Dokumentasi Panduan Crawling Data Twitter Pemilu 2024 (Penelitian ABSA Transformer)

1. Latar Belakang

Judul penelitian:

"Peningkatan Model Analisis Sentimen Berbasis Aspek terhadap Tokoh Politik Indonesia Menggunakan Transformer dan Data Pemilu 2024."

Salah satu tahap penting penelitian ini adalah **pengumpulan data opini publik dari Twitter** terkait tokoh politik Indonesia pada periode **Pemilu 2024**. Data ini digunakan untuk melengkapi dataset lama (Pemilu 2014) dan **menyeimbangkan aspek yang jarang muncul** seperti *empati* dan *kontinuitas*.

2. Tujuan Kegiatan Crawling

a. Mengambil tweet berbahasa Indonesia dari periode **September 2023 – Mei 2024**.
b. Mengelompokkan tweet berdasarkan **aspek opini politik**:
   - **Integritas** (kejujuran, moral, anti korupsi)
   - **Kapabilitas** (kemampuan memimpin, kecerdasan, profesionalitas)
   - **Empati** (kepedulian terhadap rakyat, kepekaan sosial)
   - **Akseptabilitas** (penerimaan publik, popularitas)
   - **Kontinuitas** (keberlanjutan program, visi jangka panjang)
c. Menyimpan hasil crawling ke dalam file .csv untuk setiap aspek.
d. Menyediakan dataset siap anotasi (pelabelan sentimen manual).

3. Rancangan Keyword per Aspek

Aspek	Makna	Contoh Keyword Crawling (gunakan di script)
Integritas	Kejujuran, moral, anti korupsi	"bersih dari korupsi" OR "tidak korupsi" OR "pemimpin jujur" OR "integritas tinggi" OR "amanah"
Kapabilitas	Kecakapan, kemampuan memimpin	"kompeten" OR "kapabilitas" OR "berpengalaman memimpin" OR "cerdas" OR "profesional"
Empati	Kepedulian terhadap masyarakat	"peduli rakyat" OR "turun ke rakyat" OR "bantu korban" OR "dekat dengan rakyat"

Akseptabilitas	Popularitas dan penerimaan publik	"disukai rakyat" OR "diterima masyarakat" OR "elektabilitas tinggi" OR "popularitas kandidat"
Kontinuitas	Kelanjutan program dan visi jangka panjang	"lanjutkan program" OR "keberlanjutan pembangunan" OR "visi jangka panjang" OR "melanjutkan kerja"

Catatan:
Kata kunci boleh ditambah, asal **relevan dengan aspek tersebut** dan **tidak bersifat partisan langsung** (hindari kata “Ganjar”, “Prabowo”, “Anies”, dsb. dalam keyword utama — boleh muncul alami di hasil tweet).

4. Persiapan Lingkungan

Jalankan script ini di **Google Colab** atau **VS Code**.

a. Install Library

```
pip install snscrape pandas tqdm
```

b. Buat Folder Project

```
ABSA-Transformer-Pemilu2024/
│
├─ data/
│   └─ crawling_2024/
│
├─ crawl_twitter_absa.py
```

5. Script Crawling

Modul crawling bisa download disini.

6. Output yang Diharapkan

Setelah script dijalankan, akan muncul file:

```
data/crawling_2024/
│
├─ tweets_integritas.csv
├─ tweets_kapabilitas.csv
├─ tweets_empati.csv
├─ tweets_akseptabilitas.csv
└─ tweets_kontinuitas.csv
```

7. Tugas Setelah Crawling

a. **Periksa hasil CSV**: pastikan isinya tweet berbahasa Indonesia dan relevan.
b. **Hapus noise data** (tweet iklan, spam, retweet kosong, dsb).
c. **Gabungkan semua file CSV menjadi satu dataset besar:**

```
import pandas as pd
import glob

files = glob.glob("data/crawling_2024/*.csv")
df = pd.concat([pd.read_csv(f) for f in files],
ignore_index=True)
df.to_csv("data/crawling_2024/merged_tweets_2024.csv",
index=False, encoding="utf-8-sig")
```

d. **Kirim hasil** untuk dilakukan labeling aspek & sentimen.
e. Sertakan laporan kecil berisi:
   - Jumlah tweet per aspek
   - Contoh 3 tweet untuk tiap aspek
   - Tanggal crawling dilakukan
   - Kendala jika ada

8. Evaluasi Kualitas Data

```
import pandas as pd
import matplotlib.pyplot as plt
import glob

files = glob.glob("data/crawling_2024/*.csv")
counts = {f.split("_")[-1].split(".")[0]: len(pd.read_csv(f))
for f in files}

plt.bar(counts.keys(), counts.values())
plt.title("Jumlah Tweet per Aspek (Crawling 2024)")
plt.xlabel("Aspek") plt.ylabel("Jumlah Tweet") plt.show()
```

9. Catatan Tambahan

a. Script ini **tidak membutuhkan login Twitter atau API Key**.
b. Gunakan jaringan internet yang stabil (karena prosesnya bisa lama).
c. Bila crawling berhenti mendadak, jalankan ulang — snscrape akan melanjutkan hasilnya.
d. Pastikan tweet yang dikumpulkan **berkaitan dengan opini masyarakat**, bukan promosi atau akun berita otomatis.