



Computer Based Information System Journal

ISSN (Print): 2337-8794 | E- ISSN : 2621-5292
web jurnal : <http://ejournal.upbatam.ac.id/index.php/cbis>



Penerapan Metode Machine Learning dalam Mengidentifikasi Berita Hoaks

Vijay Andika Saputra¹, Sasa Ani Arnomo²

Universitas Putera Batam, Indonesia.

INFORMASI ARTIKEL

Diterima Redaksi: Januari 2024
Diterbitkan Online: Maret 2024

KATA KUNCI

Hoax News, Machine Learning,
News Analysis, Text Classification

KORESPONDENSI

E-mail: pb201510033@upbatam.ac.id
sasaupb@gmail.com

A B S T R A C T

In the era of increasingly developing digital information, hoax or fake news is a serious challenge that can influence people's perceptions and decisions. This research aims to develop a hoax news detection method using a machine learning approach on news data. The focus of this research is to identify related hoax news. This research involves collecting news data from leading online news sources and applying machine learning algorithms, including Logistic Regression, Decision Tree, Gradient Boosting, and Random Forest, to classify hoax news. Model performance is measured using accuracy, precision, recall and F1-score metrics, with test results compared with human evaluations that have been trained to recognize hoax news. The results of this research show that the machine learning approach can be successfully used to identify hoax news with a high level of accuracy. However, this study also identified limitations in this approach, such as limited training data and language complexity. Nevertheless, this research makes an important contribution in efforts to overcome the problem of hoax news and provides a basis for further development in hoax news detection using machine learning technology.

I. Latar Belakang

Dalam konteks era informasi digital yang berkembang pesat, tantangan utama yang dihadapi adalah penyebaran berita hoaks atau palsu yang dapat mempengaruhi opini masyarakat dan merugikan individu serta institusi yang menjadi sasarannya [1]. Fenomena ini menuntut pengembangan sistem yang mampu mengidentifikasi berita hoaks secara efisien dan akurat. Penelitian ini bertujuan mengembangkan model machine learning untuk mengidentifikasi berita hoaks dengan tingkat akurasi tinggi. Melalui analisis mendalam terhadap berbagai

karakteristik data berita, seperti struktur teks, penggunaan kata kunci, dan sentimen teks, serta penerapan teknik klasifikasi dan ekstraksi fitur, penelitian ini bertujuan membedakan berita hoaks dan berita sah. Evaluasi kinerja model menggunakan metrik standar seperti akurasi, presisi, recall, dan F1-score diharapkan memberikan panduan dan wawasan bagi pihak yang terlibat dalam penyediaan dan konsumsi informasi. Pengembangan sistem yang efektif ini diharapkan dapat melindungi integritas informasi, mencegah penyebaran berita palsu, dan

membangun lingkungan informasi yang lebih dapat dipercaya di era digital ini [2].

II. Kajian Literatur

Di dalam sub kajian literatur, penulis akan membahas mengenai teori-teori umum yang berhubungan dengan pelaksanaan penelitian ini yang berupa pengertian-pengertian secara umum.

2.1 Berita Hoaks

Berita adalah penyampaian informasi terkini melalui media massa, bertujuan memberikan pemahaman kepada masyarakat tentang peristiwa atau fakta terbaru yang signifikan, mencakup isu politik, sosial, ekonomi, dan budaya. Berita hoaks (hoax news) adalah informasi palsu atau tidak benar yang disajikan dalam bentuk berita dengan tujuan untuk menyesatkan atau memanipulasi pembaca atau penonton.

Berita hoaks sering kali dirancang sedemikian rupa sehingga terlihat seperti berita yang sah dan dapat dengan mudah menipu orang yang membacanya. Tujuan dari berita hoaks bisa bermacam-macam, termasuk menyebarkan propaganda, menciptakan kebingungan, mempengaruhi opini publik, atau bahkan mencari keuntungan finansial [3].

2.2 Python

Python adalah bahasa pemrograman tingkat tinggi yang bersifat serba guna, mendukung pemrograman berorientasi objek, dan dikenal karena sintaksisnya yang mudah dibaca dan mudah dipahami. Python dirancang untuk menjadi bahasa yang lebih intuitif dan efisien, dengan fokus pada penulisan kode yang lebih sedikit dibandingkan dengan bahasa pemrograman lainnya [4].

2.3 Machine Learning

Machine Learning (ML) adalah cabang dari kecerdasan buatan (artificial intelligence) yang berfokus pada pengembangan algoritma dan model komputer yang dapat belajar dari data serta melakukan tugas-tugas tertentu tanpa perlu secara eksplisit diprogram. Dalam konsep tradisional pemrograman, akan memberikan aturan dan

instruksi kepada komputer untuk melakukan suatu tugas tertentu. Namun, dalam machine learning, komputer belajar dari data yang diberikan dan dapat meningkatkan performanya seiring bertambahnya data [5].

2.4 Algoritma Klasifikasi

Algoritma klasifikasi dalam machine learning merupakan teknik yang digunakan untuk memprediksi kelas atau label suatu data berdasarkan fitur-fitur yang dimilikinya. Tujuan utamanya adalah mengelompokkan data ke dalam kelas atau kategori tertentu berdasarkan pola-pola yang terdapat dalam data pelatihan. Proses klasifikasi melibatkan pembelajaran dari data pelatihan yang sudah dikategorikan sebelumnya, dan kemudian pengetahuan yang diperoleh digunakan untuk mengklasifikasikan data baru yang belum pernah dilihat sebelumnya. Algoritma ini mencari pola atau relasi dalam data untuk membuat keputusan tentang kategori mana yang paling cocok untuk data baru.

Aplikasi dari algoritma klasifikasi mencakup pengenalan tulisan tangan, analisis sentimen, deteksi penyakit melalui gambar medis, pemisahan spam pada email, dan prediksi harga saham. Beberapa algoritma klasifikasi populer meliputi Logistic Regression, Decision Trees, Gradient Boosting dan Random Forest [6].

Random Forest secara konsisten memberikan hasil yang lebih tinggi dibandingkan dengan rata-rata kinerja algoritma klasifikasi lainnya yang terdapat dalam gambar tersebut [7].

Rumus umum Logistic Regression adalah sebagai berikut:

$$P(Y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + \dots + b_n X_n)}}$$

di mana:

- $P(Y = 1)$ adalah probabilitas bahwa output Y adalah 1.

- e adalah basis logaritma natural.

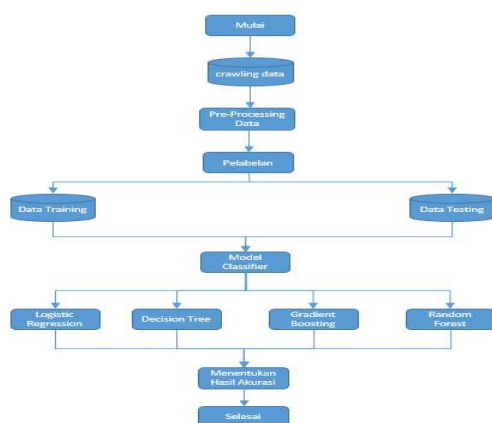
- $b_0, b_1, b_2, \dots, b_n$ adalah parameter model (koefisien).

$x_1, x_2, \dots, x_n$ adalah nilai-nilai fitur input.

Nilai-nilai fitur $x_1, x_2, \dots, x_n$ dikombinasikan dengan parameter model $b_0, b_1, b_2, \dots, b_n$ melalui fungsi logistik untuk menghasilkan probabilitas kelas positif (1). Probabilitas tersebut kemudian digunakan untuk membuat keputusan klasifikasi. Jika $P(Y = 1)$ lebih besar dari atau sama dengan 0.5, data diklasifikasikan sebagai kelas 1; sebaliknya, jika $P(Y = 1)$ kurang dari 0.5, data diklasifikasikan sebagai kelas 0.

III. Metodologi

Desain penelitian adalah rencana atau strategi yang digunakan oleh peneliti untuk menjawab pertanyaan penelitian dan mencapai tujuan penelitian.



Gambar 1. Desain Penelitian [8]

Penjelasan dari langkah-langkah yang Anda berikan adalah sebagai berikut:

1. Crawling Data

Data yang diperlukan untuk penelitian diambil dari berbagai sumber menggunakan teknik crawling. Ini dapat mencakup pengambilan data dari situs web, media sosial, atau sumber lainnya [9].

2. Preprocessing Data

Data yang telah diambil kemudian melalui tahap preprocessing, di mana dilakukan pembersihan data dari noise, normalisasi data, dan

langkah-langkah lain untuk mempersiapkan data agar siap untuk digunakan dalam model machine learning [10].

3. Pelabelan

Data yang sudah bersih akan diberi label, yaitu kategori atau klasifikasi yang sesuai dengan tujuan penelitian. Label ini biasanya mencakup klasifikasi sebagai data training atau data testing.

4. Data Training

Sebagian dari data yang telah diberi label digunakan sebagai data training. Ini adalah data yang digunakan oleh model machine learning untuk belajar dan menghasilkan pola dari fitur-fitur yang ada.

5. Data Testing

Sisanya dari data yang telah diberi label digunakan sebagai data testing. Data ini tidak digunakan dalam proses pelatihan, tetapi digunakan untuk menguji kinerja model yang telah dilatih [11].

6. Model Classifier

Model classifier adalah kerangka kerja atau struktur dasar yang digunakan untuk mengklasifikasikan data ke dalam kategori yang telah ditentukan.

7. Logistic Regression

Salah satu metode machine learning yang digunakan untuk masalah klasifikasi, khususnya ketika variabel dependen adalah biner [12].

8. Decision Tree

Pemanfaatan pohon keputusan terutama terletak pada kemampuannya untuk menyederhanakan proses pengambilan keputusan, mengurai langkah-langkahnya dari yang rinci menjadi lebih sederhana. Hal ini memungkinkan pengambil keputusan untuk lebih mudah menginterpretasikan solusi dari permasalahan yang dihadapi. Pohon keputusan juga digunakan untuk melakukan eksplorasi data dan mengungkap hubungan yang mungkin tersembunyi [13].

9. Gradient Boosting

Metode ensemble yang menggabungkan beberapa model *weak learners* menjadi satu

model yang kuat untuk meningkatkan performa [14].

10. Random Forest

Metode ensemble lain yang menggunakan banyak pohon keputusan untuk meningkatkan akurasi dan mengurangi overfitting [15].

11. Menentukan Hasil Akurasi

Setelah model-model di atas dilatih dan diuji dengan data testing, hasil akurasi dievaluasi dengan menggunakan metrik yang sesuai, seperti akurasi, presisi, recall, dan f1-score.

IV. Pembahasan

Tahapan utama dalam mengembangkan sistem kecerdasan buatan yang mampu mempelajari pola dari data melibatkan proses pelatihan model dengan memberikan akses ke dataset latih. Selama proses pelatihan, model secara iteratif disesuaikan untuk meminimalkan kesalahan dalam memprediksi output yang benar dari data yang belum pernah dilihat sebelumnya. Beberapa algoritma yang populer digunakan untuk melatih model antara lain adalah Logistic Regression, Decision Tree, Gradient Boosting, dan Random Forest. Setelah melalui tahap pelatihan, model diuji menggunakan dataset uji untuk mengevaluasi kinerjanya [16].

1. Pengumpulan Data dari Web

Pengumpulan data dari web, yang sering disebut sebagai web scraping atau web crawling, adalah proses pengambilan informasi atau data dari halaman web. Ini merupakan langkah awal penting dalam banyak proyek analisis data, penelitian, atau pembangunan model machine learning yang memerlukan data yang diperoleh dari internet. Berikut adalah langkah-langkah umum dalam pengumpulan data dari web.

A. Crawling data dari website berita Detik

```
# Detik
import requests
from bs4 import BeautifulSoup
import csv

# Initialize variables
base_url = 'https://news.detik.com/' # Base URL
page_number = 20 # Start with page 1

# Define CSV file name and field names
csv_file = 'detik_test.csv'
field_names = ['Title', 'Date', 'Link']

# Open CSV file in write mode and write header row
with open(csv_file, mode='w', newline='', encoding='utf-8') as file:
    writer = csv.writer(file)
    writer.writerow(field_names)

while True:
    # Construct the URL for the current page
    url = f'{base_url}index/{page_number}'
    response = requests.get(url)

    if response.status_code == 200:
        # Parse HTML content
        soup = BeautifulSoup(response.content, 'html.parser')

        # Find and extract tempo and authors
        title_elmt = soup.find_all(class_='grid-row list-content')
        title_elmt = title_elmt.find_all('h3', class_='mb-10 mb-15')
        title_elmt = [title_elmt.text for title_elmt in title_elmt]
        title_text = [title_elmt.strip() for title_elmt in title_elmt]

        date_elmt = soup.find_all(class_='mb-10 mb-15')
        date_elmt = [date_elmt.text for date_elmt in date_elmt]
        date_text = [date_elmt.strip() for date_elmt in date_elmt]

        link_elmt = soup.find_all(class_='mb-10 mb-15')
        link_elmt = [link_elmt.a href for link_elmt in link_elmt]
        link_text = [link_elmt.strip() for link_elmt in link_elmt]

        # If no data is found, assume we've reached the last page
        if page_number == 21:
            break

        # Write data to CSV
        for title_text, date_text, link_text in zip(title_text, date_text, link_text):
            writer.writerow([title_text, date_text, link_text])

        # Move to the next page
        page_number += 1
    else:
        print(f'Error accessing page {page_number}. Status code: {response.status_code}')
        break

print(f'Data has been saved to {csv_file}')
```

Gambar 2. Crawling Data dari Website Berita Detik

B. Crawling data dari website berita turnbackhoax

```
if response.status_code == 200:
    # Parse HTML content
    soup = BeautifulSoup(response.content, 'html.parser')

    # Find and extract tempo and authors
    title_elmt = [title_rslt1.text for title_rslt1 in
soup.find_all('h3', class_='entry-title mh-loop-title')]
    title_text = [title_rslt1.strip() for title_rslt1 in
title_elmt]

    text_text = [text_rslt.text for text_rslt in
soup.find_all('div', class_='mh-excerpt')]

    date_text = [date_rslt.text for date_rslt in
soup.find_all('span', class_='mh-meta-date updated')]

    link_text = [link_rslt.a href for link_rslt in
soup.find_all('h3', class_='entry-title mh-loop-title')]

    # If no data is found, assume we've reached the la
st page
    if page_number == 100:
        break

    # Write data to CSV
    for title_rslt, text_rslt, date_rslt, link_rslt in
zip(title_text, text_text, date_text, link_text):
        writer.writerow([title_rslt, text_rslt, date_r
slt, link_rslt])

    # Move to the next page
    page_number += 1
else:
    print(f'Error accessing page {page_number}. Status
code: {response.status_code}')
    break

print(f'Data has been saved to {csv_file}')
```

Gambar 3. Crawling Data dari Website Berita Turnbackhoax

2. Penggunaan Dataset

Dataset adalah sekumpulan data yang diperoleh melalui proses crawling menggunakan BeautifulSoup. Terdapat dua set data, satu berisi informasi yang valid dan yang lainnya berisi informasi yang terindikasi sebagai hoaks.

Title	Text	Date	Link	Category
0 Bekal "Strategic Foresight" untuk Para Capres	Pandangan jauh ke depan (foresight) perlu dimiliki oleh para Capres-Cawapres, karena bangsa kita berada pada situasi dunia yang semakin rentan dan kompleks.	detikNewsKam, 02 Nov 2023 11:15 WIB	https://news.detik.com/berita-ekonomi-bisnis/7013788/bekal-s-...	Politik
1 Mendid Zuhra "Jenderal Pengaruh Geopolitik Ter..."	Menteri Pertahanan RI Zulkri Hasan menanggapi...	detikFinanceKam, 02 Nov 2023 11:15 WIB	https://finance.detik.com/berita-ekonomi-bisnis/7015092/kapan-...	Politik
2 Kapan Pengumuman TNAU Presiden "Gawat" di Jawa...	Ketua Harian Partai Gerindra Dasco mengatakan...	detikNewsKam, 02 Nov 2023 11:11 WIB	https://news.detik.com/berita-ekonomi-bisnis/7015092/kapan-...	Politik
3 Relawan Pengusaha Muda Donor (REPMAS) Indon...	Relawan Pengusaha Muda Nasional (REPMAS) Indon...	detikNewsKam, 02 Nov 2023 11:04 WIB	https://news.detik.com/berita-ekonomi-bisnis/7015074/relawan-...	Politik
4 Pinda Nababan Sesi Cawapres yang Mubah Berhika...	Politikus Senior PDI Perjuangan Pinda Nababan...	detikDumbarKam, 02 Nov 2023 10:25 WIB	https://www.detik.com/dumbar/berita-ekonomi-bisnis/70149...	Politik
...	...	...	...	...
2155 Warga Sarungbang Tokak Pambongnan PDI, Kam...	Warga Sarungbang, Kota Pambongnan, Kota Pambongnan...	detikNewsKam, 07 Okt 2023 17:02 WIB	https://news.detik.com/berita-ekonomi-bisnis/6970285/warga-...	Outraga
2156 Pinda Minto Bantam Foto Peresmian SYL, Gm P...	Pinda Minto Bantam, Jember, dalam foto peresmian...	detikDumbarKam, 07 Okt 2023 16:49 WIB	https://news.detik.com/dumbar/berita-ekonomi-bisnis/6970285/warga-...	Outraga
2157 Gelar World Walking Day, Walikota Tangerang Kut...	Pemerintah Kota (Pemkot) Tangerang menggelar...	detikNewsKam, 07 Okt 2023 16:00 WIB	https://news.detik.com/berita-ekonomi-bisnis/6970285/warga-...	Outraga
2158 Punya Gien Pemecut Alzheimer, Chris Hemsworth Ma...	Hemsworth disebut memiliki risiko tinggi...	detikHealthKam, 07 Okt 2023 16:00 WIB	https://news.detik.com/berita-ekonomi-bisnis/6970285/warga-...	Outraga
2159 Gelpay Hari Jantung Sedunia, Bupati Majalengka...	Senam jantung sedunia, pemeriksaan kesehatan gra...	detikJatimKam, 07 Okt 2023 15:09 WIB	https://news.detik.com/jatim/berita-ekonomi-bisnis/6970138/g-...	Outraga

2160 rows * 5 columns

Gambar 4. Dataset Valid

Title	Text	Date	Link	Category
0 [SALAM] Perhelatan Puncak Bantam Formasi E Se...	Hasil Perhelatan Puncak Bantam Formasi E Se...	June 17, 2023	https://humbachhoax.id/2023/06/17/hasil-perhelatan-puncak-bantam-formasi-e-se-...	Outraga
1 [SALAM] 1480 atlet akan berangkat ke Jakarta untuk mengikuti pertandingan...	Informasi palsu, Bantam atlet akan berangkat ke Jakarta untuk mengikuti pertandingan...	September 22, 2020	https://humbachhoax.id/2020/09/22/1480-atlet-akan-berangkat-ke-jakarta-untuk-mengikuti-pertandingan-...	Outraga
2 [SALAM] Formasi E TNAU Presiden "Gawat" di Jawa...	Informasi palsu, Bantam atlet akan berangkat ke Jakarta untuk mengikuti pertandingan...	June 13, 2022	https://humbachhoax.id/2022/06/13/hasil-perhelatan-puncak-bantam-formasi-e-se-...	Outraga
3 [SALAM] Formasi E TNAU Presiden "Gawat" di Jawa...	Informasi palsu, Bantam atlet akan berangkat ke Jakarta untuk mengikuti pertandingan...	June 6, 2022	https://humbachhoax.id/2022/06/06/hasil-perhelatan-puncak-bantam-formasi-e-se-...	Outraga
4 [SALAM] Bantam atlet akan berangkat ke Jakarta untuk mengikuti pertandingan...	Bukan bantam atlet, yang diberikan kepada PDI...	April 1, 2022	https://humbachhoax.id/2022/04/01/hasil-perhelatan-puncak-bantam-formasi-e-se-...	Outraga
...	...	...	...	...
2147 [SALAM] Sarungbang Tokak Pambongnan PDI, Kam...	Debunk on berita informal yang menyebutkan b...	June 2, 2018	https://humbachhoax.id/2018/06/02/sarungbang-tokak-pambongnan-pdi-kam-...	Outraga
2148 [SALAM] Mendid Zuhra "Jenderal Pengaruh Geopolitik Ter..."	Debunk on berita bantam dari Legenda sepakba...	June 29, 2018	https://humbachhoax.id/2018/06/29/mendid-zuhra-jenderal-pengaruh-geopolitik-ter-...	Outraga
2149 [SALAM] Wapacut Pemecut Alzheimer, Chris Hemsworth Ma...	Wapacut Pemecut Alzheimer, Chris Hemsworth Ma...	June 29, 2018	https://humbachhoax.id/2018/06/29/wapacut-pemecut-alzheimer-chris-hemsworth-ma-...	Outraga
2150 [SALAM] Foto wanita dan pria yang menunjukkan foto yang...	"Seorang pengemudi Reddy yang menunjukkan foto yang...	June 25, 2018	https://humbachhoax.id/2018/06/25/seorang-pengemudi-reddy-yang-menunjukkan-foto-yang-...	Outraga
2151 [SALAM] Bupati Edikan Relawan Pengusaha Muda Donor (REPMAS) Indon...	Postingan surat edaran yang berkaitan informas...	May 26, 2018	https://humbachhoax.id/2018/05/26/bupati-edikan-relawan-pengusaha-muda-donor-repmas-indon-...	Outraga

2152 rows * 5 columns

Gambar 5. Dataset Hoax

3. Proses Pelabelan Data

Proses pelabelan data adalah tahapan dimana setiap entitas dalam dataset diberikan label atau kategori berdasarkan karakteristik atau atribut tertentu. Proses pelabelan data dapat mencakup penentuan kategori atau atribut yang sesuai dengan tujuan penelitian. Dalam identifikasi berita hoaks, setiap artikel berita dapat diberi label "valid" dengan kode "1" jika terverifikasi kebenarannya, atau "hoaks" dengan kode "0" jika terdapat indikasi informasi palsu [17].

Tabel 1. Pelabelan pada indeks berita

Berita	Is_fake
--------	---------

Bekal "Strategic Foresight" untuk Para Capres Pandangan jauh ke depan (foresight) perlu dimiliki oleh para Capres-Cawapres, karena bangsa kita berada pada situasi dunia yang semakin rentan dan kompleks.	1
Debunk ini berisi bantahan dari Legenda sepakbola Argentina, Diego Armando Maradona yang diisukan meninggal dunia akibat serangan jantung usai laga krusial antara Argentina kontra Nigeria pada perhelatan piala dunia 2018. Dirinya juga mengadakan sayembara dengan memberi imbalan sekitar 150 juta rupiah bagi siapa yang bisa mengungkap pelaku penyebaran informasi tersebut.	0

4. Preprocessing Teks

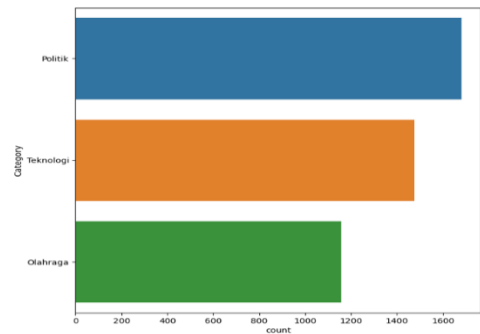
Preprocessing teks merupakan serangkaian tahapan krusial dalam menyiapkan data teks untuk analisis atau pembuatan model, dan melibatkan beberapa teknik esensial seperti case folding, tokenisasi, normalisasi, dan eliminasi stopwords. Case folding mengubah seluruh teks menjadi huruf kecil atau besar agar konsisten. Tokenisasi memecah teks menjadi unit-unit kecil seperti kata-kata atau frasa.

Tabel 2. Hasil Preprocessing

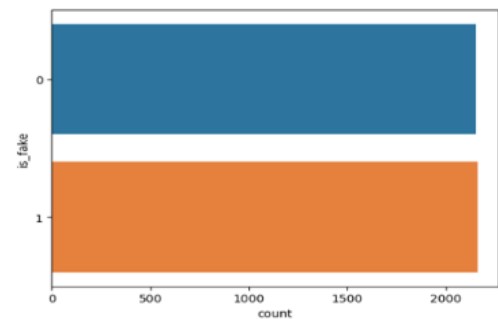
Berita	Berita Preprocessing	Is_fake
Bekal "Strategic Foresight" untuk Para Capres Pandangan jauh ke depan (foresight) perlu dimiliki oleh para Capres-Cawapres,	bekal strategic foresight capres pandangan foresight dimiliki capres cawapres bangsa situasi dunia rentan kompleks.	1

karena bangsa kita berada pada situasi dunia yang semakin rentan dan kompleks.		
Debunk ini berisi bantahan dari Legenda sepakbola Argentina, Diego Armando Maradona yang diisukan meninggal dunia akibat serangan jantung usai laga krusial antara Argentina kontra Nigeria pada perhelatan piala dunia 2018. Dirinya juga mengadakan sayembara dengan memberi imbalan sekitar 150 juta rupiah bagi siapa yang bisa mengungkap pelaku penyebaran informasi tersebut.	debunk berisi bantahan legenda sepakbola argentina diego armando maradona yang diisukan meninggal dunia serangan jantung argentina nigeria perhelatan piala dunia 2018 mengadakan sayembara memberi imbalan sekitar 150 juta rupiah mengungkap pelaku penyebaran informasi.	0

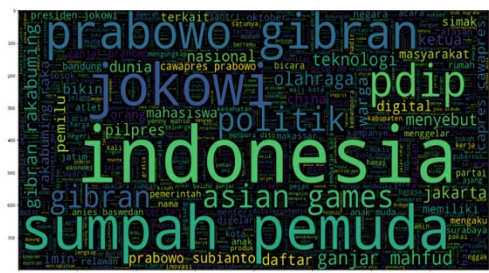
lanjut. EDA melibatkan berbagai metode eksplorasi data, termasuk visualisasi dan ringkasan statistik, untuk mengidentifikasi pola, tren, anomali, serta hubungan antar variabel.



Gambar 6. Jumlah Setiap Category



Gambar 7. Jumlah Data Hoax dan Valid



Gambar 8. Wordcloud Berita Valid

5. Analisis Data Eksploratif

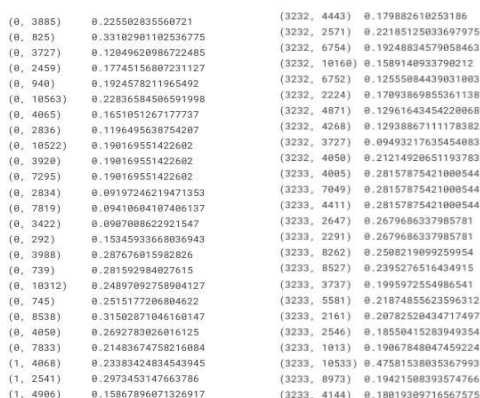
Analisis Data Eksploratif (EDA) adalah suatu pendekatan untuk merinci dan memahami struktur dan karakteristik data sebelum melakukan analisis statistik atau pemodelan lebih



Gambar 9. Wordcloud Berita Hoax

6. Ekstraksi Fitur dengan TF-IDF

Proses ekstraksi fitur dengan TF-IDF pada berita hoaks menghasilkan matriks numerik yang mencerminkan signifikansi setiap kata dalam setiap dokumen. Bobot ini dapat dijadikan fitur input dalam model machine learning untuk tugas seperti deteksi berita palsu, klasifikasi hoaks, atau analisis kecenderungan penyebaran informasi palsu.



Gambar 10. Hasil pembobotan dengan TF-IDF

7. Hasil model machine learning

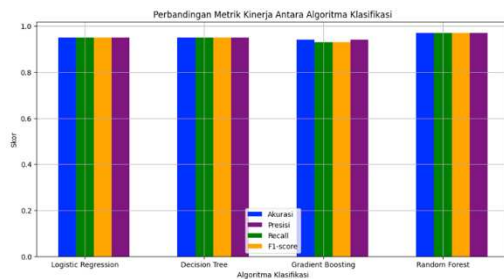
Beberapa algoritma populer untuk pelatihan model termasuk. Logistic Regression, Decision Tree, Gradient Boosting dan Random Forest. Setelah pelatihan, model diuji dengan menggunakan dataset uji untuk mengevaluasi kinerjanya. Pelatihan model machine learning adalah langkah penting dalam pengembangan solusi untuk berbagai tugas seperti klasifikasi, regresi, clustering, dan lainnya, dan

memungkinkan aplikasi yang berdasarkan pemahaman pola dari data yang diberikan.

Tabel 3. Evaluasi Model

Model	Label	Precision	Recall	F1-score	Akurasi
Logistic Regression	0	0.99	0.91	0.94	0.95
	1	0.92	0.99	0.95	
Decision Tree	0	0.93	0.96	0.95	0.95
	1	0.96	0.94	0.95	
Gradient Boosting	0	0.97	0.89	0.93	0.94
	1	0.91	0.97	0.94	
Random Forest	0	0.96	0.97	0.96	0.97
	1	0.97	0.97	0.97	

Berdasarkan evaluasi model dalam table 3 evaluasi model, dapat disimpulkan bahwa setiap model mengalami peningkatan akurasi yang signifikan. Model Random Forest menunjukkan kinerja terbaik dengan akurasi mencapai 0.97. Saat mempertimbangkan nilai precision dan recall pada kelas hoax, model Logistic Regression terbukti lebih superior daripada model lainnya. Dalam konteks klasifikasi berita hoax, perhatian khusus perlu diberikan pada nilai recall untuk kelas hoax. Model Logistic Regression menonjol dengan nilai recall sebesar 0.99. Oleh karena itu, dapat disimpulkan bahwa model Random Forest memimpin dalam akurasi dan F1-score, sementara model Logistic Regression unggul dalam Precision dan Recall.



Gambar 11. Perbandingan rata-rata algoritma klasifikasi

Berdasarkan analisis pada gambar 11 yang menunjukkan perbandingan rata-rata performa algoritma klasifikasi, dapat disimpulkan bahwa algoritma terbaik adalah Random Forest dengan rata-rata mencapai 0.97. Nilai ini menunjukkan bahwa.

```
# Model Testing
def output_label(x):
    if x == 0:
        return "Fake News"
    elif x == 1:
        return "Not A Fake News"

def manual_testing(news):
    testing_news = ("Class: {}".format(news))
    new_def_test = pd.DataFrame(testing_news)
    new_x_test = new_def_test["Class"].values
    new_y_test = vectorization.transform(new_x_test)
    pred_LR = LR.predict(new_y_test)
    pred_DT = DT.predict(new_y_test)
    pred_GB = GB.predict(new_y_test)
    pred_RF = RF.predict(new_y_test)

    return print("Logit Prediction: ( ) \nDT Prediction: ( ) \nGB Prediction: ( ) \nRF Prediction: ( )".format(output_label(pred_LR[0]),
    output_label(pred_DT[0]),
    output_label(pred_GB[0]),
    output_label(pred_RF[0])))

news = str(input())
manual_testing(news)

# Hasil "Strategic Foresight" untuk Para Capres Pandangan jauh ke depan (foresight) perlu dilakukan oleh para Capres-Capres, karena bangsa kita berada di
ada ancaman dunia yang semakin nyata dan kompleks.
```

Gambar 12. Model testing data valid

Berdasarkan pada gambar 12 model testing data valid, dengan melakukan test dengan salah satu indeks data untuk menguji seberapa akurat model yang di uji, dapat dilihat bahwa sampel data yang digunakan adalah data valid, dengan penjelasan dari setiap algoritma memberikan hasil 'NOT A FAKE NEWS'. Menandakan bahwa hasil uji yang kita lakukan telah berhasil.

V. Kesimpulan

Pengembangan model machine learning untuk mengidentifikasi berita hoaks adalah langkah krusial dalam mengatasi penyebaran informasi palsu di era digital. Metodologi melibatkan pengumpulan data, preprocessing teks, ekstraksi fitur dengan TF-IDF, dan

penerapan algoritma klasifikasi seperti Logistic Regression, Decision Tree, Gradient Boosting, dan Random Forest.

Proses tersebut mencakup pengumpulan dataset, pelabelan data, prapemrosesan teks, analisis data eksploratif, dan ekstraksi fitur menggunakan TF-IDF. Penerapan algoritma klasifikasi dan evaluasi dengan metrik standar seperti akurasi, presisi, recall, dan F1-score memberikan gambaran kinerja model dalam mengidentifikasi berita hoaks.

Analisis hasil bertujuan untuk memahami kinerja model, mengidentifikasi kesalahan, dan menganalisis fitur-fitur berpengaruh. Keseluruhan proses ini diharapkan dapat memberikan kontribusi dalam mengembangkan solusi cerdas untuk melawan penyebaran berita hoaks dan membangun lingkungan informasi yang lebih dapat dipercaya di era digital.

Ucapan Terima Kasih

Penulis ingin menyampaikan kan ucapan terima kasih yang tulus kepada Bapak Sasa Ani Armono atas bimbingan dan dukungan luar biasa dalam penyusunan skripsi ini. Bapak Sasa Ani Armono telah memberikan pandangan yang berharga, arahan yang jelas, serta wawasan mendalam terkait Penerapan Metode Machine Learning dalam Mengidentifikasi Berita Hoaks. Bimbingan Bapak sangat berarti dalam memandu langkah-langkah penelitian, memberikan solusi terhadap kendala-kendala yang muncul, dan memberikan inspirasi untuk terus berkembang dalam dunia pengetahuan. Terima kasih atas kesabaran, dedikasi, dan ilmu yang telah Bapak bagikan selama proses penulisan skripsi ini. Semoga hasil penelitian ini dapat memberikan kontribusi positif dan menjadi bahan referensi yang bermanfaat.

Daftar Pustaka

- [1] A. Rahmadhany, A. Aldila Safitri, and I. Irwansyah, "Fenomena Penyebaran Hoax

- dan Hate Speech pada Media Sosial,” *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 3, no. 1, pp. 30–43, 2021, doi: 10.47233/jteksis.v3i1.182.
- [2] F. Rahutomo, I. Y. R. Pratiwi, and D. M. Ramadhani, “Eksperimen Naïve Bayes Pada Deteksi Berita Hoax Berbahasa Indonesia,” *J. Penelit. Komun. DAN OPINI PUBLIK*, vol. 23, no. 1, Jul. 2019, doi: 10.33299/jpkop.23.1.1805.
- [3] S. D. Winarta, G. J. Aqsha, and F. A. P. Anoraga, “Komunikasi Perubahan Sosial Masyarakat Dalam Menerima Berita Hoax Di Era Digital,” *Pros. Semin. Nas. Ilmu Ilmu Sos.*, pp. 194–202, 2022.
- [4] A. Martelli, A. M. Ravenscroft, S. Holden, and P. McGuire, *Python in a Nutshell*. “O’Reilly Media, Inc.,” 2015. doi: 10.1002/9781119038016.app1.
- [5] R. Diana, H. Warni, and T. Sutabri, “Penggunaan Teknoogi Machine Laerning untuk Pelayanan Monitoring Kegiatan Belajar Mengajar pada SMK Bina Sriwijaya Palembang,” *J. Tek. Inform.*, vol. 5, no. 1, pp. 41–50, 2017, [Online]. Available: <https://jurnal.stmik-dci.ac.id/index.php/jutekin/article/view/709/630>
- [6] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.
- [7] M. A. Abubakar, M. Muliadi, A. Farmadi, R. Herteno, and R. Ramadhani, “Random Forest Dengan Random Search Terhadap Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung,” *J. Inform.*, vol. 10, no. 1, pp. 13–18, 2023, doi: 10.31294/inf.v10i1.14531.
- [8] E. Ditendra, S. Suryani, S. Romelah, M. H. Arsyiddik Tanjung, and M. Sarah, “Perbandingan Algoritma Klasifikasi untuk Analisis Sentimen Islam Nusantara di Indonesia,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 2, no. 1, pp. 71–77, 2022, doi: 10.57152/malcom.v2i1.199.
- [9] I. M. A. Purniawan, G. M. A. Sasmita, and I. P. A. E. Pratama, “Clustering Berita Menggunakan Algoritma Tf-Idf Dan K-Means Dengan Memanfaatkan Sumber Data Crawling Pada Situs Detik.Com,” *JITTER J. Ilm. Teknol. dan Komput.*, vol. 3, no. 1, p. 821, 2022, [Online]. Available: <http://dx.doi.org/10.24843/jtrti.2022.v03.i01.p18%0Ahttps://ojs.unud.ac.id/index.php/jitter/article/download/82983/43030>
- [10] J. E. Widayaya and S. Budi, “Pengaruh Preprocessing Terhadap Klasifikasi Diabetic Retinopathy dengan Pendekatan Transfer Learning Convolutional Neural Network,” *J. Tek. Inform. dan Sist. Inf.*, vol. 7, no. 1, pp. 110–124, 2021, doi: 10.28932/jutisi.v7i1.3327.
- [11] Baiq Nurul Azmi, Arief Hermawan, and Donny Avianto, “Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver,” *JTIM J. Teknol. Inf. dan Multimed.*, vol. 4, no. 4, pp. 281–290, 2023, doi: 10.35746/jtim.v4i4.298.
- [12] Q. R. Cahyani *et al.*, “Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using Logistic Regression Algorithm Article Info ABSTRAK,” *JOMLAI J. Mach. Learn. Artif. Intell.*, vol. 1, no. 2, pp. 2828–9099, 2022, doi: 10.55123/jomlai.v1i2.598.
- [13] S. A. Arnomo, “Analisa Decision Tree untuk Kepuasan Penggunaan Sinyal dari Base Transceiver Station (BTS),” *J. Sist. dan Teknol. Inf.*, vol. 9, no. 2, p. 199, 2021, doi: 10.26418/justin.v9i2.43425.
- [14] Sri Diantika, Hiya Nalatissifa, Riki Supriyadi, Nurlaelatul Maulidah, and Ahmad Fauzi, “Implementation of Multi-Class Gradient Boosting To Classify Animal Species in Zoos,” *Antivirus J. Ilm. Tek. Inform.*, vol. 17, no. 1, pp. 33–40, 2023, doi: 10.35457/antivirus.v17i1.2812.
- [15] N. G. Ramadhan, F. D. Adhinata, A. J. T. Segara, and D. P. Rakhmadani, “Deteksi Berita Palsu Menggunakan Metode Random Forest dan Logistic Regression,” *JURIKOM (Jurnal Ris. Komputer)*, vol. 9,

- no. 2, p. 251, Apr. 2022, doi: 10.30865/jurikom.v9i2.3979.
- [16] D. Agustriawan, "Penerapan Pendekatan Machine Learning Pada Pengembangan Basis Data Herbal Sebagai Sumber Informasi Kandidat Obat Kanker," *J. Teknol. Ind. Pertan.*, vol. 29, no. 2, pp. 175–182, 2019, doi: 10.24961/j.tek.ind.pert.2019.29.2.175.
- [17] P. Klasifikasi *et al.*, "JEPIN (Jurnal Edukasi dan Penelitian Informatika)".