LAPORAN MODUL 14
PRAKTIKUM PRAKTIKUM BIG DATA ANALYTICS

Disusun oleh :

Nama : Marchelo Imanuel Salhuteru

NIM : 225410046

Kelas : Informatika2

PROGRAM STUDI INFORMATIKA

PROGRAM SARJANA

FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS TEKNOLOGI DIGITAL INDONESIA

YOGYAKARTA

2025

LAPORAN PRAKTIKUM PRAKTIKUM BIG DATA ANALYTICS
MODUL 14

A. TUJUAN PRAKTIKUM

Mampu memahami dan mengimplementasikan Teknik pengumpulan, pengolahan dan penyajian data

B. DASAR TEORI

-

C. PEMBAHASAN LISTING

Berisi pembahasan listing yang telah dipraktikkan saat jam praktikum

Listing yang dikerjakan saat praktikum di lab (capture praktik + latihan)

Praktik

1. **Praktik 1**

   Mencari dataset => Mahasiswa bebas mencari dataset yang siap diolah (jumlah data minimal 100 dan sebutkan sumber link datanya. Bisa data teks (twitter) atau angka)

   Dataset => Indonesian-Twitter-Emotion-Dataset
   Sumber dataset => https://notabug.org/id/indonesian-twitter-emotion-dataset.git

2. Yang harus dilakukan :
   - Jelaskan crawling data
   - Tentukan proses preprosesingnya
   - Lakukan visualisasi dataset

   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()
```

pembahasan :

Mengimpor Library: Mengimpor library yang diperlukan seperti re untuk regular expression, nltk untuk Natural Language Toolkit, dan stopwords dari nltk.corpus.

Mengunduh Stopwords: Mengunduh daftar kata-kata umum (stopwords) dalam Bahasa Indonesia jika belum ada. Stopwords adalah kata-kata yang sering muncul tetapi tidak memiliki makna yang signifikan untuk analisis teks (contoh: "yang", "dan", "di").

Menetapkan Stopwords: Membuat set dari stopwords Bahasa Indonesia untuk pencarian yang lebih efisien.

Fungsi preprocess: Mendefinisikan sebuah fungsi bernama preprocess yang mengambil teks sebagai input dan melakukan serangkaian operasi:

Mengubah teks menjadi huruf kecil (text.lower()).

Menghapus URL, mention (@username), dan hashtag (#hashtag) menggunakan regular expression.

Menghapus tanda baca.

Membagi teks menjadi kata-kata (text.split()).

Menghapus stopwords dari daftar kata-kata.

Mengembalikan daftar kata-kata yang sudah diproses (tokenized).

Menerapkan Preprocessing: Menerapkan fungsi preprocess ke setiap baris dalam kolom 'tweet' dan menyimpan hasilnya dalam kolom baru bernama 'tokenized_tweet'.
Membuat Kolom Teks Bersih: (Opsional) Menggabungkan kembali token-token dalam 'tokenized_tweet' menjadi satu string teks bersih dan menyimpannya di kolom baru bernama 'cleaned_text'. Ini berguna jika Anda memerlukan versi teks bersih tanpa tokenisasi.
Menampilkan Hasil: Menampilkan beberapa baris pertama dari kolom 'tweet' asli, 'cleaned_text', dan 'tokenized_tweet' untuk melihat hasil pra-pemrosesan.
Output:

	tweet	cleaned_text	tokenized_tweet
0	Soal jln Jatibaru,polisi tdk bs GERTAK gubernu...	jln jatibarupolisi tdk bs gertak gubernur eman...	[jln, jatibarupolisi, tdk, bs, gertak, gubernu...
1	Sesama cewe lho (kayaknya), harusnya bisa lebi...	cewe lho kayaknya rasain sibuk jaga rasain sak...	[cewe, lho, kayaknya, rasain, sibuk, jaga, ras...
2	Kepingin gudeg mbarek Bu hj. Amad Foto dari go...	kepingin gudeg mbarek bu hj amad foto google s...	[kepingin, gudeg, mbarek, bu, hj, amad, foto, ...
3	Jln Jatibaru,bagian dari wilayah Tn Abang.Peng...	jln jatibarubagian wilayah tn abangpengaturan ...	[jln, jatibarubagian, wilayah, tn, abangpengat...
4	Sharing pengalaman aja, kemarin jam 18.00 bata...	sharing pengalaman aja kemarin jam 1800 batali...	[sharing, pengalaman, aja, kemarin, jam, 1800,...

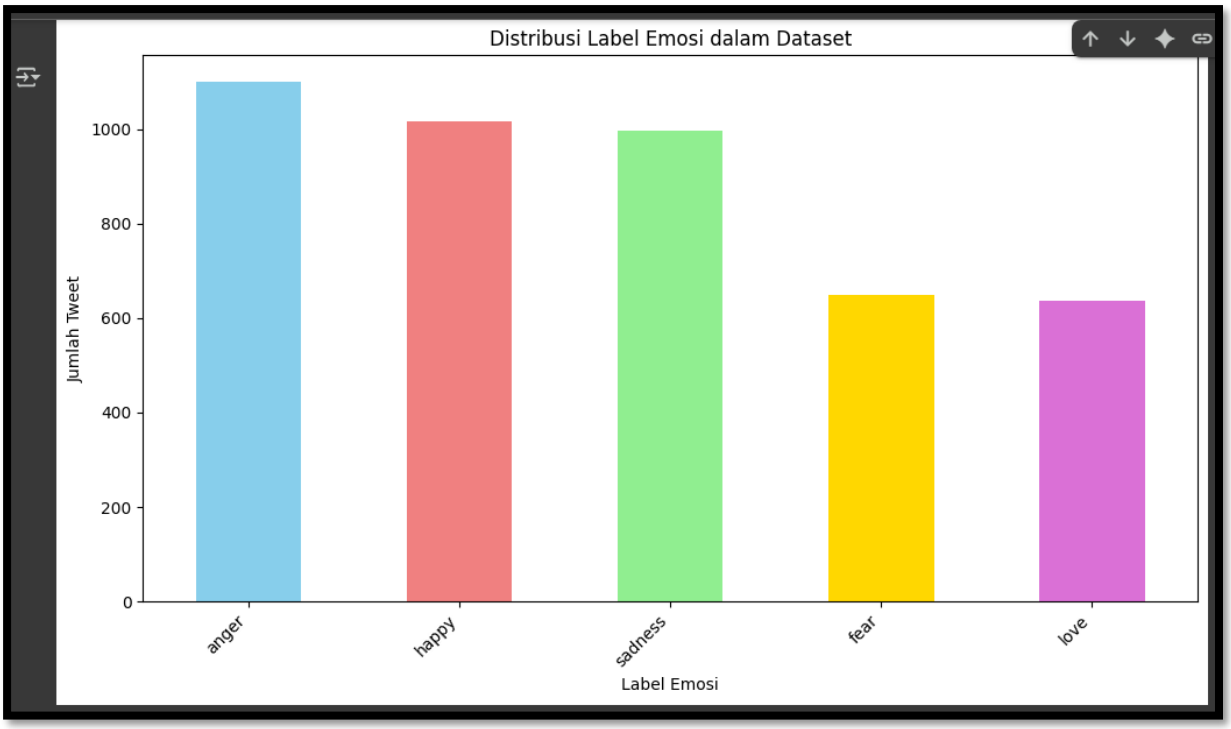
Visualisasi data

Subtask:

Membuat visualisasi untuk memahami distribusi label emosi dalam dataset.

```
# 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.

D. PEMBAHASAN TUGAS

-

E. KESIMPULAN

kita belajar bahwa data mentah perlu dibersihkan, penting untuk memahami struktur dan sebaran data, dan kita menemukan tantangan potensial (ketidakseimbangan kelas) yang perlu diatasi di tahap berikutnya. Ini adalah proses dasar yang penting dalam banyak tugas analisis data dan machine learning yang melibatkan data teks.

F. LAMPIRAN

Terlampir.

NIM :
Nama :
Kelas :

LISTING PRAKTIKUM BIG DATA ANALYTICS MODUL 1
PERTEMUAN 1

PRAKTIK
(Listing digunakan untuk memasukkan gambar dari hasil praktik dan latihan di kelas pada saat praktikum dilaksanakan)