TUGAS MANDIRI

(Fundamental Of Data Mining)

(ANALISIS SENTIMEN ULASAN APLIKASI GOJEK MENGGUNAKAN NAIVE BAYES)

Nama : Sarah Nur Anisa
NPM : 231510063
Dosen : Erlin Elisa, S.Kom., M.Kom.

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNIK DAN KOMPUTER
UNIVERSITAS PUTERA BATAM
2026

1. Deskripsi Dataset

- Sumber dataset : Kaggle.com – Gojek App Reviews (Ari Zidane)
- Jumlah record : 225.002 data ulasan
- Jumlah atribut : 5 atribut
- Dafar atribut :

No	Atribut	Deskripsi
1	userName	Nama pengguna
2	content	Teks ulasan pengguna
3	score	Rating (1–5)
4	at	Waktu ulasan
5	appVersion	Versi aplikasi

- Tipe data : teks (text mining)
- Target/label (jika supervised) : sentiment ulasan (positive, neutral, negative)
- Permasalahan yang ingin diselesaikan :
  "Melakukan analisis sentimen untuk mengklasifikasikan ulasan pengguna aplikasi Gojek ke dalam sentimen positif, netral, dan negatif secara otomatis menggunakan algoritma Data Mining."

---

2. Persiapan Data & Preprocessing

- **Data cleaning**
  - Menghapus data kosong pada kolom content
  - Memastikan seluruh data teks valid
- **Label Encoding**
  Label sentimen dibentuk berdasarkan nilai rating (score):

Score	Sentimen
4 – 5	Positive
3	Neutral
1 – 2	Negative

- **Text Processing**
  - Mengubah teks menjadi huruf kecil
  - Menghilangkan karakter non-alfabet
  - Transformasi teks ke bentuk numerik menggunakan TF-IDF Vectorizer
- **Split Data**

  Data dibagi menggunakan metode train-test split:

Tipe Data	Persentase
Training	80%
Testing	20%

Sebelum & sesudah preprocessing

Tahap	Kondisi
Sebelum preprocessing	Data teks mentah, belum memiliki label sentimen
Sesudah preprocessing	Data bersih, berlabel sentimen, berbentuk numerik (TF-IDF)

Analisis Statistik & Visualisasi

- Statistik deskriptif dataset

Dataset yang digunakan pada penelitian ini memiliki total 225.002 data ulasan dengan tiga kelas sentimen, yaitu positif, negatif, dan netral. Dataset kemudian dibagi menjadi data training (80%) dan data testing (20%) untuk keperluan evaluasi model.

- Distribusi sentiment data testing

Sentimen	Jumlah Data
Positive	32.030
Negative	15.527
Neutral	2.231

Insight:

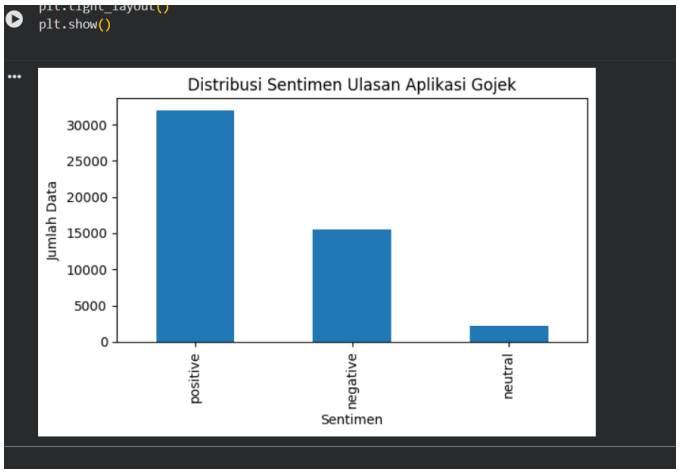
Distribusi tersebut menunjukkan bahwa kelas sentimen positif tetap mendominasi, sedangkan kelas sentimen netral memiliki jumlah data paling sedikit. Hal ini menandakan bahwa data testing bersifat tidak seimbang (imbalanced).

- Visualisasi

Visualisasi distribusi sentimen dibuat dalam bentuk grafik batang (bar chart) menggunakan library Matplotlib pada Google Colab. Grafik menunjukkan perbandingan jumlah data pada masing-masing kelas sentimen. Berdasarkan grafik distribusi sentimen, terlihat bahwa kelas sentimen positif

memiliki jumlah data paling besar, diikuti oleh sentimen negatif, dan sentimen netral sebagai kelas minoritas. Kondisi ini berpotensi mempengaruhi performa model klasifikasi, khususnya dalam memprediksi kelas sentimen netral.

Gambar 3.1 Distribusi Sentimen Ulasan Aplikasi Gojek (Data Testing)



Insight:

- Dataset bersifat **imbalanced**
- Sentimen netral memiliki jumlah data paling sedikit
- Ketidakseimbangan kelas mempengaruhi performa model

---

4. Pemilihan dan Penerapan Algoritma

Tuliskan:

- Nama algoritma : multinominal naïve bayes
- Alasan pemilihan :
    1. Cocok untuk klasifikasi data teks
    2. Efisien untuk dataset besar
    3. Sering digunakan dalam analisis sentimen

📌 algoritma yang diuji :

Algoritma	Library Python	Tujuan
Naive Bayes	sklearn.naive_bayes	Klasifikasi sentimen

5️⃣ Pengujian dan Evaluasi Model

- **Metode Evaluasi**

  Jenis tugas: **Klasifikasi**

  Metrik evaluasi yang digunakan:

  - Accuracy
  - Precision
  - Recall
  - F1-Score

- **Hasil Evaluasi Model**

Metrik	Nilai
Accuracy	0.883

- **Classification Report**

Sentimen	Precision	Recall	F1-Score
Negative	0.73	0.91	0.81
Neutral	0.42	0.00	0.01

Sentimen	Precision	Recall	F1-Score
Positive	0.95	0.93	0.94

6️⃣ Analisis & Interpretasi Hasil

“Hasil pengujian menunjukkan bahwa model Multinomial Naive Bayes mampu mencapai akurasi sebesar 88,3%. Model menunjukkan performa yang sangat baik dalam mengklasifikasikan sentimen positif dan negatif, yang ditunjukkan oleh nilai precision dan recall yang tinggi pada kedua kelas tersebut.

Namun, performa model pada kelas sentimen netral masih rendah. Hal ini disebabkan oleh jumlah data sentimen netral yang jauh lebih sedikit dibandingkan kelas lainnya, baik pada data training maupun data testing. Ketidakseimbangan data (imbalanced dataset) menyebabkan model kesulitan dalam mempelajari pola sentimen netral secara optimal.”

7️⃣ Kesimpulan & Rekomendasi

- **Kesimpulan**

1. Data Mining berhasil digunakan untuk analisis sentimen ulasan aplikasi Gojek.
2. Algoritma Naive Bayes memberikan performa klasifikasi yang baik.
3. Akurasi model mencapai 88,3%.

- **Rekomendasi**

1. Melakukan balancing data untuk meningkatkan performa kelas netral.
2. Mencoba algoritma lain seperti SVM atau Random Forest.

3. Menambahkan preprocessing lanjutan seperti stemming dan stopword removal Bahasa Indonesia.

Lampiran (Opsional)

- Cuplikan kode Python (jika tidak ditaruh di bagian utama)
  a. Memuat dataset

```python
import pandas as pd

data = pd.read_csv('gojekreview.csv')
data.head()
```

Kode di atas digunakan untuk memuat dataset ulasan aplikasi Gojek ke dalam DataFrame menggunakan library Pandas.

  b. Pembentukan label sentimen

```python
def label_sentiment(score):
    if score >= 4:
        return 'positive'
    elif score == 3:
        return 'neutral'
    else:
        return 'negative'

data['sentiment'] = data['score'].apply(label_sentiment)
```

Kode ini digunakan untuk membentuk label sentimen berdasarkan nilai rating pengguna.

  c. Transformasi teks menggunakan TF-IDF

```python
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf = TfidfVectorizer(max_features=5000)
X = tfidf.fit_transform(data['content'])
y = data['sentiment']
```

TF-IDF digunakan untuk mengubah data teks menjadi representasi numerik agar dapat diproses oleh algoritma klasifikasi.

d. Pembagian data dan pelatihan model

```
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

model = MultinomialNB()
model.fit(X_train, y_train)
```

Dataset dibagi menjadi data training dan data testing dengan rasio 80:20, kemudian model Multinomial Naive Bayes dilatih menggunakan data training.

e. Evaluasi model

```
from sklearn.metrics import accuracy_score, classification_report

y_pred = model.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

Evaluasi model dilakukan untuk mengukur performa klasifikasi menggunakan metrik accuracy, precision, recall, dan F1-score.

- Link repository (GitHub/Drive/Colab) : https://drive.google.com/drive/folders/12YNYLNzsUV608-8iGDCPKGp3wj5YRu1B?usp=sharing