



**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	: A12.2021.06612	Nama	: Maulida Aristia Tsani

1. Nama Dataset

Indonesian Covid19 Vaccination

Link Dataset : <https://www.kaggle.com/datasets/kharisma4792/indonesian-covid19-vaccination-tweets>

2. Tujuan Sentiment Analysis

Untuk mengetahui opini atau tanggapan masyarakat terhadap program vaksin covid-19 dan mengidentifikasi sentimen positif, negative dan netral. Informasi yang dihasilkan dapat membantu pemerintah, Lembaga kesehatan untuk merancang strategi komunikasi yang lebih efektif dan mengukur tingkat penerimaan masyarakat terhadap vaksin covid-19.

3. Library yang digunakan

- a. Numpy : untuk operasi numerik dan manipulasi array.
- b. Pandas : untuk manipulasi dan analisis data tabular.
- c. Matplotlib.pyplot : untuk membuat visualisasi data seperti grafik dan plot.
- d. Seaborn : untuk membuat visualisasi data statistic yang lebih menarik dan informatif.
- e. Sklearn : menyediakan alat dan fungsionalitas untuk analisis data, pemodelan machine learning, dan evaluasi model.
- f. Spacy : menyediakan alat untuk melakukan tugas seperti tokenisasi, pengenalan entitas, penguraian sintaksis dan analisis sentiment. Selain itu, library spacy merupakan library pemrosesan Bahasa alami (NLP) untuk analisis teks.



FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO SEMARANG

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

- g. Nltk : untuk membantu melakukan pemrosesan Bahasa alami dengan mudah

4. Exploratory Data Analysis (EDA)

a. Menampilkan dataset

```
df_raw = pd.read_csv("/content/data_berlabel.csv")

df_raw.head()
```

Tampilan hasil:

	date	id	text	username	like_count	displayname	lang	sentimen	year
0	2021-12-31 23:33:04+00:00	1477060265057214466	Indonesia punya stok \nVaksin covid 19, 3 bula...	ahimcakep	1	ahimcakep ðððððððð	in	netral	2021
1	2021-12-31 23:28:02+00:00	1477059000378163202	Vaksinasi Covid-19 Dosis Keempat Sudah Boleh d...	haluanharian	0	Harian Haluan	in	netral	2021
2	2021-12-31 23:03:37+00:00	1477052854753906688	Dari awal covid 2020-2021-2022. Sukur kagak ke...	Tebobotee1	0	Tebobotee	in	netral	2021
3	2021-12-31 22:57:02+00:00	1477051199438286853	2021 :\nVaksin covidâððððððð\nKB3 bubar, pindah ...	nursetiandi	0	Nur Setiandi	in	netral	2021
4	2021-12-31 22:45:49+00:00	1477048376294207488	Satu kalimat utk tahun 2021: "Alhamdulillah mas...	FauziAdri	2	Fauzi Adrianto	in	positif	2021

b. Deskripsi dataset

```
df_raw.info()
```

Tampilan hasil:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10727 entries, 0 to 10726
Data columns (total 9 columns):
#   Column          Non-Null Count  Dtype
---  -
0   date            10727 non-null  object
1   id              10727 non-null  int64
2   text            10727 non-null  object
3   username        10727 non-null  object
4   like_count      10727 non-null  int64
5   displayname     10722 non-null  object
6   lang            10727 non-null  object
7   sentimen        10727 non-null  object
8   year            10727 non-null  int64
dtypes: int64(3), object(6)
memory usage: 754.4+ KB
```

c. Perbandingan jumlah data tiap kelas

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

Tampilan hasil:

	index	sentimen
0	positif	4576
1	netral	4354
2	negatif	1797

d. Feature shape

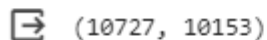
```
from sklearn.feature_extraction.text import TfidfVectorizer

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_raw.text).toarray()
labels = df_raw.sentimen

print(features.shape)
```

Tampilan hasil:



```
(10727, 10153)
```

e. Word Cloud

```
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator

# Most famous nouns used in Tweet Bapak Jokowi

wordcloud = WordCloud(max_font_size=50, max_words=100,
    background_color="black").generate(' '.join(noun))
plt.figure(figsize=(12, 10))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```





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])
```



Tampilan hasil:

→ netral

positif

→ positif



FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO SEMARANG

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

b. Modeling

```
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))

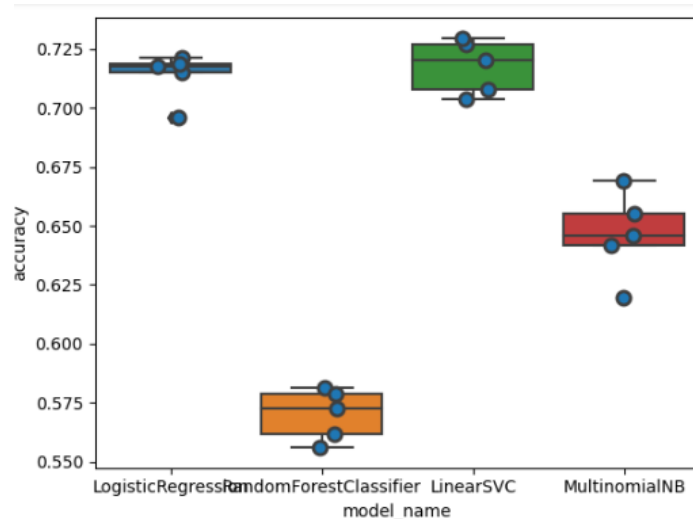
cv_df = pd.DataFrame(entries, columns=['model_name', 'fold_idx',
'accuracy'])

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

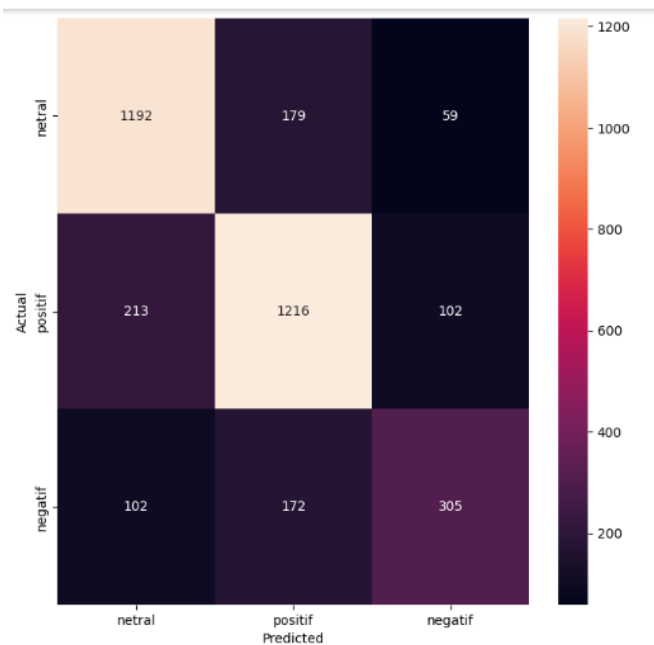
conf_mat = confusion_matrix(y_test, y_pred)
fig, ax = plt.subplots(figsize=(8,8))
sns.heatmap(conf_mat, annot=True, fmt='d',
xticklabels=sentimen_id_df.sentimen.values,
yticklabels=sentimen_id_df.sentimen.values)
plt.ylabel('Actual')
plt.xlabel('Predicted')
plt.show()
```



**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()))
```

Tampilan hasil:

	precision	recall	f1-score	support
netral	0.79	0.83	0.81	1430
positif	0.78	0.79	0.79	1531
negatif	0.65	0.53	0.58	579
accuracy			0.77	3540
macro avg	0.74	0.72	0.73	3540
weighted avg	0.76	0.77	0.76	3540



**FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO SEMARANG**

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

6. Kesimpulan

Berdasarkan visualisasi wordcloud sentimen analisis vaksin Covid-19 mayoritas kata yang muncul menunjukkan sentimen positif terhadap vaksinasi. Masyarakat masih aktif dalam mendukung program vaksinasi, khususnya vaksin jenis “Astrazeneca” dan “Pfizer”. Model yang paling baik digunakan di sentimen analisis vaksin Covid-19 adalah model Linear SVC dengan nilai akurasi sebesar 0.718. Dari confusion matrix diatas, nilai tertinggi terdapat diantara actual positif dan predicted positif atau TP (True Positif) sebesar 1216 yang berarti jumlah komentar yang benar-benar positif dan di prediksi positif dengan benar. Performa yang didapat dari sentimen analisis vaksin Covid-19 yaitu menghasilkan sentimen positif dengan nilai precision 78%, recall 79% dan f1-score 79%. Model memiliki akurasi sekitar 77% dimana model mungkin memiliki kinerja yang cukup baik dalam mengklasifikasikan ke kelas masing-masing. Secara keseluruhan, mencerminkan sikap positif dan kesadaran masyarakat terhadap pentingnya vaksinasi Covid-19 sebagai bagian dari strategi pencegahan yang berkelanjutan.