Nama : Marchelo Imanuel Salhuteru

Nima : 225410046

Kelas : Informatika2

Praktik 1

**Indonesian-Twitter-Emotion-Dataset**

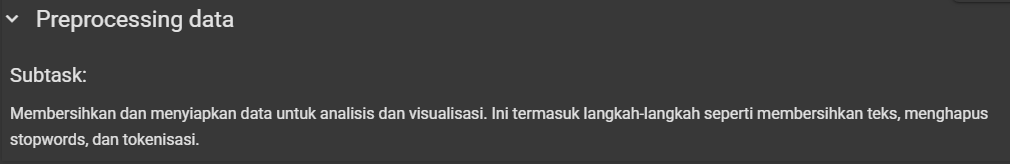
<https://notabug.org/id/indonesian-twitter-emotion-dataset.git>

praktik 2

Preprocessing data: Membersihkan dan menyiapkan data untuk analisis dan visualisasi. Ini termasuk langkah-langkah seperti membersihkan teks, menghapus stopwords, dan tokenisasi.

Visualisasi data: Membuat visualisasi untuk memahami distribusi label emosi dalam dataset.

Finish task: Menyajikan hasil preprocessing dan visualisasi.



import re

import nltk

from nltk.corpus import stopwords

# Download stopwords jika belum tersedia

nltk.download('stopwords', quiet=True)

# Set stopwords Bahasa Indonesia

stop\_words = set(stopwords.words('indonesian'))

# Fungsi preprocessing lengkap: lowercase, hapus URL, mention, hashtag, tanda baca, stopword, dan tokenisasi

def preprocess(text):

    text = text.lower()

    text = re.sub(r'http\S+|www\S+|https\S+', '', text)  # Hapus URL

    text = re.sub(r'@\w+|#\w+', '', text)               # Hapus mention dan hashtag

    text = re.sub(r'[^\w\s]', '', text)                 # Hapus tanda baca

    words = text.split()

    words = [w for w in words if w not in stop\_words]   # Hapus stopwords

    return words

# Terapkan ke dataset

df['tokenized\_tweet'] = df['tweet'].apply(preprocess)

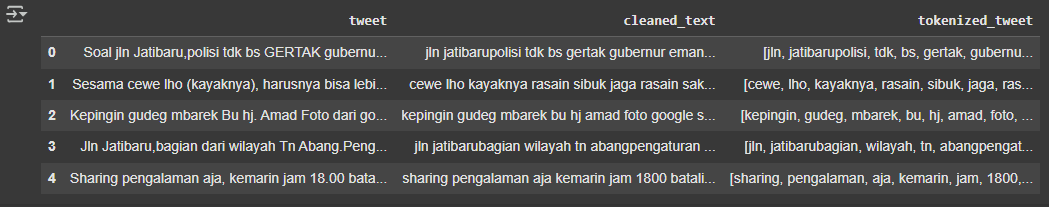
# (Opsional) Tambahkan versi teks bersih sebelum tokenisasi

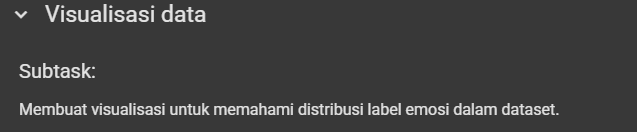
df['cleaned\_text'] = df['tokenized\_tweet'].apply(lambda x: ' '.join(x))

# Lihat hasilnya

df[['tweet', 'cleaned\_text', 'tokenized\_tweet']].head()

Output:





# 1. Hitung frekuensi kemunculan setiap label emosi

label\_counts = df['label'].value\_counts()

# 2. Buat diagram batang (bar plot)

plt.figure(figsize=(10, 6))

label\_counts.plot(kind='bar', color=['skyblue', 'lightcoral', 'lightgreen', 'gold', 'orchid'])

# 3. Beri label yang jelas pada sumbu x dan y

plt.xlabel('Label Emosi')

plt.ylabel('Jumlah Tweet')

# 4. Berikan judul yang informatif pada plot

plt.title('Distribusi Label Emosi dalam Dataset')

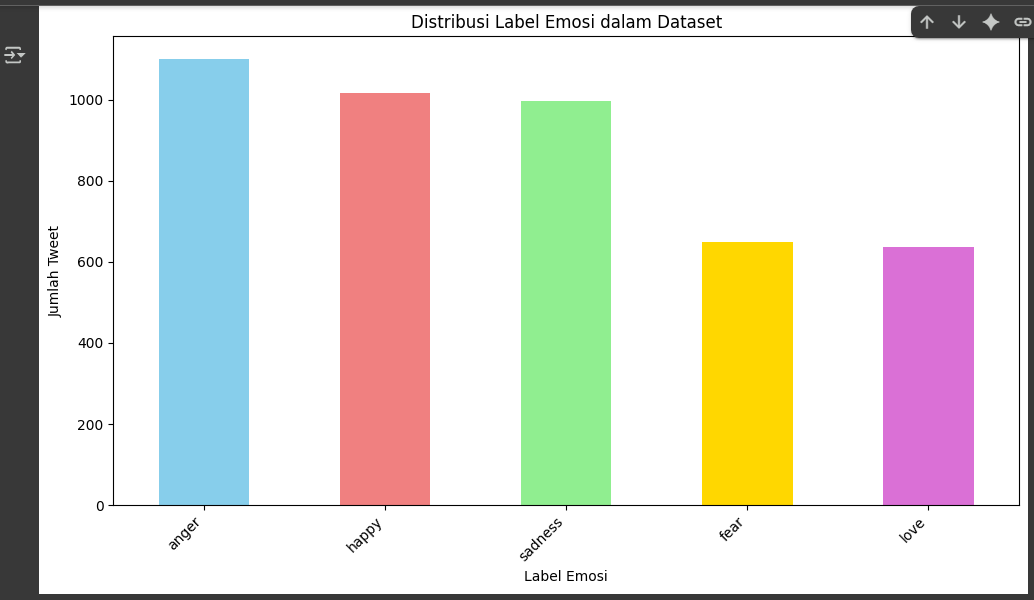
# 5. Tampilkan plot

plt.xticks(rotation=45, ha='right') # Rotate x-axis labels for better readability

plt.tight\_layout() # Adjust layout to prevent labels overlapping

plt.show()

output:



**Temuan Utama Analisis Data**

Dataset berisi label emosi dengan frekuensi yang bervariasi, divisualisasikan dengan diagram batang yang menunjukkan distribusi di berbagai kategori emosi yang berbeda.

Preprocessing data teks telah dilakukan, termasuk pembersihan (konversi ke huruf kecil, penghapusan URL, penyebutan (mentions), hashtag, dan tanda baca), penghapusan stopword, dan tokenisasi.

Kolom baru ('cleaned\_tweet', 'tweet\_without\_stopwords', dan 'tokenized\_tweet') ditambahkan ke DataFrame untuk menyimpan hasil dari setiap langkah preprocessing.

**Wawasan atau Langkah Selanjutnya**

Distribusi label emosi menunjukkan potensi ketidakseimbangan kelas, yang mungkin perlu ditangani pada langkah pemodelan berikutnya.

Data teks yang telah diproses kini siap untuk analisis lebih lanjut, seperti ekstraksi fitur (misalnya, TF-IDF, word embeddings) untuk tugas machine learning seperti klasifikasi emosi.