EVALUASI PERBANDINGAN PERFORMA MODEL BERT DAN NAÏVE BAYES PADA SENTIMEN ANALISIS PEMILIHAN PRESIDEN INDONESIA 2024

Laporan Project

Pemrosesan Bahasa Alami
Teknik Informatika
Kelas B

OUR TEAM

KELOMPOK 8

Izzat Ikhwan
225150200111010

M Husain F
225150207111027

Gratia Yudika
225150207111015

M Hasan F
225150207111026

Outline

Pembahasan

01 Project Overview

02 Data Understanding

03 Metodologi Penyelesaian

04 Perancangan Flowchart

05 Implementasi dan Pengujian

06 Kesimpulan dan Saran

PROJECT OVERVIEW

Deskripsi

Studi Kasus

Di era digital yang semakin berkembang pesat, **media sosial telah menjadi platform utama** bagi masyarakat dalam **menyuarakan opini dan pandangan** mereka, termasuk dalam konteks politik. **Twitter,** sebagai salah satu platform media sosial terbesar, menjadi wadah **diskusi signifikan** terkait isu-isu politik, termasuk **Pemilihan Presiden Indonesia 2024.** **Analisis sentimen** terhadap tweet-tweet ini menjadi **sangat penting** untuk memahami persepsi publik, mengidentifikasi tren opini, dan memberikan wawasan berharga bagi berbagai pemangku kepentingan dalam proses demokrasi Indonesia.

Dalam studi kasus ini, kami mengembangkan **sistem analisis sentimen** otomatis yang mampu **mengklasifikasikan** tweet terkait kandidat presiden 2024 ke dalam kategori **sentimen positif dan negatif.** Sistem ini dirancang menggunakan pendekatan deep learning dengan model **BERT (Bidirectional Encoder Representations from Transformers),** yang telah terbukti efektif dalam tugas pemrosesan bahasa alami, khususnya analisis sentimen. Sebagai baseline, model tradisional Naive Bayes juga diimplementasikan untuk mengevaluasi keunggulan pendekatan modern dibandingkan metode konvensional.

Rumusan Masalah

1. Bagaimana cara mengembangkan sistem analisis sentimen yang akurat untuk mengklasifikasikan sentimen masyarakat terhadap kandidat presiden 2024 di media sosial Twitter menggunakan model BERT?

2. Seberapa efektif perbandingan performa antara model deep learning BERT dengan model baseline Naive Bayes dalam analisis sentimen tweet?

3. Bagaimana karakteristik dan pola sentimen masyarakat terhadap masing-masing kandidat presiden berdasarkan hasil analisis sentimen menggunakan model yang dikembangkan?

Tujuan Penelitian

1. Mengembangkan dan mengimplementasikan sistem analisis sentimen berbasis deep learning menggunakan model BERT untuk mengklasifikasikan tweet terkait kandidat presiden 2024.

2. Melakukan analisis komparatif performa antara model BERT dan model baseline Naive Bayes dalam tugas klasifikasi sentimen tweet politik.

3. Menganalisis dan menginterpretasikan pola sentimen masyarakat terhadap masing-masing kandidat presiden berdasarkan hasil klasifikasi model yang dikembangkan.

Manfaat Penelitian

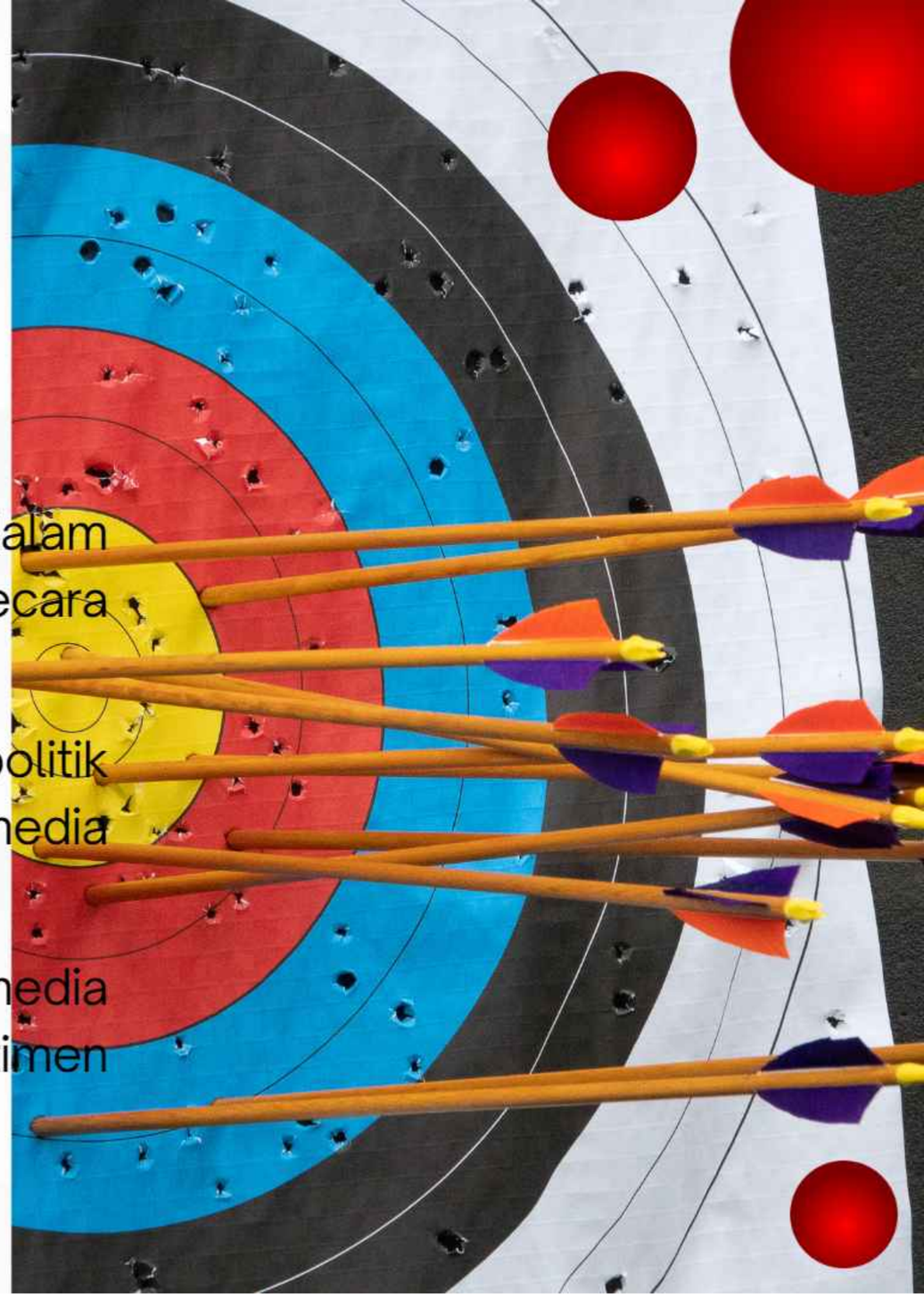
Manfaat Akademis:

- Memberikan kontribusi pada pengembangan metode analisis sentimen, khususnya dalam konteks politik, dengan mengeksplorasi efektivitas model BERT.
- Menyediakan benchmark perbandingan antara pendekatan deep learning modern dengan metode klasik dalam analisis sentimen.
- Menghasilkan dataset tervalidasi yang dapat digunakan untuk penelitian lanjutan dalam bidang NLP.

Manfaat Penelitian

Manfaat Praktis:

- Membantu tim kampanye dan strategi politik dalam memahami persepsi publik terhadap kandidat secara real-time dan objektif.
- Memberikan wawasan bagi media dan analis politik dalam menginterpretasikan tren opini publik di media sosial.
- Mendukung pengembangan sistem monitoring media sosial yang dapat digunakan untuk analisis sentimen dalam berbagai konteks politik di Indonesia.



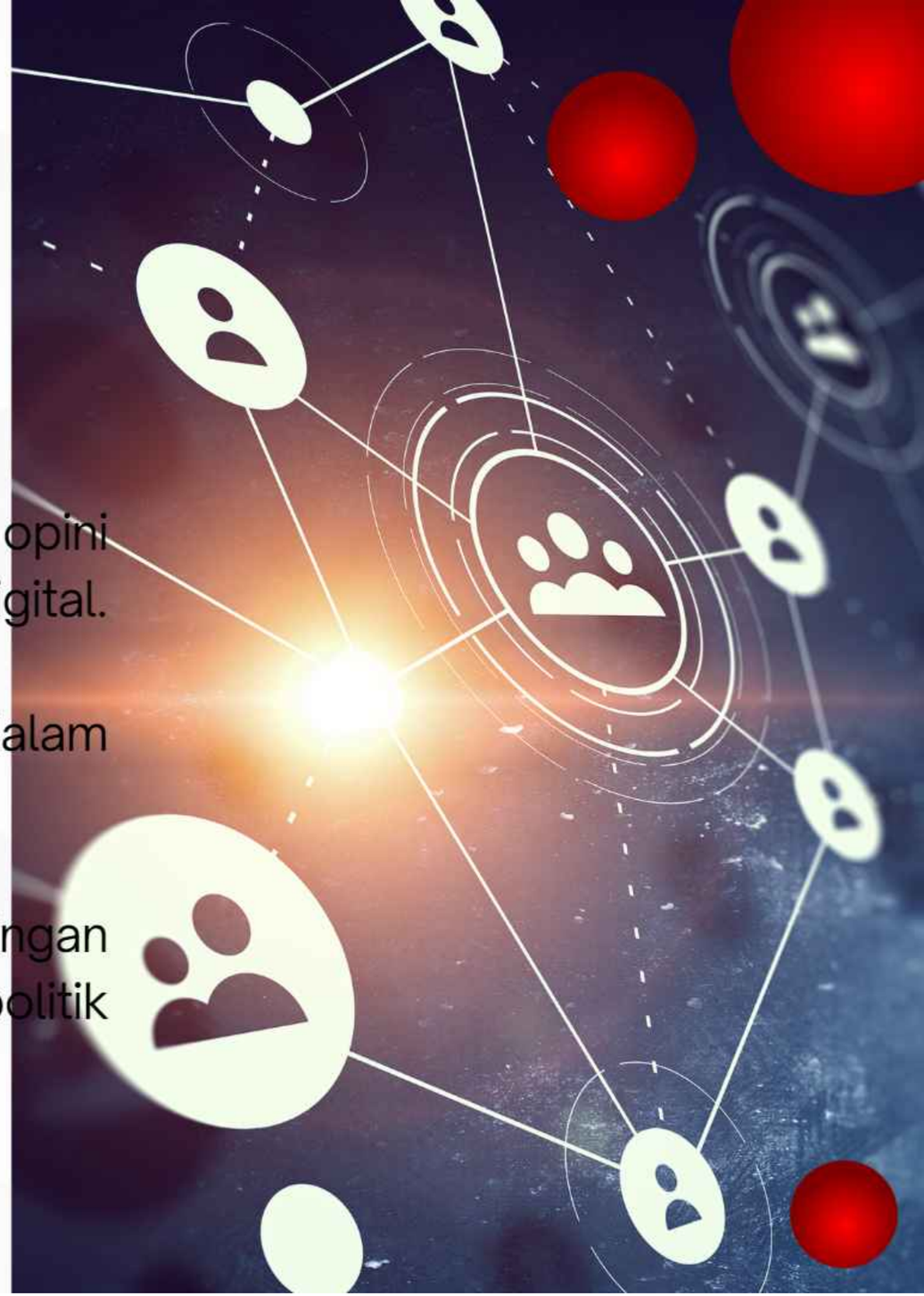
Manfaat Penelitian

Manfaat Sosial:

- Meningkatkan pemahaman tentang dinamika opini publik dalam konteks pemilihan presiden di era digital.

- Mendorong transparansi dan keterbukaan dalam diskusi politik di media sosial.

- Membantu masyarakat dan pemangku kepentingan dalam memahami polarisasi dan tren sentimen politik di media sosial.



FILKOM
FAKULTAS ILMU KOMPUTER

DATA UNDER-STANDING

Data dan Karakteristiknya

Dataset yang digunakan dalam penelitian ini berasal dari "Indonesia Presidential Candidates Dataset, 2024" yang tersedia di Kaggle dan Mendeley. Dataset ini memiliki karakteristik sebagai berikut:

1. Sumber Data

- Platform: Twitter
- Periode pengumpulan:
  - Ganjar Pranowo: Oktober 2022 - April 2023
  - Prabowo Subianto: Desember 2022 - April 2023
  - Anies Baswedan: Januari - April 2023
- Jumlah total tweet: 30.000 (10.000 per kandidat)

2. Cakupan Konten

Dataset mencakup berbagai aspek diskusi publik, seperti:

- Visi dan misi kandidat
- Track record dan pengalaman
- Kebijakan yang diusung
- Respon terhadap isu-isu nasional
- Dukungan dari partai politik dan tokoh publik
- Interaksi dengan masyarakat
- Kampanye dan program kerja
- Kritik dan tanggapan publik

3. Nilai Penelitian

Dataset ini memiliki nilai penting untuk:

- Pemahaman dinamika opini publik
- Identifikasi isu-isu kunci yang mempengaruhi
- persepsi publik
- Evaluasi efektivitas komunikasi politik
- Prediksi tren dukungan publik
- Studi komparatif pasca-pemilihan

Struktur
Dataset

Dataset disimpan dalam format CSV dengan struktur yang kaya informasi:

Kolom-kolom Utama:

- Date: Waktu tweet dibuat
- Created: Waktu akun dibuat
- User ID: Identifikasi unik pengguna
- Followers: Jumlah pengikut akun
- Following: Jumlah yang diikuti akun
- Tweet Count: Jumlah tweet akun
- TweetLocation: Lokasi pengguna (jika tersedia)
- Text: Isi tweet
- label: Kategori sentimen (positif/negatif)
- Candidate: Nama kandidat yang dibahas

Distribusi Dataset

- **Total Records: 30.000 tweet**
- **Distribusi per Kandidat: 10.000 tweet**
- **Distribusi Sentimen (setelah preprocessing):**
  - Positif: ±73% (21.654 tweet)
  - Negatif: ±27% (8.074 tweet)

Link Dataset

JOCELYN DUMLAO · UPDATED 10 MONTHS AGO

32 | New Notebook | Download

Indonesia Presidential Candidate's Dataset, 2024

2024 Indonesia Presidential Candidates: Profiles and Policies Dataset

Data Card | Code (2) | Discussion (0) | Suggestions (0)

About Dataset

Description

The raw data was downloaded by using Python programming with Twitter API. Data comes from the Twitter platform that discusses the issue of Indonesia's 2024 presidential candidate. There are a total of 30,000 data with each candidate breakdown as follows:

- Ganjar Pranowo: October 2022 - April 2023
- Prabowo Subianto: December 2022 - April 2023
- Anies Baswedan: January - April 2023

The data used is data before the determination of presidential candidates, but the topic of the Indonesian presidential election on social media Twitter has been widely discussed. The data obtained will be useful for future research as a comparison in determining the results of the Indonesian Presidential election when compared with data at the time of determining candidates and campaigns as well as data after being elected President (actual results)

Usability
10.00

License
CC0: Public Domain

Expected update frequency
Never

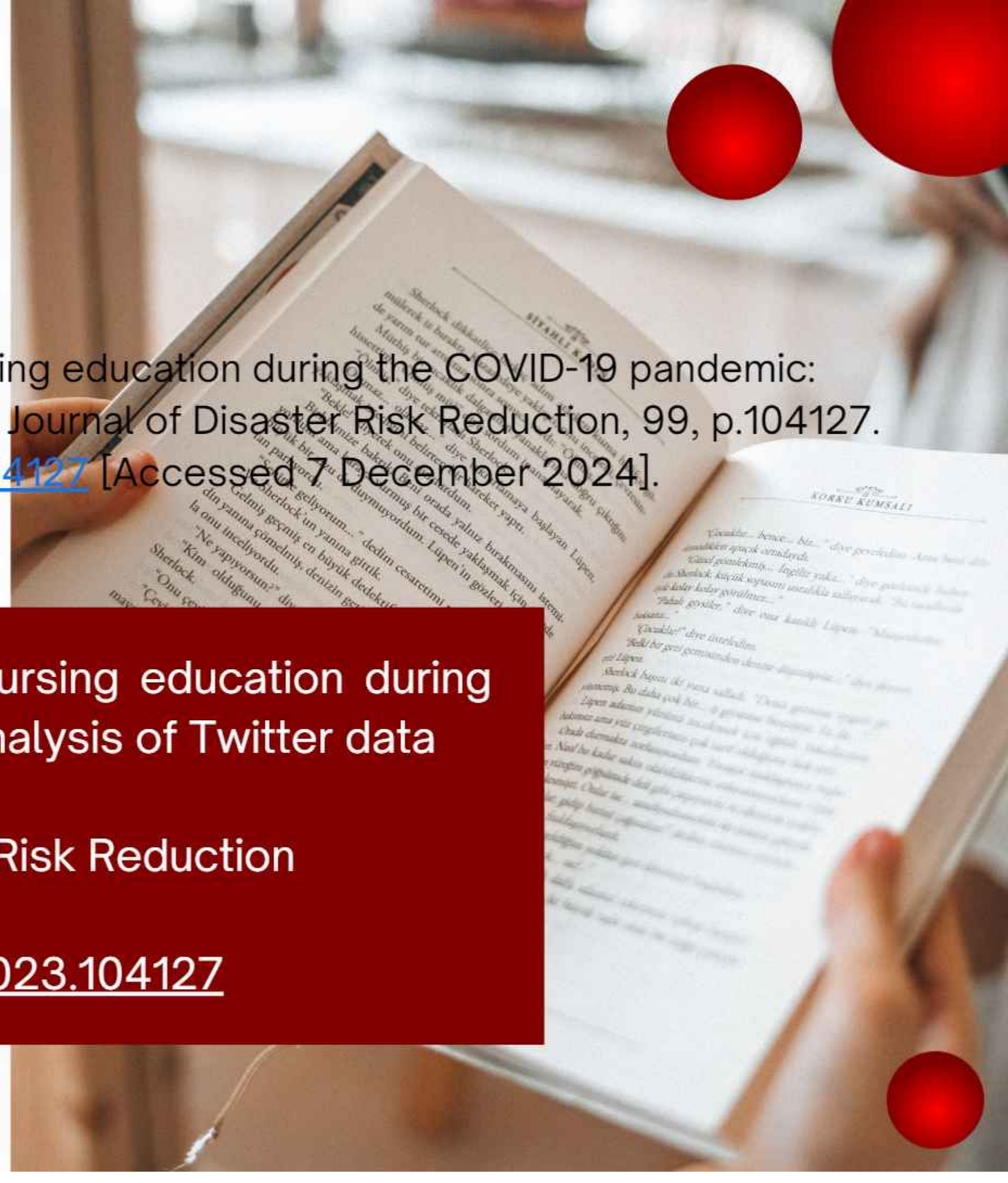
Tags
Politics

Dataset dapat diakses melalui:

- **Kaggle:** https://www.kaggle.com/datasets/jocelyndumlao/indonesia-presidential-candidates-dataset-2024/
- **Mendeley:** https://data.mendeley.com/datasets/7w5zvr8jgp/5

Paper
Rujukan

Korkmaz, A. Ç., 2023. Public's perception on nursing education during the COVID-19 pandemic: SENTIMENT analysis of Twitter data. International Journal of Disaster Risk Reduction, 99, p.104127. Available at: https://doi.org/10.1016/j.ijdrr.2023.104127 [Accessed 7 December 2024].



- **Judul Paper:** Public's perception on nursing education during the COVID-19 pandemic: SENTIMENT analysis of Twitter data
- **Penulis:** Ayşe Çiçek Korkmaz
- **Jurnal:** International Journal of Disaster Risk Reduction
- **Volume:** 99, Desember 2023
    - **URL:** https://doi.org/10.1016/j.ijdrr.2023.104127

FILKOM
FAKULTAS ILMU KOMPUTER

METODOLOGI PENYELESAIAN STUDI KASUS

Metodologi

Penyelesaian

1. Instalasi dan Persiapan Lingkungan

- Menginstal pustaka dan dependencies yang dibutuhkan.
- Melakukan konfigurasi environment analisis.

2. Import Library

- Pustaka untuk data manipulation: Digunakan untuk pengolahan data (contoh: pandas, numpy).
- Pustaka untuk visualisasi: Digunakan untuk analisis visual (contoh: matplotlib, seaborn).
- Pustaka untuk machine learning dan NLP: Digunakan untuk modeling dan analisis teks (contoh: sklearn, spaCy, transformers).

3. Download dan Ekstraksi Data

- Mendownload dataset dari Google Drive.
- Mengekstrak file dataset ke dalam folder kerja.

Metodologi

Penyelesaian

4. Load Dataset

- Memuat tiga dataset kandidat presiden.
- Memeriksa struktur data awal untuk memastikan integritas dataset.

5. Kombinasi Dataset

- Menambahkan identitas kandidat ke dalam dataset.
- Menggabungkan dataset menjadi satu kesatuan.

6. Seleksi Kolom dan Penanganan Missing Value

- Memilih kolom yang relevan untuk analisis.
- Membersihkan data kosong atau missing value.

7. Eksplorasi Data Awal (EDA)

- Membuat visualisasi distribusi data.
- Menganalisis panjang tweet dan pola temporal.
- Menganalisis frekuensi kata untuk masing-masing kandidat.

Metodologi

Penyelesaian

8. Data Processing dan Preprocessing Pipeline

- Membersihkan teks menggunakan cleantext.
- Melakukan tokenisasi dan lemmatization menggunakan spaCy.
- Menghapus emoji dari teks.
- Melakukan normalisasi kata slang.
- Membangun pipeline preprocessing yang terintegrasi.

9. EDA Setelah Preprocessing

- Menampilkan preview data setelah preprocessing.
- Membuat visualisasi frekuensi kata.
- Membuat WordCloud berdasarkan kandidat dan sentimen.

10. Data Sampling & Post-Processing

- Menganalisis distribusi label pada dataset.
- Mengonversi label ke format biner.
- Menangani ketidakseimbangan data dengan oversampling.
- Membagi dataset ke dalam train, validation, dan test set.
- Melakukan one-hot encoding pada label.

Metodologi

Penyelesaian

11. Modelling dan Evaluasi Model

a) Implementasi Baseline Model (Naive Bayes)

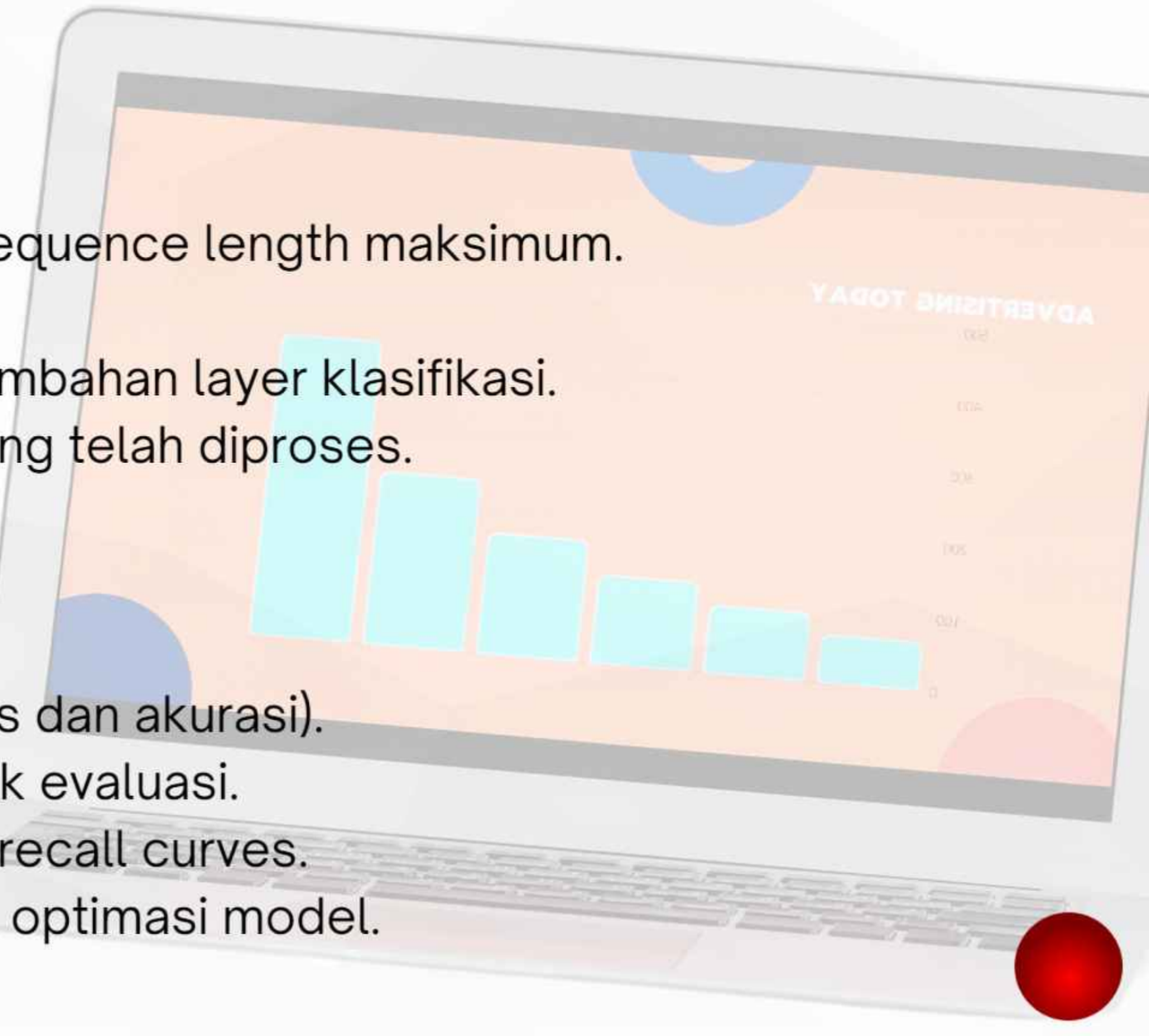
- Tokenisasi dengan CountVectorizer.
- Mengonversi teks ke dalam representasi TF-IDF.
- Melatih model Naive Bayes.
- Melakukan prediksi dan menghitung probabilitas.

b) **Evaluasi Model Naive Bayes**

- Membuat evaluator untuk visualisasi evaluasi.
- Menganalisis confusion matrix dan classification report.
- Membuat visualisasi ROC curve dan precision-recall curve.
- Menganalisis misclassifications untuk perbaikan model.

Metodologi

Penyelesaian



c) Implementasi Model BERT

- Menyiapkan tokenizer BERT pre-trained.
- Melakukan tokenisasi dataset dengan sequence length maksimum.
- Membuat model BERT pre-trained.
- Membangun arsitektur model dengan tambahan layer klasifikasi.
- Melatih model menggunakan dataset yang telah diproses.

d) Evaluasi Model BERT

- Membuat prediksi pada data test.
- Memvisualisasikan history pelatihan (loss dan akurasi).
- Menganalisis confusion matrix dan metrik evaluasi.
- Membuat visualisasi ROC dan precision-recall curves.
- Menganalisis kesalahan klasifikasi untuk optimasi model.

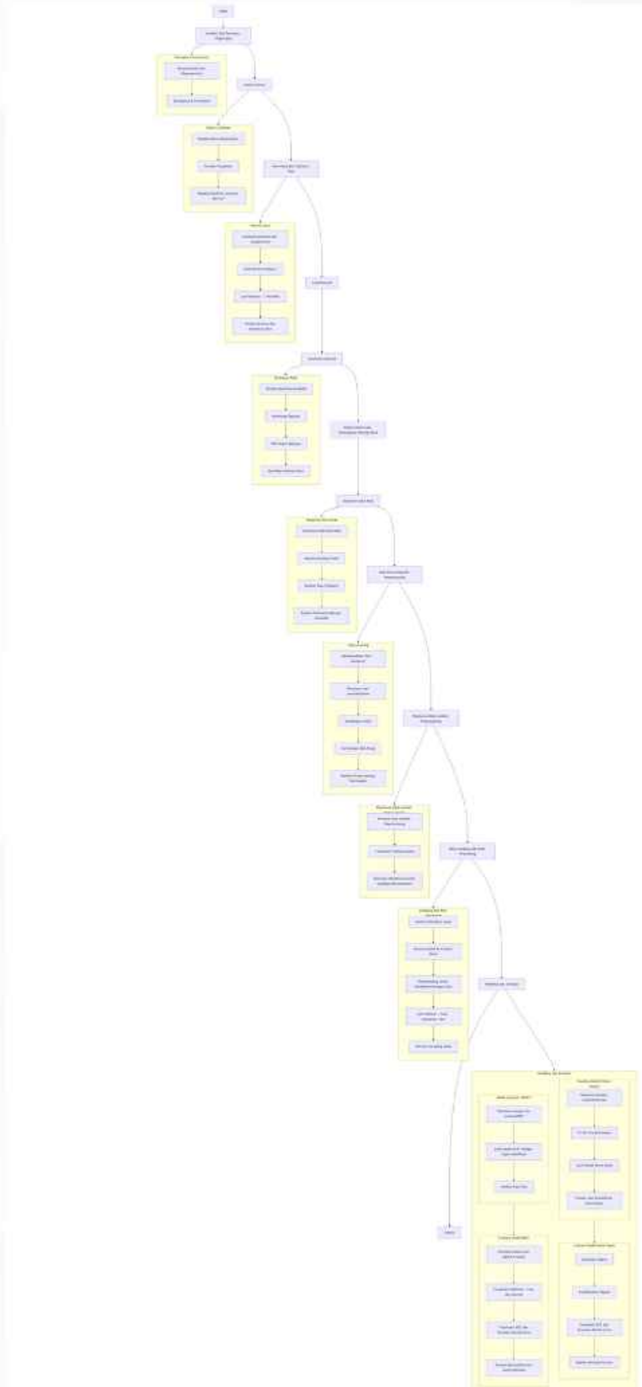
PERANCANGAN

FLOWCHART

Perancangan
Flowchart

Untuk lebih detail dan jelas dapat diakses melalui tautan berikut:

klik disini



FILKOM
FAKULTAS ILMU KOMPUTER

IMPLEMENTASI DAN PENGUJIAN

untuk sesi ini kami akan bahas lebih lanjut melalui google colab lembar kerja kami

https://colab.research.google.com/drive/1ZwaPPy9D_FAAb5tCxV4s57eWBqtfsHTq?usp=sharing

FILKOM
FAKULTAS ILMU KOMPUTER

KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan analisis perbandingan model, **BERT secara signifikan mengungguli model baseline Naive Bayes** dalam klasifikasi sentimen. Peningkatan akurasi dari 85% menjadi 95% menunjukkan kemampuan **BERT memahami konteks bahasa secara mendalam.** Dengan tingkat misklasifikasi yang rendah (4.57% dibandingkan 15.14% pada Naive Bayes) dan metrik evaluasi yang seimbang seperti precision, recall, serta F1-score (95% untuk kedua kelas), **BERT terbukti andal dan tidak bias terhadap kelas tertentu.** Model ini juga menunjukkan kemampuan prediksi yang konsisten dengan precision hingga 97% dan recall hingga 97%. Keunggulan BERT semakin terlihat pada metrik reliabilitas, seperti AUC ROC 0.98 dan precision-recall curve 0.99, yang menegaskan kemampuan diskriminatifnya. **Stabilitas performa dan distribusi data yang seimbang** menjadikan BERT solusi yang dapat diandalkan untuk implementasi dalam skala yang lebih luas.

Saran

Untuk pengembangan sistem analisis sentimen ke depan, disarankan membangun **sistem monitoring performa real-time** dan melakukan **fine-tuning model secara berkala** dengan data terbaru agar tetap relevan. **Perluasan dataset** dengan variasi bahasa dan konteks juga penting untuk meningkatkan akurasi prediksi. Dalam praktiknya, pengembangan **dashboard real-time dengan sistem peringatan dini perlu diutamakan,** didukung oleh optimasi efisiensi waktu pemrosesan dan pengembangan API yang scalable. Keberhasilan implementasi model BERT dalam analisis sentimen kandidat presiden Indonesia 2024 **menunjukkan potensi besar untuk memperluas penerapan teknologi ini** dalam memahami sentimen publik pada berbagai dinamika sosial dan politik.

Terima Kasih!

Kelompok 8
Pemrosesan Bahasa Alami