

FAKE NEWS PREDICTION

Fake News & True News dataset

PRESENTED BY I MADE BAYU S.W

About Me

I Made Bayu Satriai Wadhana

Data Science Enthusiast

Computer Engineering graduate with hands-on experience in data analysis, cloud computing, and web development. Passionate about uncovering insights through data and transforming information into impactful decisions.

Education Background :

Telkom University (2020-2024)

Computer engineering graduate (GPA : 3.39/4.00)

Dbimbing Data Scientist Bootcamp (2025-Now)

Learned data science methodologies, Python, and data visualization.

Working Experience :

Web Developer Intern PT.Indonusa Multijaya (june 2023 - august 2023)

Create a website for company profiling





Background

Penyebaran fake news di era digital semakin meningkat dan berdampak luas pada masyarakat. Berita palsu sering kali sulit dibedakan dengan berita asli karena menggunakan gaya bahasa yang mirip, sehingga analisis manual menjadi tidak efisien.

Untuk itu, diperlukan pendekatan otomatis yang mampu mengklasifikasikan berita secara cepat dan akurat. Dengan memanfaatkan Natural Language Processing (NLP) dan Machine Learning, proyek ini bertujuan membangun sistem deteksi berita palsu sehingga informasi dapat diverifikasi secara lebih efektif dan andal.

Tabel of Content

O1 Business Problem

O2 Project Vision And Mission

O3 Data Understanding

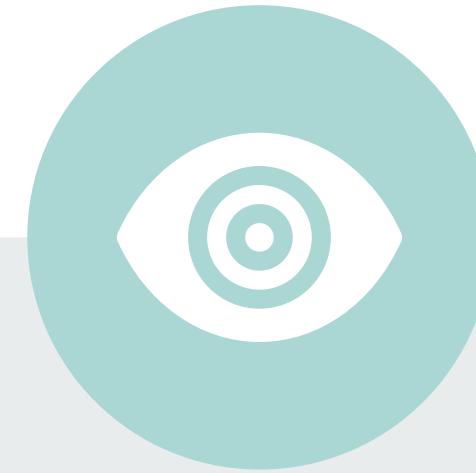
O4 Data Cleaning & Reprocessing

O5 Exploratory Data Analyst

O6 Machine Learning Model

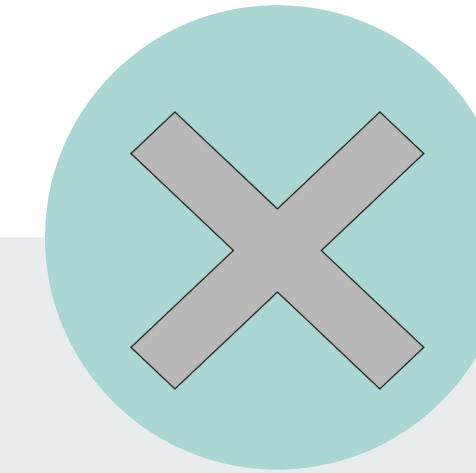
O7 Conclusion & Business Recommendation

Business Problem



Penyebaran Berita Palsu

Maraknya fake news di media online menyesatkan masyarakat, memengaruhi opini publik, dan berpotensi menimbulkan instabilitas sosial maupun politik.



Verifikasi Manual Tidak Efisien

Pemeriksaan berita satu per satu membutuhkan waktu lama dan sumber daya besar, sehingga sulit diimplementasikan pada arus informasi yang sangat cepat.

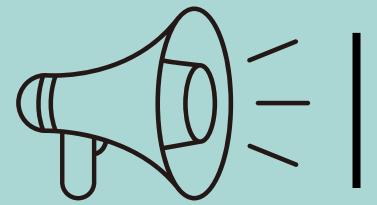


Dampak pada Kepercayaan Publik

Tingginya jumlah berita palsu menurunkan kredibilitas media dan platform digital, serta mengikis kepercayaan masyarakat terhadap informasi yang mereka terima.

Project vision and mission

Menciptakan sistem cerdas berbasis NLP dan Machine Learning untuk mendeteksi berita palsu secara cepat, akurat, dan dapat diandalkan, sehingga membantu menjaga kualitas informasi di era digital.



Mengklasifikasikan berita menjadi True dan Fake secara otomatis.



Membangun model Machine Learning yang optimal melalui preprocessing, feature engineering, dan hyperparameter tuning.



Memberikan insight praktis untuk membantu masyarakat, media, dan platform digital dalam mengurangi penyebaran hoaks.

Data Understanding

Fake News & True News dataset

- **Jumlah data:**

1. Fake news: 23.481 artikel
2. True news: 21.417 artikel

Kolom	Deskripsi
Title	Judul Berita
Text	Isi Berita
Subject	topik berita (misalnya: politics, world, government)
Date	Tanggal publikasi

DATA CLEANING & PREPROCESSING

proses pembersihan dan persiapan data agar siap digunakan dalam analisis dan pemodelan. Langkah ini penting untuk memastikan data yang digunakan berkualitas, konsisten, dan dapat merepresentasikan pola yang ingin dipelajari.



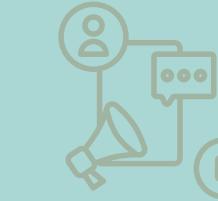
Cek data ada missing atau tidak

Handling Missing Values



melabel berita true = 0 dan fake = 1

Labelling



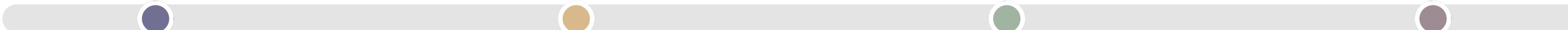
Gabung 2 dataset menjadi 1 (dataset true dan fake)

Concat



Mengacak data agar tidak berurutan

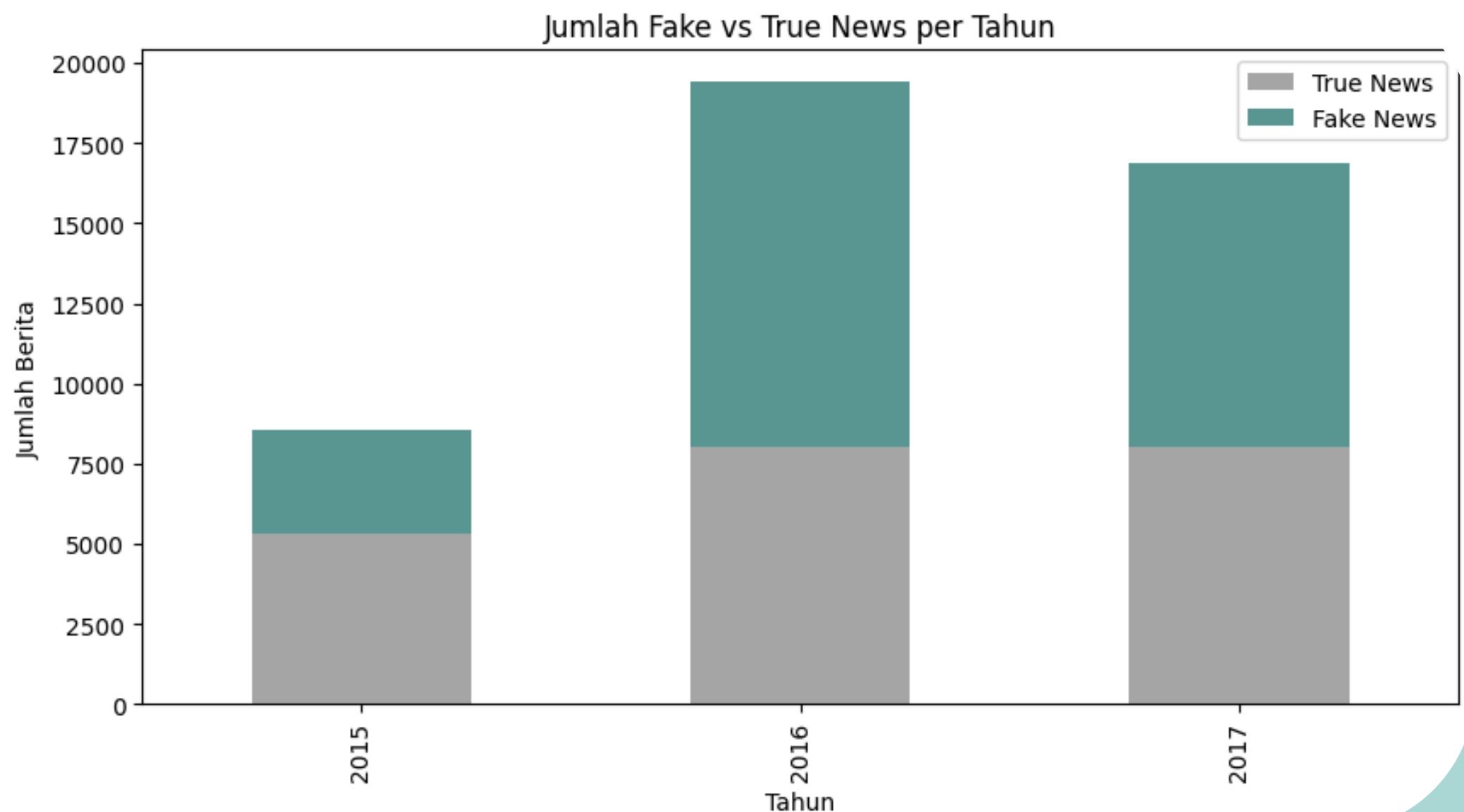
Shuffling



Exploratory Data Analyst

Fake News vs Real News

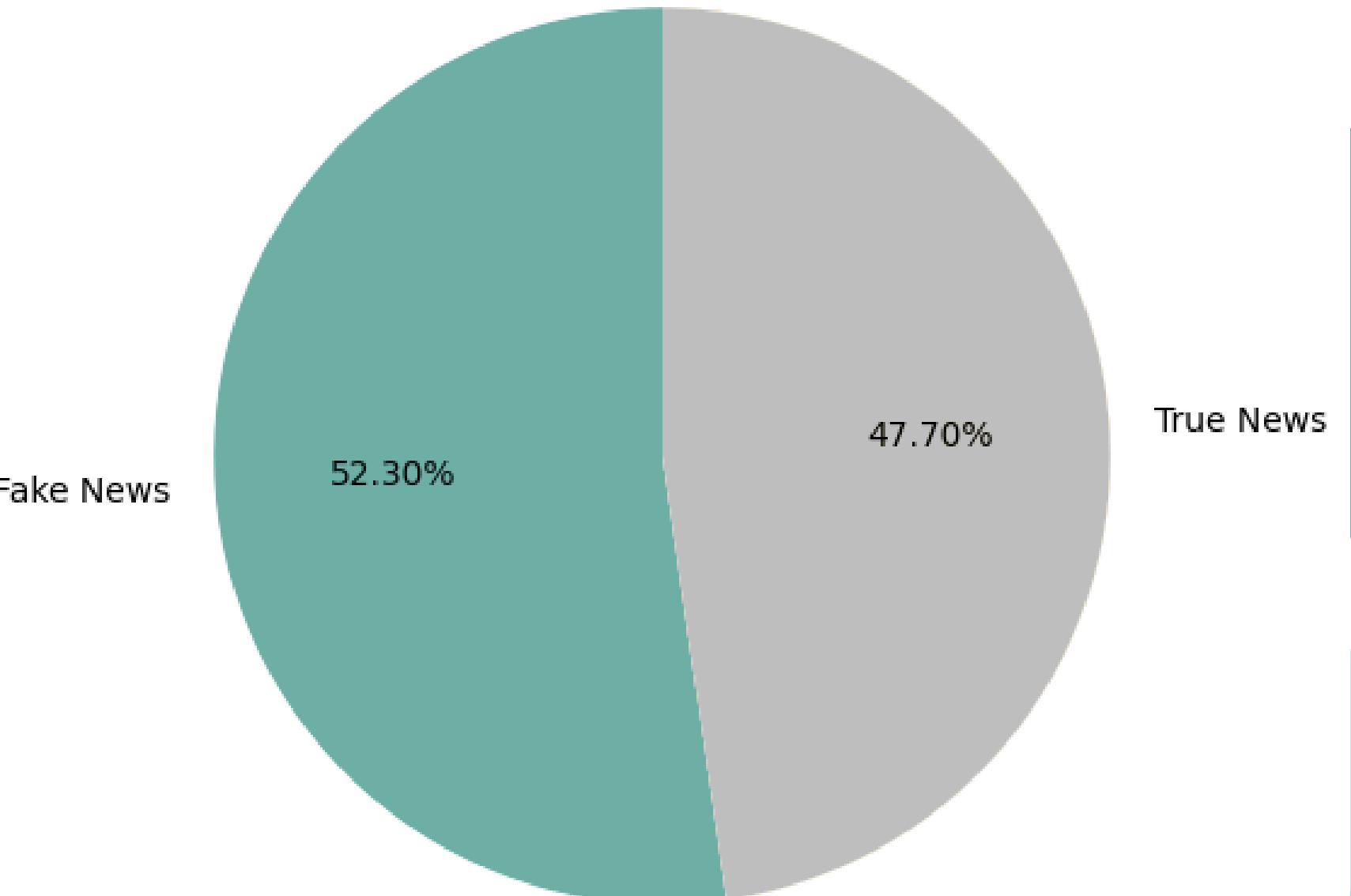
- Data periode 2015–2017 menunjukkan peningkatan signifikan jumlah berita, terutama pada tahun 2016.
- Distribusi True dan Fake relatif seimbang, tetapi terjadi lonjakan besar berita palsu pada 2016–2017.
- Fenomena ini selaras dengan Amerika mengalami maraknya fake news yang berkaitan dengan isu politik dan pemilu presiden 2016.



News Distribution

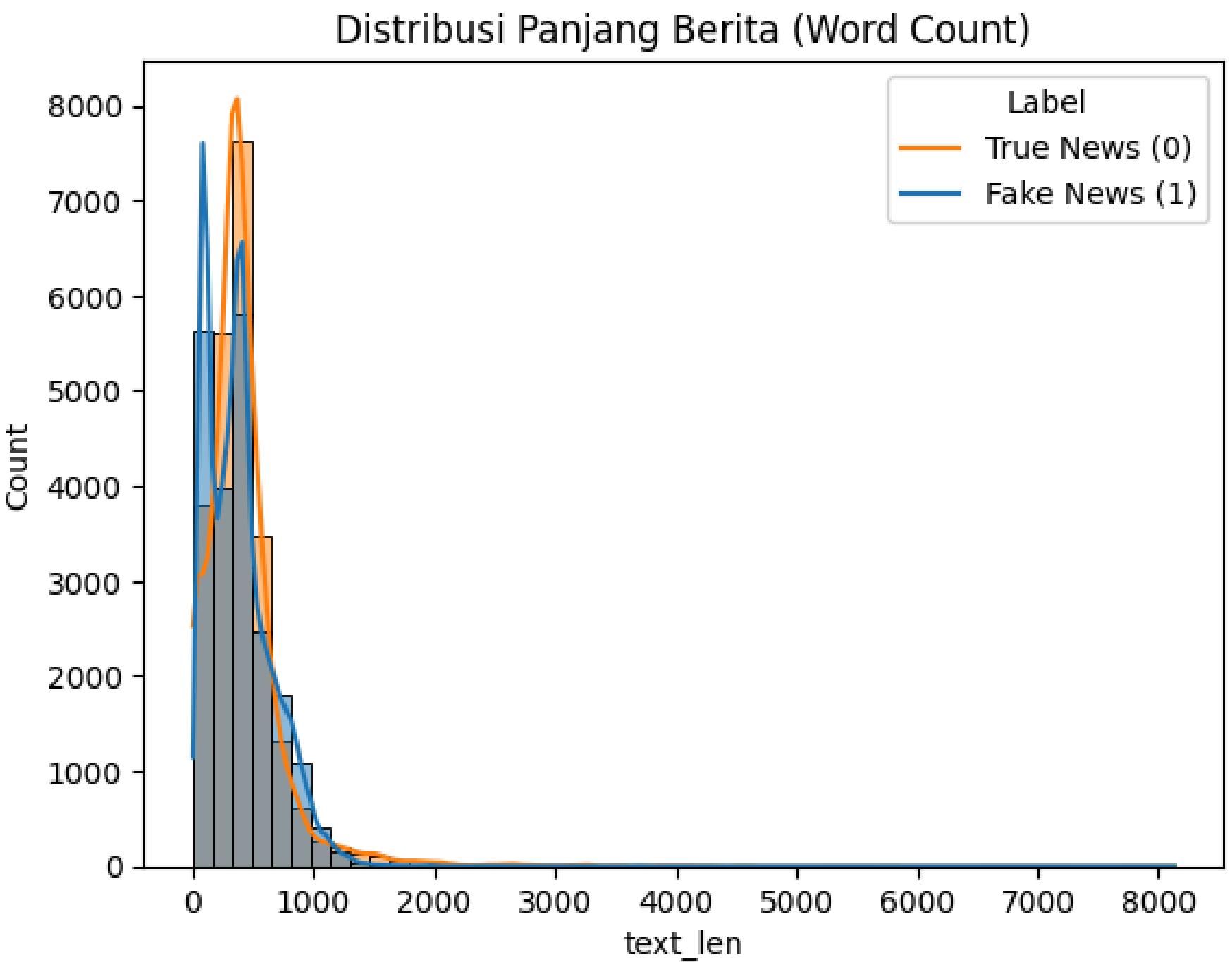
- Dataset berisi 23.481 berita Fake (52.3%) dan 21.417 berita True (47.7%).
- Distribusi cukup seimbang, sehingga model tidak bias ke salah satu kelas.
- Proporsi ini mendukung pelatihan yang adil untuk mendeteksi keduanya.

Distribusi Fake vs True News

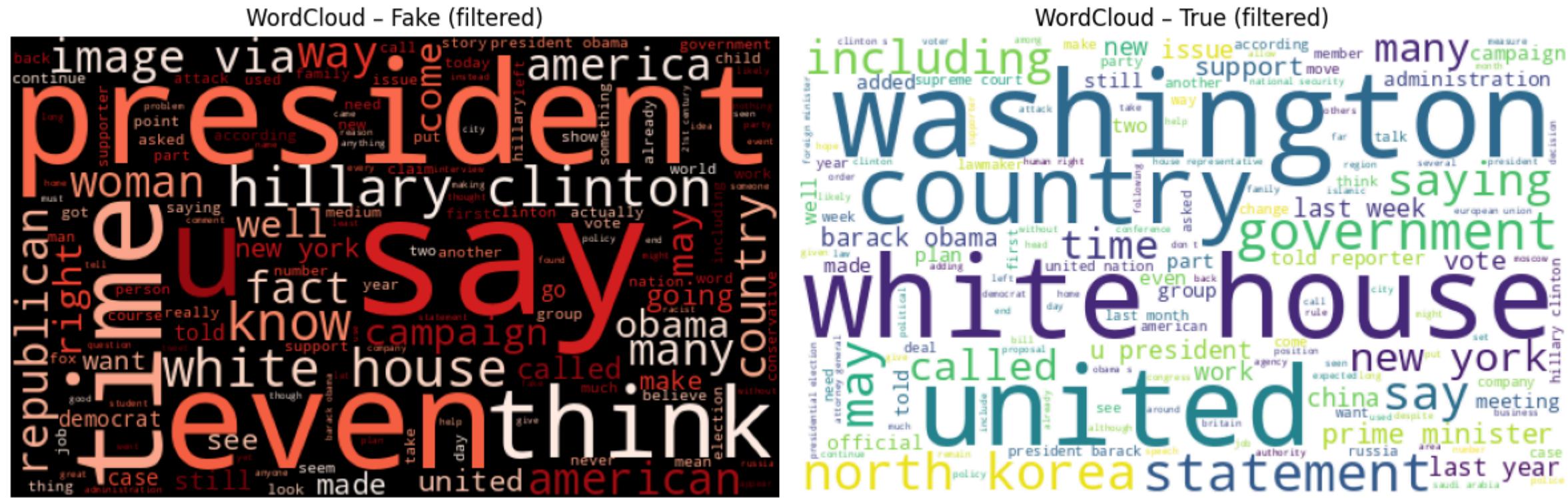


News Word count

- Panjang berita bervariasi dari <100 hingga ribuan kata.
- Mayoritas berada di rentang 200–800 kata.
- Distribusi Fake dan True News mirip, meski Fake News cenderung sedikit lebih pendek.
- Word count dapat membantu model mengenali pola penulisan.



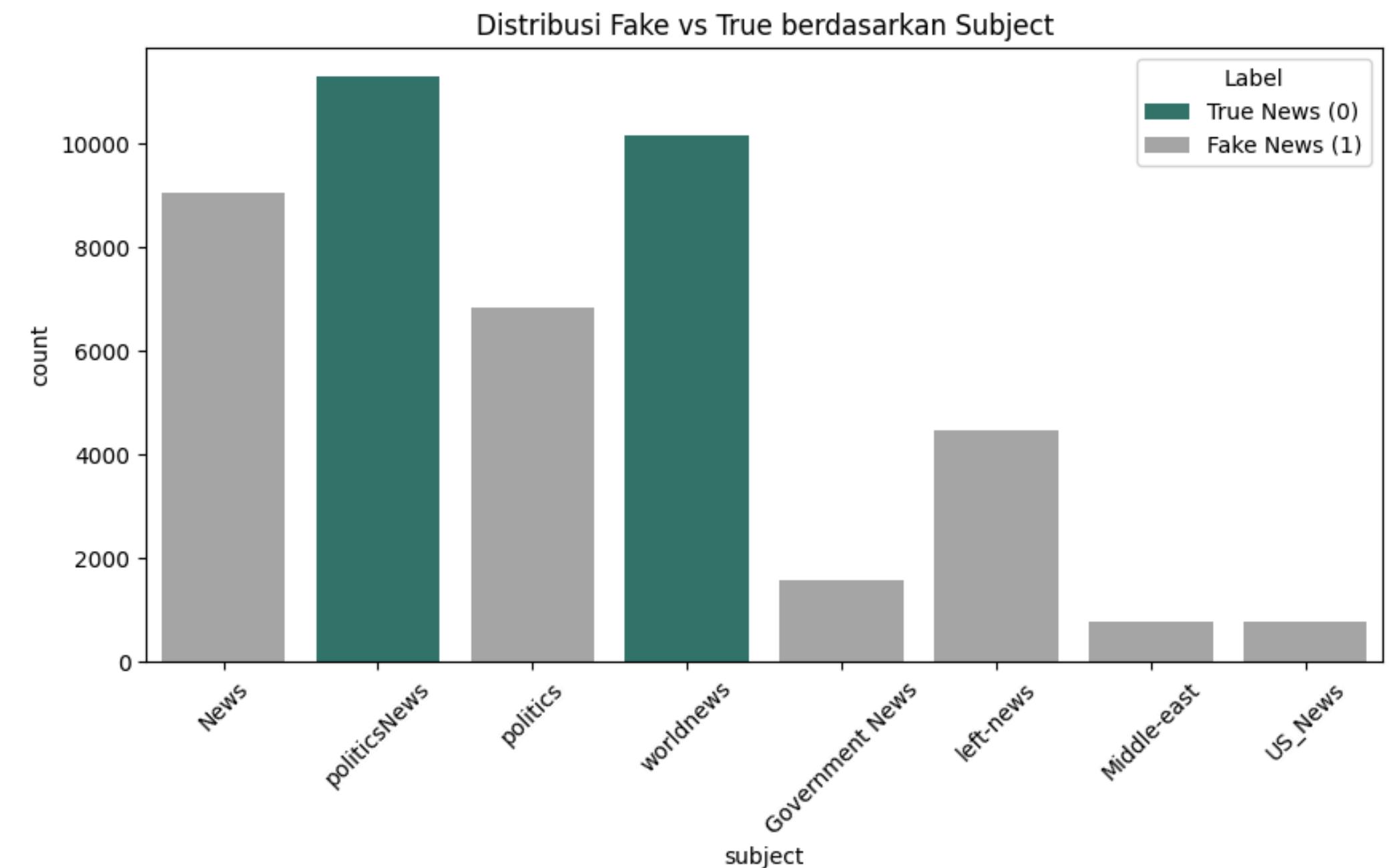
Dominant word



- Fake News banyak menonjolkan kata politik & tokoh (misalnya president, hillary, clinton).
 - True News lebih sering memuat kata institusional & geografis (misalnya washington, united, white house).
 - Ada tumpang tindih kata, tapi konteks pemakaiannya berbeda.
 - Isu politik tampak lebih kuat melekat pada Fake News.

Subject News Distribution

- Mayoritas berita berasal dari kategori Politics dan World News.
- Fake News paling sering muncul pada Political News dan World News.
- True News lebih dominan di kategori News umum.
- Kategori lain (Government, Middle East, US News) jumlahnya relatif kecil.



Feature Engineering

Deleting

Menghapus Stopword, ubah uper case menjadi lowe case, hapus tanda baca

Lemmatization

Mengsimplekan ke bentuk kata pertama kata (running to run)

TF-DIF

Mengurangi bobot kata yang terlalu sering muncul di semua dokumen

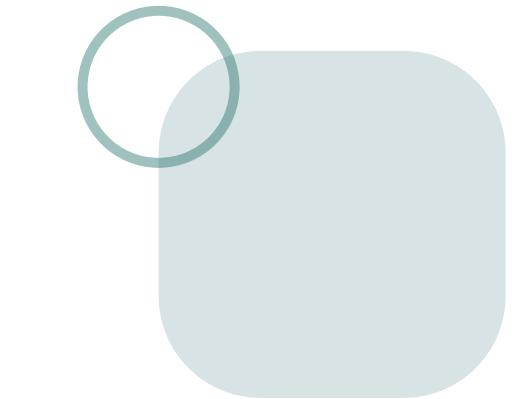
Numerical Feature

karakteristik berita dalam bentuk angka yang bisa membantu model membedakan True vs Fake news.



MECHINE LEARNING

- 01** Logistic Regression
- 02** Naive Bayes
- 03** Linear SVM



MODEL EVALUATION

Test Prediction				
Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	99%	99%	99%	99%
Naive Bayes	89%	94%	85%	90%
Linear SVM	98%	98%	98%	98%

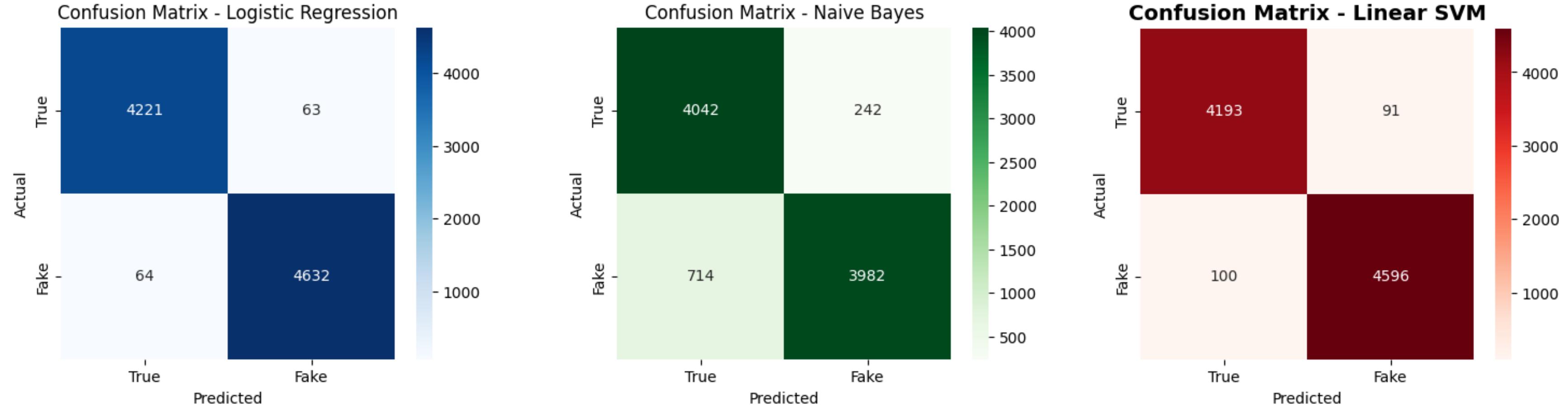
- Logistic Regression, performa terbaik, semua metrik 99%.
- Linear SVM, stabil, seluruh metrik 98%.
- Naive Bayes, lebih rendah (89% akurasi, 94% precision, 85% recall, 90% F1), cocok sebagai baseline ringan.

MODEL EVALUATION

Train Prediction				
Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	99%	99%	99%	99%
Naive Bayes	90%	95%	85%	90%
Linear SVM	98%	98%	98%	98%

- Logistic Regression, performa tertinggi, semua metrik 99%.
- Linear SVM, stabil dengan metrik 98%.
- Naive Bayes, lebih rendah (90% akurasi, 95% precision, 85% recall, 90% F1), cocok sebagai baseline.

CONFUSION MATRIX



Logistic Regression

- Kesalahan sangat kecil = 63 FN, 64 FP
- Prediksi True & Fake seimbang dan akurat

Naive Bayes

- Lebih banyak salah prediksi = 242 FN, 714 FP
- Cepat & ringan, tapi akurasi lebih rendah

Linear SVM

- Stabil dengan kesalahan sedikit lebih tinggi = 91 FN, 100 FP
- Masih konsisten di kedua kelas

LOGISTIC REGRESSION EVALUATION

GridSearch CV

Grindsearch CV(logreg)				
News	F1 Score	Accuracy	Precision	Recall
1 (Fake)	0.9921	0.9918	0.9911	0.9932
0 (True)	0.9914	0.9918	0.9925	0.9902

GridSearchCV pada Logistic Regression menunjukkan performa yang sangat konsisten dengan akurasi 99.18%. Parameter terbaik yang dipilih adalah: C=10, class_weight=balanced, penalty=L2, solver=liblinear. Confusion Matrix menunjukkan distribusi prediksi yang hampir sempurna dengan hanya sedikit kesalahan klasifikasi (42 True = Fake dan 32 Fake = True dari total 8.980 data uji).

LOGISTIC REGRESSION EVALUATION

Probabilitas, ROC/PR, dan tuning threshold

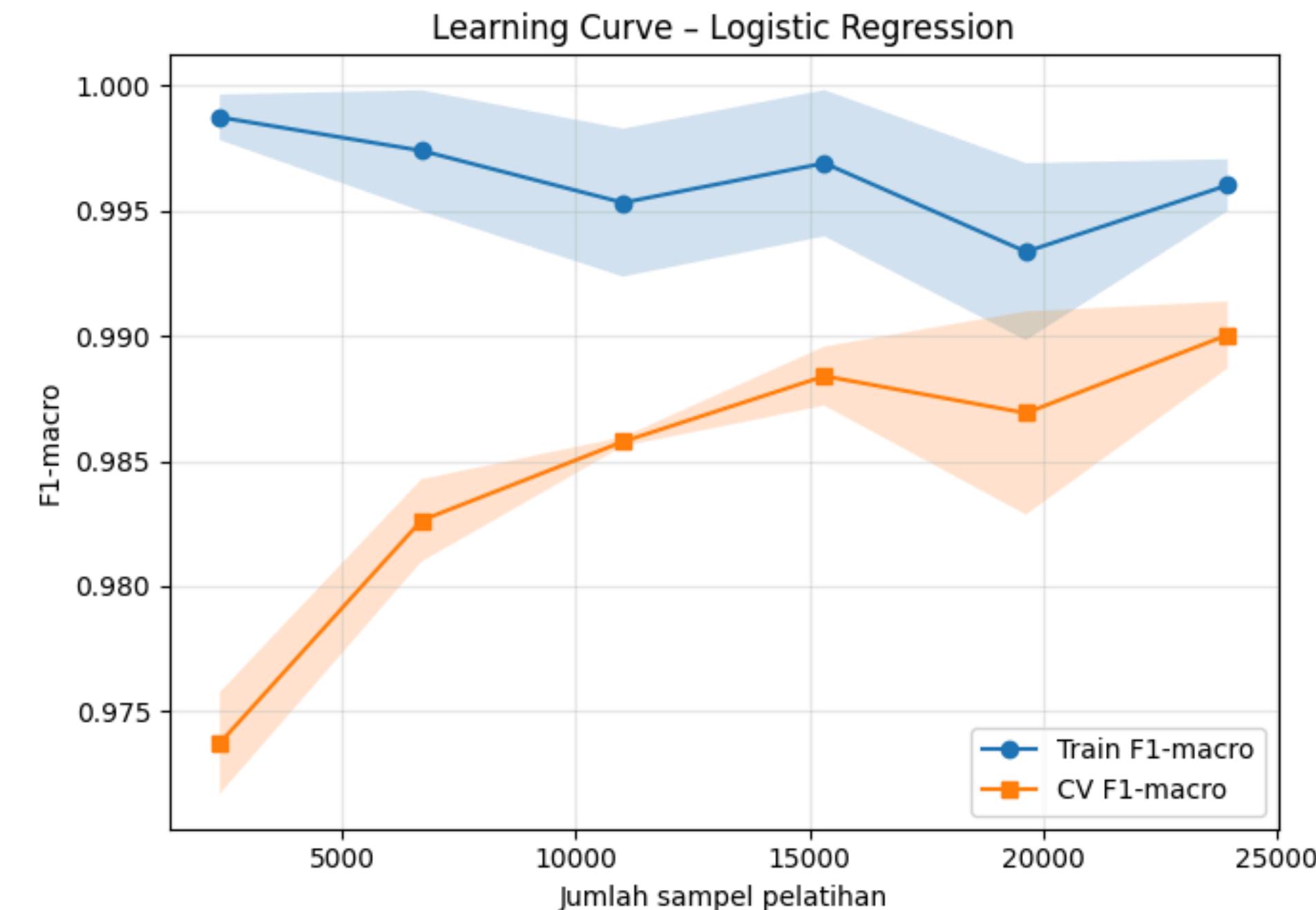
Probabilitas, ROC/PR, dan tuning threshold		
ROC - AUC	PR-AUC	Tuning Threshold
0.99937	0.9994	0.50794

Hasil evaluasi Logistic Regression menunjukkan performa sangat baik, dengan ROC-AUC 0.99937 dan PR-AUC 0.9994 yang menandakan kemampuan hampir sempurna membedakan berita Fake dan True. Threshold optimal di sekitar 0.5079 memberi keseimbangan terbaik antara precision dan recall.

LOGISTIC REGRESSION EVALUATION

Learning Curve

Learning curve menunjukkan Logistic Regression konsisten dengan performa tinggi. F1-macro train stabil mendekati 1.0, sedangkan validasi silang (CV) meningkat hingga stabil di kisaran 0.99. Hal ini menandakan model tidak overfitting maupun underfitting, serta mampu belajar baik meski data diperbesar.

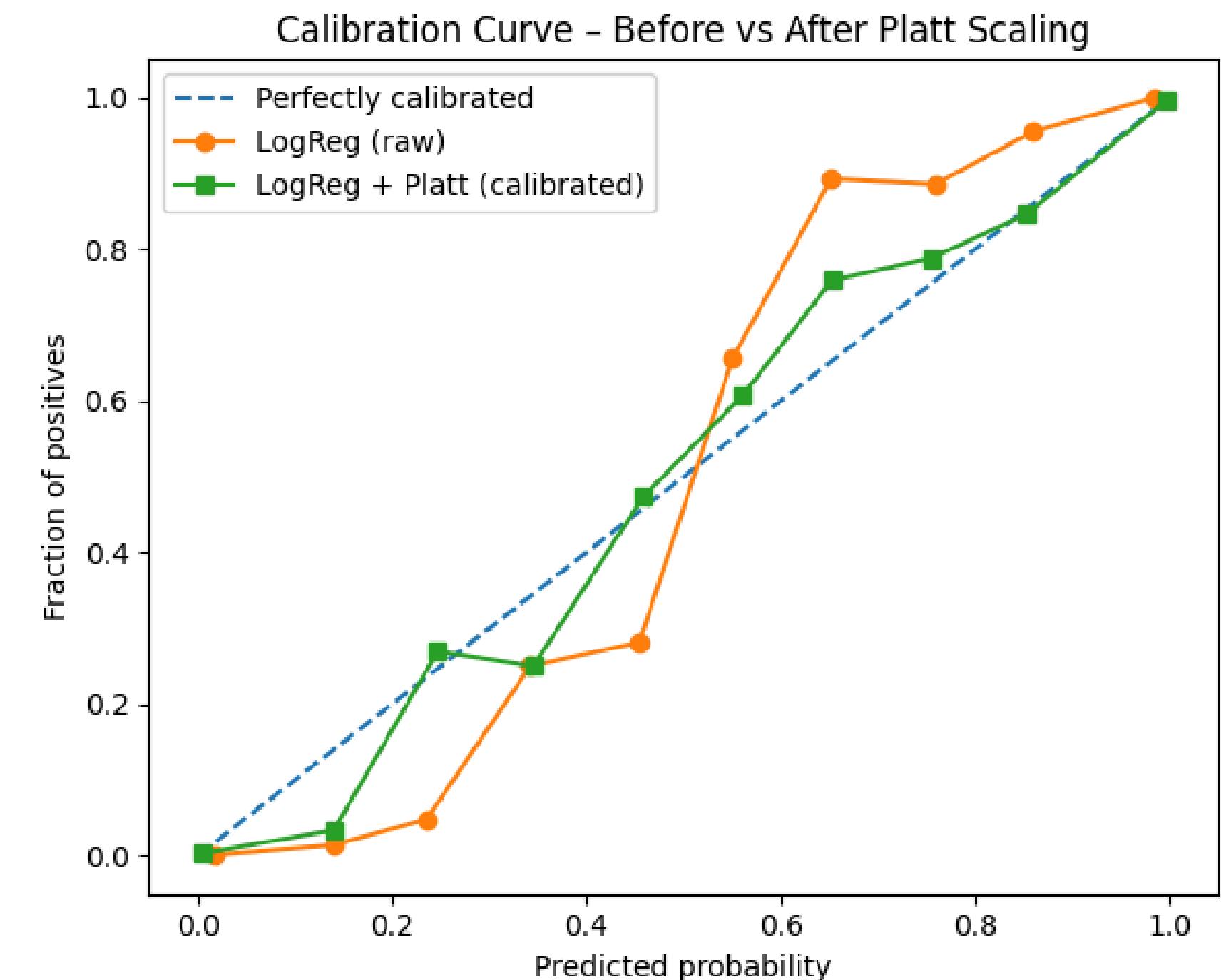


LOGISTIC REGRESSION EVALUATION

Calibration Curve

Condition	Barrier Score
LogReg (Raw)	0.0123
Logreg + Platt	0.0098

Hasil calibration curve menunjukkan bahwa LogReg (Raw) masih agak overconfident dengan Brier Score 0.0123, sehingga probabilitas prediksi sering terlalu tinggi dibandingkan frekuensi aktual. Setelah dilakukan Platt Scaling, performa kalibrasi membaik dengan Brier Score turun menjadi 0.0098, dan kurva hijau terlihat lebih mendekati garis biru (perfectly calibrated).



CONCLUSION

Model we chose and why?

Berdasarkan evaluasi, Logistic Regression menjadi model terbaik dengan akurasi, precision, recall, dan F1 score konsisten di 99%. Nilai ROC-AUC 0.999 dan PR-AUC 0.999 menegaskan kemampuannya membedakan berita asli dan palsu hampir sempurna. Calibration curve juga menunjukkan peningkatan reliabilitas setelah Platt Scaling dengan penurunan Brier Score. Dibandingkan Linear SVM (98%) dan Naive Bayes (90%), Logistic Regression unggul tidak hanya dalam performa klasifikasi, tetapi juga pada probabilitas yang lebih selaras dengan kenyataan, sehingga dipilih sebagai model utama untuk deteksi fake news.

RECOMENDATION BUSINESS ACTION

Implementasi Sistem Deteksi Otomatis

- Integrasikan model Logistic Regression (hasil terbaik) ke platform berita atau media sosial untuk memfilter konten lebih cepat.

Monitoring & Updating Model

- Lakukan retraining model secara berkala dengan data terbaru agar tetap akurat menghadapi pola bahasa dan isu yang terus berubah.

Kolaborasi dengan Stakeholder

- Berikan insight dari model ini kepada jurnalis, platform digital, dan pembuat kebijakan untuk membantu mengurangi penyebaran hoaks.

THANK YOU

FOR YOUR ATTENTION

September 2025



Project



imadebayusatriawardhana@gmail.com



I Made Bayu Satria Wardhana

