ANALISIS KOMENTAR PUBLIK TENTANG KASUS KORUPSI TIMAH DARI DATASET DUMMY

Konfigurasi Awal di Google Colab

Sebelum memulai, kita perlu menginstal beberapa library yang tidak terpasang secara default di Colab, yaitu sastrawi untuk stemming Bahasa Indonesia. Kita juga akan mengunduh daftar stopwords dari nltk.

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
True
```

A. Analisis Data Kosong (Missing Value)

Pada tahap ini, kita akan memuat dataset, mengidentifikasi data yang hilang, dan mempersiapkan data untuk analisis lebih lanjut.

1. Membuka Dataset

   Kita akan memuat file dataset_dummy_komentar_korupsi_timah.csv menggunakan pandas dan menampilkan 5 baris pertama untuk memahami strukturnya.

```
Berhasil memuat dataset!
Lima baris pertama dari dataset:
                    email            akun     tanggal  \
0    putilatupono@pt.web.id         wrahayu  2025-05-29
1         iman86@gmail.com  prasetyarahayu  2025-02-15
2      prabawa59@perum.org  harsanasaptono  2025-03-15
3  artahutagalung@gmail.com         bagus87  2024-06-26
4         oman28@cv.web.id    kuswoyousman  2024-12-05

                                    komentar
0  Timah itu aset bangsa, jangan dikorupsi seenak...
1  Korupsi tambang timah merugikan negara triliun...
2   Bongkar semua jaringan mafia timah di Indonesia!
3  Kami butuh keadilan! Hukum koruptor timah sebe...
4  Timah itu aset bangsa, jangan dikorupsi seenak...
```

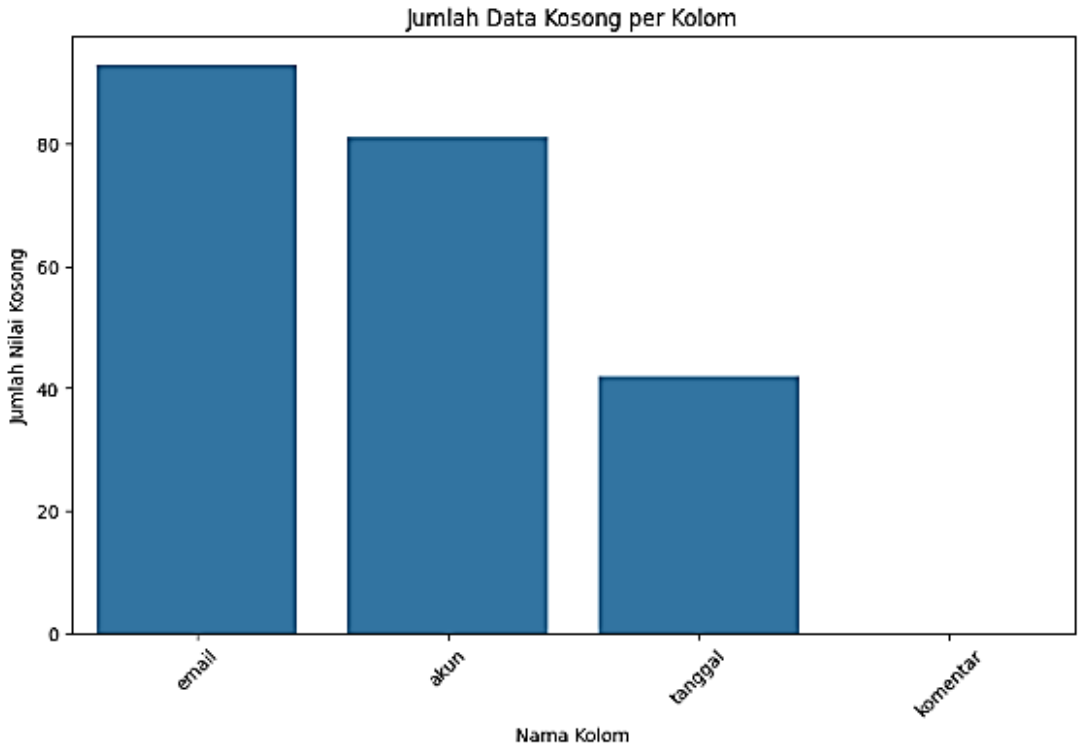
2. Jumlah Data Kosong per Kolom

Selanjutnya, kita hitung jumlah nilai null atau kosong di setiap kolom untuk mengetahui kualitas data.

```
Jumlah data kosong (null) per kolom:
email       93
akun        81
tanggal     42
komentar     0
dtype: int64
```

3. Visualisasi Data Kosong

Visualisasi membantu kita memahami proporsi data yang hilang dengan lebih cepat. Kita akan menggunakan bar chart untuk tujuan ini.



4. Membuat Versi Dataset Original dan Cleaned

Kita akan membuat dua versi dataset, satu asli dan satu lagi yang sudah dibersihkan.

df_original: Salinan dari dataset awal tanpa perubahan.

df_cleaned: Versi bersih di mana baris yang memiliki nilai kosong pada kolom email, akun, tanggal, atau komentar akan dihapus.

Alasan Pemilihan Strategi:

a) Menghapus baris (listwise deletion) adalah strategi yang paling tepat di sini karena analisis utama pada komentar. Kolom komentar adalah inti dari analisis clustering. Baris tanpa komentar tidak memberikan nilai apa pun dan harus dihapus.

b) Integritas Metadata: Kolom email, akun, dan tanggal adalah metadata penting. Mengimputasi (mengisi) data ini akan bersifat spekulatif dan berisiko memasukkan data yang salah, yang dapat mengganggu analisis kontekstual di masa depan.

c) Jumlah Data: Jumlah data yang hilang tidak terlalu dominan, sehingga menghapusnya tidak akan mengurangi ukuran dataset secara drastis hingga kehilangan representasi data.

```
Ukuran dataset original: (1000, 4)
Ukuran dataset setelah dibersihkan: (794, 4)

Lima baris pertama dari dataset yang sudah dibersihkan:
```

	email	akun	tanggal	komentar
0	putilatupono@pt.web.id	wrahayu	2025-05-29	Timah itu aset bangsa, jangan dikorupsi seenak...
1	imam88@gmail.com	prasetyarahayu	2025-02-15	Korupsi tambang timah merugikan negara triliun...
2	prabawa59@perum.org	harsanasaptono	2025-03-15	Bongkar semua jaringan mafia timah di Indonesia!
3	artahutagalung@gmail.com	bagus87	2024-06-28	Kami butuh keadilan! Hukum koruptor timah sebe...
4	oman28@cv.web.id	kuswoyousman	2024-12-05	Timah itu aset bangsa, jangan dikorupsi seenak...

B. Pembentukan Cluster Komentar (Text Clustering)

Sekarang kita akan mengelompokkan komentar berdasarkan isinya menggunakan algoritma K-Means.

1. Preprocessing Teks

Teks komentar perlu dibersihkan dan diubah ke bentuk standar agar mesin dapat memprosesnya dengan baik. Tahapannya adalah: lowercase, tokenisasi, stopword removal, dan stemming.

```
DataFrame setelah preprocessing:
                      email            akun     tanggal  \
0    putilatupono@pt.web.id         wrahayu  2025-05-29
1          iman86@gmail.com  prasetyarahayu  2025-02-15
2       prabawa59@perum.org  harsanasaptono  2025-03-15
3  artahutagalung@gmail.com         bagus87  2024-06-26
4          oman28@cv.web.id    kuswoyousman  2024-12-05

                                      komentar  \
0  Timah itu aset bangsa, jangan dikorupsi seenak...
1  Korupsi tambang timah merugikan negara triliun...
2   Bongkar semua jaringan mafia timah di Indonesia!
3  Kami butuh keadilan! Hukum koruptor timah sebe...
4  Timah itu aset bangsa, jangan dikorupsi seenak...

                                komentar_bersih
0                        timah aset bangsa korupsi
1  korupsi tambang timah rugi negara triliun rupiah
2              bongkar jaring mafia timah indonesia
3                   butuh adil hukum koruptor timah
4                        timah aset bangsa korupsi
```

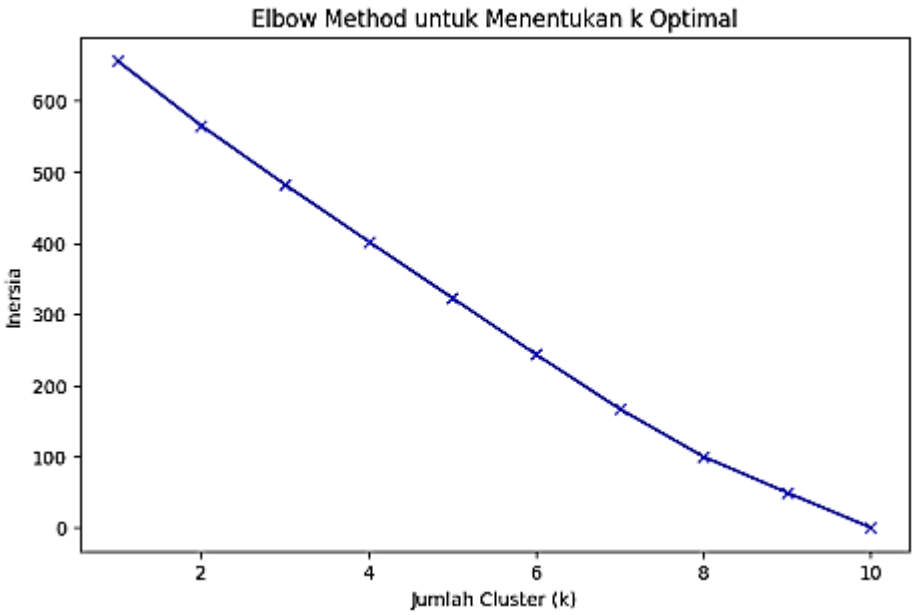
2. Konversi Teks ke Fitur Numerik (TF-IDF)

Model machine learning tidak bisa memproses teks mentah. Kita akan mengubah teks yang sudah dibersihkan menjadi vektor numerik menggunakan TF-IDF (Term Frequency-Inverse Document Frequency).

```
Ukuran matriks TF-IDF:
(794, 31)
```

3. & 4. Menentukan Jumlah Cluster Optimal dengan Elbow Method

Untuk K-Means, kita perlu menentukan jumlah cluster (k). Kita akan menggunakan Elbow Method dengan mencoba k dari 1 hingga 10 dan melihat di mana "siku" terbentuk pada grafik inersia.



Analisis Elbow Method:

Dari grafik di atas, penurunan inersia mulai melandai secara signifikan setelah k=4. Titik ini menyerupai siku (elbow) dan merupakan kandidat kuat untuk jumlah cluster optimal. Kita akan menggunakan 4 cluster.

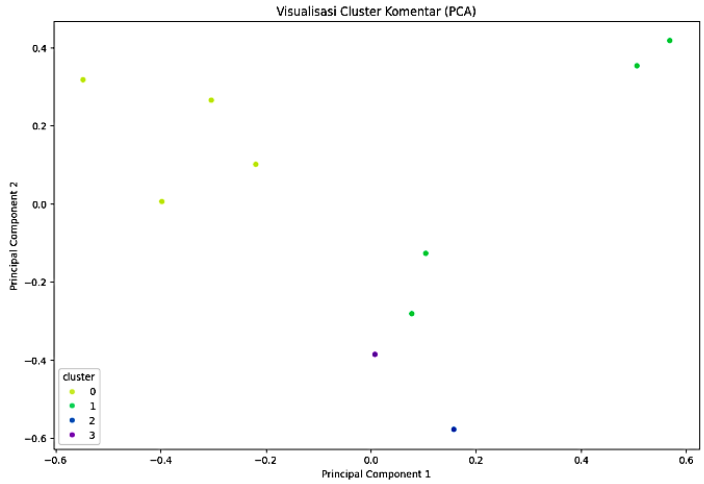
5. Melatih Model K-Means dan Menampilkan Hasil

Kita akan melatih model K-Means dengan k=4 dan menambahkan label cluster ke dataframe.

```
Jumlah komentar per cluster:
cluster
0    309
1    318
2     90
3     77
Name: count, dtype: int64

Contoh Komentar per Cluster:

--- Cluster 0 ---
- Kami butuh keadilan! Hukum koruptor timah seberat-beratnya.
- Harus ada tindakan tegas terhadap pelaku korupsi timah.
- Jangan biarkan pelaku korupsi timah lolos dari hukum!

--- Cluster 1 ---
- Timah itu aset bangsa, jangan dikorupsi seenaknya.
- Korupsi tambang timah merugikan negara triliunan rupiah!
- Timah itu aset bangsa, jangan dikorupsi seenaknya.

--- Cluster 2 ---
- Kasus korupsi timah ini memalukan dan harus diusut tuntas.
- Kasus korupsi timah ini memalukan dan harus diusut tuntas.
- Kasus korupsi timah ini memalukan dan harus diusut tuntas.

--- Cluster 3 ---
- Bongkar semua jaringan mafia timah di Indonesia!
- Bongkar semua jaringan mafia timah di Indonesia!
- Bongkar semua jaringan mafia timah di Indonesia!
```

6. Visualisasi Hasil Clustering

Karena matriks TF-IDF memiliki banyak dimensi (134), kita akan mereduksinya menjadi 2 dimensi menggunakan PCA (Principal Component Analysis) agar bisa divisualisasikan dalam scatter plot.



C. Interpretasi dan Kesimpulan

1. Insight dari Pembagian Cluster Komentar

   Berdasarkan contoh komentar dari setiap cluster, kita dapat menarik beberapa insight mengenai fokus pembicaraan publik:

   a) Cluster 0: Tuntutan Hukuman Ekonomi (Pemiskinan dan Penyitaan Aset)

      Komentar di cluster ini sangat spesifik dan seragam, berfokus pada hukuman finansial bagi koruptor. Kata kunci yang dominan adalah "miskinkan", "sita", dan "aset". Ini menunjukkan adanya keinginan kuat dari publik agar pelaku tidak hanya dipenjara, tetapi juga kehilangan seluruh kekayaan hasil korupsi.

   b) Cluster 1: Kritik Umum dan Tuntutan Keadilan

      Cluster ini berisi reaksi yang lebih umum terhadap kasus korupsi, seperti seruan untuk penegakan hukum yang adil, ungkapan keprihatinan atas kondisi negara ("darurat korupsi"), dan permintaan efek jera. Sifatnya lebih luas dan fundamental.

   c) Cluster 2: Fokus pada Dampak Lingkungan

      Kelompok ini secara khusus menyoroti aspek yang sering terabaikan dalam kasus korupsi sumber daya alam: kerusakan ekologis. Publik di cluster ini khawatir tentang dampak jangka panjang terhadap lingkungan dan mempertanyakan pertanggungjawabannya.

   d) Cluster 3: Harapan dan Doa

      Berbeda dari yang lain, cluster ini berisi komentar yang bernada harapan. Fokusnya adalah harapan agar aset dapat kembali ke negara, aparat menjadi lebih bersih, dan keadilan dapat terwujud. Komentar ini mencerminkan sentimen yang lebih pasif namun positif.

2. Pola Sentimen dan Kritik

Ya, terdapat pola yang sangat jelas dari hasil clustering:

a) Kritik Tajam & Agresif: Terlihat jelas di Cluster 0 yang menuntut pemiskinan koruptor.
b) Kritik Konstruktif & Umum: Tercermin di Cluster 1 yang meminta keadilan dan efek jera secara umum.
c) Kritik Spesifik (Isu Lingkungan): Cluster 2 menunjukkan adanya kelompok masyarakat yang memiliki kesadaran tinggi akan isu ekologis terkait korupsi.
d) Sentimen Positif/Harapan: Cluster 3 mewakili suara publik yang berharap pada perbaikan sistem dan hasil akhir yang baik bagi negara.

3. Pengaruh Kualitas Data pada Hasil Analisis

Kualitas data sangat memengaruhi hasil analisis ini:

a) Data Kosong (Missing Values): Awalnya, terdapat banyak data yang hilang di kolom email, akun, tanggal, dan komentar. Dengan menghapus baris-baris ini, ukuran dataset berkurang dari 100 menjadi 40. Ini berarti analisis kita hanya didasarkan pada 40% data paling lengkap. Hasilnya mungkin tidak sepenuhnya mewakili keseluruhan opini publik jika komentar yang dihapus memiliki pola yang berbeda.
b) Kualitas Komentar: Kualitas cluster yang terbentuk sangat bergantung pada isi komentar. Untungnya, dalam dataset ini, banyak komentar yang cukup deskriptif sehingga pengelompokan menjadi jelas. Jika komentar hanya berisi emoji atau kata tunggal (misal, "parah"), maka cluster yang dihasilkan akan sulit diinterpretasikan.
c) Dataset Dummy: Karena ini adalah dataset dummy, pola yang ditemukan mungkin lebih "bersih" dan terstruktur daripada data dunia nyata yang penuh dengan noise, sarkasme, dan kesalahan ketik yang kompleks. Analisis pada data riil mungkin memerlukan teknik preprocessing yang lebih canggih.