KELOMPOK 3

CLUBDEV AI

KLASIFIKASI BERITA HOAX DARI DOKUMEN BERITA US MENGGUNAKAN PENDEKATAN BOOSTING DAN NATURAL LANGUAGE PROCESSING

PROBLEM DESCRIPTION

- Penyebaran berita hoax di Amerika Serikat semakin masif → ribuan artikel/berita palsu beredar di berbagai platform digital.
- Berita hoax sering kali menyerupai berita faktual → masyarakat sulit membedakan informasi benar dengan yang menyesatkan.
- Dampak → keputusan publik, opini politik, dan perilaku sosial dapat terdistorsi.
- Tantangan utama → diperlukan sistem otomatis untuk mengidentifikasi dan mengklasifikasikan berita palsu dengan akurat dan cepat.

TUJUAN

Membangun model machine learning berbasis NLP yang dapat mengklasifikasikan berita sebagai berita asli atau berita hoax secara akurat berdasarkan informasi-informasi yang ada dalam berita.

EDA

Dataset Summary

HOAX

23481 Baris

FACT

21417 Baris

Fitur
title
text
subject
date

Target

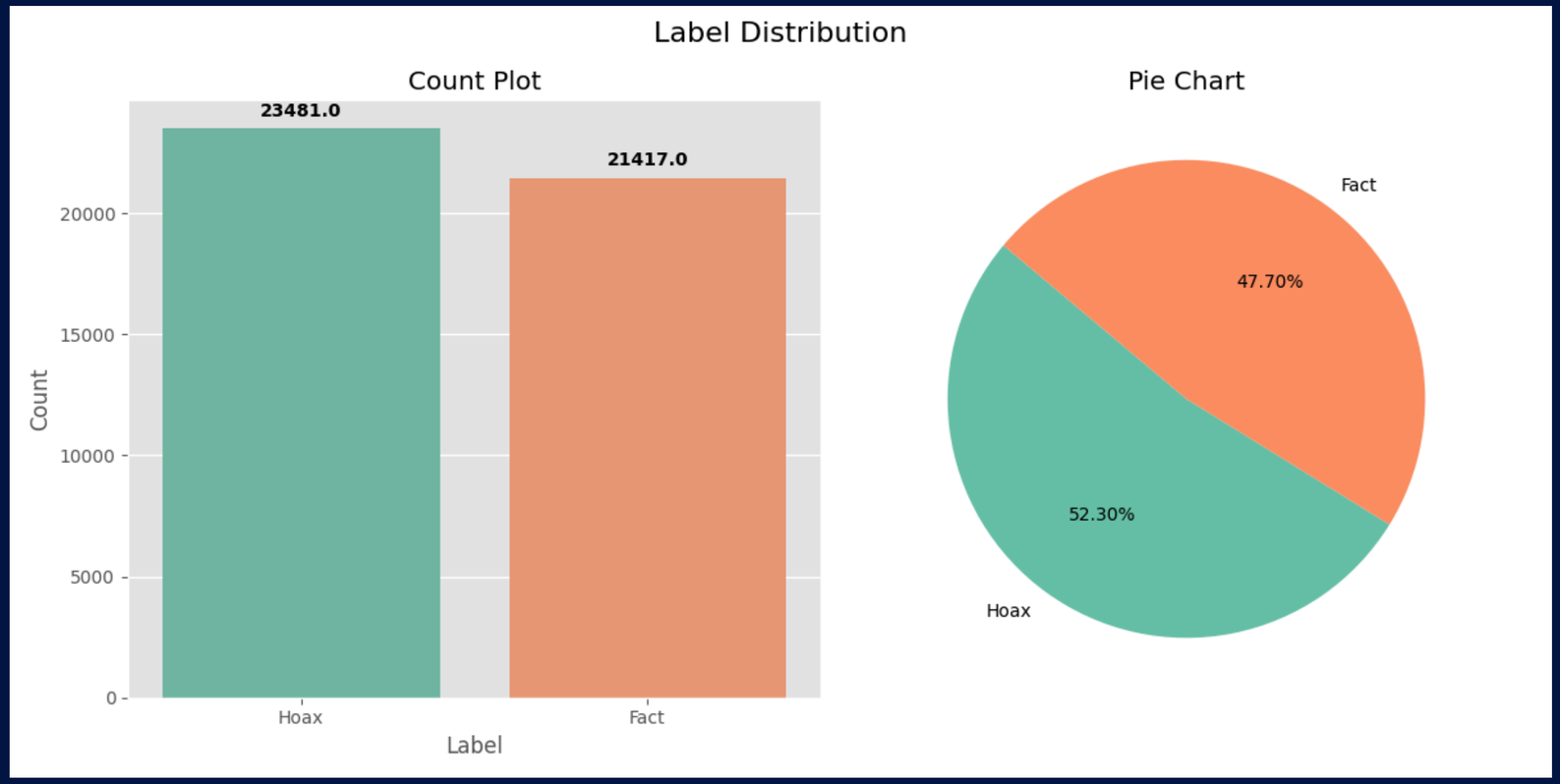
LABEL

(HOAX/TRUE)

EDA

Distribusi Label

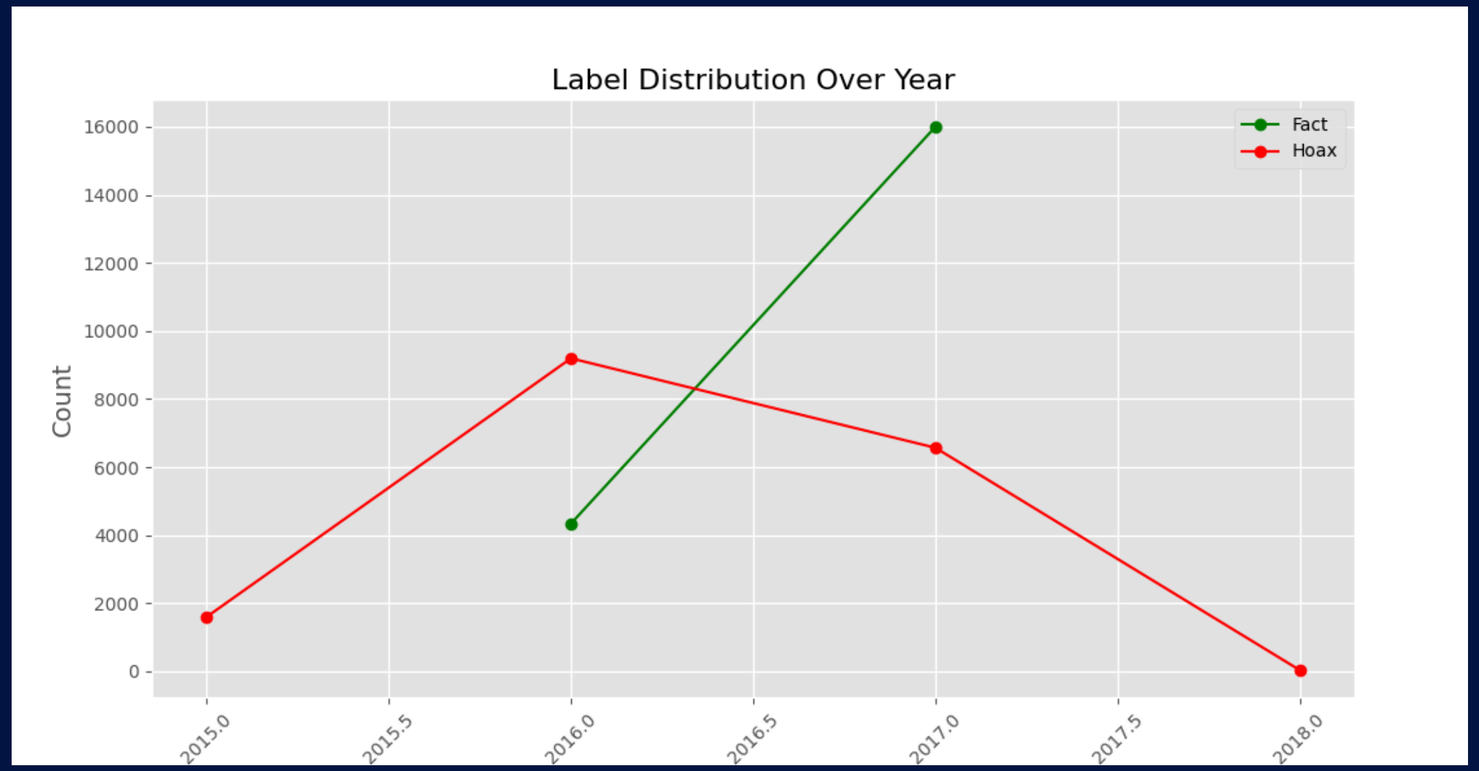
Distribusi label terlihat sedikit imbalance antara label **Hoax** dan label **Fact** di mana label **Hoax** lebih tinggi **4.60%** dibandingkan label **Fact**



EDA

Tren Berita Hoax dan Fact

Berita **hoax** mencapai jumlah tertinggi pada tahun **2016.** Namun, mengalami penurunan signifikan pada tahun-tahun berikutnya hingga tahun **2018**. Sebaliknya, berita **fact** mengalami peningkatan mulai tahun **2016** dan mencapai jumlah tertinggi pada tahun **2017**. Hal ini menunjukkan adanya pergeseran tren seiring waktu

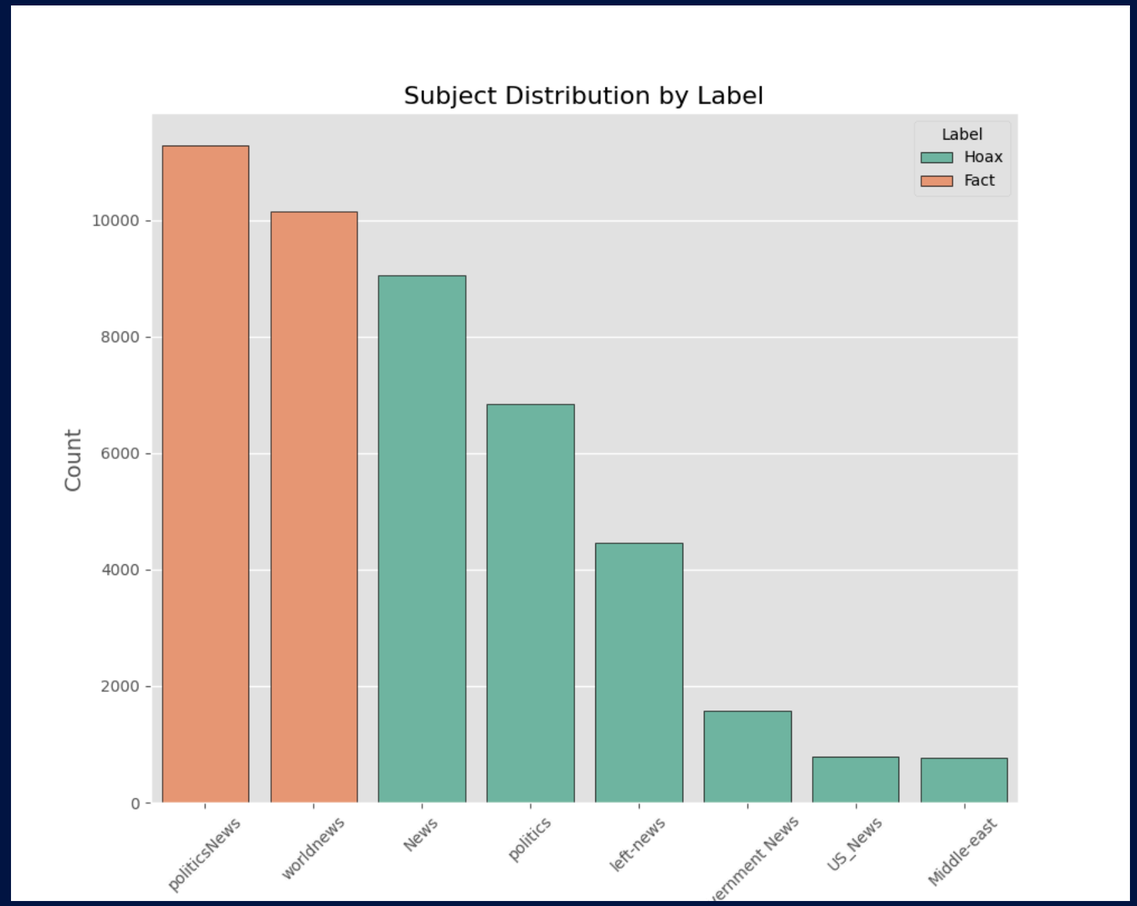


EDA

Distribusi Subject

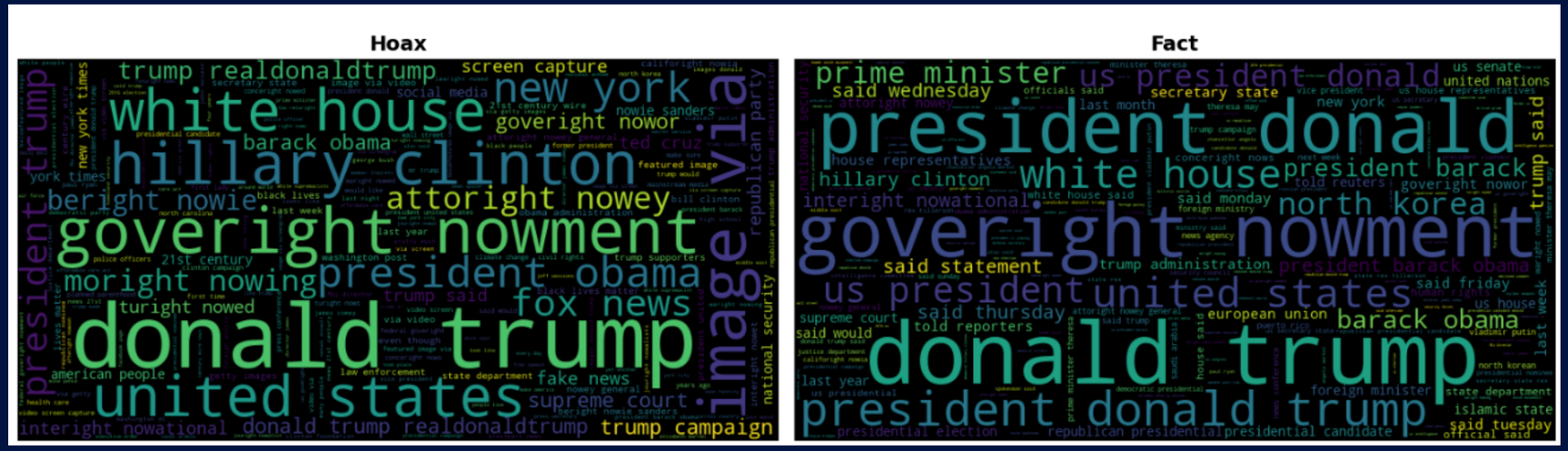
Terlihat bahwa artikel dengan label **Fact** hanya muncul pada kategori **politicsNews** dan **worldnews**.
Hal ini menunjukkan bahwa isu politik dan berita dunia merupakan topik yang lebih sering diberitakan secara faktual. Sementara itu, kategori sisanya berisi artikel dengan label **Hoax**.
Oleh karena fitur ini dapat membuat model machine learning menjadi bias

SOLUTION ⟶ **DROP FITUR**



EDA

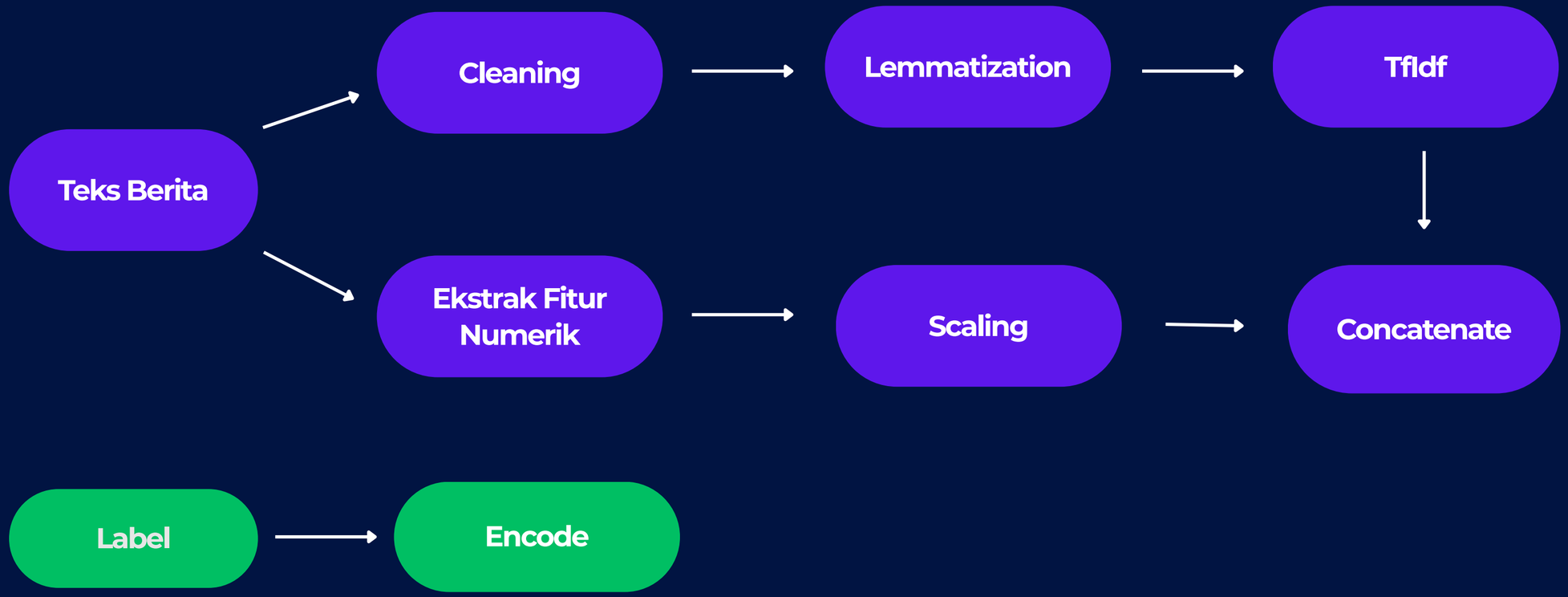
Worcloud Teks per Label



Berita **hoax** maupun **fact** sama-sama banyak membahas topik politik seperti Donald Trump, Hillary Clinton, dan White House. Namun, perbedaan terlihat pada gaya penyajian: **hoax** lebih sering memunculkan istilah yang bersifat sensasional, sedangkan **fact** didominasi oleh bigram formal, seperti *said statement* atau *told reporters* yang menunjukkan kutipan resmi.

DATA PREPARATION

Alur Preprocessing



FEATURE ENGINEERING

Ekstraksi Fitur Numerik

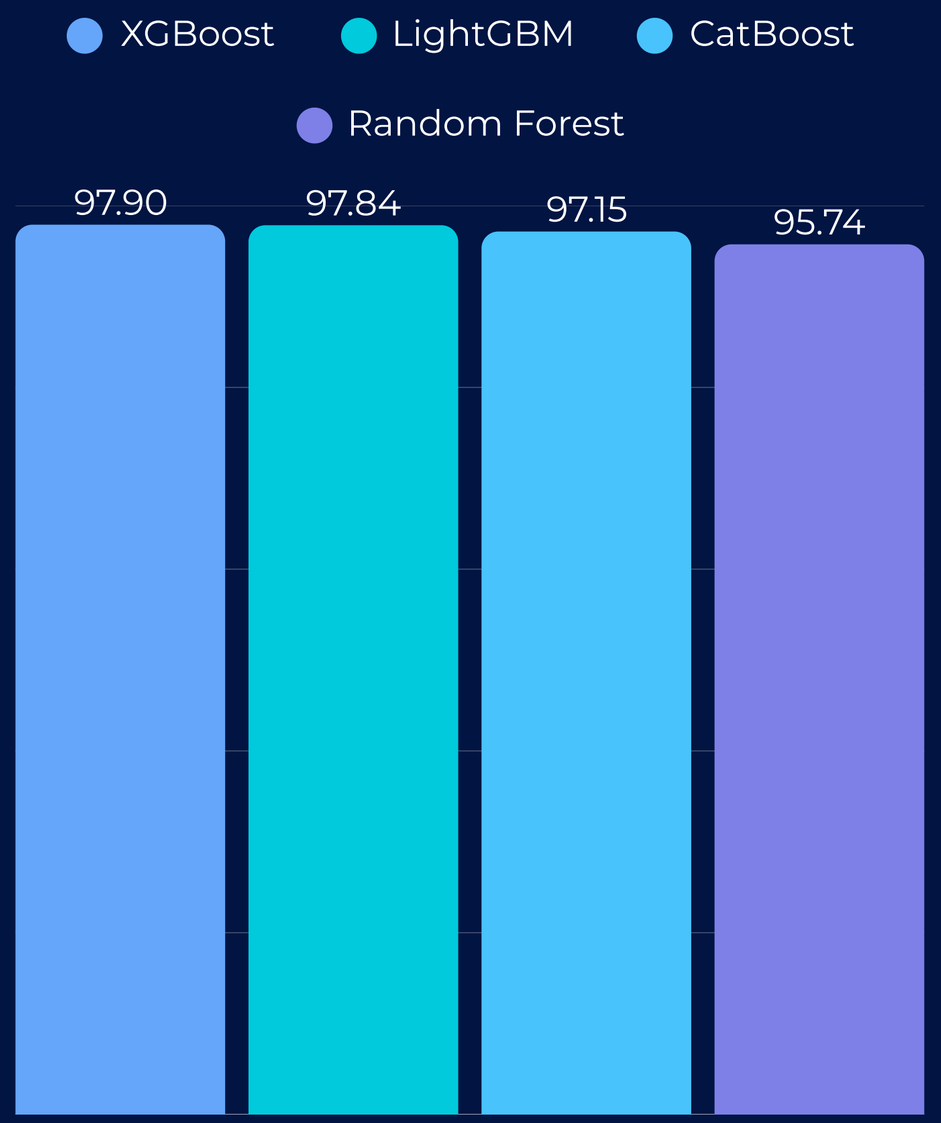
Fitur hasil ekstrak fitur text:

Fitur	Penjelasan	P-Value (T-test)
word_count	Jumlah total kata dalam teks	0
sentence_count	Jumlah total kalimat dalam teks	0
lexical_diversity	Rasio antara jumlah kata unik dengan total jumlah kata	0.017
polarity	Skor sentimen dari -1 (negatif) sampai +1 (positif)	0
subjectivity	Skor dari 0 (objektif) sampai 1 (sangat subjektif)	0

MODELLING

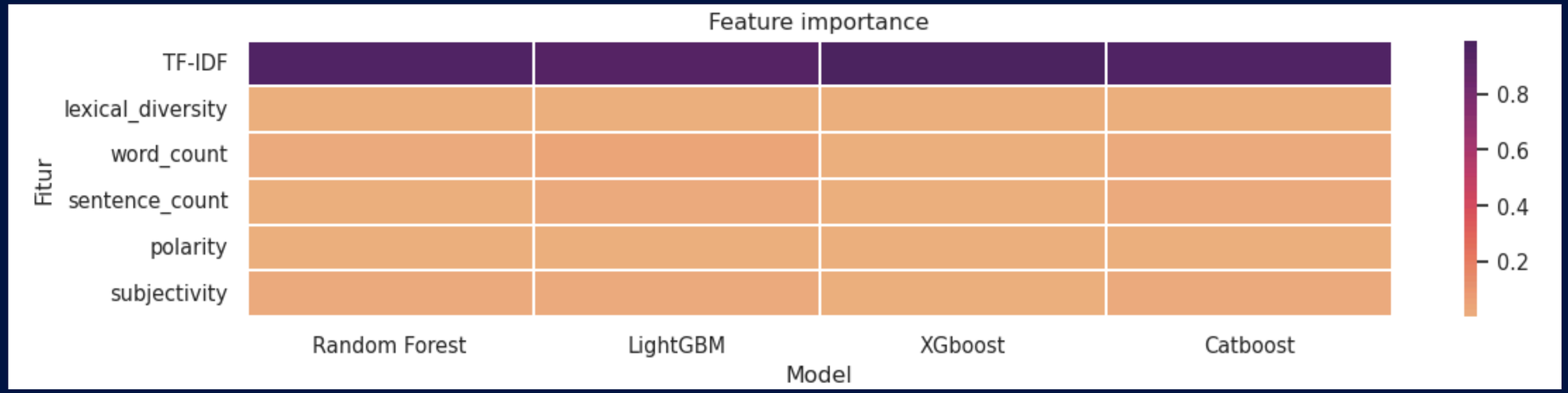
Baseline Comparison

XGBoost memiliki performa metriks terbaik dibandingkan yang lain dengan skor recall **97.90**



EVALUATION

Feature Importance



TF-IDF memiliki pengaruh yang s**ignifikan** untuk keempat model, sementara **Fitur numerik** tidak memiliki pengaruh yang signifikan

CONCLUSION

KEY INSIGHT

- **Semua Model** yang ditrain memiliki performa yang cukup baik dalam membedakan berita **Hoax** dibuktikan dengan skor recall di atas **95%**
- **TF-IDF** memberikan pengaruh **signifikan** untuk semua model.

FUTURE WORK

- **Ekstraksi fitur** lebih bervariasi
- Menggunakan **Embedding** untuk p**reprocessing teks**
- Melakukan **hyperparameter tuning** untuk memperoleh konfigurasi model yang lebih optimal
- Menggunakan teknik ensemble seperti **stacking** atau **voting** untuk meningkatkan stabilitas model
- **Uji coba integrasi** ke dalam platform deteksi berita daring

THANK YOU