Terbit online pada laman web jurnal: **https://rjti.upp.ac.id/index.php/rjti**

RJTI
Riau Jurnal Teknik Informatika

Vol. 4 No. 1 (2025) 52 – 64 | ISSN: 2829-2529 (Media Online)

Perbandingan Algoritma Naive Bayes Dan Support Vector Machine Dalam Deteksi Berita Hoax Berbahasa Indonesia

Keren Aprilia Mumbunan[1], Michyta Marchantia Betsi Bawata[2], Miracle Prayer kusen[3], Victor Tarigan[4], Ade Yusupa[5]
[1,2,3,4,5]Fakultas Teknik, Prodi Teknik Informatika, Universitas Sam Ratulangi Manado
[1]kerenmumbunan026@student.unsrat.ac.id, [2]michytabawata026@student.unsrat.ac.id, [3]miraclekusen026@student.unsrat.ac.id, [4]victortarigan@unsrat.ac.id, [5]ade@unsrat.ac.id

Abstract

Fake news has become a major challenge in the digital age, particularly in Indonesia, where unverified information spreads rapidly through social media platforms. This study evaluates the effectiveness of Naïve Bayes (NB) and Support Vector Machine (SVM) algorithms in detecting fake news written in Indonesian. The dataset consists of 4,599 news articles collected from Twitter and GitHub repositories, categorized as either hoax or legitimate news. Several preprocessing techniques were applied to enhance classification accuracy, including tokenization, stopword removal, stemming, and TF-IDF vectorization. The models were assessed using accuracy, precision, recall, and F1-score metrics. Experimental results indicate that SVM outperforms Naïve Bayes, achieving an accuracy of 70.87%, while Naïve Bayes attains 66.52%. SVM also demonstrates higher precision (72%) and an F1-score of 82%, whereas Naïve Bayes has a higher recall of 99% but a lower F1-score (80%). These findings suggest that SVM is more effective for Indonesian fake news classification, while Naïve Bayes is preferable in scenarios requiring faster training. Future research should explore deep learning-based approaches such as BERT or LSTM, expand the dataset with diverse sources, and develop hybrid models integrating Naïve Bayes and SVM to optimize classification performance.

Keywords: Fake News Detection, Machine Learning, Naïve Bayes, Support Vector Machine, Text Classification

Abstrak

Berita hoaks menjadi tantangan besar di era digital, terutama di Indonesia, di mana informasi yang belum terverifikasi dapat menyebar dengan cepat melalui media sosial. Penelitian ini mengevaluasi efektivitas algoritma Naïve Bayes (NB) dan Support Vector Machine (SVM) dalam mendeteksi berita hoax berbahasa Indonesia. Dataset yang digunakan terdiri dari 4.599 artikel berita dari Twitter dan repositori GitHub, yang dikategorikan sebagai hoax atau valid. Berbagai teknik pra pemrosesan teks diterapkan untuk meningkatkan akurasi, termasuk tokenisasi, penghapusan stopword, stemming, dan vektorisasi TF-IDF. Model dievaluasi dengan metrik akurasi, presisi, recall, dan F1-score. Hasil eksperimen menunjukkan bahwa SVM lebih unggul dengan akurasi 70,87%, sementara Naïve Bayes memperoleh 66,52%. SVM juga memiliki presisi lebih tinggi (72%) dan F1-score (82%), sedangkan Naïve Bayes memiliki recall lebih tinggi (99%) tetapi F1-score lebih rendah (80%). Hasil ini menunjukkan bahwa SVM lebih efektif dalam klasifikasi berita hoaks, sementara Naïve Bayes lebih sesuai untuk kondisi yang membutuhkan waktu pelatihan yang cepat. Penelitian mendatang disarankan untuk mengeksplorasi metode deep learning seperti BERT atau LSTM, memperluas dataset dengan sumber yang lebih beragam, serta mengembangkan model hybrid yang menggabungkan Naïve Bayes dan SVM guna meningkatkan performa klasifikasi.

Kata kunci: *Deteksi Berita Hoax, Machine Learning, Naïve Bayes, Support Vector Machine, Klasifikasi Teks*

1. Pendahuluan

Dalam era digital yang berkembang pesat, informasi dapat dengan mudah tersebar melalui berbagai platform media sosial seperti WhatsApp, Facebook, Twitter, dan Instagram. Namun, kemudahan ini juga membawa tantangan besar, salah satunya adalah penyebaran berita hoaks yang semakin sulit dikendalikan. Sebagai contoh, berita palsu mengenai kesehatan, seperti klaim keliru tentang obat COVID-19, pernah beredar luas di Indonesia dan menyebabkan kepanikan di masyarakat. Selain itu, hoaks politik, seperti informasi menyesatkan

terkait pemilu, juga sering mempengaruhi opini publik secara signifikan. Berita hoaks semacam ini dapat membentuk persepsi masyarakat, mempengaruhi kebijakan, menciptakan keresahan sosial, hingga mengancam stabilitas politik dan ekonomi. Penyebaran hoaks semakin diperburuk oleh algoritma media sosial yang cenderung memprioritaskan interaksi pengguna daripada validitas informasi, sehingga berita yang menarik perhatian, termasuk hoaks, lebih sering muncul di linimasa pengguna.

Fenomena ini mendorong berbagai pihak, termasuk pemerintah, organisasi media, dan akademisi, untuk mengembangkan metode deteksi berita hoax yang lebih cepat, akurat, dan otomatis. Metode konvensional seperti verifikasi manual oleh fact-checker seringkali kurang efektif karena membutuhkan waktu dan tenaga yang besar. Oleh karena itu, pendekatan berbasis Artificial Intelligence (AI), khususnya Machine Learning (ML) dan Natural Language Processing (NLP), menjadi solusi yang menjanjikan dalam menangani penyebaran berita hoaks.

Berbagai penelitian telah mengeksplorasi penggunaan algoritma Machine Learning dalam mendeteksi berita hoaks. Naive Bayes (NB) dan Support Vector Machine (SVM) merupakan dua algoritma yang sering digunakan dalam klasifikasi teks, termasuk dalam deteksi berita palsu. Naive Bayes berbasis probabilistik dan mampu menangani data dalam skala besar dengan efisiensi tinggi, sementara SVM memiliki kemampuan generalisasi yang lebih baik dalam memisahkan kategori berita berdasarkan karakteristik teksnya. Beberapa studi sebelumnya menunjukkan bahwa SVM sering kali memiliki akurasi lebih tinggi dibandingkan Naive Bayes, terutama dalam analisis teks dengan dimensi fitur yang kompleks. Namun, masih terdapat tantangan dalam optimalisasi performa kedua algoritma ini, terutama dalam bahasa Indonesia, yang memiliki struktur dan kosakata yang berbeda dibandingkan bahasa Inggris.

Penelitian ini bertujuan untuk menganalisis dan membandingkan performa Naive Bayes dan SVM dalam mendeteksi berita hoax berbahasa Indonesia. Dataset yang digunakan terdiri dari 4.599 berita yang dikategorikan sebagai hoax atau valid. Berbagai teknik pra pemrosesan diterapkan, seperti tokenisasi, penghapusan stopword, stemming, dan vektorisasi TF-IDF, guna meningkatkan kualitas data sebelum diklasifikasikan. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score untuk mengukur efektivitas kedua algoritma dalam mengidentifikasi berita hoax.

Dengan penelitian ini, diharapkan dapat diperoleh pemahaman yang lebih mendalam mengenai efektifitas metode Machine Learning dalam deteksi berita hoax. Selain itu, hasil penelitian ini juga diharapkan dapat berkontribusi dalam pengembangan sistem deteksi otomatis yang lebih andal, akurat, dan efisien sehingga dapat digunakan dalam skala yang lebih luas, termasuk oleh platform media sosial dan lembaga pemeriksa fakta.

2. Metode Penelitian

Penelitian ini bertujuan untuk membandingkan performa algoritma Machine Learning dalam mendeteksi berita hoax berbahasa Indonesia. Metode yang digunakan mencakup pemilihan pendekatan eksperimen, pengumpulan dan pengolahan dataset, penerapan model klasifikasi, serta evaluasi kinerja berdasarkan berbagai metrik performa.

Gambar 1. Alur Penelitian

Dalam penelitian ini, metode eksperimen diterapkan untuk menganalisis efektivitas algoritma Naive Bayes dan Support Vector Machine (SVM) dalam mengklasifikasikan berita hoaks. Eksperimen ini memungkinkan evaluasi berbasis data yang telah dikategorikan sebelumnya. Proses penelitian terdiri dari beberapa tahap utama, yaitu pengumpulan data, preprocessing, pembangunan model, evaluasi, serta analisis hasil.

Dataset yang digunakan terdiri dari 4.617 berita yang dikategorikan sebagai hoax atau valid. Data diperoleh dari media sosial Twitter serta repositori GitHub yang menyediakan kumpulan berita untuk kepentingan penelitian. Dari total dataset, 3.042 sampel diklasifikasikan sebagai hoaks, 730 sebagai berita valid, dan 845 sampel dengan label tidak jelas (‘?’) yang kemudian di filter agar tidak mengganggu analisis.

DOI : https://doi.org/10.30606/rjti.v4i1.3262

Setiap berita dalam dataset memiliki atribut seperti kategori topik, kata kunci utama, teks berita, gambar (jika tersedia), sumber URL, serta label hoaks atau valid. Mengingat distribusi data yang tidak seimbang, analisis dilakukan dengan mempertimbangkan kemungkinan bias dalam pelatihan model.

Sebelum model dilatih, dataset melalui tahap preprocessing untuk meningkatkan kualitas data. Langkah-langkah yang dilakukan mencakup penghapusan elemen tidak relevan seperti URL dan mention, konversi teks menjadi huruf kecil guna menghindari perbedaan karena kapitalisasi, serta penghapusan tanda baca, angka, dan kata-kata umum (stopwords) yang tidak berkontribusi signifikan terhadap klasifikasi. Selain itu, dilakukan tokenisasi dan stemming untuk menyederhanakan kata menjadi bentuk dasarnya. Hasil preprocessing kemudian disimpan dalam dataset yang lebih bersih untuk tahap ekstraksi fitur dan pelatihan model.

Tabel 1. Tabel Distribusi Dataset

Label	Jumlah Data
Valid	730
Hoax	3.042
?	845
Total	4.617

Penelitian ini menggunakan dua algoritma utama, yaitu Naive Bayes dan Support Vector Machine (SVM), yang telah terbukti efektif dalam klasifikasi teks. Naive Bayes menggunakan pendekatan probabilistik berdasarkan Teorema Bayes dengan asumsi bahwa setiap fitur bersifat independen, sementara Multinomial Naive Bayes diterapkan dalam penelitian ini karena cocok untuk data berbasis teks dengan representasi TF-IDF. Sementara itu, model SVM dengan kernel linear digunakan dalam klasifikasi ini karena terbukti efisien dalam pemrosesan teks. Pengaturan hyperparameter, seperti nilai parameter C, diuji menggunakan teknik grid search untuk menemukan konfigurasi optimal dalam membedakan berita hoaks dan valid.

Dataset dibagi menjadi dua bagian utama, yaitu 80% untuk pelatihan dan 20% untuk pengujian. Pembagian ini dilakukan menggunakan metode stratifikasi untuk memastikan distribusi label tetap seimbang, sehingga model dapat belajar secara proporsional dari data yang tersedia.

Kinerja model dievaluasi menggunakan beberapa metrik utama, termasuk akurasi untuk mengukur persentase prediksi yang benar dari keseluruhan data uji, presisi untuk menilai ketepatan model dalam mengidentifikasi berita hoax, serta recall yang mengukur sejauh mana model berhasil mendeteksi berita hoax secara benar. Selain itu, digunakan F1-score untuk menyeimbangkan presisi dan recall, terutama ketika terdapat ketidakseimbangan data, serta AUC-ROC untuk menilai kemampuan model dalam membedakan antara berita hoax dan berita valid berdasarkan probabilitas prediksi.

Model dikembangkan menggunakan bahasa pemrograman Python, dengan pustaka Scikit-Learn, NLTK, Pandas, dan TF-IDF Vectorizer. Seluruh proses pelatihan dan evaluasi dilakukan menggunakan Google Collab, yang memungkinkan pemrosesan dataset dalam jumlah besar.

Setelah pelatihan model selesai, evaluasi dilakukan dengan membandingkan hasil prediksi model terhadap label asli dalam data uji. Perbandingan antara Naive Bayes dan SVM bertujuan untuk menentukan algoritma yang lebih unggul dalam mendeteksi berita hoax serta memahami karakteristik masing-masing model dalam menangani data berbasis teks.

Dengan menerapkan pendekatan eksperimen yang sistematis dan teknik evaluasi yang komprehensif, penelitian ini diharapkan dapat memberikan wawasan mengenai efektivitas Machine Learning dalam mendeteksi berita hoax berbahasa Indonesia serta menjadi dasar pengembangan sistem deteksi hoaks yang lebih andal.

3. Hasil dan Pembahasan

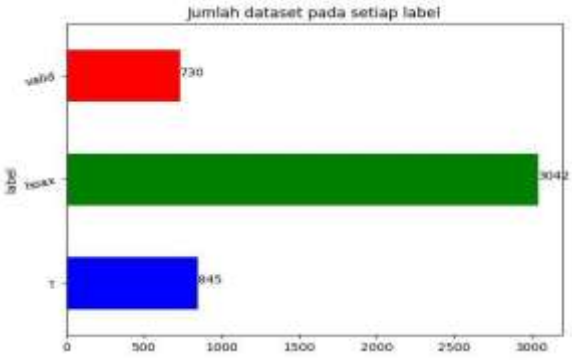
Setelah melakukan preprocessing data, termasuk pembersihan teks, tokenisasi, dan representasi fitur menggunakan TF-IDF, dataset dibagi menjadi 80% data latih dan 20% data uji. Model kemudian dilatih menggunakan Naive Bayes dan SVM dengan kernel linear.

Penelitian ini menggunakan dataset yang terdiri dari kumpulan berita berbahasa Indonesia yang telah diklasifikasikan ke dalam dua kategori utama, yaitu hoax dan non-hoax. Dataset ini diperoleh dari berbagai sumber berita daring, termasuk situs berita yang kredibel serta platform yang sering menyebarkan informasi yang kurang valid.

DOI : https://doi.org/10.30606/rjti.v4i1.3262

Dataset ini memiliki beberapa karakteristik utama. Jumlah berita yang dikategorikan sebagai valid sebanyak 730 sampel, sementara jumlah berita hoax mencapai 3.040 sampel. Selain itu, terdapat 845 berita yang tidak termasuk dalam kategori hoax maupun valid, yang juga menjadi bagian dari dataset ini. Secara rata-rata, setiap berita dalam dataset terdiri dari sekitar 250 kata.

Sumber utama dataset ini berasal dari kumpulan berita yang dikumpulkan dari berbagai situs berita daring serta dataset publik yang telah melalui proses kurasi. Format data yang digunakan dalam penelitian ini mencakup teks berita serta label yang menunjukkan apakah berita tersebut merupakan hoax atau non-hoax.



Gambar 3. Jumlah dataset pada setiap tabel

Sebelum digunakan untuk pelatihan model, dataset ini terlebih dahulu melewati proses preprocessing guna meningkatkan kualitas data dan memastikan model dapat mengolahnya dengan optimal. Proses ini mencakup beberapa tahap penting, mulai dari pembersihan teks hingga representasi dalam bentuk numerik.

Tahap pertama dalam preprocessing adalah pembersihan teks. Pada tahap ini, semua URL dan mention (@username) dihapus agar elemen yang tidak relevan tidak mengganggu analisis. Selain itu, tanda baca, angka, dan karakter spesial juga dihapus untuk memastikan teks lebih bersih dan mudah diolah. Seluruh teks kemudian dikonversi menjadi huruf kecil (case folding) guna menghindari perbedaan yang tidak diperlukan akibat kapitalisasi huruf.

Selanjutnya, dilakukan proses tokenisasi, yaitu pemisahan teks menjadi kata-kata individual agar dapat dianalisis dengan lebih terstruktur. Setelah itu, dilakukan penghapusan stopwords, di mana kata-kata umum yang tidak memiliki makna signifikan dalam klasifikasi, seperti "dan", "yang", dan "di", dihilangkan untuk mengurangi noise dalam data.

Tahap berikutnya adalah stemming, yaitu proses mengubah kata ke bentuk dasarnya menggunakan algoritma stemming dalam Bahasa Indonesia. Langkah ini bertujuan agar kata-kata yang memiliki makna serupa dapat diidentifikasi sebagai satu entitas yang sama, sehingga analisis lebih efektif.

Setelah teks dibersihkan, dilakukan proses ekstraksi fitur agar dapat digunakan dalam model klasifikasi. Untuk model Machine Learning seperti Naive Bayes dan SVM, teks dikonversi menjadi representasi numerik menggunakan metode TF-IDF (Term Frequency-Inverse Document Frequency). Sementara itu, untuk model berbasis Deep Learning seperti LSTM dan IndoBERT, teks direpresentasikan dalam bentuk vektor berdimensi tinggi menggunakan Word Embeddings, seperti FastText atau Word2Vec.

Setelah seluruh tahapan preprocessing selesai, dataset yang telah dibersihkan disimpan dalam file dataset_cleaned.csv, yang selanjutnya akan digunakan dalam proses pelatihan model klasifikasi.

Setelah melakukan preprocessing berupa lower case, remove punctuation, tokenization, normalization, stemming, dan stopwords removal, proses yang dilakukan selanjutnya yaitu TF-IDF (Term Frequency – Inverse Document Frequency). TF-IDF dapat digunakan untuk mengetahui frekuensi dari istilah tertentu yang relatif terhadap sebuah kata dalam kumpulan dokumen dan melihat seberapa umum atau tidak umum sebuah kata yang ada di antara corpus (sekumpulan teks yang terstruktur). Pada proses TF-IDF ini menggunakan urutan token berupa unigram dalam implementasinya sehingga jumlah token dari TF-IDF hanya satu kata saja.

Tabel 2. Tabel Distribusi Dataset

Term	TF	IDF
data	0.4423258684646914	1.0
dan	0.4423258684646914	1.0
menggunakan	0.2948839123097942	1.0

DOI : https://doi.org/10.30606/rjti.v4i1.3262

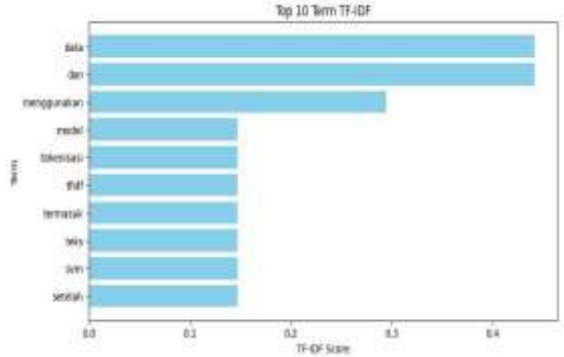
model	0.14744195615489714	1.0	dilatih	0.14744195615489714	1.0
tokenisasi	0.14744195615489714	1.0	dibagi	0.14744195615489714	1.0
tfidf	0.14744195615489714	1.0	dengan	0.14744195615489714	1.0
termasuk	0.14744195615489714	1.0	dataset	0.14744195615489714	1.0
teks	0.14744195615489714	1.0	uji	0.14744195615489714	1.0
svm	0.14744195615489714	1.0			
setelah	0.14744195615489714	1.0			
representasi	0.14744195615489714	1.0			
preprocessing	0.14744195615489714	1.0			
pembersihan	0.14744195615489714	1.0			
naive	0.14744195615489714	1.0			
bayes	0.14744195615489714	1.0			
menjadi	0.14744195615489714	1.0			
melakukan	0.14744195615489714	1.0			
linear	0.14744195615489714	1.0			
latih	0.14744195615489714	1.0			
kernel	0.14744195615489714	1.0			
kemudian	0.14744195615489714	1.0			
fitur	0.14744195615489714	1.0			

Tabel 3 menampilkan hasil TF-IDF pada bagian "Hasil dan Pembahasan" yang dapat dihitung dengan mengalikan nilai TF dan IDF secara value by value, kemudian disimpan ke dalam dataframe. Kemudian, hasil perhitungan TF-IDF ini diubah menjadi bentuk Sparse Matrix yang menampilkan semua term dengan nilai TF-IDF terbesar. Berdasarkan tabel di atas, kata "data" dan "dan" memiliki nilai TF tertinggi sebesar 0.442326, menunjukkan bahwa kata-kata ini sering muncul dalam teks yang dianalisis. Selain itu, kata "menggunakan" memiliki bobot TF yang cukup tinggi (0.294884), yang menunjukkan relevansi yang signifikan dalam dokumen. Sebagian besar kata lain memiliki bobot TF yang lebih kecil tetapi tetap memiliki IDF sebesar
1.0, yang berarti kata-kata tersebut muncul di semua dokumen dalam corpus. Kata-kata seperti "model", "tokenisasi", "tfidf", dan "teks" memiliki nilai TF yang lebih rendah tetapi tetap relevan dalam konteks analisis teks. Dengan demikian, analisis TF-IDF ini menunjukkan bahwa beberapa kata memiliki bobot yang lebih tinggi dalam dokumen yang dianalisis, yang dapat membantu dalam proses klasifikasi teks atau pemrosesan lebih lanjut dalam Machine Learning. Selanjutnya, hasil TF-IDF ini dapat divisualisasikan menggunakan library matplotlib dalam Python untuk menampilkan seluruh term berdasarkan nilai TF-IDF dalam bentuk grafik atau tabel.

Gambar 1 menunjukkan visualisasi dari hasil pembobotan TF-IDF dalam bentuk diagram batang. Grafik ini menampilkan 10 term dengan nilai TF-IDF tertinggi. Dari hasil grafik, kata "data" dan "dan" mendominasi dengan bobot tertinggi, diikuti oleh "menggunakan".
Sementara kata-kata lainnya memiliki nilai yang lebih rendah tetapi tetap berkontribusi dalam analisis dokumen. Visualisasi ini membantu dalam memahami distribusi kata dalam teks serta dapat digunakan untuk

DOI : https://doi.org/10.30606/rjti.v4i1.3262

fitur ekstraksi dalam model pembelajaran mesin atau analisis teks lebih lanjut.



Gambar 4. Top 10 Term TF-IDF

Dalam penelitian ini, dataset yang digunakan dibagi menjadi dua bagian utama, yaitu data pelatihan dan data pengujian. Data pelatihan digunakan untuk melatih model dengan tujuan mengenali pola yang dapat membedakan antara berita hoax, berita valid, dan berita dengan label tidak jelas ("?"). Model klasifikasi yang digunakan dalam penelitian ini adalah Multinomial Naïve Bayes dan Support Vector Machine (SVM). Sementara itu, data pengujian digunakan untuk mengevaluasi performa model yang telah dilatih. Dengan menggunakan data yang sebelumnya belum pernah diproses oleh model, evaluasi ini bertujuan untuk mengukur tingkat akurasi model dalam mengklasifikasikan berita.

Dataset yang digunakan dalam penelitian ini memiliki total 4.617 entri, yang terdiri dari tiga kategori utama. Kategori pertama adalah berita hoax, yang berjumlah 3.053 entri atau 66,12% dari keseluruhan dataset. Kategori kedua adalah berita valid, yang mencakup 720 entri atau sekitar 15,59% dari total data. Kategori terakhir adalah berita dengan label tidak jelas ("?"), yang terdiri dari 844 entri atau 18,29% dari keseluruhan dataset.

Pembagian dataset dilakukan dengan menggunakan rasio 80:20, di mana 80% dari data digunakan sebagai data pelatihan, sementara 20% sisanya digunakan sebagai data pengujian. Teknik stratified sampling diterapkan dalam proses ini untuk memastikan distribusi label dalam data pelatihan dan pengujian tetap seimbang serta representatif terhadap dataset asli.

Setelah proses pembagian dataset dilakukan, data pelatihan mencakup 3.693 entri. Dari jumlah tersebut, 2.440 entri merupakan berita hoax, yang mencakup 66,11% dari total data pelatihan. Selanjutnya, sebanyak 562 entri termasuk dalam kategori berita valid, yang setara dengan 15,60% dari total data. Selain itu, 691 entri memiliki label tidak jelas ("?"), yang mencakup 18,29% dari data pelatihan.

Sementara itu, data pengujian terdiri dari 924 entri. Dari jumlah tersebut, 613 entri merupakan berita hoax, yang mencakup 66,16% dari total data pengujian. Selanjutnya, sebanyak 158 entri merupakan berita valid, yang setara dengan 15,56% dari total data pengujian. Selain itu, 169 entri memiliki label tidak jelas ("?"), yang mencakup 18,28% dari data pengujian.

Untuk memberikan gambaran lebih jelas mengenai distribusi data, dilakukan visualisasi dalam bentuk tiga diagram pie. Diagram pertama menunjukkan distribusi label dalam dataset asli. Diagram kedua menggambarkan distribusi label dalam data pelatihan, sedangkan diagram ketiga menampilkan distribusi label dalam data pengujian. Dari visualisasi tersebut, dapat terlihat bahwa pembagian dataset tetap menjaga keseimbangan proporsi antar kelas. Dengan demikian, model dapat menerima data yang representatif selama proses pelatihan dan evaluasi, sehingga diharapkan mampu mengklasifikasikan berita dengan tingkat akurasi yang optimal.

Tabel 3. Tabel Data Latih

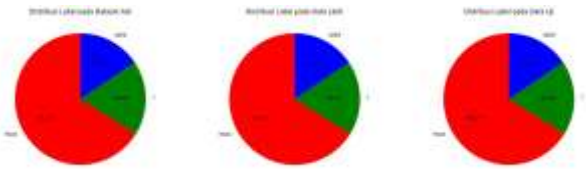
	Teks	Label
1011	petua minum air rebusan bawang putih mengobati virus covid wanita menepuk jidat	hoax
858	apa sih virus corona virus corona sejenis jamur wkwkwkwkwk virus ambek jamur ae wes bedo keluarga wajah menangis kencang wajah menangis kencang wajah menangis kencang	hoax
1396	cek kesehatan jantung paru paru segelas air mineral read more cek paru paru jantung agan	hoax
1210	baca jurnal dr universitas windsor canada ekstrak akar bunga dandelion bs jd obat kanker	hoax
1473	kram perut terinjak korban luka kulit melepuh akibat tembakan gas air mata bengawan memanggil bengawan melawan	hoax

DOI : https://doi.org/10.30606/rjti.v4i1.3262

1479	nagih hahaha kalo minum kopi pongo kopi cegah kejang anak	?
3237	sampo bayi formulanya lembut utk rambut tdk merusak warna rambut hair care tip	hoax
632	turki kalah erdogan cium tangan biden hahahha mampusss	hoax
3925	manfaat mandi pagi badan segar wajah fresh badan terasa bergairah mata ga ngantuk pagi semangat	valid
2071	burung rangkong gading statusnya ditetapkan terancam punah disebabkan maraknya perburuan liar	valid
1254	cek fakta asap batok kelapa obat covid	valid
4112	manfaat rebusan daun sirsak salah satunya mengobati kanker kelenjar getah bening	valid
2455	disinformasi nasib pesantren terancam uu ciptaker omnibus law follback nkri cc	hoax
4103	minggu gelar lapak semoga yg berminat beli minuman herbal dari rempah alami membantu imun tubuh masa pandemirsa enak harga bersahabat bnyak manfaat yang sayang orang tua untuk darah tinggi diabet asam urat tungkak lambung yang lagi insomania juga cocok lo jahe marah nya tinggal seduh wajah tersenyum mata tersenyum	valid

Tabel 2. Tabel Data Uji

	Teks	Label
2754	ah golongan darah o disukai nyamuk golongan darah o manis dibanding golongan darah	hoax
4335	melancarkan pencernaan bayam sayuran hijau kandungan serat berdampak kesehatan pencernaan rutin mengkonsumsi bayam	valid
619	musuh jantung air es heartattack allyxgray hblpslv kookv earthquake upunionconcertxsbfive imgforลน์ pslopeningceremony coronavirus coronaviruesue meanphiravich btscomeback valimaifirstlook fifa fifa	hoax
699	who confuted that tokek could heal hiv aids hoax	valid
2837	sukaberita wortel ubi bagus kesehatan mata kebiasaan bermain	hoax
860	scientist yg berkecimpung biochemistry biotechnology paham yg virus corona virus sejenis jamur mould	hoax



Gambar 5. Distribusi pada dataset asli, data latih dan data uji

Naive Bayes merupakan salah satu algoritma klasifikasi yang berbasis probabilitas dan didasarkan pada Teorema Bayes. Algoritma ini bekerja dengan mengasumsikan bahwa setiap fitur dalam suatu data bersifat independen terhadap fitur lainnya, yang dikenal sebagai naive assumption. Dalam penelitian ini, metode yang digunakan adalah Multinomial Naive Bayes karena algoritma ini sangat sesuai untuk tugas klasifikasi teks berdasarkan distribusi kata yang terdapat dalam dokumen.

Rumus Naive Bayes

Naive Bayes menghitung probabilitas suatu dokumen X termasuk dalam kategori C_{k} dengan rumus berikut:

$$P(C_{\mathrm{k}}/X) = \frac{P(X/C_{\mathrm{k}})P(C_{\mathrm{k}})}{P(X)}$$

Pada rumus ini, $P(C_{\mathrm{k}}/X)$ merepresentasikan probabilitas bahwa suatu dokumen X termasuk dalam kategori C_{k}. Probabilitas $P(C_{\mathrm{k}}/X)$ menunjukkan kemungkinan suatu dokumen X muncul dengan syarat bahwa dokumen tersebut termasuk dalam kategori

DOI : https://doi.org/10.30606/rjti.v4i1.3262

C_kSelanjutnya, $P(C_k)$ menggambarkan seberapa besar kemungkinan kategori C_k dalam dataset secara keseluruhan. Sementara itu, $P(X)$ merupakan probabilitas keseluruhan dari fitur yang muncul dalam dataset.

Dalam Multinomial Naïve Bayes, probabilitas suatu kata dalam dokumen dihitung menggunakan rumus berikut:

$$P(wi/C_k) = \frac{count(wi,C_k) + a}{\Sigma jcount(wj,C_k) + aV}$$

Pada rumus ini, $P(wi/C_k)$ menunjukkan probabilitas kemunculan kata wi dalam kategori C_k. Nilai count (wi, C_k) text merepresentasikan jumlah kemunculan kata wi dalam dokumen yang termasuk dalam kategori C_k. Selain itu, $\Sigma j\ count(wj, C_k$ merupakan total jumlah kata dalam kategori C_k. Untuk menghindari kemungkinan nilai probabilitas nol, parameter smoothing α digunakan dalam perhitungan ini. Terakhir, V menunjukkan jumlah total kata unik yang terdapat dalam seluruh dataset.

Dengan pendekatan ini, algoritma Naive Bayes dapat melakukan klasifikasi teks berdasarkan distribusi kata dalam dokumen secara efisien dan efektif.

Pada persamaan yang digunakan, $count(wi, C_k)$merepresentasikan jumlah kemunculan kata wi dalam kategori C_k. Simbol $/V/$ menunjukkan ukuran kosakata unik yang terdapat dalam dataset. Selain itu, parameter α digunakan sebagai nilai smoothing untuk menghindari probabilitas nol dalam perhitungan.

Dalam proses klasifikasi menggunakan Multinomial Naive Bayes, tahap pertama yang dilakukan adalah ekstraksi fitur. Pada tahap ini, data teks dikonversi menjadi representasi numerik menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). Perhitungan TF-IDF dilakukan dengan menggunakan rumus berikut:

$$TF(w) = \frac{fw}{\Sigma_k f_k}, \quad IDF(w) = log\frac{N}{dfw}$$

Pada persamaan tersebut, nilai fw menyatakan frekuensi kata dalam suatu dokumen, sedangkan $\Sigma_k f_k$ merepresentasikan jumlah total kata dalam dokumen tersebut. Variabel N menunjukkan jumlah total dokumen dalam dataset, sementara dfw adalah jumlah dokumen yang mengandung kata tertentu.

Setelah fitur TF-IDF berhasil diekstraksi, model Multinomial Naive Bayes kemudian dilatih dengan pendekatan probabilistik agar mampu mengklasifikasikan teks secara akurat.

Model ini menghitung probabilitas kemunculan setiap kata dalam setiap kategori untuk menentukan probabilitas bahwa sebuah dokumen termasuk dalam kategori tertentu.

Pada tahap prediksi, probabilitas sebuah dokumen masuk dalam suatu kategori dihitung menggunakan rumus:

$$P(C_k/X) = \prod_{i=1}^{n} P(xi/C_k)$$

Setelah itu, dokumen diklasifikasikan ke dalam kategori dengan probabilitas tertinggi.

Untuk mengevaluasi performa model, terdapat beberapa metrik yang digunakan. Metrik pertama adalah akurasi, yang mengukur sejauh mana model dapat mengklasifikasikan dokumen dengan benar. Akurasi dihitung dengan membandingkan jumlah prediksi benar dengan jumlah total prediksi yang dibuat oleh model, menggunakan rumus:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Metrik kedua adalah presisi, yang menunjukkan seberapa banyak prediksi positif yang benar dibandingkan dengan seluruh prediksi positif yang dibuat oleh model. Nilai presisi dihitung dengan membagi jumlah prediksi positif yang benar dengan jumlah total prediksi positif, menggunakan rumus:

$$Precision = \frac{TP}{TP + FP}$$

Metrik ketiga adalah recall, yang mengukur sejauh mana model dapat menemukan semua contoh positif yang sebenarnya ada dalam data. Recall dihitung dengan membagi jumlah prediksi positif yang benar dengan jumlah total data positif yang ada, menggunakan rumus:

$$Recall = \frac{TP}{TP + FN}$$

Metrik terakhir adalah F1-Score, yang merupakan rata-rata harmonik dari presisi dan recall. Metrik ini digunakan untuk menyeimbangkan keduanya dalam

DOI : https://doi.org/10.30606/rjti.v4i1.3262

satu nilai yang lebih representatif. Nilai F1-Score dihitung menggunakan rumus:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

dengan pendekatan ini, algoritma Naive Bayes dapat digunakan sebagai metode yang cepat dan efisien dalam mendeteksi berita hoax. Namun, performa model ini masih dapat ditingkatkan dengan teknik tambahan, seperti pemrosesan teks yang lebih lanjut atau kombinasi dengan model lain, seperti Support Vector Machine (SVM).

Gambar 6. Perbandingan Performa Naive Bayes pada setiap Label Dataset

Support Vector Machine (SVM) adalah algoritma pembelajaran mesin yang digunakan untuk mendeteksi berita hoax dengan mencari *hyperplane* optimal yang memisahkan kelas berita hoaks dan valid. SVM bekerja dengan cara menentukan batas keputusan (*decision boundary*) yang paling baik untuk memisahkan dua kelas data dalam ruang berdimensi tinggi.

Dalam SVM, terdapat beberapa jenis *kernel* yang digunakan untuk mengubah ruang fitur agar lebih optimal dalam proses pemisahan data. Jenis *kernel* pertama adalah Linear Kernel, yang merupakan *kernel* dasar yang digunakan ketika data dapat dipisahkan secara linear dalam ruang fitur. Linear Kernel didefinisikan dengan persamaan:

$$K\,(x, xi) = x\,.\,x_i^T$$

Jenis *kernel* kedua adalah Polynomial Kernel, yang mempertimbangkan interaksi antara fitur dengan menaikkan derajatnya ke pangkat tertentu. Dengan demikian, *Polynomial Kernel* memungkinkan pemetaan fitur ke dimensi yang lebih tinggi sesuai dengan kebutuhan pemisahan data. Persamaan yang digunakan dalam *Polynomial Kernel* adalah sebagai berikut:

$$K\,(x, xi) = (1 + x\,.\,x_i^T)^{\ d}$$

Jenis *kernel* ketiga adalah Sigmoid Kernel, yang bekerja seperti fungsi aktivasi dalam jaringan saraf tiruan (*neural network*) dan digunakan dalam pemetaan non-linear.

Jenis *kernel* keempat adalah Radial Basis Function (RBF) Kernel, yang bekerja dengan memetakan data ke dimensi yang lebih tinggi. *Kernel* ini berguna untuk data yang tidak dapat dipisahkan secara linear dan didefinisikan dengan persamaan:

$$K\,(x, xi) = exp(-\gamma\,//x - xi//^2)$$

Dalam penelitian ini, model SVM dilatih menggunakan *Linear Kernel*, karena dataset berita hoaks memiliki karakteristik yang lebih sesuai dengan pemisahan linear. Model SVM dengan *Linear Kernel* didefinisikan dengan persamaan berikut:

$$f(x) = w^T x + b$$

Dalam persamaan tersebut, $\boldsymbol{w}$ adalah vektor bobot, $\boldsymbol{x}$ adalah vektor fitur, dan $\boldsymbol{b}$ adalah bias.

Akurasi Model

Dalam penelitian ini, model yang digunakan untuk melakukan klasifikasi berita hoaks dan valid adalah Multinomial Naïve Bayes dan Support Vector Machine (SVM). Model Naïve Bayes bekerja berdasarkan probabilitas bersyarat dengan asumsi bahwa setiap fitur bersifat independen. Sementara itu, SVM menggunakan konsep *hyperplane* untuk memisahkan kelas secara optimal dalam ruang berdimensi tinggi.

Berdasarkan hasil pengujian, tingkat akurasi yang diperoleh untuk masing-masing model adalah sebagai berikut. Model Naïve Bayes memiliki akurasi sebesar 66,52%, sedangkan model SVM mencapai akurasi sebesar 70,87%. Dari hasil ini, dapat disimpulkan bahwa model SVM memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan model Naïve Bayes. Hal ini menunjukkan bahwa SVM lebih efektif dalam mengklasifikasikan berita hoaks dan valid dalam dataset yang digunakan.

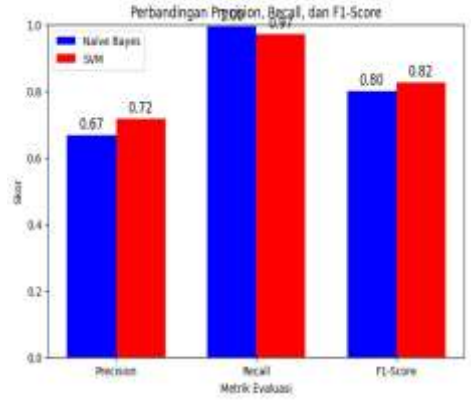
DOI : https://doi.org/10.30606/rjti.v4i1.3262

Gambar 7. Perbandingan Akurasi Model (Bold, 11pt)

Hasil pengujian ini dipengaruhi oleh beberapa faktor yang berkontribusi terhadap performa model. Salah satu faktor utama adalah kemampuan SVM dalam menangani data yang tidak terstruktur, yang lebih baik dibandingkan dengan Naïve Bayes. Hal ini disebabkan oleh penggunaan *hyperplane* yang memungkinkan pemisahan kelas data secara lebih optimal. Selain itu, asumsi Naïve Bayes yang menganggap setiap fitur bersifat independen dapat menjadi kelemahan dalam kasus data teks, karena dalam banyak situasi fitur-fitur tersebut saling berkorelasi, yang berpotensi menurunkan performa model.

Faktor lain yang mempengaruhi hasil prediksi adalah jumlah fitur yang digunakan dalam representasi TF-IDF. Penggunaan lebih banyak fitur dapat meningkatkan akurasi model, namun di sisi lain juga meningkatkan kompleksitas komputasi. Oleh karena itu, pemilihan jumlah fitur yang optimal menjadi aspek penting dalam meningkatkan kinerja model klasifikasi.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa SVM lebih unggul dalam mendeteksi berita hoax dibandingkan dengan model Naive Bayes dalam dataset yang digunakan. Analisis Metrik Evaluasi



Gambar 8. Perbandingan Precision, Recall, dan F1-Score

Dalam penelitian ini, model Support Vector Machine (SVM) dan Naïve Bayes dievaluasi menggunakan metrik precision, recall, dan F1-score untuk mengukur performa dalam mendeteksi berita hoax.

Pada aspek precision, model SVM memiliki nilai 72%, yang lebih tinggi dibandingkan dengan model Naïve Bayes yang hanya mencapai 67%. Nilai precision yang lebih tinggi menunjukkan bahwa SVM lebih baik dalam mengklasifikasikan berita hoaks dengan lebih sedikit menghasilkan false positives dibandingkan dengan Naïve Bayes.

Pada metrik recall, model Naïve Bayes memiliki nilai 99%, yang sedikit lebih tinggi dibandingkan dengan model SVM yang memiliki nilai 97%. Hal ini mengindikasikan bahwa Naïve Bayes lebih mampu mendeteksi hampir semua berita hoaks yang ada dalam dataset. Namun, keunggulan dalam recall ini juga dapat berakibat pada peningkatan jumlah false positives, di mana model lebih cenderung mengklasifikasikan berita valid sebagai hoaks.

Untuk metrik F1-score, model SVM memiliki nilai 82%, sedikit lebih tinggi dibandingkan dengan model Naïve Bayes yang memiliki nilai 80%. Metrik F1-score merupakan ukuran yang menggabungkan precision dan recall untuk memberikan gambaran keseimbangan antara keduanya. Dengan nilai F1-score yang lebih tinggi, model SVM menunjukkan stabilitas yang lebih baik dalam mendeteksi berita hoax dengan keseimbangan antara ketepatan (precision) dan kelengkapan deteksi (recall) yang lebih baik dibandingkan Naïve Bayes.

Secara keseluruhan, meskipun model Naïve Bayes unggul dalam aspek recall, model SVM menunjukkan performa yang lebih baik dalam precision dan F1-score. Oleh karena itu, model SVM lebih efektif dalam menangani deteksi berita hoax dengan tingkat akurasi yang lebih seimbang.

Dalam eksperimen ini, kami mengukur kinerja dua model, yaitu Naïve Bayes dan SVM, menggunakan metrik akurasi, precision, recall, dan F1-score. Untuk memperjelas perbandingan kinerja antara kedua model, hasil evaluasi disajikan dalam bentuk tabel komparatif berikut:

Tabel 4. Tabel Komparatif

Metrik	Naive Bayes	SVM
Akurasi	66%	70%
Precision	67%	72%

DOI : https://doi.org/10.30606/rjti.v4i1.3262

Recall	99%	97%
F1-Score	80%	82%

Tabel di atas menunjukkan bahwa meskipun kedua model memiliki kinerja yang cukup baik, SVM sedikit unggul dalam hal precision, recall, dan F1-score dibandingkan dengan Naïve Bayes. Metrik-metrik ini memberikan gambaran yang lebih komprehensif tentang kemampuan kedua model dalam mengklasifikasikan berita sebagai hoax atau valid.

Dari hasil evaluasi, dapat dilihat bahwa model Naïve Bayes memberikan akurasi yang cukup baik, namun sedikit lebih rendah dibandingkan dengan model SVM. SVM cenderung memberikan hasil yang lebih stabil, terutama pada metrik precision dan recall, yang mengindikasikan bahwa model ini lebih tepat dalam mengklasifikasikan berita hoaks dan lebih sensitif terhadap kelas hoaks. F1-score, yang menggabungkan precision dan recall, juga menunjukkan performa yang lebih baik pada SVM, menandakan bahwa model ini lebih seimbang dalam menangani kedua kelas.

Sementara itu, Naïve Bayes meskipun memiliki akurasi yang cukup tinggi, menunjukkan kinerja yang kurang optimal pada metrik recall, yang berarti model ini cenderung menghasilkan banyak false negatives (berita hoaks yang terklasifikasi sebagai non-hoaks). Hal ini mungkin disebabkan oleh asumsi independensi antar fitur yang digunakan oleh Naïve Bayes, yang tidak selalu berlaku pada data teks yang kompleks seperti berita hoaks.

Salah satu faktor yang mempengaruhi kinerja kedua model adalah masalah ketidakseimbangan dataset. Dataset berita hoaks cenderung memiliki jumlah data hoaks yang lebih sedikit dibandingkan dengan berita non-hoax. Hal ini dapat menyebabkan model lebih mudah untuk memprediksi kelas mayoritas (non-hoaks) dan mengabaikan kelas minoritas (hoaks). Hal ini tercermin dalam rendahnya nilai recall pada model Naïve Bayes.

Untuk mengatasi masalah ini, salah satu pendekatan yang dapat dilakukan adalah dengan melakukan teknik resampling seperti undersampling pada kelas mayoritas atau oversampling pada kelas minoritas. Selain itu, pemberian bobot yang lebih besar pada kelas minoritas dalam algoritma SVM dapat membantu memperbaiki kinerja model terhadap berita hoaks.

Batasan eksperimen ini harus diperhatikan agar hasil yang diperoleh dapat lebih dipahami dalam konteks yang lebih luas. Salah satu hal yang perlu diperhatikan adalah adanya berita hoaks dengan tema kesehatan yang mungkin lebih mendominasi dalam dataset ini, yang dapat mempengaruhi kinerja model. Data yang sangat spesifik ini mungkin mengarah pada overfitting, di mana model hanya mampu mengenali pola tertentu yang ada dalam dataset, tetapi tidak generalize dengan baik ke berita hoaks dengan tema lain yang mungkin muncul di dunia nyata.

Selain itu, jumlah data yang terbatas juga bisa menjadi faktor yang mempengaruhi performa model. Semakin banyak data yang digunakan, semakin baik model dapat belajar mengenali pola yang lebih beragam. Namun, keterbatasan sumber daya dalam mengumpulkan data atau ketidakseimbangan dalam jenis data yang tersedia dapat menjadi tantangan dalam penelitian ini.

Selain faktor dataset dan teknik model, ada faktor eksternal lain yang dapat mempengaruhi hasil eksperimen ini, seperti pemilihan fitur. Pemilihan fitur yang digunakan dalam preprocessing teks, seperti penggunaan TF-IDF atau Word2Vec, dapat sangat mempengaruhi kinerja model dalam mengklasifikasikan berita hoaks. Menggunakan representasi kata yang lebih canggih seperti Word Embeddings atau teknik deep learning seperti BERT dapat meningkatkan kualitas model dalam memahami konteks berita.

Hasil eksperimen menunjukkan bahwa model Support Vector Machine (SVM) memiliki performa yang lebih unggul dibandingkan dengan Naïve Bayes (NB) dalam mendeteksi berita hoax. Perbedaan ini disebabkan oleh beberapa faktor utama yang mempengaruhi efektivitas masing-masing model dalam mengolah data teks.

Salah satu keunggulan utama SVM adalah kemampuannya dalam menangani data dengan dimensi tinggi. Model ini bekerja dengan memetakan teks yang telah dikonversi menjadi vektor TF-IDF ke dalam ruang berdimensi tinggi, memungkinkan pemisahan kelas yang lebih optimal. Sebaliknya, Naïve Bayes mengandalkan asumsi independensi fitur, yang dalam konteks data teks tidak selalu akurat. Hal ini menyebabkan keterbatasan dalam menangkap hubungan antara kata-kata dalam dokumen, sehingga dapat berdampak pada akurasi klasifikasi yang lebih rendah dibandingkan SVM.

Berdasarkan hasil pengujian, SVM menunjukkan akurasi sebesar 70,87%, yang lebih tinggi dibandingkan dengan Naïve Bayes, yang hanya mencapai 66,52%. Perbedaan ini menunjukkan bahwa SVM lebih andal dalam memprediksi berita hoaks, dengan tingkat

DOI : https://doi.org/10.30606/rjti.v4i1.3262

kesalahan klasifikasi yang lebih rendah dibandingkan dengan Naïve Bayes.

Evaluasi berdasarkan metrik precision, recall, dan F1-score juga menunjukkan pola yang serupa. Model SVM memiliki precision sebesar 72%, yang berarti model ini lebih akurat dalam mengklasifikasikan berita hoaks dengan jumlah false positives yang lebih sedikit dibandingkan dengan Naïve Bayes, yang hanya memiliki precision sebesar 67%. Namun, dalam metrik recall, Naïve Bayes lebih unggul dengan skor 99%, dibandingkan dengan SVM, yang memiliki recall sebesar 97%. Meskipun Naïve Bayes lebih baik dalam mendeteksi berita hoax secara keseluruhan, model ini cenderung menghasilkan lebih banyak kesalahan dalam bentuk false positives, sehingga berdampak pada ketidakakuratan klasifikasi. Sementara itu, F1-score, yang merupakan keseimbangan antara precision dan recall, menunjukkan bahwa SVM memiliki nilai 82%, lebih tinggi dibandingkan dengan Naïve Bayes, yang hanya mencapai 80%. Dengan demikian, SVM lebih stabil dalam mendeteksi berita hoax karena memiliki keseimbangan yang lebih baik antara ketepatan dan kelengkapan klasifikasi.

Selain akurasi, efisiensi model dalam hal waktu pelatihan juga menjadi faktor penting. Berdasarkan hasil pengujian, Naïve Bayes memiliki waktu pelatihan yang lebih cepat dibandingkan dengan SVM. Hal ini disebabkan oleh metode perhitungan Naïve Bayes, yang hanya memerlukan operasi probabilitas sederhana, sedangkan SVM harus mengoptimalkan hyperplane dalam ruang berdimensi tinggi, yang lebih kompleks secara komputasi. Walaupun Naïve Bayes unggul dalam efisiensi waktu, SVM tetap lebih unggul dalam hal akurasi, sehingga lebih cocok digunakan dalam skenario yang mengutamakan performa dibandingkan kecepatan pemrosesan.

Penelitian ini secara khusus diterapkan pada berita hoaks yang berkaitan dengan isu kesehatan. Berita hoaks dalam bidang ini sering kali memiliki pola linguistik yang khas, seperti penggunaan istilah medis secara berlebihan atau klaim yang tidak didukung oleh bukti ilmiah. Oleh karena itu, penggunaan TF-IDF dalam ekstraksi fitur teks terbukti sangat membantu dalam mengidentifikasi kata-kata yang memiliki bobot informasi tinggi, sehingga dapat meningkatkan efektivitas klasifikasi berita hoaks.

Secara keseluruhan, meskipun Naïve Bayes unggul dalam aspek kecepatan pelatihan dan recall, model SVM tetap menjadi pilihan yang lebih baik dalam mendeteksi berita hoax karena memiliki akurasi yang lebih tinggi, precision yang lebih baik, serta keseimbangan performa yang lebih stabil.

4. Kesimpulan

Penelitian ini membandingkan performa algoritma Naïve Bayes dan Support Vector Machine (SVM) dalam mendeteksi berita hoax berbahasa Indonesia. Hasil evaluasi menunjukkan bahwa SVM memiliki akurasi lebih tinggi (70,87%) dibandingkan Naïve Bayes (66,52%), dengan keunggulan pada precision dan keseimbangan antara precision dan recall. Sementara itu, Naïve Bayes unggul dalam recall, yang menunjukkan sensitivitas lebih tinggi dalam mendeteksi berita hoax meskipun dengan precision yang lebih rendah.

Dari segi efisiensi, Naïve Bayes lebih cepat dalam waktu pelatihan, sedangkan SVM lebih optimal dalam pemisahan kelas pada data teks berdimensi tinggi. Oleh karena itu, SVM lebih direkomendasikan untuk aplikasi yang membutuhkan akurasi tinggi, sedangkan Naïve Bayes cocok untuk aplikasi dengan pemrosesan cepat dan toleransi terhadap akurasi lebih rendah.

Untuk penelitian selanjutnya, pendekatan deep learning seperti BERT atau LSTM disarankan. BERT dan LSTM menjanjikan karena dapat memahami konteks lebih baik daripada model berbasis TF-IDF, yang memungkinkan deteksi berita hoax yang lebih akurat. Penelitian selanjutnya juga bisa mengeksplorasi kombinasi Naïve Bayes dan SVM dalam model hybrid untuk mengoptimalkan keseimbangan antara kecepatan dan akurasi.

Daftar Rujukan

[1] Raza, M. N. (2024). Sistem Deteksi Berita Hoax Menggunakan Algoritma Naïve Bayes Dan Random Forest Pada Machine Learning. *Pondasi: Journal of Applied Science Engineering*, *1*(2), 43-57.

[2] Gulo, E. S., Gulo, Y. R., & Marbun, S. F. (2022). PERBANDINGAN EFEKTIFITAS ALGORITMA DECISSION TREE, NAÏVE BAYES, K-NEAREST NEIGHBOR DAN SUPPORT VECTOR MACHINE DALAM MELAKUKAN KLASIFIKASI. *JURNAL TEKNOLOGI DAN ILMU KOMPUTER PRIMA (JUTIKOMP)*, *5*(2), 54-59.

[3] Ripa'i, A., Santoso, F., & Lazim, F. (2024). Deteksi Berita Hoax dengan Perbandingan Website Menggunakan Pendekatan Deep Learning Algoritma BERT. *G-Tech: Jurnal Teknologi Terapan*, *8*(3), 1749-1758.

[4] Imran, B., Karim, M. N., & Ningsih, N. I. (2024). Klasifikasi Berita Hoax Terkait Pemilihan Umum Presiden Republik Indonesia Tahun 2024 Menggunakan Naïve Bayes Dan Svm. *Jurnal Ilmiah Dinamika Rekayasa*, *20*(1), 1-9.

DOI : https://doi.org/10.30606/rjti.v4i1.3262

[5] Pasaribu, V. R. (2021). Penerapan Algoritme Naïve Bayes Classifiers pada Sistem Pendeteksi Berita Hoax Berbahasa Indonesia. Universitas Sriwijaya.

[6] Febriyanty, N. E. (2023). *Deteksi berita Hoax dari media Online Indonesia menggunakan Algoritma Naive Bayes dan Support Vector Machine* (Doctoral dissertation, Universitas Islam Negeri Maulana Malik Ibrahim).

[7] Sani, R. R., Pratiwi, Y. A., Winarno, S., Udayanti, E. D., & Alzami, F. (2022). Analisis Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Berita Hoax pada Berita Online Indonesia. *Jurnal Masyarakat Informatika*, *13*(2), 85-98.

[8] Fadhilahsari, S., & Ajie, H. PERBANDINGAN ANALISIS EMOSIONAL PENGGUNA TWITTER PADA PEMINDAHAN IBU KOTA INDONESIA MENGGUNAKAN ALGORITMA NAÏVE BAYES DAN SUPPORT VECTOR MACHINE.

[9] Maulana, M. I., Martanto, M., & Hayati, U. (2024). Perbandingan Algoritma Naïve Bayes Dan K-Nearest Neighbors Untuk Klasifikasi Topik Berita Pada Situs Detik. Com. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(3), 3733-3742.

[10] Tuhenay, D. (2021). *Perbandingan Klasifikasi Bahasa Menggunakan Metode Naïve Bayes Classifier (NBC) Dan Support Vector Machine (SVM). JIKO (Jurnal Informatika dan Komputer), 4 (2), 105-111*.

[11] Pinjaman, K. P., Bagja, A., & Kusrini, M. Komparasi Algoritma Naïve Bayes Dan Support Vector Machine (SVM) Untuk Klasifikasi.

[12] Kalua, A. L., Yusupa, A., Tarigan, V., & Komansilan, R. (2024). Pemetaan Hutan Mangrove Di Sulawesi Utara. *Jurnal Infomedia: Teknik Informatika, Multimedia & Jaringan*, *9*(1), 22-29.

[13] Tarigan, V. T., & Yusupa, A. (2024). Perbandingan Algoritma Maching Learning dalam Analisis Sentimen Mobil Listrik di Indonesia pada Media Sosial Twitter/X. *Jurnal Informatika Polinema*, *10*(4), 479-490.

[14] Tarigan, V. (2023). Pembuatan aplikasi data mining untuk memperediksi masa studi mahasiswa menggunakan algoritma Naive Bayes. *Informatika*, *11*(1), 54-62.

[15] Tarigan, V. (2023). Seleksi Fitur Dengan Menggunakan Metode Entropy Pada Algoritma Klasifikasi Naive Bayes Untuk Penyakit Diabetes. *Jurnal Infomedia: Teknik Informatika, Multimedia, dan Jaringan*, *8*(2), 66-77.

DOI : https://doi.org/10.30606/rjti.v4i1.3262