PEMBELAJARAN MESIN

LAPORAN UTS

Disusun oleh :

- ✓ Afif Qomarul Ghunlam / 2041720176
- ✓ Iqhsan Bagus Prasetyo / 2041720096
- ✓ Muhammad Al Kausar Ramadhan / 2041720193
- ✓ Muhamad Alif Rizmi / 2041720196
- ✓ R. Muhamad Azmi Herdi Shofiyullah / 2041720079

PROGRAM STUDI D4 TEKNIK INFORMATIKA

JURUSAN TEKNOLOGI INFORMASI

POLITEKNIK NEGERI MALANG

2022

LANGKAH – LANGKAH PRATIKUM

1. Kode program dibawah merupakan library – library yang digunakan

```
import numpy as np
import pandas as pd
import re

import nltk
nltk.download('wordnet')
from nltk.corpus import stopwords
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_extraction.text import CountVectorizer,TfidfTransformer
from sklearn.model_selection import train_test_split
from scipy.stats import itemfreq
```

[1] ✓ 7.2s

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\user\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

2. Kode program dibawah digunakan untuk mengambil file “tweet_emotions.csv”, lalu dibawahnya untuk menampilkan datanya sebanyak 50

```
df = pd.read_csv('data/tweet_emotions.csv')
df.head(50)
```

[2] ✓ 0.4s

	tweet_id	sentiment	content
0	1956967341	empty	@tiffanylue i know i was listenin to bad habi...
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...
2	1956967696	sadness	Funeral ceremony...gloomy friday...
3	1956967789	enthusiasm	wants to hang out with friends SOON!
4	1956968416	neutral	@dannycastillo We want to trade with someone w...

3. Lalu variable data digunakan untuk menampung isian dari variable df. Lalu varable data sentiment digunakan untuk mencari data disgust, dan hate pada column sentiment, kemudian isin digunakan untuk memberikan nilai true pada data yang ada di column sentiment selain disgust dan hate

```
data = df
data.head()
```

[3] ✓ 0.6s

	tweet_id	sentiment	content
0	1956967341	empty	@tiffanylue i know i was listenin to bad habi...
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...
2	1956967696	sadness	Funeral ceremony...gloomy friday...
3	1956967789	enthusiasm	wants to hang out with friends SOON!
4	1956968416	neutral	@dannycastillo We want to trade with someone w...

```
data.sentiment = np.where((data.sentiment == 'disgust') |(data.sentiment == 'hate'),'hate',data.sentiment)
```

[4] ✓ 0.8s

```
data=data[data.sentiment.isin(['sadness','worry','joy'])]
```

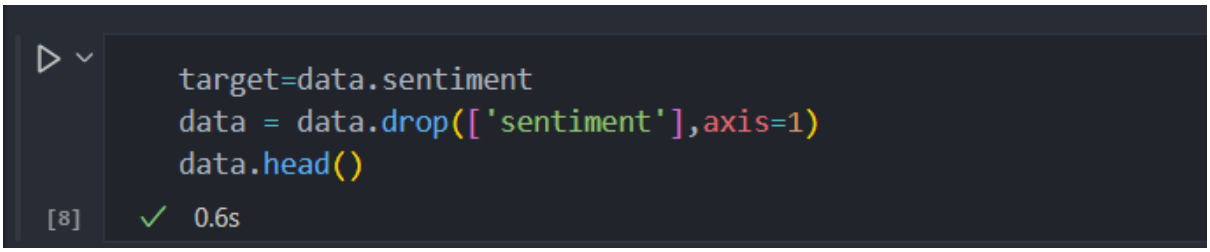
[5] ✓ 0.7s

```
data.sentiment.value_counts()
```

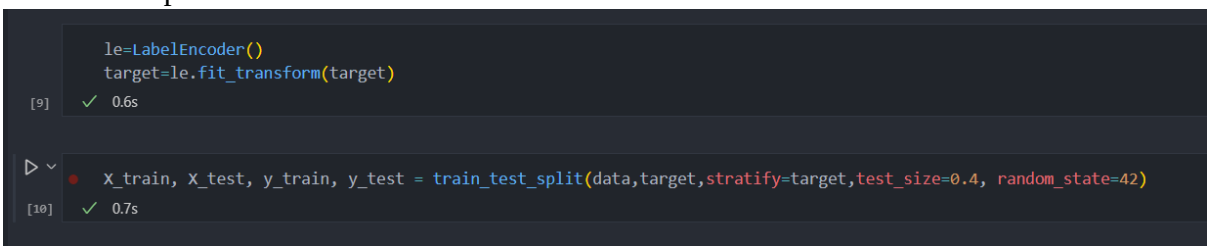
[6] ✓ 0.7s

4. Kode program dibawah digunakan untuk membersihkan data pada column content untuk menghilangkan special character dan beberapa situs website seperti http atau www

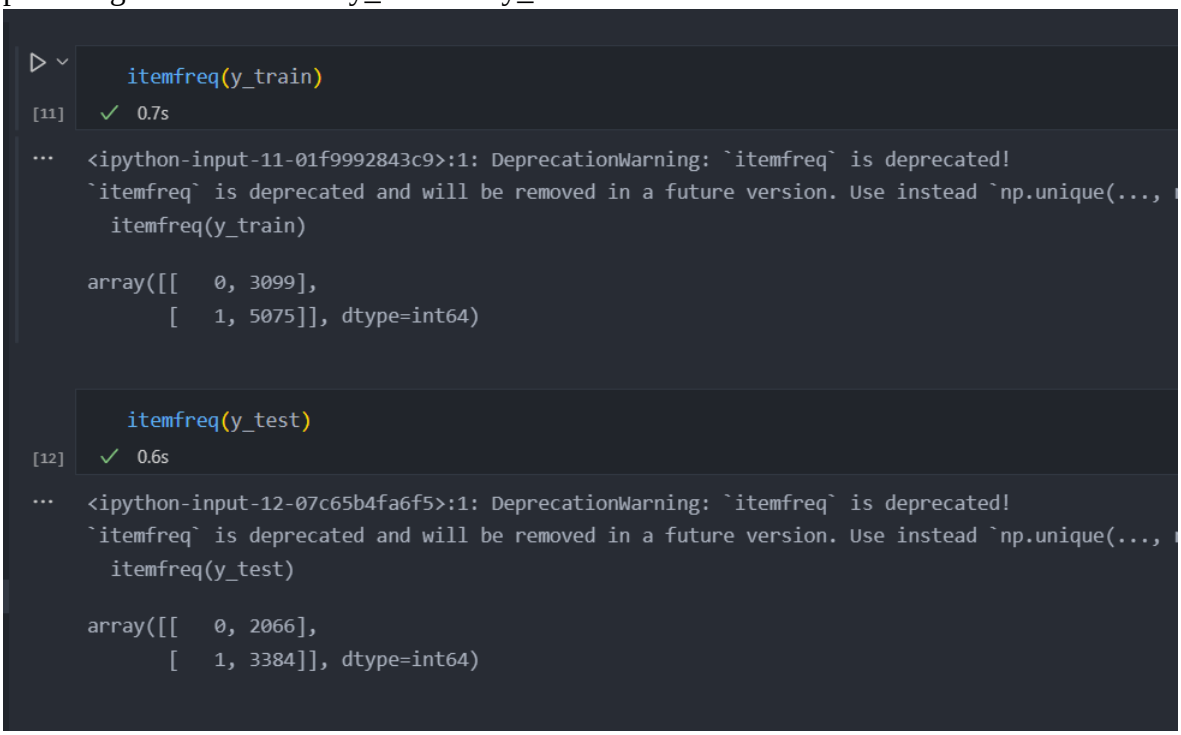
5. Kode program dibawah digunakan untuk memindahkan data yang ada di column sentiment ke variable target, lalu variable data terdapat fungsi untuk drop column sentiment



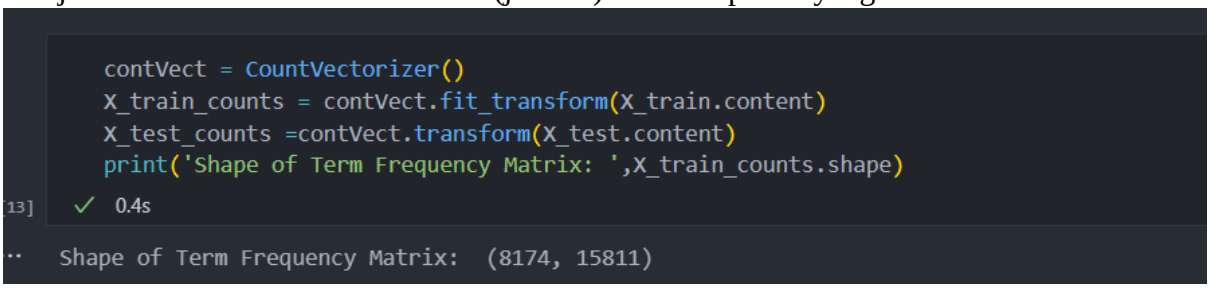
6. Kode program dibawah digunakan untuk melakukan encoder pada variable target yang berisikan data column sentiment. Lalu dibawahnya terdapat kode program untuk melakukan split test



7. Kedua kode program yang ada digambar bawah adalah untuk melakukan melakukan perhitungan frekuensi dari y_train dan y_test



8. Lalu pada kode program dibawah digunakan untuk m engubah teks yang diberikan menjadi vektor berdasarkan frekuensi (jumlah) dari setiap kata yang muncul

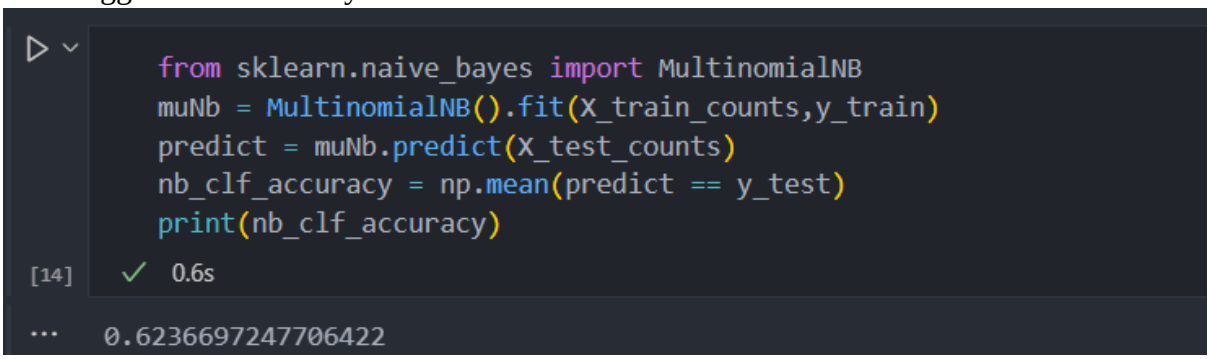


```
contVect = CountVectorizer()
X_train_counts = contVect.fit_transform(X_train.content)
X_test_counts =contVect.transform(X_test.content)
print('Shape of Term Frequency Matrix: ',X_train_counts.shape)
```

[13] ✓ 0.4s

```
Shape of Term Frequency Matrix:  (8174, 15811)
```

9. Pada kode program dibawah digunakan untuk menghitung akurasi, pada perhitungan ini menggunakan NaiveBayes

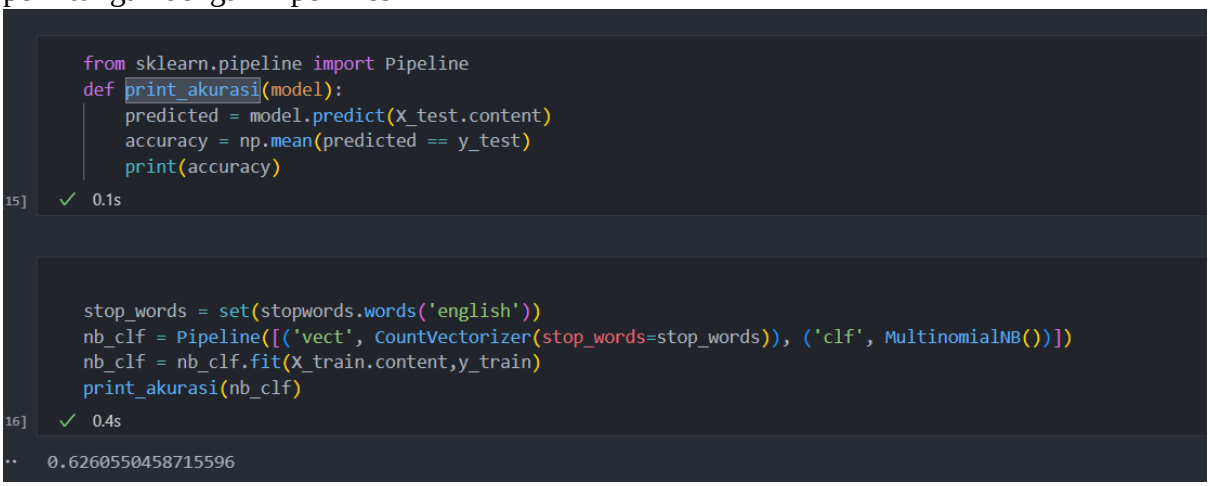


```
from sklearn.naive_bayes import MultinomialNB
muNb = MultinomialNB().fit(X_train_counts,y_train)
predict = muNb.predict(X_test_counts)
nb_clf_accuracy = np.mean(predict == y_test)
print(nb_clf_accuracy)
```

[14] ✓ 0.6s

```
0.6236697247706422
```

10. Pada kode program dibawah digunakan untuk menghitung akurasi menggunakan metode PipeLines. Pada perhitungan ini data yang ada di column test masih blm bersih oleh karena itu pada perhitungan ini kami mencoba menggunakan stop_words untuk menghilangkan kata kerja yang ada di column content, dan mencoba melakukan perhitungan dengan PipeLines



```
from sklearn.pipeline import Pipeline
def print_akurasi(model):
    predicted = model.predict(X_test.content)
    accuracy = np.mean(predicted == y_test)
    print(accuracy)
```

15] ✓ 0.1s

```
stop_words = set(stopwords.words('english'))
nb_clf = Pipeline([('vect', CountVectorizer(stop_words=stop_words)), ('clf', MultinomialNB())])
nb_clf = nb_clf.fit(X_train.content,y_train)
print_akurasi(nb_clf)
```

16] ✓ 0.4s

```
0.6260550458715596
```