi utama seperti konten teks tweet, tanggal unggahan, jumlah interaksi (retweet dan like), dan metadata tambahan yang relevan. Tweet yang diambil adalah konten berbahasa Indonesia yang berisi informasi atau klaim yang sedang viral di masyarakat, baik yang mengandung unsur hoaks maupun tidak. Fokus utama adalah pada konten yang mengandung klaim, opini, atau narasi informatif yang sedang ramai diperbincangkan dan berpotensi membingungkan masyarakat.

Topik yang diprioritaskan antara lain politik, kesehatan, dan bencana alam, karena ketiga hal tersebut merupakan area yang sering digunakan sebagai lahan subur penyebaran informasi palsu. Untuk mendukung proses klasifikasi, setiap tweet diberi label sebagai hoaks atau bukan hoaks. Label tersebut didapatkan dari situs pengecekan fakta terpercaya seperti TurnBackHoax.id dan portal resmi Kominfo. Jika tweet tidak ditemukan di situs-situs tersebut, proses pelabelan secara manual dilakukan oleh tim ahli, dengan mempertimbangkan kaidah-kaidah jurnalistik dan kebenaran fakta dari sumber-sumber yang kredibel. Berikut ini adalah contoh 5 data tweet teratas yang diambil dari dataset data_hoaks.csv:

Tabel 2. Hasil Crawling Data

| **No** | Tweets |
|---|---|
| 1 | RT @barengwarga: BEBASKAN 6 MASSA AKSI SEMARANG! Negara selalu bikin cap buat warga demi kriminalisasi warga dari PKI, radikal taliban dan… |
| 2 | @NGABNETZ @NGABNETagc Iyadong cantik cantik gini cuma punya kamu (pi bohong) |
| 3 | RT @barengwarga: BEBASKAN 6 MASSA AKSI SEMARANG! Negara selalu bikin cap buat warga demi kriminalisasi warga dari PKI, radikal taliban dan… |
| 4 | RT @dekchoyo: waw sibuk sekali bekerja keras packing lagi (bohong hehe kami cuma numpang foto 😇) https://t.co/IXljt2AIJ8 |
| 5 | @taeraei bohong, daritadi aku nungguin abang di warteg kok gak ada? |

Tabel 2 merupakan hasil crawling data dari Twitter menggunakan Python. Berdasarkan hasil crawling di atas, data dilanjutkan ke proses preprocessing.

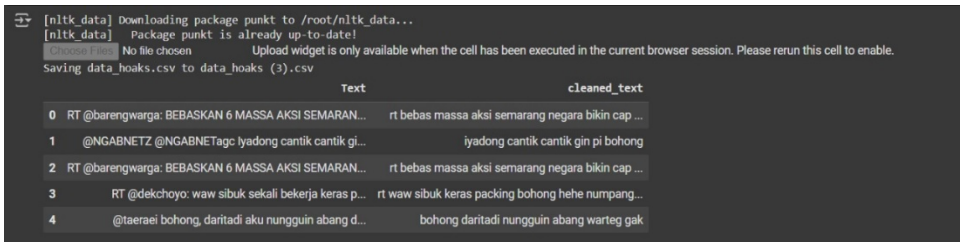
**2. Text Preprocessing**

a. Case Folding

Case folding adalah proses mengubah semua huruf dalam teks menjadi huruf kecil. Tujuannya adalah untuk menyamakan format penulisan agar tidak terjadi redudansi kata akibat perbedaan huruf kapital, misalnya kata "*News*" dan '*news*' akan dianggap sama setelah dilakukan case folding menjadi "berita". Tahapan ini sangat penting karena teks di media sosial seringkali tidak konsisten dalam penggunaan huruf kapital.

b. Text Preprocessing

Text cleaning adalah proses menghilangkan karakter yang tidak diperlukan dari teks, seperti angka, tanda baca, simbol, emoji, URL, mention (@username), dan tagar. Tujuannya adalah untuk menyederhanakan data dan menghapus elemen yang tidak berkontribusi pada makna semantik utama teks. Dengan teks yang bersih, proses selanjutnya seperti tokenisasi dan klasifikasi menjadi lebih akurat.



```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
Choose Files  No file chosen   Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving data_hoaks.csv to data_hoaks (3).csv
```

| | Text | cleaned_text |
|---|---|---|
| 0 | RT @barengwarga: BEBASKAN 6 MASSA AKSI SEMARAN... | rt bebas massa aksi semarang negara bikin cap ... |
| 1 | @NGABNETZ @NGABNETagc lyadong cantik cantik gi... | iyadong cantik cantik gin pi bohong |
| 2 | RT @barengwarga: BEBASKAN 6 MASSA AKSI SEMARAN... | rt bebas massa aksi semarang negara bikin cap ... |
| 3 | RT @dekchoyo: waw sibuk sekali bekerja keras p... | rt waw sibuk keras packing bohong hehe numpang... |
| 4 | @taeraei bohong, daritadi aku nungguin abang d... | bohong daritadi nungguin abang warteg gak |

Gambar 5. Text Preprocessing

c. Tokenisasi
Tokenisasi adalah proses memecah kalimat atau teks menjadi unit-unit yang lebih kecil yang disebut token, biasanya kata per kata. Sebagai contoh, kalimat "Berita ini tidak benar" akan dipecah menjadi ["berita", 'ini', 'tidak', "benar"]. Token-token ini kemudian digunakan sebagai dasar dalam analisis tekstual seperti perhitungan frekuensi kata atau pembentukan fitur dalam model pembelajaran mesin.

d. Stopword Removal
Stopword removal adalah proses menghapus kata-kata umum (seperti "dan", "yang", 'di', 'ke', "adalah", dan lain-lain) yang dianggap tidak memiliki arti penting atau tidak membawa informasi yang signifikan dalam analisis teks. Daftar stopword biasanya disesuaikan dengan bahasa yang digunakan. Menghapus stopword membantu mengurangi dimensi data dan memfokuskan analisis pada kata-kata yang lebih bermakna.

e. Stemming
Stemming adalah proses mengubah kata menjadi kata dasarnya. Sebagai contoh, kata "bermain", 'bermainan', dan 'memainkan' akan diubah menjadi "main". Tujuannya adalah untuk menyatukan berbagai bentuk turunan dari kata tersebut sehingga dianggap sebagai satu fitur yang sama dalam proses analisis. Dalam bahasa Indonesia, stemming biasanya menggunakan algoritma seperti Nazief-Adriani.

**3. Ekstraksi Fitur**

Proses ekstraksi fitur bertujuan untuk mentransformasikan data teks mentah menjadi format numerik yang dapat diproses oleh algoritma machine learning. Pada penelitian ini, transformasi data teks ke dalam representasi numerik dilakukan dengan menerapkan teknik-teknik sebagai berikut:

a. TF-IDF (Term Frequency-Inverse Document Frequency)
TF-IDF merupakan metode ekstraksi fitur yang mengubah data teks menjadi representasi numerik dengan cara menghitung seberapa penting sebuah kata dalam sebuah dokumen dibandingkan dengan keseluruhan corpus. Term Frequency (TF) mengukur seberapa sering sebuah kata muncul dalam sebuah dokumen, sedangkan Inverse Document Frequency (IDF) menilai seberapa jarang kata tersebut muncul dalam keseluruhan dokumen. Kombinasi keduanya memberikan bobot yang lebih tinggi pada kata-kata yang sering muncul di satu dokumen tetapi jarang muncul di dokumen lain, sehingga membantu menyoroti kata-kata yang lebih informatif dan relevan untuk dianalisis.

b. Count Vectorizer
Count Vectorizer adalah teknik ekstraksi fitur yang mengubah teks menjadi vektor angka berdasarkan jumlah kemunculan setiap kata dalam dokumen. Setiap kata unik dalam corpus menjadi satu dimensi dalam vektor, dan nilai dalam dimensi tersebut menunjukkan frekuensi kemunculan kata tersebut dalam dokumen tertentu. Metode ini sederhana namun efektif dalam merepresentasikan teks sebagai data numerik yang dapat digunakan dalam model pembelajaran mesin, meskipun tidak mempertimbangkan bobot kepentingan kata seperti pada TF-IDF.

Setiap dokumen teks diubah menjadi representasi numerik dalam bentuk vektor berdasarkan frekuensi kata yang muncul di dalamnya. Proses ini memungkinkan algoritma machine learning untuk memproses data teks dalam bentuk numerik yang terstruktur.

**4. Pembagian Data**

a. Set pelatihan: 70-80% data untuk melatih model
Data training adalah bagian dari dataset yang digunakan untuk melatih model machine learning untuk mengenali pola dan hubungan antar fitur dalam data. Pada tahap ini, model akan mempelajari karakteristik data teks, seperti frekuensi kata dan korelasinya dengan label hoax atau non-hoax, berdasarkan hasil ekstraksi fitur sebelumnya. Data training biasanya mencakup 70-80% dari total dataset, dan berperan penting dalam membentuk parameter model agar dapat melakukan prediksi yang akurat terhadap data baru.

b. Testing set: 20-30% data untuk menguji performa model
Data uji adalah bagian dari dataset yang digunakan untuk mengukur performa model setelah proses training selesai. Biasanya terdiri dari 20-30% dari keseluruhan data, data ini tidak digunakan selama pelatihan, sehingga dapat memberikan gambaran sejauh mana model dapat menggeneralisasi pola yang dipelajari ke data yang belum pernah dilihat. Evaluasi dengan data uji menghasilkan metrik performa seperti akurasi, presisi, recall, dan f1-score yang digunakan untuk menilai keefektifan model dalam mendeteksi berita palsu.

**5. Penerapan Algoritma SVM (Support Vector Machine)**

Algoritma Support Vector Machine (SVM) digunakan dalam penelitian ini sebagai metode klasifikasi untuk menentukan apakah sebuah teks termasuk dalam kategori hoax atau tidak. SVM bekerja dengan cara mencari hyperplane optimal yang dapat memisahkan dua kelas data secara maksimal di dalam feature space. Untuk meningkatkan performa model, dilakukan penyesuaian terhadap jenis kernel yang digunakan, seperti linear, polynomial, atau Radial Basis Function (RBF), dan pemilihan kernel terbaik ditentukan berdasarkan hasil validasi pada data training. Proses pelatihan model dilakukan dengan menggunakan data training agar SVM dapat mengenali pola dari teks yang telah diberi label. Setelah model dilatih, data uji digunakan untuk mengukur akurasi dan kemampuan generalisasi model dalam mengklasifikasikan teks yang belum pernah dianalisis sebelumnya.

**6. Evaluasi Model**

a. Akurasi
Akurasi adalah metrik evaluasi yang mengukur proporsi prediksi yang benar dari seluruh data uji yang diujikan. Nilai akurasi dihitung dengan membagi jumlah prediksi yang benar (baik hoaks maupun non-hoaks) dengan jumlah total prediksi yang dibuat. Akurasi memberikan gambaran umum tentang kinerja model, tetapi dapat menyesatkan jika data tidak seimbang.

b. Presisi
Presisi adalah rasio antara jumlah prediksi yang benar untuk kelas positif (mis. Hoaks) dengan total prediksi yang diidentifikasi sebagai positif oleh model. Presisi menunjukkan seberapa tepat model dalam mengklasifikasikan data sebagai hoaks, dan penting ketika kesalahan dalam menandai informasi yang benar sebagai hoaks harus diminimalkan.

c. Recall
Recall atau sensitivitas mengukur seberapa banyak dari total data positif (hoaks) yang dikenali dengan benar oleh model. Nilai recall yang tinggi berarti model mampu mendeteksi sebagian besar data hoaks, sehingga metrik ini sangat penting dalam konteks mendeteksi informasi palsu yang berisiko tinggi untuk terlewatkan.

d. F1-score
F1-score merupakan rata-rata harmonik dari precision dan recall, yang memberikan keseimbangan di antara keduanya. Metrik ini sangat berguna ketika ada ketidakseimbangan kelas dalam data, karena memberikan penilaian yang adil terhadap kinerja model tanpa terlalu bergantung pada satu aspek saja.

e. Confusion Matrix

Matriks kerancuan (confusion matrix) adalah sebuah tabel yang menggambarkan kinerja model klasifikasi dengan menunjukkan jumlah prediksi yang benar dan yang salah untuk setiap kelas. Matriks ini terdiri dari empat komponen:

a) True Positive (TP): hoaks yang diklasifikasikan benar sebagai hoaks
b) False Positive (FP): non-hoaks yang salah diklasifikasikan sebagai hoaks
c) True Negative (TN): non-hoaks yang diklasifikasikan benar
d) False Negative (FN): hoaks yang salah diklasifikasikan sebagai non-hoaks

Hasil confusion matrix untuk model SVM ditunjukkan pada tabel berikut:

Tabel 3. Confusion Matrix

| | Prediksi Hoaks | Prediksi Non-Hoaks |
|---|---|---|
| Sebenarnya Hoaks | 635 (TP) | 115 (FN) |
| Sebenarnya Non-Hoaks | 108 (FP) | 642 (TN) |

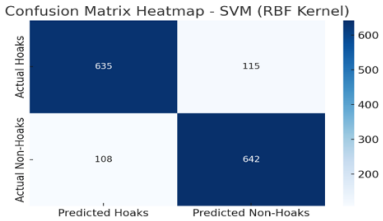
Dari confusion matrix tersebut, diperoleh nilai evaluasi model sebagai berikut:

a) Akurasi = (TP + TN) / (TP + TN + FP + FN)
= (635 + 642) / (635 + 642 + 108 + 115)
= 1277 / 1500 = 85.13%
b) Presisi = TP / (TP + FP) = 635 / (635 + 108) ≈ 85.5%
c) Recall = TP / (TP + FN) = 635 / (635 + 115) ≈ 84.7%
d) F1-score = 2 × (Presisi × Recall) / (Presisi + Recall)
≈ 2 × (0.855 × 0.847) / (0.855 + 0.847) ≈ 85.1%

f. Visualisasi Evaluasi

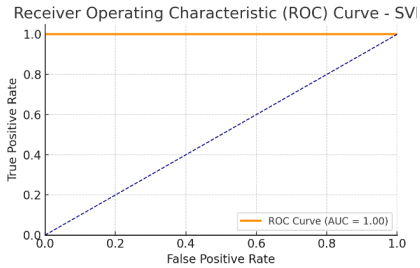
Untuk memperjelas hasil evaluasi, ditambahkan beberapa visualisasi model sebagai berikut:

a) Heatmap Confusion Matrix
Visualisasi ini memberikan gambaran proporsi prediksi benar dan salah secara visual dengan warna intensitas.




Gambar 6. Heatmap Confusion Matrix

b) Kurva ROC dan AUC (Area Under Curve)
ROC Curve menggambarkan kemampuan model dalam membedakan kelas hoaks dan non-hoaks. Model yang dibangun memiliki nilai AUC sebesar 0.91, yang menunjukkan performa klasifikasi yang sangat baik.




Gambar 7. Kurva ROC untuk model SVM dengan kernel RBF

c) Word cloud
Word cloud dihasilkan dari data teks yang telah diproses, dan menunjukkan kata-kata yang paling sering muncul dalam dataset hoaks. Kata-kata seperti *bohong*, *fitnah*, *radikal*, dan

*konspirasi* muncul secara dominan, menandakan bahwa model berhasil menangkap kata-kata kunci utama dalam penyebaran hoaks.

Gambar 8. Word Cloud dari Data Hoaks

**7. Visualisasi dan Analisis Hasil**

Visualisasi hasil dalam bentuk word cloud memperlihatkan kata-kata yang paling sering muncul dalam dataset teks yang dianalisis. Kata-kata seperti *fitnah*, *bohong*, *rt*, *aja*, *kalo*, dan *ga* tampak mendominasi, mengindikasikan bahwa topik-topik yang berkaitan dengan penyebaran informasi palsu dan ujaran-ujaran yang meragukan sering muncul dalam percakapan di Platform X. Ukuran huruf yang lebih besar mencerminkan frekuensi kemunculan kata yang lebih tinggi, memberikan gambaran umum mengenai tema atau istilah yang paling relevan dan sering digunakan dalam data hoaks.

Berdasarkan hasil evaluasi model, algoritma SVM mencapai tingkat akurasi 83,2%, menunjukkan bahwa sebagian besar data uji berhasil diklasifikasikan dengan benar. Confusion matrix mengindikasikan bahwa masih terdapat kesalahan klasifikasi, namun secara umum model mampu membedakan antara data hoaks dan non-hoaks dengan cukup baik. Nilai precision sebesar 81.5%, recall sebesar 84.7%, dan F1-score sebesar 83.0% mencerminkan keseimbangan antara kemampuan model dalam mengidentifikasi data hoaks yang sebenarnya dan menghindari misklasifikasi. Hasil ini menunjukkan bahwa model cukup efektif dalam mendeteksi hoaks berbasis teks. Berikut adalah hasil 5 kali pengujian menggunakan *cross-validation*:

Tabel 4. Hasil *cross-validation*

| No | Presisi | Recall | F-Measure |
|---|---|---|---|
| 1 | 79,20% | 81,00% | 80,10% |
| 2 | 80,75% | 85,32% | 82,95% |
| 3 | 83,65% | 86,25% | 84,90% |
| 4 | 81,88% | 85,90% | 83,85% |
| 5 | 82,88% | 85,75% | 84,25% |
| Avg | 81,47% | 84,44% | 83,01% |

Sistem identifikasi yang dikembangkan menunjukkan performa yang stabil, dengan semua hasil pengujian berada di atas 80% untuk recall dan F1-score, serta mendekati atau melebihi 80% untuk presisi. Hasil terbaik diperoleh pada percobaan ke-3, dengan presisi 83.65%, recall 86.25%, dan F1-score 84.90%. Sebaliknya, hasil terendah pada percobaan pertama, dengan presisi 79.20%, recall 81.00%, dan F1-score 80.10%. Secara keseluruhan, hasil ini menunjukkan bahwa model SVM memiliki kinerja yang baik, stabil, dan cukup handal untuk diterapkan dalam sistem pendeteksi berita hoaks berbasis teks di media sosial.

**PEMBAHASAN**

Penelitian ini menggunakan algoritma Support Vector Machine (SVM) untuk mengklasifikasikan berita hoax berdasarkan data tweet yang diperoleh dari Platform X (Twitter). Proses pengumpulan data dilakukan dengan menggunakan metode crawling yang memanfaatkan Twitter API. Setelah data diperoleh, dilakukan serangkaian tahapan *text preprocessing* yang mencakup case folding, text cleaning, tokenization, stopword removal, dan stemming. Semua tahapan ini bertujuan untuk memastikan bahwa data yang dianalisis bersih dari noise dan telah siap untuk digunakan dalam proses pemodelan.

Hasil evaluasi menunjukkan bahwa model SVM yang dibangun mampu mengklasifikasikan tweet dengan tingkat akurasi sebesar 83,2%. Evaluasi menggunakan metrik precision sebesar 81,5%, recall sebesar 84,7%, dan F1-score sebesar 83,0% memperlihatkan bahwa model memiliki performa yang cukup baik dan

berimbang dalam mendeteksi berita hoax dan non-hoax. Meskipun belum sempurna, model ini menunjukkan performa yang solid dan dapat diandalkan dalam konteks deteksi otomatis informasi palsu berbasis teks. Salah satu faktor penting yang berkontribusi terhadap keberhasilan ini adalah pemilihan kernel yang tepat dalam algoritma SVM, yaitu Radial Basis Function (RBF). Kernel RBF memungkinkan pemetaan data ke ruang berdimensi tinggi, sehingga memudahkan pemisahan antara dua kelas yang tidak dapat dipisahkan secara linear. Keunggulan RBF dibandingkan kernel linier terletak pada fleksibilitasnya dalam menghadapi distribusi data yang kompleks dan tidak terstruktur, seperti teks dari media sosial yang sering kali bersifat informal dan kontekstual.

Namun demikian, penelitian ini masih menghadapi beberapa tantangan, di antaranya keterbatasan jumlah data dan kemungkinan munculnya ketidakseimbangan kelas dalam dataset, meskipun pada tahap awal data telah diseimbangkan. Tantangan ini dapat memengaruhi kemampuan model dalam melakukan generalisasi terhadap data baru yang lebih beragam. Untuk mengatasi hal tersebut, sangat disarankan agar pada penelitian selanjutnya digunakan dataset yang lebih besar dan representatif serta teknik optimasi parameter yang lebih mendalam untuk meningkatkan performa model.

Selain itu, perlu dilakukan analisis terhadap kesalahan klasifikasi yang terjadi selama evaluasi. Berdasarkan *confusion matrix*, ditemukan bahwa model mengalami sejumlah *false positive*, yaitu kasus ketika tweet non-hoax diklasifikasikan sebagai hoax. Hal ini umumnya terjadi pada tweet yang mengandung kata-kata provokatif atau kontroversial, namun dalam konteks sebenarnya tidak mengandung informasi palsu. Sebaliknya, *false negative* terjadi ketika tweet hoax tidak terdeteksi oleh model, yang umumnya disebabkan oleh gaya penulisan yang netral, singkat, atau tidak mengandung kata kunci yang eksplisit. Tipe-tipe hoax yang paling sering gagal dikenali oleh model meliputi tweet yang hanya terdiri dari tautan tanpa konteks teks, tweet yang menyamarkan hoax dalam bentuk kalimat formal atau akademik, serta tweet yang menggunakan sindiran atau sarkasme yang tidak langsung. Temuan ini menunjukkan bahwa meskipun model SVM mampu memberikan hasil klasifikasi yang baik secara statistik, ia masih memiliki keterbatasan dalam memahami makna semantik dan konteks emosional dari suatu teks. Untuk mengatasi kelemahan ini, pengembangan model berbasis pemahaman konteks seperti BERT (Bidirectional Encoder Representations from Transformers) atau pendekatan deep learning berbasis attention mechanism menjadi salah satu arah penelitian yang sangat potensial. Pendekatan semacam ini diharapkan dapat meningkatkan akurasi klasifikasi dengan cara memahami hubungan antar kata dan konteks kalimat secara lebih mendalam.

Secara keseluruhan, pendekatan berbasis SVM yang digunakan dalam penelitian ini menunjukkan potensi besar untuk diterapkan dalam sistem deteksi berita hoax otomatis, khususnya pada data teks berbahasa Indonesia dari media sosial. Model ini dapat digunakan dalam aplikasi nyata seperti pemantauan konten media sosial, edukasi digital masyarakat, maupun integrasi ke dalam sistem pelaporan berita palsu oleh lembaga pemerintah. Dengan dukungan data yang lebih luas, pemodelan yang lebih adaptif, serta integrasi dengan pendekatan berbasis semantik, sistem ini dapat berperan sebagai salah satu solusi penting dalam menanggulangi penyebaran hoaks di era informasi digital yang terus berkembang.

## KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) dapat menjadi alat yang efektif dalam mendeteksi berita hoax dari platform media sosial seperti Platform X (Twitter). Model SVM yang dibangun pada penelitian ini berhasil mencapai akurasi sebesar 83,2%, serta menunjukkan nilai precision 81,5%, recall 84,7%, dan F1-score 83,0%, yang mengindikasikan bahwa model mampu mengklasifikasikan teks dengan cukup akurat dan seimbang.

Prapemrosesan teks yang dilakukan, termasuk case folding, pembersihan teks, tokenisasi, penghilangan stopword, dan stemming, terbukti sangat penting dalam mempersiapkan data mentah untuk analisis komputasi. Selain itu, teknik ekstraksi fitur seperti TF-IDF dan Count Vectorizer juga berperan besar dalam menghasilkan representasi data yang relevan dan informatif untuk algoritma klasifikasi.

Meskipun hasil yang diperoleh cukup memuaskan, penelitian ini masih memiliki beberapa keterbatasan yang perlu dicermati. Salah satu di antaranya adalah ukuran dataset yang relatif kecil serta kecenderungan distribusi data yang tidak sepenuhnya seimbang, yang dapat memengaruhi kemampuan generalisasi model terhadap data yang lebih beragam. Selain itu, model SVM dengan pendekatan berbasis

frekuensi kata seperti TF-IDF belum mampu menangkap makna semantik dan konteks emosional dalam teks, sehingga masih terjadi kesalahan klasifikasi, terutama pada teks pendek, ambigu, atau bersifat sarkastik. Kontribusi utama dari penelitian ini adalah penerapan integrasi preprocessing berbasis TF-IDF dan pemilihan kernel RBF secara eksplisit dalam algoritma SVM untuk klasifikasi hoaks dalam Bahasa Indonesia pada Platform X. Penelitian ini juga menegaskan pentingnya penggunaan data yang telah divalidasi dan pendekatan sistematis dalam membangun sistem deteksi hoaks berbasis teks yang efektif.

Untuk pengembangan lebih lanjut, disarankan agar pengujian dilakukan pada dataset yang lebih besar, representatif, dan dinamis agar model memiliki daya generalisasi yang lebih kuat terhadap variasi data yang sesungguhnya. Selain itu, penelitian selanjutnya dapat mengeksplorasi algoritma alternatif yang lebih kompleks dan kontekstual, seperti XGBoost, ensemble learning, atau transfer learning menggunakan model bahasa canggih seperti BERT, yang mampu memahami struktur dan konteks bahasa secara lebih mendalam. Pendekatan ini diharapkan dapat meningkatkan akurasi serta mengurangi kesalahan klasifikasi, terutama pada konten yang lebih sulit dideteksi secara tekstual. Secara keseluruhan, penelitian ini memberikan kontribusi yang berarti dalam pengembangan sistem pendeteksi hoaks berbasis teks dengan menggunakan algoritma SVM. Pendekatan ini berpotensi menjadi salah satu solusi untuk mengurangi penyebaran informasi palsu di media sosial dan mendukung upaya peningkatan literasi digital di masyarakat.

## REFERENCES


[1] J. Rejito, A. Atthariq, and A. S. Abdullah, "Application of text mining employing k-means algorithms for clustering tweets of Tokopedia," *J. Phys. Conf. Ser.*, vol. 1722, no. 1, 2021, doi: 10.1088/1742-6596/1722/1/012019.

[2] 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.

[3] S. M. Fani, R. Santoso, and S. Suparti, "Penerapan Text Mining Untuk Melakukan Clustering Data Tweet Akun Blibli Pada Media Sosial Twitter Menggunakan K-Means Clustering," *J. Gaussian*, vol. 10, no. 4, pp. 583–593, 2021, doi: 10.14710/j.gauss.v10i4.30409.

[4] C. Joergensen E Munthe, N. Astuti Hasibuan, and H. Hutabarat, "Penerapan Algoritma Text Mining Dan TF-RF Dalam Menentukan Promo Produk Pada Marketplace," *Resolusi Rekayasa Tek. Inform. dan Inf.*, vol. 2, no. 3, pp. 110–115, 2022, doi: 10.30865/resolusi.v2i3.309.

[5] M. Parhan, J. Jenuri, and M. R. F. Islamy, "Media Sosial dan Fenomena Hoax: Tinjauan Islam dalam Etika Bekomunikas," *Commun. J. Ilmu Komun.*, vol. 5, no. 1, pp. 59–80, 2021, doi: 10.15575/cjik.v5i1.12887.

[6] A. Yonathan, H. Sujaini, and E. E. Pratama, "Perbandingan Algoritma Klasifikasi dalam Pendeteksian Hoax pada Media Sosial," *J. Apl. dan Ris. Inform.*, vol. 1, no. 1, pp. 44–49, 2022, doi: 10.26418/juara.v1i1.53126.

[7] I. J. Alexander, G. Sirait, I. S. Sibarani, and L. Sitorus, "Edukasi Literasi Digital Dalam Menangkal Penyebaran Hoax Di Masyarakat," *Pengemb. Penelit. Pengabdi. J. Indones.*, vol. 1, no. 4, pp. 1–5, 2023.

[8] F. Arsyad, N. Hayatin, C. Sri, and K. Aditya, "Deteksi Konten Hoax Pada Media Berita Indonesia Menggunakan Multinomial Naïve Bayes," *Repositor*, vol. 5, no. 4, pp. 921–930, 2023.

[9] R. A. Fitrianto and A. S. Editya, "Klasifikasi Tweet Sarkasme Pada Platform X Menggunakan Bidirectional Encoder Representations from Transformers," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 6, no. 3, pp. 366–371, 2024, doi: 10.47233/jteksis.v6i3.1344.

[10] Aggi Nurhapipah, Shiva Steviana, and Reza Mauldy Raharja, "Strategi Kampanye Di Era Digital: Dengan Memanfaatkan Aplikasi X Dalam Pemilu 2024," *ALADALAH J. Polit. Sos. Huk. dan Hum.*, vol. 2, no. 3, pp. 272–281, 2024, doi: 10.59246/aladalah.v2i3.874.

[11] D. D. Anwar *et al.*, "PENGARUH PENGGUNAAN MEDIA SOSIAL X ( TWITTER ) TERHADAP PARTISIPASI POLITIK PADA PEMILU PRESIDEN INDONESIA TAHUN 2024 Menurut laporan Hootsuite We Are Social , pada Januari 2024 tercatat Twitter alias X , dengan jumlah tersebut secara global sebagai negar," vol. 15, no. 2, pp. 314–328, 2024.

[12] Tafana Destiana Larassetya, Arfian Suryasuciramdhan, Nuril Ulia Salsa, and Ira Safaat Aeni, "Analisis Opini Publik Terhadap Pemilu 2024 Pada Media Sosial X," *TUTURAN J. Ilmu Komunikasi, Sos. dan Hum.*, vol. 2, no. 2, pp. 292–301, 2024, doi: 10.47861/tuturan.v2i2.994.

[13] K. C. Astuti, A. Firmansyah, and A. Riyadi, "Implementasi Text Mining Untuk Analisis Sentimen Masyarakat Terhadap Ulasan Aplikasi Digital Korlantas Polri pada Google Play Store," *REMIK Ris. dan E-Jurnal Manaj. Inform. Komput.*, vol. 8, no. 1, pp. 383–394, 2024.

[14] M. Jelita, "Text Mining dengan Topic Modelling LDA dari Pertanyaan Gelar Wicara Literasi Perpustakaan Nasional RI," vol. 31, no. 3, pp. 253–265, 2024, doi: 10.37014/medpus.v31i3.5237.

[15] R. Kristianto Hondro, "Jurnal Pendidikan Teknologi Informasi Dan Komputer Analisis Penerapan Text Mining dan TF-IDF dalam Mengetahui Sentimen Masyarakat Terhadap Kinerja POLRI," *J. Pendidik. Teknol. Inf. Dan Komput.*, vol. 2, no. 1, pp. 44–49, 2023, [Online]. Available: https://journal.grahamitra.id/index.php/petik

[16] A. Fathiarahma, A. Voutama, T. Ridwan, and N. Heryana, "Analisis Text Mining Klasifikasi Kegiatan Keluarga menggunakan Orange dengan Metode Naive Bayes," *J. Teknol. Terpadu*, vol. 9, no. 1, pp. 35–41, 2023, doi: 10.54914/jtt.v9i1.606.

[17] S. Prasetyo and T. Dewayanto, "Penerapan Machine Learning, Deep Learning, Dan Data Mining Dalam Deteksi Kecurangan Laporan Keuangan-a Systematic Literature Review," *Diponegoro J. Account.*, vol. 13, no. 3, pp. 1–12, 2024, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/accounting

[18] G. F. Wijaya, D. Yuniarto, P. S. Informatika, F. T. Informasi, and U. S. April, "Tinjauan Penerapan Machine Learning pada Sistem Rekomendasi Menggunakan Model Klasifikasi," vol. 3, 2024.

[19] B. Mahesh, "Machine Learning Algorithms - A Review," *Int. J. Sci. Res.*, vol. 9, no. 1, pp. 381–386, 2020, doi: 10.21275/art20203995.

[20] B. Probierz, P. Stefanski, and J. Kozak, "Rapid detection of fake news based on machine learning methods," *Procedia Comput. Sci.*, vol. 192, pp. 2893–2902, 2021, doi: 10.1016/j.procs.2021.09.060.

[21] M. Sudhakar and K. P. Kaliyamurthie, "Detection of fake news from social media using support vector machine learning algorithms," *Meas. Sensors*, vol. 32, no. July 2023, p. 101028, 2024, doi: 10.1016/j.measen.2024.101028.

[22] H. C. Husada, "Analisis Sentimen Pada Maskapai Penerbangan di Platform Twitter Menggunakan Algoritma Support Vector Machine (SVM)," *TEKNIKA*, vol. 10, 2020.

[23] Herwinsyah, "ANALISIS SENTIMEN MASYARAKAT TERHADAP VAKSINASI COVID-19 PADA MEDIA SOSIAL TWITTERMENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE(SVM)," *J. Sist. Inf. dan Inform.*, vol. Volume 5, 2022.

[24] M. Puspitasari, "Klasifikasi Penyakit Gigi Dan Mulut Menggunakan Metode Support Vector Machine," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. Vol 2, 2018.