



SISTEM DETEKSI BERITA HOAX MENGGUNAKAN ALGORITMA NAÏVE BAYES DAN RANDOM FOREST PADA MACHINE LEARNING

Muh Nailar Raza

**Program Studi Teknik Informatika, Fakultas Teknik
Universitas Muhammadiyah Ponorogo
nailarraza16@gmail.com**

ABSTRACT

In the digital era, the spread of fake news or disinformation through social media has become a serious challenge. This research designs and implements a fake news detection system using Naïve Bayes and Random Forest algorithms in machine learning. The aim of this research is to improve the accuracy of online news information recognition. This research limits detection to Indonesian-language news and uses data from specific news sources. The methods used include data collection, data cleaning, and algorithm testing. Accuracy evaluation is conducted on the detection results using the accuracy of the Naïve Bayes and Random Forest algorithms. The research results show the superiority of the Random Forest algorithm in improving the accuracy of fake news detection. The contribution of this research is expected to provide a foundation for the development of a better fake news detection system, with the potential to enhance the accuracy of online news information recognition.

Keywords: Fake News, Naïve Bayes Algorithm, Random Forest Algorithm, Machine Learning, Social Media.

ABSTRAK

Dalam era digital, penyebaran berita hoaks atau disinformasi melalui media sosial menjadi tantangan serius. penelitian ini merancang dan mengimplementasikan sistem deteksi berita hoaks menggunakan algoritma Naïve Bayes dan Random Forest pada machine learning. Tujuan penelitian ini adalah untuk meningkatkan keakuratan pengenalan informasi berita online. Penelitian ini membatasi deteksi terhadap berita berbahasa Indonesia dan menggunakan data dari sumber-sumber berita tertentu. Metode yang digunakan mencakup pengumpulan data, pembersihan data, dan pengujian algoritma. Evaluasi keakuratan dilakukan terhadap hasil deteksi menggunakan akurasi algoritma Naïve Bayes dan Random Forest. Hasil penelitian menunjukkan keunggulan algoritma Random Forest dalam meningkatkan akurasi deteksi berita hoaks. Kontribusi penelitian ini diharapkan dapat memberikan landasan untuk pengembangan sistem deteksi berita hoaks yang lebih baik, dengan potensi meningkatkan keakuratan pengenalan informasi berita online.

Kata Kunci: Berita Hoaks, Algoritma Naïve Bayes, Algoritma Random Forest, Machine Learning, Media Sosial.



PENDAHULUAN

Dalam era digital dan kemajuan teknologi informasi, media sosial telah menjadi bagian tak terpisahkan dari kehidupan sehari-hari masyarakat modern. Fenomena ini memberikan kemudahan bagi individu atau kelompok untuk berinteraksi, berkomunikasi, dan berbagi informasi dengan cepat dan efisien. Media sosial memungkinkan pengguna untuk mengunggah berbagai aktivitas, pemikiran, dan pandangan mereka secara real-time, menciptakan ekosistem informasi yang dinamis. Namun, bersamaan dengan manfaatnya, penggunaan media sosial juga membawa tantangan serius terkait penyebaran berita palsu atau hoax. Di era digital yang serba cepat ini, di mana setiap individu dapat menjadi sumber informasi, tantangan untuk membedakan antara fakta dan fiksi menjadi semakin kompleks. Kecepatan dan luasnya jangkauan media sosial membuat berita palsu dapat dengan mudah menyebar dan memengaruhi opini publik. Berita palsu, atau hoaks, tidak hanya menyebar melalui platform web dan media sosial dengan kecepatan yang mengkhawatirkan, tetapi juga memiliki potensi untuk memanipulasi opini publik, merusak reputasi, dan bahkan mempengaruhi hasil pemilihan umum. Hal ini menimbulkan ancaman terhadap keberlanjutan informasi yang akurat dan dapat dipercaya di tengah masyarakat. Oleh karena itu, penanganan masalah penyebaran berita palsu di media sosial menjadi semakin mendesak.

Penggunaan Naive Bayes Classifier dan Random Forest telah muncul sebagai dua pilar utama dalam perang melawan berita palsu. Kedua algoritma pembelajaran mesin ini telah menunjukkan kemampuan yang luar biasa dalam mengklasifikasikan teks dan menganalisis sentimen, dengan Naive Bayes yang mengandalkan asumsi independensi antar fitur, dan Random Forest yang memanfaatkan kekuatan metode ensemble dengan menggabungkan banyak pohon keputusan untuk meningkatkan akurasi prediksi. Penelitian terbaru telah mengungkapkan bahwa Naive Bayes Classifier dapat mencapai tingkat akurasi yang mengesankan dalam mendeteksi keaslian berita pada platform seperti Twitter, dengan angka yang mencapai hingga 92%. Di sisi lain, Random Forest telah dibandingkan dengan berbagai metode klasifikasi lain dan ditemukan memiliki tingkat prediksi yang lebih unggul, terutama dalam hal akurasi. Ini menunjukkan potensi besar dari kedua metode ini dalam memerangi penyebaran informasi yang menyesatkan.

Penelitian terkait media sosial, seperti yang dilakukan oleh Yunanto et al. (2021), telah menyoroti urgensi mendeteksi berita hoax dengan cepat. Studi ini menekankan bahwa integrasi konten pengguna media sosial dari berbagai wilayah dapat memberikan informasi penting secara efisien dan efektif. Namun, masalah deteksi berita palsu di media sosial tetap menjadi fokus perhatian utama, dan perlu adanya penelitian yang lebih lanjut untuk mengatasi kompleksitas dan dinamika penyebaran informasi di platform tersebut.

Penelitian terdahulu oleh Ramadhan et al. (2022), memberikan kontribusi penting dalam pemahaman terhadap berita palsu di media sosial. Meskipun demikian, perkembangan terbaru menunjukkan perlunya fokus pada pengembangan metode deteksi berita palsu yang lebih canggih dan responsif terhadap strategi baru yang digunakan oleh pelaku penyebaran berita palsu. Dalam konteks Indonesia, yang memiliki tingkat penetrasi internet yang tinggi, khususnya melalui media sosial, penanganan masalah berita palsu di platform tersebut menjadi semakin mendesak.

Dengan memahami latar belakang ini, penelitian ini mencoba memberikan kontribusi melalui pengembangan sistem deteksi berita hoax menggunakan algoritma machine learning. Dalam konteks teknologi saat ini yang terus berkembang, penelitian ini berusaha tidak hanya untuk meningkatkan akurasi deteksi berita palsu berbasis web, tetapi juga untuk mengoptimalkan penerapan Naive Bayes Classifier dan Random Forest. Dengan mengintegrasikan teknik-teknik terbaru dalam Natural Language Processing (NLP) dan pembelajaran mesin, penelitian ini bertujuan untuk mengembangkan model yang lebih tangguh dan adaptif, yang mampu mengatasi berbagai jenis data dan format berita palsu yang semakin canggih.

Oleh karena itu, penelitian ini diharapkan tidak hanya memberikan kontribusi yang signifikan dalam memerangi penyebaran berita palsu, tetapi juga dalam mendukung pembentukan ekosistem informasi yang lebih sehat dan terpercaya. Dengan demikian, kita dapat memastikan bahwa informasi yang tersedia untuk publik adalah akurat dan dapat dipercaya, sehingga memungkinkan pengambilan keputusan yang lebih tepat dan informasi yang lebih transparan bagi semua pihak.



KAJIAN TEORI

A. Penelitian Terdahulu

Penelitian pertama deteksi Berita Palsu dengan Random Forest. Dalam penelitian ini, fokus utama adalah mencegah penyebaran berita palsu yang dapat menyebabkan informasi yang salah diterima oleh masyarakat. Metode yang digunakan adalah supervised learning dengan menggunakan model random forest. Dataset berupa 6256 baris judul berita dengan kelas fake atau real. Melalui proses cleaning, tokenisasi, dan stemming, dataset disiapkan untuk pelatihan model. Hasil yang diperoleh menggunakan model random forest mencapai akurasi sebesar 84%, mengungguli model logistic regresi yang hanya mencapai 77%. (Ramadhan et al., 2022)

Penelitian kedua klasifikasi Berita Hoaks dengan Naïve Bayes dan Seleksi Maju. Penelitian ini menangani dampak serius dari penyebaran berita palsu, seperti kerugian materiil dan dampak psikologis yang dapat menciptakan rasa takut. Metode klasifikasi yang digunakan adalah naïve bayes dengan menggunakan data miming. Dengan memanfaatkan seleksi maju, penelitian ini mencapai akurasi sebesar 84%, dengan peningkatan signifikan pada presisi hingga 91,19%. Pendekatan ini diharapkan dapat meningkatkan tingkat akurasi pada klasifikasi berita hoaks Bahasa Indonesia. (Bagus et al., 2023)

Penelitian ketiga deteksi Berita Hoaks dengan Naïve Bayes Classifier. Berita hoaks di internet menjadi permasalahan global yang dapat mengancam ketertiban di masyarakat. Penelitian ini menggunakan Naïve Bayes Classifier untuk mendeteksi berita palsu di media sosial. Dengan melibatkan 700 data pelatihan dan 300 data uji dari situs Kumparan, model ini mencapai keakuratan sebesar 81%. Proses analisis menggunakan algoritma Naïve Bayes Classifier dengan penerapan CRISP-DM model. Hasil penelitian juga memperlihatkan kemudahan penerapan model ini melalui antarmuka pengguna di platform Heroku. Dengan pendekatan yang berbeda, ketiga penelitian ini secara kumulatif membuktikan efektivitas berbagai metode, mulai dari supervised learning dengan random forest hingga klasifikasi dengan Naïve Bayes, dalam mendeteksi dan mengatasi penyebaran berita palsu secara efisien. (Agustina & Hermawati, 2021)

Penelitian keempat menyatakan bahwa banyaknya berita online yang tersedia seringkali menarik minat masyarakat, namun, keberlimpahan informasi tersebut kadang membuat sulit bagi orang untuk mendapatkan informasi yang dapat dipercaya. Berita palsu, yang mencakup kata atau kalimat dengan informasi yang tidak benar, menjadi isu yang berusaha membohongi atau memanipulasi pembaca atau pendengarnya agar mendukung atau percaya pada kontennya. Para penyebar berita palsu biasanya mengetahui ketidakbenaran isi berita yang mereka sebar. Tujuan dari penelitian ini adalah untuk mendeteksi berita palsu yang menyebar di media sosial. Dalam upaya mengklasifikasi berita palsu, deteksi validitas berita menggunakan algoritma naïve bayes sebagai metode kategorisasi teks berbasis pembelajaran mesin. Penelitian ini juga melibatkan pembangunan sebuah situs web yang menyediakan fitur layanan web, pencarian berita di Twitter, dan klasifikasi berita secara manual. Antarmuka pengguna dari situs web ini, yang berbasis PHP, memungkinkan pengguna untuk berinteraksi secara langsung seperti memberikan komentar, login, atau melihat artikel-artikel yang sudah diklasifikasi. Sementara itu, backend dari situs web ini adalah program klasifikasi teks berbasis Python. Dari hasil percobaan, terbukti bahwa algoritma Naïve Bayes dapat efektif digunakan untuk mengklasifikasi berita palsu, mencapai nilai akurasi terbaik sebesar 92% pada 309 artikel uji. (Setiawan et al., 2021)

Penelitian kelima menyatakan bahwa kemajuan teknologi informasi yang terus berkembang memberikan dampak besar pada kehidupan manusia, khususnya dalam hal akses informasi. Saat ini, masyarakat dapat dengan mudah mengakses berbagai informasi melalui perangkat teknologi informasi. Namun, seiring dengan kemudahan akses tersebut, muncul tantangan baru, yaitu penyebaran informasi yang belum terverifikasi, yang dikenal dengan istilah berita palsu atau hoax. Kecepatan dalam menyajikan berita oleh media online dapat menjadi celah bagi pihak yang tidak bertanggung jawab untuk menyebarkan informasi yang tidak akurat. Hal ini menjadi lebih krusial, terutama dalam menghadapi isu-isu penting seperti kenaikan harga Bahan Bakar Minyak (BBM). Kenaikan harga BBM sering menjadi fokus utama di berbagai media berita dan platform sosial media di Indonesia. Sayangnya, sebagian besar informasi yang tersebar sulit diverifikasi kebenarannya, menciptakan ketidakpastian di kalangan masyarakat. Dalam



menghadapi tantangan ini, penting untuk mengembangkan mekanisme atau sistem yang dapat membantu mengklasifikasikan berita palsu terkait kenaikan harga BBM. Salah satu solusi yang diusulkan adalah menggunakan algoritma Support Vector Machine (SVM) dalam pembuatan sistem klasifikasi berita palsu. SVM, sebagai salah satu algoritma machine learning yang efektif, digunakan untuk mengklasifikasikan berita menjadi fakta dan bukan fakta. Penelitian ini mencakup beberapa tahapan, dimulai dari pengumpulan data untuk memastikan representativitas dataset terhadap isu kenaikan harga BBM, preprocessing data, hingga pengujian model SVM. Hasil pengujian menunjukkan bahwa algoritma SVM mampu mengklasifikasikan berita dengan tingkat akurasi yang dapat diterima. Oleh karena itu, penggunaan SVM dalam sistem klasifikasi berita palsu terkait kenaikan harga BBM diharapkan dapat menjadi solusi efektif untuk mengatasi penyebaran informasi yang tidak akurat. Langkah selanjutnya setelah pengujian adalah deployment model, yaitu menerapkan sistem klasifikasi ini dalam skala yang lebih luas, memberikan masyarakat akses pada informasi yang telah melalui proses verifikasi dengan lebih percaya diri. (Elisabeth et al., 2023).

B. Dasar Teori

1. Berita

Berita adalah sebuah laporan peristiwa aktual yang faktual, penting dan menarik bagi sebagian besar pembaca, serta menyangkut kepentingan mereka. Menurut Kusumaningrat, berita memiliki 6 unsur yang sering di ingat sebagai 5W 1H, antara lain yaitu :

- What* : apa yang telah terjadi ?
- Who* : siapa yang terlibat dalam kejadian ?
- When* : kapan peristiwa itu terjadi ?
- Why* : kenapa peristiwa itu bisa terjadi ?
- Where* : dimana peristiwa itu terjadi ?
- How* : Bagaimana peristiwa itu bisa terjadi ?

Berita juga memiliki beberapa sifat yaitu :

- Aktual, berarti harus memuat informasi yang terbaru dan terkini
- Faktual, berarti berdasarkan fakta dan data yang akurat
- Objektif, berarti harus disampaikan secara objektif dan tidak memihak
- Menarik, berarti harus menarik untuk dibaca dan disimak
- Penting, berarti harus memuat informasi yang penting bagi khalayak

Berita memiliki fungsi utama untuk menginformasikan, namun selain itu berita juga berfungsi untuk menghibur dan mendidik, bahkan bisa untuk memengaruhi dan mempersatukan sesuatu hal yang tadinya dipisahkan.

2. Hoax / Berita Palsu

Menurut Muhammad Alwi Dahlan, seorang ahli komunikasi dari Universitas Indonesia. Hoax atau berita palsu adalah sebuah manipulasi berita yang sengaja dilakukan bertujuan untuk memberikan pengakuan dan pemahaman yang salah. Hoax dapat berupa teks, gambar, audio atau video yang dimanipulasi untuk menipu pembaca atau pendengar atau penonton. Hoax dapat dibuat dengan berbagai tujuan khusus seperti untuk keuntungan politik, ekonomi atau bisa saja keuntungan pribadi.

3. Naive Bayes

Klasifikasi Bayes Naif adalah algoritma klasifikasi probabilistik yang didasarkan pada Teorema Bayes dan asumsi kemerdekaan fitur, algoritma ini digunakan untuk memprediksi probabilitas suatu data baru milik suatu kelas tertentu.

Teorema Bayes adalah rumus matematika yang digunakan untuk memperbarui probabilitas suatu peristiwa berdasarkan informasi terbaru. Teorema ini menyatakan bahwa suatu peristiwa A yang terjadi setelah mengetahui peristiwa B dapat dihitung dengan rumus berikut:



$$\frac{P(A|B) = (P(B|A) * P(A))}{P(B)}$$

Dimana :

$P(A|B)$ adalah probabilitas peristiwa A terjadi setelah mengetahui peristiwa B

$P(B|A)$ adalah probabilitas peristiwa B terjadi jika peristiwa A sudah terjadi

$P(A)$ adalah probabilitas apriori peristiwa A

$P(B)$ adalah probabilitas apriori peristiwa B

Naïve Bayes mengasumsikan bahwa semua fitur dalam suatu data saling independen satu sama lain. Yang artinya nilai suatu fitur tidak mempengaruhi nilai fitur lain. Naive Bayes bekerja dengan menghitung probabilitas suatu data baru milik suatu kelas tertentu berdasarkan probabilitas apriori kelas dan probabilitas fitur-fitur dalam data baru. Probabilitas dihitung menggunakan Teorema Bayes. Kelebihan Naive Bayes adalah Sederhana, Efisien, Robus (dapat menangani data bising dan tak lengkap), serta Efektif. Kemudian Kekurangan Naive Bayes adalah Asumsi Fitur tidak selalu akurat, Sensitif terhadap Outlier, dan membutuhkan banyak data.

4. *Random Forest*

Klasifikasi Random Forest adalah algoritma klasifikasi ensemble yang menggunakan pohon keputusan acak untuk menghasilkan prediksi yang lebih akurat dan stabil. Algoritma ini dikembangkan oleh Leo Breiman pada tahun 2001 dan telah menjadi salah satu klasifikasi yang populer dan banyak digunakan. Random Forest bekerja dengan membangun banyak pohon keputusan acak dari subset data pelatihan yang berbeda. Setiap pohon keputusan dilatih menggunakan subset data yang dipilih secara acak dengan penggantian (bootstrap sampling). Selain itu pada setiap node pohon keputusan, hanya subset fitur acak yang dipertimbangkan untuk pemisah. Kelebihan dari Random Forest adalah memiliki akurasi yang tinggi, Stabilitas tinggi, mudah diinterpretasikan, dapat menangani data multi dimensi. Klasifikasi ini juga memiliki kekurangan yaitu, kompleksitas komputasi, overfitting, serta membutuhkan banyak data. Penerapan Random Forest dalam sebuah aplikasi dapat berupa klasifikasi teks, gambar, data medis, dan data keuangan.

Random Forest menggunakan beberapa rumus matematika untuk membangun dan memprediksi data, berikut adalah rumus yang paling umum digunakan dalam algoritma Random Forest :

- a) Entropy, digunakan untuk mengukur ketidakpastian dalam kumpulan data

$$\text{Rumus : } \text{Entropy}(S) = - \sum p(i) * \log_2(p(i))$$

Dimana : S adalah kumpulan data, $p(i)$ adalah probabilitas dari kelas i

- b) Information Gain, digunakan untuk mengukur seberapa banyak entropy yang berkurang ketika sebuah fitur dibagi

$$\text{Rumus : } \text{Gain}(S, A) = \text{Entropy}(S) - \sum \left(\frac{|S_v|}{|S|} \right) * \text{Entropy}(S_v)$$

Dimana : S adalah kumpulan data, A adalah fitur yang dibagi, S_v adalah subset dari S yang dibagi berdasarkan nilai A

- c) Out-Of-Bag Error atau OOB, digunakan untuk memperkirakan akurasi model random forest

$$\text{Rumus : } \text{OOB error} = \frac{(\text{sum of misclassifications})}{(\text{number of out-of-bag samples})}$$

Dimana : sum of misclassifications adalah jumlah sampel yang salah klasifikasi, dan number of OOB samples adalah yang tidak digunakan untuk membangun pohon Keputusan



- d) Voting, digunakan untuk menggabungkan prediksi dari beberapa decision tree dalam Random Forest. Prediksi akhir adalah kelas yang paling banyak dipilih dari pohon keputusan.

Rumus : $Prediction = \operatorname{argmax}(\sum votes(i))$

Dimana : Prediction adalah prediksi akhir, argmax adalah fungsi mengambil nilai maksimal, $\sum votes(i)$ adalah jumlah suara untuk kelas i, votes(i) adalah suara dari satu pohon keputusan untuk sebuah kelas.

5. Klasifikasi

Menurut Dahlan, Klasifikasi adalah suatu proses pengelompokan benda-benda atau data ke dalam kelas berdasarkan ciri-ciri persamaan dan perbedaan. Klasifikasi ini bertujuan untuk memudahkan dalam memahami, mengorganisir dan mencari informasi.

6. Python

Menurut Guido van Rossum pencipta bahasa Python, adalah bahasa program tujuan umum yang berorientasi pada objek, interaktif, dan tingkat tinggi. Python dirancang untuk mudah dibaca dan dipelajari, dengan penekanan pada keterbacaan kode.

Berbeda dengan Penulis buku Python Programming Language Mark Lutz, Python adalah bahasa pemrograman yang kuat dan serbaguna yang dapat digunakan untuk berbagai macam tugas seperti pengembangan web dan program sistem hingga analisis data dan pembelajaran mesin.

7. Data Mining

Adalah sebuah proses analitik yang dirancang untuk mengeksplorasi data dalam mencari sebuah pola atau hubungan yang konsisten. Proses ini sering menggunakan teknik statistika, matematika, kecerdasan buatan dan pembelajaran mesin. Data mining digunakan untuk mengubah data mentah menjadi informasi yang bermanfaat dengan mencari pola dalam kumpulan data besar.

Manfaat data mining dapat dilihat dari dua sudut pandang, yaitu secara komersial dan keilmuan. Dari sudut pandang komersial, data mining dapat digunakan untuk menangani peledakan volume data, menyimpannya, mengekstraknya, dan memanfaatkannya untuk menghasilkan informasi yang dibutuhkan. Dalam konteks keilmuan data mining dapat membantu mengumpulkan data informasi berita dari internet sebanyak-banyaknya.

8. Machine Learning

Teknologi machine learning (ML) merujuk pada sistem mesin yang dirancang untuk dapat belajar secara mandiri tanpa memerlukan arahan langsung dari penggunaannya. Konsep pembelajaran mesin ini dibangun berdasarkan prinsip-prinsip dari disiplin ilmu lain seperti statistika, matematika, dan data mining, yang memungkinkan mesin untuk belajar dengan menganalisis data tanpa perlu direprogram atau diperintah.

Dalam konteks ini, machine learning memiliki kapabilitas untuk mengakses dan memahami data yang ada tanpa perlu petunjuk eksternal. ML juga mampu mempelajari baik data yang sudah ada maupun data yang diperolehnya sendiri, memungkinkannya untuk melaksanakan berbagai tugas sesuai dengan pemahaman yang telah diperoleh.

Istilah machine learning pertama kali diperkenalkan oleh beberapa ilmuwan matematika seperti Adrien Marie Legendre, Thomas Bayes, dan Andrey Markov pada tahun 1920-an, yang membahas dasar-dasar machine learning dan konsepnya. Sejak saat itu, ML terus mengalami perkembangan pesat. Salah satu contoh terkenal penerapan ML adalah Deep Blue yang dikembangkan oleh IBM pada tahun 1996.(Dicoding, 2020)



METODE PENELITIAN

Penelitian ini menggunakan metode kuantitatif dengan pendekatan eksperimental dengan alasan bertujuan untuk meningkatkan akurasi yang dapat di ukur dengan data numerik, serta penelitian ini melibatkan penerapan dua algoritma klasifikasi (Naive Bayes dan Random Forest) dan Membandingkan performanya dalam mendeteksi berita palsu.

A. Teknik Pengumpulan Data

Data dikumpulkan berdasarkan Data Teks dan Data Label, dimana data teks didapat dari berita yang telah beredar di internet serta data label yang digunakan untuk memberikan label "asli" dan "palsu" pada setiap berita.

B. Teknik Analisis Data

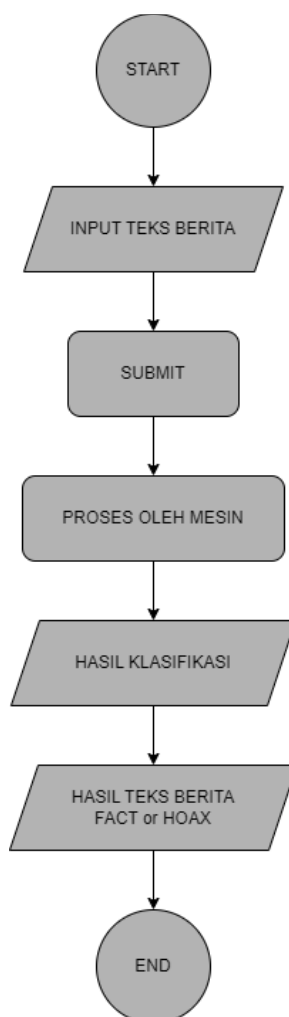
Menggunakan Naive Bayes dan Random Forest untuk mengklasifikasikan berita, serta melakukan evaluasi model dengan menghitung kinerja seperti akurasi, presisi, dan recall untuk membandingkan performa kedua algoritma.

C. Alat Penelitian

Penelitian ini menggunakan Bahasa Pemrograman Python yang disambungkan ke interface pengguna melalui bahasa pemrograman web yaitu HTML dan CSS.

D. Flowchart

Berikut ini adalah struktur susunan dalam proses machine learning pengujian berita palsu :

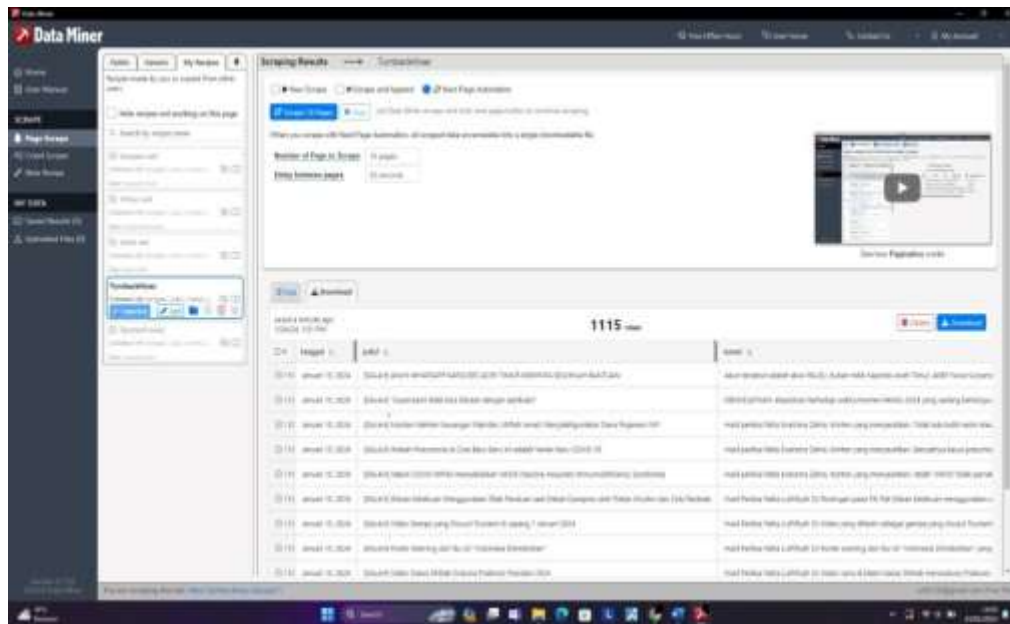


Gambar 1. Flowchart

HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Data yang terkandung dalam pelatihan dapat dikategorikan menjadi dua bagian, yaitu data berupa fakta yang valid dan data palsu. Untuk memperoleh data yang digunakan sebagai data training pada pengujian mesin learning.



Gambar 2. Proses Data Training

1. Sumber Data

Data diambil dari pencarian situs web berita seperti detik.com, Kompas.com, dan TribunNews dengan tujuan mendapatkan informasi yang komprehensif, penulis mengakses berbagai sumber untuk mengumpulkan data yang relevan. Proses ini merupakan bagian integral dari penelitian, di mana keberagaman situs web memberikan sudut pandang yang berbeda dan mendalam terhadap topik yang dibahas. Pemanfaatan detik.com, Kompas.com, dan TribunNews sebagai sumber informasi mencerminkan pendekatan holistik dalam menggali wawasan. Dengan demikian, hasil pencarian data diharapkan dapat memberikan kerangka pemahaman yang lebih lengkap dan akurat terkait dengan topik penelitian yang sedang dijalan.

2. Data Mining

Proses penambangan data atau data mining dari sumber-sumber seperti detik.com, Kompas.com, dan TribunNews melibatkan ekstraksi informasi yang berharga dan terkait dengan topik penelitian. Langkah awal mencakup pengumpulan data dari berbagai artikel, berita, atau konten yang relevan dengan tujuan penelitian. Setelah data terkumpul, analisis dilakukan untuk mengidentifikasi pola, tren, dan informasi tersembunyi. Metode ini memungkinkan peneliti untuk merinci dan memahami konten yang ada, serta mengungkap wawasan yang mungkin tidak langsung terlihat. Teknik penambangan data dapat mencakup penggunaan algoritma dan model statistik untuk mengelompokkan data, membuat prediksi, atau mengeksplorasi hubungan antarvariabel. Dengan menggunakan sumber-sumber berita tersebut, penambangan data bertujuan untuk mengoptimalkan pemahaman terhadap topik tertentu, mendukung analisis, dan menyajikan temuan yang berarti dalam kerangka

penelitian yang sedang dilakukan.

3. Total Data

Dalam fase pencarian data melalui situs-situs web seperti detik.com, Kompas.com, dan TribunNews, penelitian ini berhasil mengumpulkan dataset yang signifikan. Proses penambangan data menghasilkan total 1736 data fakta dan 2728 data hoax dari berbagai sumber berita tersebut. Penelitian ini memanfaatkan keberagaman situs web sebagai sumber informasi untuk mendapatkan sudut pandang yang komprehensif terkait topik yang diteliti. Dengan memperoleh dataset sebanyak ini, analisis lebih lanjut dapat dilakukan untuk mengidentifikasi pola-pola, tren, dan perbedaan antara informasi fakta dan hoax. Total keseluruhan data yang diperoleh mencapai 4466, mencerminkan pendekatan holistik dalam penelitian untuk memastikan keterwakilan yang baik dari berbagai perspektif dan sumber informasi. Hasil penambangan data ini diharapkan dapat memberikan pemahaman yang mendalam dan akurat terhadap fenomena yang sedang diteliti.

B. Tabel Data Training

Berikut beberapa sampel data yang digunakan untuk membangun machine learning deteksi berita hoax :

Tabel 1. Contoh 6 Data Training ber-Label		
No	Narasi	Label
1	Hasil periksa fakta Siti Lailatul Fitriyah Video ribuan kader Gerindra dukung Ganjar sebagai Presiden 2024 adalah konten tidak benar. Faktanya tidak ada bukti kredibel terkait	0
2	Hoaks seputar Timnas Indonesia kerap beredar di saat ada turnamen atau pertandingan besar. Hoaks ini menyebar melalui media sosial maupun aplikasi percakapan	0
3	Beredar di media sosial postingan foto pelajar sedang memasang foto Anies Baswedan dan Muhaimin Iskandar di ruang kelas	0
4	Kasus Covid-19 kembali meningkat, termasuk di Indonesia. Meski begitu, aktivitas fisik atau olahraga anak maupun remaja tetap bisa dilakukan. Ini tipsnya	1
5	Kementerian Koperasi dan UKM telah memanggil 12 pihak yang 5 diduga melanggar penyaluran Kredit Usaha Rakyat (KUR). Diantaranya 3 Bank BUMN,Bank	1
6	Tahun 2023, KAI Logistik berhasil mencatatkan pertumbuhan bisnis dengan ragam capaian baik pada aspek kinerja keuangan maupun kinerja operasional.	1

Keterangan : 0 untuk parameter berita Hoax
 1 untuk parameter berita Fakta

C. Pengujian Algoritma

1. Naive Bayes

Hasil pengujian akurasi menggunakan metode Naïve Bayes mencapai tingkat keseluruhan sebesar 96,88 %. Pada konteks ini, terdapat perbandingan dengan algoritma pembanding, yakni algoritma random forest, di mana terlihat bahwa metode Naïve Bayes menunjukkan kinerja



```
***
              precision    recall  f1-score   support

      0         0.96      0.97      0.97         553
      1         0.97      0.96      0.97         538

   accuracy              0.97         1091
  macro avg              0.97      0.97      0.97         1091
 weighted avg              0.97      0.97      0.97         1091

Accuracy Naive Bayes: 96.88%
```

Gambar 3. Hasil Pengujian Akurasi Metode Naive Bayes

2. Random Forest

Algoritma Random Forest diimplementasikan sebagai alat pembanding terhadap algoritma sebelumnya, yaitu algoritma Naïve Bayes. Hasil uji akurasi dari Algoritma Random Forest menunjukkan angka sebesar 98,99%, menandakan bahwa algoritma ini mencapai tingkat akurasi yang lebih tinggi dibandingkan dengan algoritma Naïve Bayes yang sebelumnya digunakan.

```
              precision    recall  f1-score   support

      0         1.00      0.98      0.99         553
      1         0.98      1.00      0.99         538

   accuracy              0.99         1091
  macro avg              0.99      0.99      0.99         1091
 weighted avg              0.99      0.99      0.99         1091

Accuracy Random Forest: 98.99%
```

Gambar 4. Hasil Pengujian Akurasi Metode Random Forest

D. Pembuatan Aplikasi

1. Flask

Flask adalah suatu rangkaian kerja web yang ditulis dalam bahasa pemrograman Python dan dikategorikan sebagai kerangka mikro. Sebagai kerangka aplikasi dan antarmuka web, Flask bersama dengan bahasa Python memfasilitasi pengembang dalam pembuatan situs web yang terstruktur dan pengelolaan perilaku situs web dengan lebih sederhana. Flask dianggap sebagai microframework karena tidak mengharuskan penggunaan alat atau pustaka spesifik. Fitur dan komponen umum, seperti validasi formulir dan database, tidak terpasang secara otomatis di Flask. Hal ini terjadi karena fitur dan komponen tersebut diperoleh dari pihak ketiga, dan Flask dapat menggunakan ekstensi yang membuatnya terlihat seolah-olah diimplementasikan oleh Flask sendiri. Walaupun disebut sebagai microframework, Flask tidak kekurangan fitur. Istilah "microframework" di sini menunjukkan bahwa Flask dirancang untuk membuat inti aplikasi sesederhana mungkin, namun tetap mudah untuk diperluas. Oleh karena itu, tingkat fleksibilitas dan ekstensibilitas Flask relatif tinggi dibandingkan dengan beberapa kerangka kerja lainnya

2. Google Colaboratory

Google Colaboratory atau Google Colab adalah dokumen yang dapat dieksekusi yang memungkinkan pengguna menyimpan, membuat, dan berbagi program melalui Google Drive. Dengan beragam keuntungan dan manfaatnya, software ini dapat menjadi solusi bisnis yang efektif. Fitur-fitur tersebut bahkan tersedia secara gratis. Untuk memahami

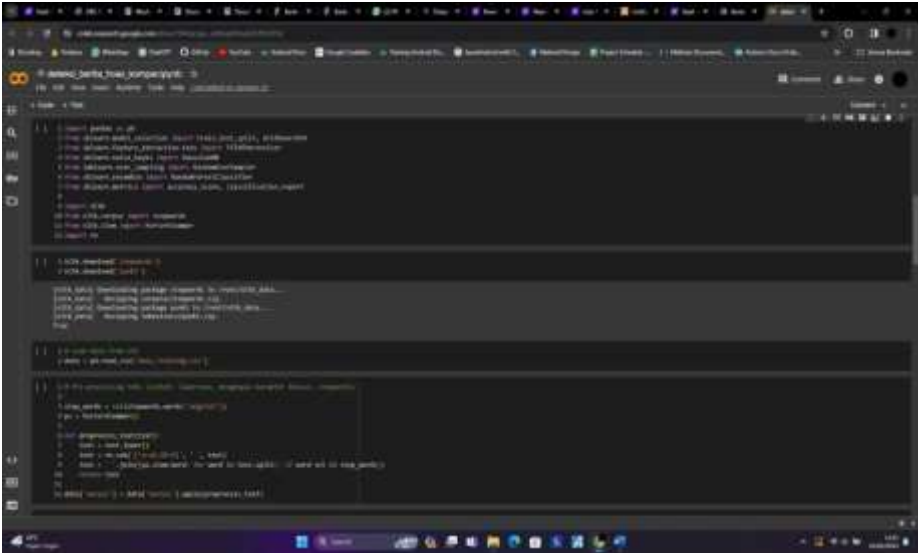
esensi Google Colab dengan lebih mudah, dapat dianggap sebagai alat yang disediakan oleh Google untuk mendukung pengguna dalam proses pemrograman dan pengolahan data. Pengguna dapat menggunakan alat ini tanpa perlu menginstal perangkat lunak tambahan, dan hanya perlu login dengan akun Google serta menyimpan file Python di Google Drive pribadi mereka. Keunikan lainnya adalah pengguna tidak perlu melakukan instalasi atau konfigurasi apa pun untuk menjalankan kode Python, karena semua pengaturan dan penyesuaian diserahkan ke cloud. Oleh karena itu, Google 19 Colab dapat menjadi pilihan yang tepat, terutama bagi mereka yang belum terbiasa atau sedang mempelajari Python.

E. Hasil Uji

1. Pembuatan Aplikasi berbasis web

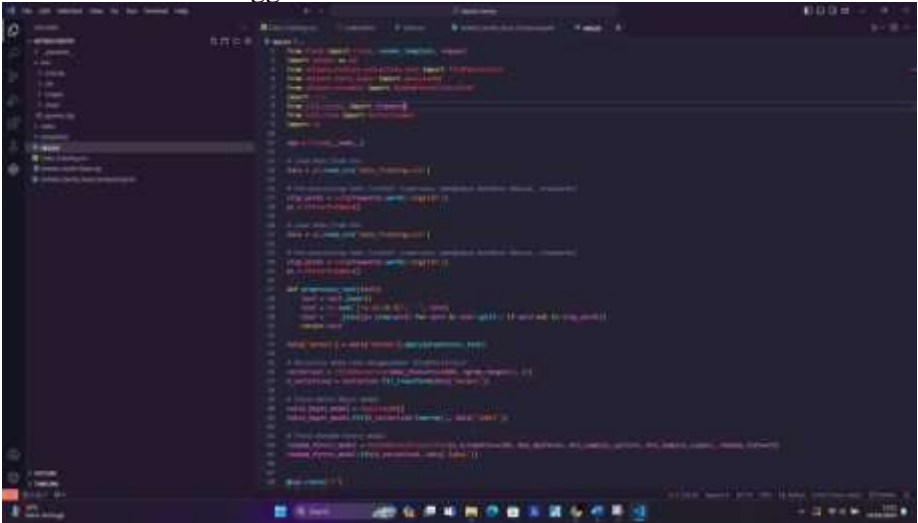
Berikut ini adalah Proses Pembuatan Aplikasi Deteksi Berita Hoaks,

- a) Pembuatan Formula di Google Colabs



Gambar 5. Tampilan Projek Google Colabs

- b) Pembuatan situs web menggunakan Flask

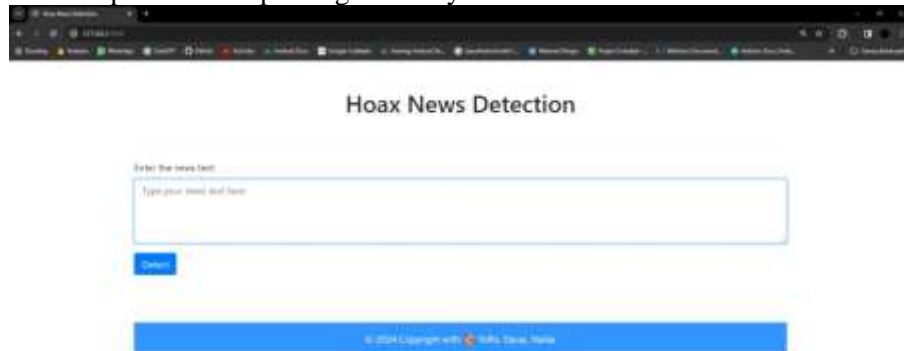


Gambar 6. Tampilan Projek Kodingan Flask di Visual Code Studio



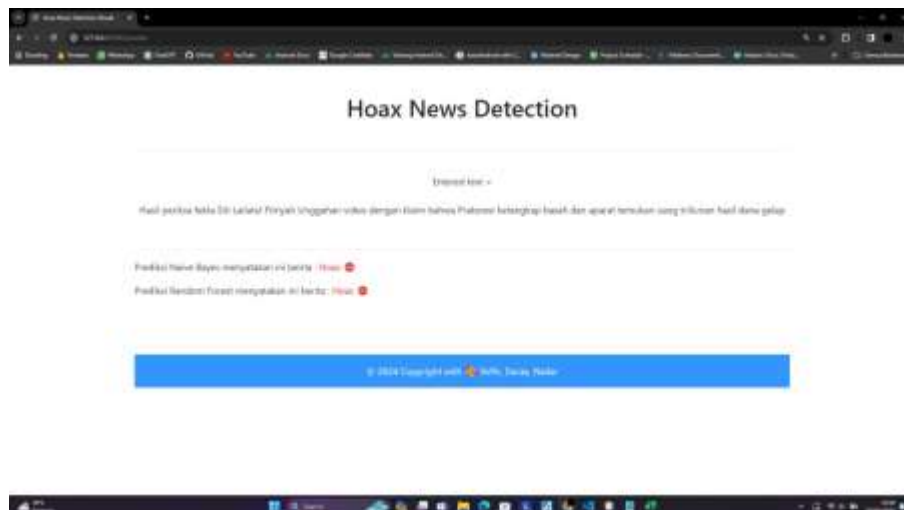
c) Uji Coba di situs web lokal

Berikut ini adalah User Interface dari Aplikasi Deteksi Berita Hoaks menggunakan metode Flask pada bahasa pemrograman Python.



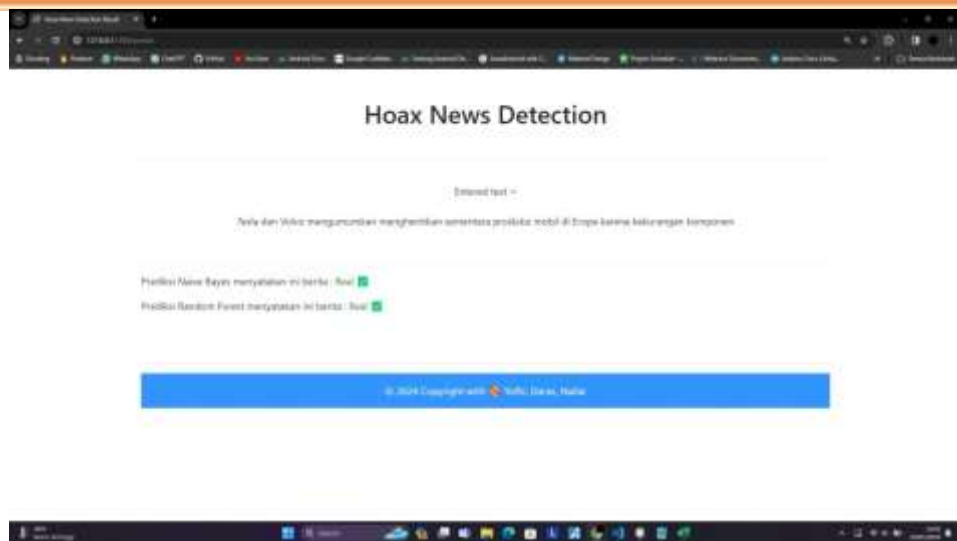
Gambar 7. Tampilan Awal Web Deteksi Berita Palsu

Ketika User memasukkan kalimat yang merupakan hasil training dari data Palsu, maka sistem akan menampilkan prediksi dari kedua metode Naive Bayes dan Random Forest dengan Label warna Merah atau berarti itu merupakan berita palsu.



Gambar 8. Tampilan Hasil Ketika diberi masukan berita Palsu

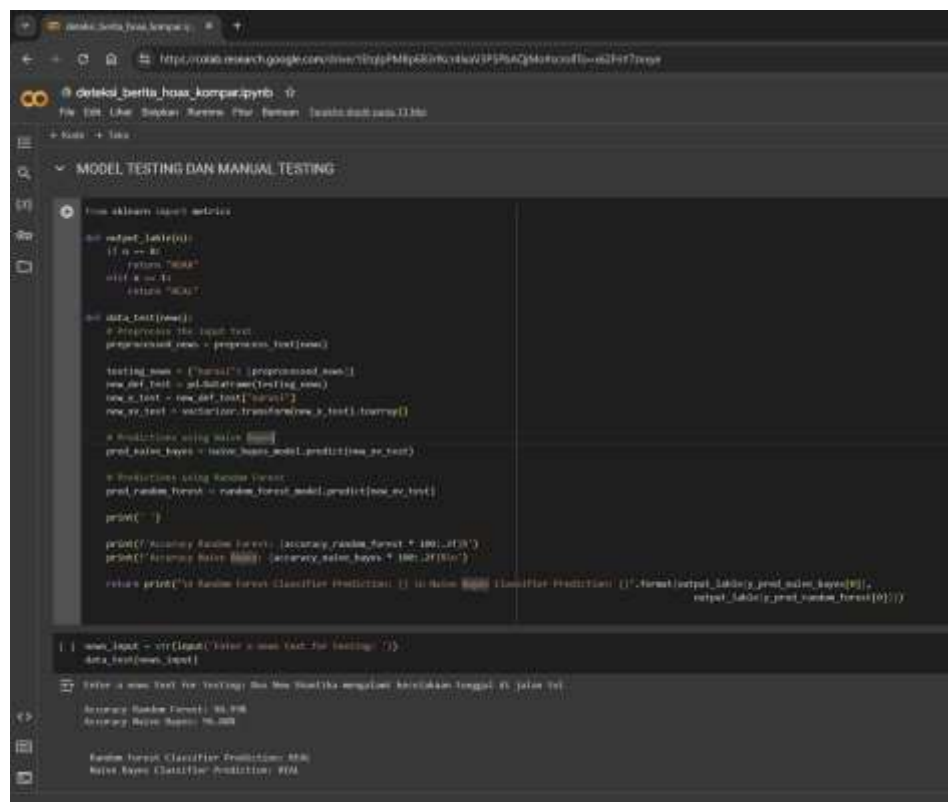
Ketika user memasukkan kalimat yang berpotensi dalam data training adalah berlabel Fakta maka hasil Prediksi pada tampilan akan berwarna hijau



Gambar 9. Tampilan Hasil Ketika diberi masukan berita Fakta

2. Akurasi Algoritma Naive Bayes

Dalam membuktikan tingkat akurasi Algoritma Naive Bayes sesuai dengan yang dituliskan sebelumnya. Maka dibuatlah uji coba seperti pada gambar dibawah ini, menggunakan aplikasi google colab.



Gambar 10. Tampilan Hasil Ketika diberi masukan berita Fakta

Terlihat pada gambar hasil akurasi nya mirip dengan apa yang diimplementasikan pada situs web. Alasannya karena memang hasil pada saat implementasi adalah hasil kloningan dari bahan uji.



3. Akurasi Algoritma Random Forest

Selain pembuktian untuk Naive Bayes, juga dibutuhkan pembuktian untuk Algoritma Random Forest yang hasilnya dapat di lihat di poin sebelum ini, yaitu pada gambar 10

KESIMPULAN

Dalam konteks penelitian ini, berhasil dikembangkan sistem deteksi berita palsu menggunakan algoritma machine learning, terutama Naïve Bayes dan Random Forest. Algoritma Naïve Bayes mencapai tingkat akurasi sebesar 96,88%, menunjukkan kinerja yang baik, meskipun ada ruang untuk peningkatan lebih lanjut. Sebagai pembanding, Random Forest menunjukkan tingkat akurasi yang lebih tinggi, yakni sebesar 98,99%, menandakan efektivitasnya dalam menangani penyebaran berita palsu di lingkungan media sosial. Aplikasi deteksi berita palsu berhasil dibangun dengan menggunakan Flask sebagai kerangka kerja web yang memberikan fleksibilitas tinggi.

Untuk penelitian selanjutnya, disarankan bisa meningkatkan akurasi algoritma Naïve Bayes, melalui pemilihan fitur yang lebih optimal dan tuning parameter yang lebih cermat. Penelitian lebih lanjut dengan dataset yang lebih luas dapat memperkuat generalisasi model. Pengembangan antarmuka pengguna yang lebih baik di aplikasi deteksi berita palsu dapat meningkatkan user experience dan adopsi oleh pengguna. Integrasi dengan platform media sosial seperti Twitter dan Facebook juga dapat memperluas cakupan deteksi. Evaluasi terhadap metode deteksi lainnya, seperti deep learning, dapat menjadi penelitian lanjutan yang menarik. Dengan implementasi saran-saran ini, diharapkan penelitian ini dapat terus berkembang dan memberikan kontribusi yang lebih besar dalam menangani permasalahan berita palsu di era media sosial.

DAFTAR PUSATAKA

- Agustina, N., & Hermawati, M. (2021). Implementasi Algoritma Naïve Bayes Classifier untuk Mendeteksi Berita Palsu pada Sosial Media. *Factor Exacta*, 14(4), 1979–276. <https://doi.org/10.30998/faktorexacta.v14i4.11259>
- Amal, I., Pamungkas, E. W., Kom, S., & Kom, M. (2023). *Aplikasi Pendeteksi Berita Palsu Bahasa Indonesia Menggunakan Framework Flask dan Streamlit serta Algoritma Machine Learning*. 1–18. [https://eprints.ums.ac.id/id/eprint/116531%0Ahttps://eprints.ums.ac.id/116531/1/Naskah Publikasi_Ikhlasul Amal.pdf](https://eprints.ums.ac.id/id/eprint/116531%0Ahttps://eprints.ums.ac.id/116531/1/Naskah%20Publikasi_Ikhlasul%20Amal.pdf)
- Badriyah, Siti., “Pengertian Hoaks: Sejarah, Jenis, Contoh, Penyebab dan Cara Menghindarinya”, Gramedia Blog, Gramedia, di akses tanggal 7 Juni 2024, <https://www.gramedia.com/literasi/pengertian-hoaks/>
- Bagus, D., Prasetyo, C., Andono, N., & Supriyanto, C. (2023). *JURNAL MEDIA INFORMATIKA BUDIDARMA Metode Naive Bayes Classifier dan Forward Selection Untuk Deteksi Berita Hoaks Bahasa Indonesia*. 7, 1541–1550. <https://doi.org/10.30865/mib.v7i3.6459>
- Dicoding. (2020). *Apa itu Machine Learning? Beserta Pengertian dan Cara Kerjanya*. Dicoding. <https://www.dicoding.com/blog/machine-learning-adalah/>
- Elisabeth, G., Rahma Salsa Bilah, Ardini, S. N., Agustina, N., & Rismayadi, D. A. (2023). Klasifikasi Berita Palsu Kenaikan Harga Bahan Bakar Minyak (Bbm) Menggunakan Algoritma Support Vector Machine (Svm). *Naratif: Jurnal Nasional Riset, Aplikasi Dan Teknik Informatika*, 5(2), 104–109. <https://doi.org/10.53580/naratif.v5i2.188>
- Hikmat Kusumaningrat & Purnama Kusumaningrat. *Jurnalistik : Teori Dan Praktik / Hikmat Kusumaningrat , Purnama Kusumaningrat*. 2005
- Noviah, S. R. (2023). *Mengenal Google Colab dan Manfaatnya*. PT. Exabytes Network Indonesia. <https://www.exabytes.co.id/blog/google-colab-adalah/>
- Pambudi, I. (2018). Analisis Framing Video Breaking News Jatuhnya pesawat Lion Air JT-610 pada Official Youtube Kompas TV. *Repostory.IAIN Tulungagung*, 13–20. [http://repo.iain-tulungagung.ac.id/16051/5/Bab II.pdf](http://repo.iain-tulungagung.ac.id/16051/5/Bab%20II.pdf)
- Ramadhan, N. G., Adhinata, F. D., Segara, A. J. T., & Rakhmadani, D. P. (2022). Deteksi Berita Palsu Menggunakan Metode Random Forest dan Logistic Regression. *JURIKOM (Jurnal Riset Komputer)*, 9(2), 251. <https://doi.org/10.30865/jurikom.v9i2.3979>



- Rumah, B. D. S. di. (2021). *Mengenal Flask, Library Machine Learning Python Idaman Developer*. Yayasan Multimedia Nusantara & Xeratic.
<https://dqlab.id/mengenal-flask-library-machine-learning-python-idamandevloper>
- Setiawan, E. I., Johanes, S., Hermawan, A. T., & Yamasari, Y. (2021). Deteksi Validitas Berita pada Media Sosial Twitter dengan Algoritma Naive Bayes. *Journal of Intelligent System and Computation*, 3(2), 55–60. <https://doi.org/10.52985/insyst.v3i2.164>
- Wicaksono, A. E. (2016). Implementasi Data Mining Dalam Pengelompokan Peserta Didik Di Sekolah. *Jurnal Teknologi Rekayasa*, 21(3), 206–216.
- Yunanto, R., Purfini, A. P., & Prabuwisesa, A. (2021). Survei Literatur: Deteksi Berita Palsu Menggunakan Pendekatan Deep Learning. *Jurnal Manajemen Informatika (JAMIKA)*, 11(2), 118–130. <https://doi.org/10.34010/jamika.v11i2.5362>