



**FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO SEMARANG**

JL. IMAM BONJOL NO. 207 SEMARANG TELP. 024-3575915, 024-3575916

JAWABAN UJIAN AKHIR SEMESTER GANJIL 2023/2024

Mata Kuliah : Analitika Media Sosial	Sifat : Take Home
Hari/tanggal : Jumat, 12 Januari 2024	Waktu : 10.20 – 12.00
Kelompok : A12.6503	Dosen : Ika Novita Dewi, MCS
NIM : A12.2021.06620	Nama : Chusnuut Tacharri

1. Nama Dataset

Sentiment Intensif Mobil Listrik Indonesia

Link Dataset: <https://www.kaggle.com/datasets/billycemerson/analisis-sentimen-terkait-intensif-mobil-listrik>

2. Tujuan Sentiment Analysis

- a. Untuk mengetahui opini masyarakat terhadap adanya mobil listrik
- b. Untuk menganalisis umpan balik masyarakat untuk memahami kepuasan atau ketidakpuasan masyarakat terkait dengan adanya mobil listrik.
- c. Untuk memberikan informasi yang dapat digunakan oleh perusahaan dalam pengambilan keputusan terkait dengan mobil listrik.

3. Library yang digunakan

- a. Numpy: Digunakan untuk operasi numerik dan manipulasi array.
- b. Pandas: Digunakan untuk manipulasi dan analisis data, membaca dan menulis data berbagai format (CSV, Excel, SQL, dan lainnya).
- c. Matplotlib: Digunakan untuk visualisasi data, menggambar grafik, plot, histogram, dan visualisasi lainnya.
- d. Seaborn: Pustaka untuk membuat grafisk statistik yang menarik dan informative, dibangun di atas Matplotlib dan menyediakan antarmuka yang lebih tinggi untuk membuat visualisasi yang informative.
- e. Sklearn: Menyediakan alat dan fungsionalitas untuk analisis data, pemodelan machine learning dan evaluasi model.



FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO SEMARANG

Jl. IMAM BONJOL NO. 207 SEMARANG TELP. 024-3575915, 024-3575916

- f. Spacy: Pustaka pemrosesan bahasa alami (NLP) yang kuat, digunakan untuk tugas-tugas seperti tokenisasi, pemodelan bahasa, ekstraksi entitas, dan pemahaman konteks teks.
- g. Natural Language Toolkit (NLTK): Pustakan pemrosesan bahasa alami yang menyediakan alat dan sumber daya untuk tugas NLP, termasuk korpus, tokenizer, dan Algoritma pemrosesan bahasa alami.

4. Exploratory Data Analysis (EDA)

a. Menampilkan dataset

```
df = pd.read_csv("//content/mobil_listrik.csv")  
df.head()
```

Tampilan hasil:

	id_komentar	nama_akun	tanggal	text_cleaning	sentimen
0	UgzblI5eyrly3-gdUUJ4AaABAg	Sqn Ldr	2023-08-06 12:54:49+00:00	saran sih bikin harga ionic sama kayak brio ...	positif
1	UgzEDUIv3OTrV943p8p4AaABAg	lushen ace	2023-08-04 12:16:23+00:00	problem subsidi kualitas diturunin harga dinai...	negatif
2	UgwqJqu6JMF4EH2CsVV4AaABAg	Fatih Al-Ayyubi	2023-08-04 10:17:57+00:00	baik kualitas kembang dulu baik kualitas motor...	positif
3	UgyYicCMR1rKwuOj2Y14AaABAg	yp office	2023-08-04 08:29:54+00:00	model jelek kwalitas buruk harga mahal croot	negatif
4	UgxKAcLuAwZQK6es-x4AaABAg	Lembur Kuring	2023-08-04 07:55:37+00:00	syarat ngaco woy anak muda blom punya ruma...	negatif

b. Deskripsi dataset

```
df.info()
```

Tampilan hasil:

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 1517 entries, 0 to 1516  
Data columns (total 5 columns):  
#   Column          Non-Null Count  Dtype  
---  ---  
0   id_komentar     1517 non-null   object  
1   nama_akun       1516 non-null   object  
2   tanggal         1517 non-null   object  
3   text_cleaning   1515 non-null   object  
4   sentimen        1517 non-null   object  
dtypes: object(5)  
memory usage: 59.4+ KB
```




FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO SEMARANG

Jl. IMAM BONJOL NO. 207 SEMARANG Telp. 024-3575915, 024-3575916

c. Perbandingan jumlah data tiap kelas

```
df_counts = df["sentimen"].value_counts().reset_index()
df_counts.head()
```

Tampilan hasil:



	index	sentimen
0	negatif	869
1	positif	504
2	netral	144

d. Feature shape


```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf = TfidfVectorizer(
    sublinear_tf=True,
    min_df=5,
    norm='l2',
    encoding='latin-1',
    ngram_range=(1,2),
    stop_words='english'
)

features = tfidf.fit_transform(df.text).toarray()
labels = df.sentimen_id

features.shape
```

Tampilan hasil:



(1515, 977)



e. Word Cloud

Tampilan hasil:





FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO SEMARANG

Jl. IMAM BONJOL NO. 207 SEMARANG TELP. 024-3575915, 024-3575916

5. Sentiment Analysis

a. Text Processing

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer

from sklearn.naive_bayes import MultinomialNB

X_train, X_test, y_train, y_test = train_test_split(
    df['text'],
    df['sentimen_id'],
    random_state=0
)

count_vect = CountVectorizer()
X_train_counts = count_vect.fit_transform(X_train)

tfidf_transformer = TfidfTransformer()
X_train_tfidf = tfidf_transformer.fit_transform(X_train_counts)

clf = MultinomialNB().fit(X_train_tfidf, y_train)

sample1 = df.sample(1)
print(sample1.sentimen)
print(df.text[sample1.index[0]])

pred =
clf.predict(count_vect.transform([df.text[sample1.index[0]]]))
print(mapping_index[pred][0])

sample2 = df.sample(1)
print(sample2.sentimen)
print(df.text[sample2.index[0]])

pred =
clf.predict(count_vect.transform([df.text[sample2.index[0]]]))
print(mapping_index[pred][0])

pred = clf.predict(count_vect.transform([df.text[1500]]))

print(mapping_index[pred][0])
```



negatif



FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO SEMARANG

Jl. IMAM BONJOL NO. 207 SEMARANG TELP. 024-3575915, 024-3575916

b. Modeling

```
# find the best model

from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import LinearSVC

from sklearn.model_selection import cross_val_score

models = [
    LogisticRegression(random_state=0),
    RandomForestClassifier(n_estimators=200,max_depth=3,random_state=0),
    LinearSVC(),
    MultinomialNB()
]

CV = 5
cv_df = pd.DataFrame(index=range(CV * len(models)))

entries = []
for model in models:
    model_name = model.__class__.__name__
    accuracies = cross_val_score(model, features, labels,
    scoring='accuracy', cv=CV)

    for fold_idx, accuracy in enumerate(accuracies):
        entries.append((model_name, fold_idx, accuracy))

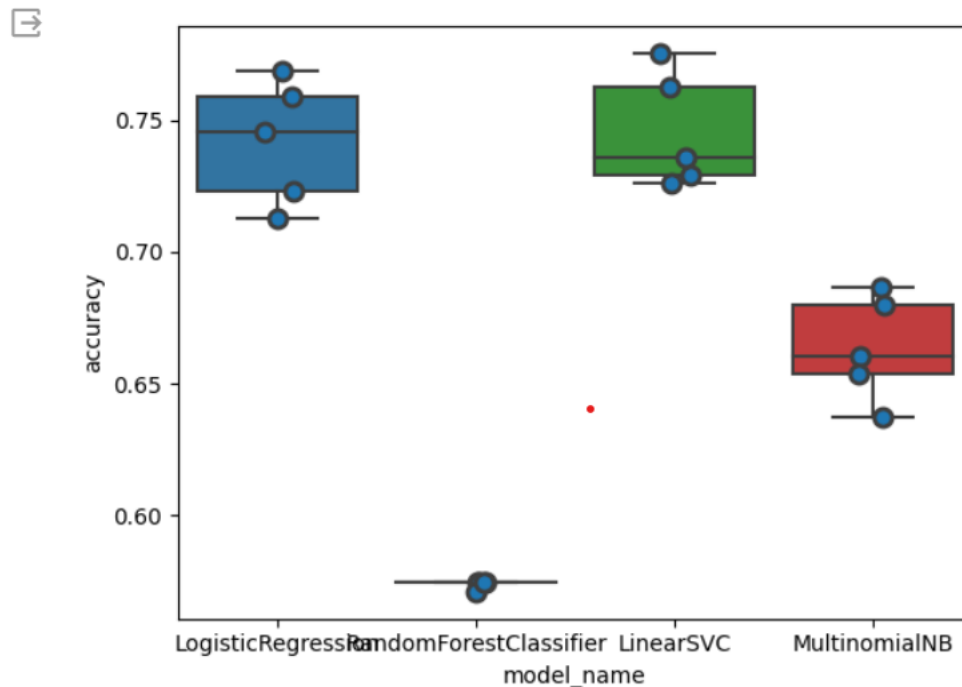
cv_df = pd.DataFrame(entries, columns=['model_name', 'fold_idx',
'accuracy'])

import seaborn as sns

sns.boxplot(x='model_name', y='accuracy', data=cv_df)
sns.stripplot(x='model_name', y='accuracy', data=cv_df,
              size=8, jitter=True, edgecolor="gray", linewidth=2)

plt.show()
```

Tampilan hasil:



c. Confusion Matrix

```
from sklearn.svm import LinearSVC
import seaborn as sns

model = LinearSVC()
X_train, X_test, y_train, y_test, indices_train, indices_test =
train_test_split(features, labels, df.index, test_size=0.33,
random_state=0)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

from sklearn.metrics import confusion_matrix

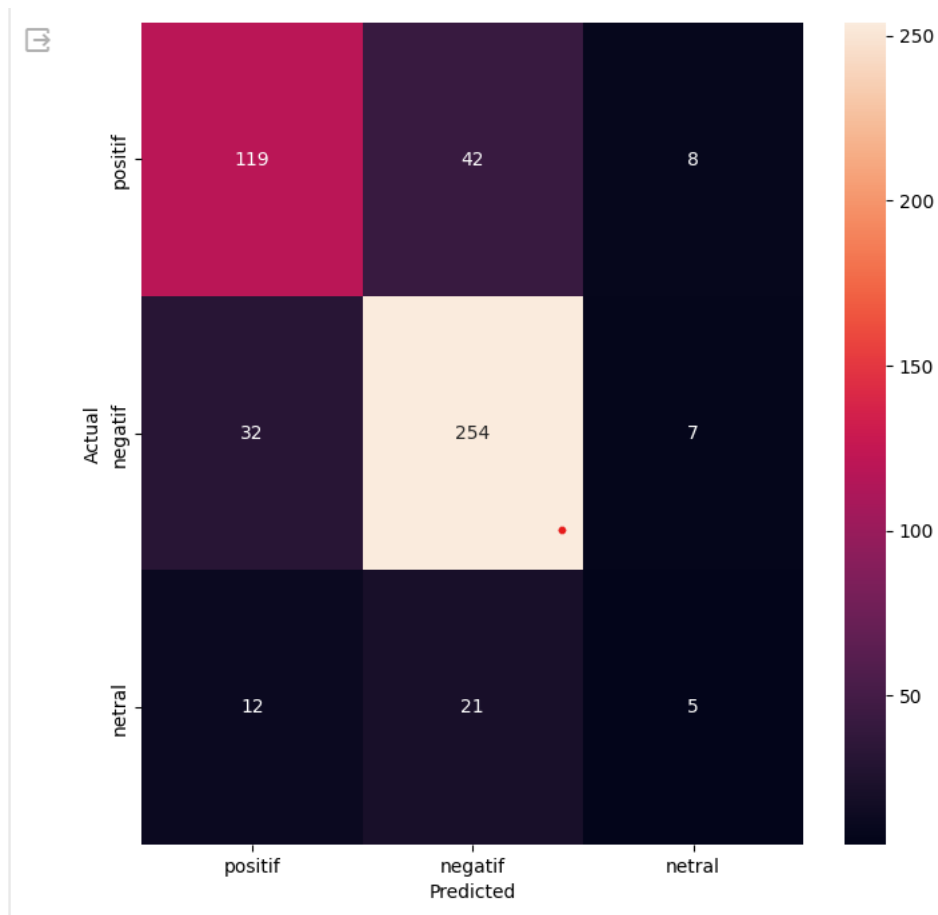
conf_mat = confusion_matrix(y_test, y_pred)
fig, ax = plt.subplots(figsize=(8,8))
sns.heatmap(conf_mat, annot=True, fmt='d',
            xticklabels=sentimen_id_df.sentimen.values,
            yticklabels=sentimen_id_df.sentimen.values)
plt.ylabel('Actual')
plt.xlabel('Predicted')
plt.show()
```




FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO SEMARANG

Jl. IMAM BONJOL NO. 207 SEMARANG TELP. 024-3575915, 024-3575916

Tampilan hasil:



d. Performance

```
from sklearn import metrics  
  
print(metrics.classification_report(y_test, y_pred,  
target_names=df['sentimen'].unique()))
```



FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO SEMARANG

Jl. IMAM BONJOL NO. 207 SEMARANG TELP. 024-3575915, 024-3575916

Tampilan hasil:

	precision	recall	f1-score	support
positif	0.73	0.70	0.72	169
negatif	0.80	0.87	0.83	293
netral	0.25	0.13	0.17	38
accuracy			0.76	500
macro avg	0.59	0.57	0.57	500
weighted avg	0.74	0.76	0.74	500

•

6. Kesimpulan

Berdasarkan analisis word cloud dari visualisasi data sentimen analisis mobil listrik mayoritas kata yang muncul menunjukkan sentimen negatif terkait mobil listrik, subsidi, harga, station pengisian, dengan kekhawatiran terhadap dampaknya terhadap kesejahteraan rakyat. Model yang paling baik digunakan di sentimen analisis mobil listrik ini yaitu model LinearSVC dengan *accuracy* nilai 0.745875. Dari confusion matrix di atas nilai tertinggi berada di antara *actual negative* dan *predicted negative* (TN) dimana menghasilkan sebesar 254, bahwa jumlah komentar dan diprediksi negatif dengan benar. *Performace* yang didapat berdasarkan visualisasi di atas yaitu menghasilkan sentimen negatif dengan *precision*, *recall*, *F1-score* sebesar 80%, 87%, 83% dimana model memiliki presisi yang tinggi dalam mengidentifikasi kelas negatif, dan recall yang cukup tinggi menunjukkan bahwa model secara efektif mengenali sebagian besar *instance* yang sebenarnya negatif. *F1-score* yang tinggi menunjukkan keseimbangan yang baik antara presisi dan recall. Akurasi model yang didapat sebesar 76% menunjukkan seberapa baik model dapat mengklasifikasikan secara benar pada semua kelas.