

UAS

Deteksi Jenis Komentar Pengguna Twitter

Mata Kuliah : Pembelajaran Mesin

Dosen Pembimbing :

Adevian Fairuz Pratama, S.ST, M.Eng



Oleh :

TI 3E

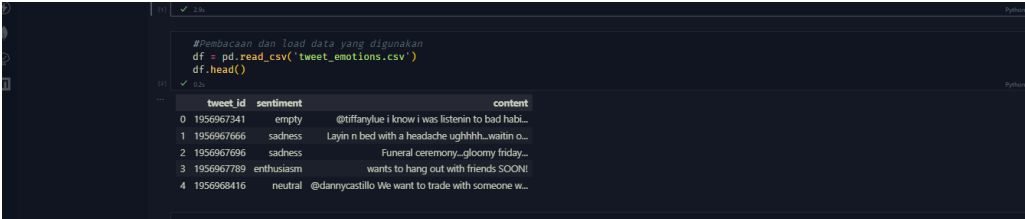
Muhammad Lazuardi Timur 2041720114

Firgi Sotya Izzuddin 2041720207

PROGRAM STUDI D4 TEKNIK INFORMATIKA

**JURUSAN TEKNOLOGI INFORMASI
POLITEKNIK NEGERI MALANG**

2022/2023

Langkah	Keterangan
1	<p>Proses import library-library yang dibutuhkan untuk pengerjaan kali ini</p> <pre data-bbox="352 472 1380 952">import numpy as np import pandas as pd import re import matplotlib.pyplot as plt import seaborn as sns from sklearn.preprocessing import LabelEncoder from sklearn.preprocessing import StandardScaler from sklearn.cluster import KMeans from sklearn.model_selection import train_test_split from sklearn.feature_extraction.text import TfidfVectorizer from sklearn.ensemble import RandomForestClassifier from sklearn.metrics import accuracy_score from sklearn.metrics import classification_report from sklearn.metrics import confusion_matrix import nltk from nltk.tokenize import sent_tokenize, word_tokenize from nltk.stem.porter import PorterStemmer from nltk.corpus import stopwords from sklearn.cluster import KMeans</pre>
2	<p>Proses load data dari tweet_emotions.csv yang akan digunakan</p> 

PREPROCESSING DATA

Case Folding : Konversi kalimat yang ada menjadi aturan lowercase

```
#CASE FOLDING :Pengubahan Huruf besar ke huruf kecil
def clean_lower(lwr):
    lwr = lwr.lower() # mengubah menjadi lowercase text
    return lwr

# Buat kolom tambahan untuk data content yang telah dicasefolding
```

Case Folding : Menghapus karakter special seperti simbol-simbol

```
#CASE FOLDING : penghapusan karakter (@, ', ", dll)
#Remove Punctuation
clean_spcl = re.compile('[/(){}[\]\|@,;]')
clean_symbol = re.compile('[^0-9a-z]')
def clean_punct(text):
    text = clean_spcl.sub('', text)
    text = clean_symbol.sub(' ', text)
    return text
```

Case Folding: Penghapusan spasi yang berlebihan seperti enter dan lain lain

```
#CASE FOLDING : Penghapusan Spasi yang berlebihan
def _normalize_whitespace(text):
    corrected = str(text)
    corrected = re.sub(r'//t',r'\t', corrected)
    corrected = re.sub(r'\s+',r'\s', corrected)
    corrected = re.sub(r'\n',r'\n', corrected)
    corrected = re.sub(r'\r',r'\r', corrected)
    corrected = re.sub(r'\t',r'\t', corrected)
    return corrected.strip(" ")
```

Tokenizing : Memisahkan fitur

Tokenizing

```
#tokenize pada index ke 1

tweet=pd.DataFrame(data)
token=nltk.tokenize.WhitespaceTokenizer().tokenize(data[0])
token

[0]: ✓ 0.3s Python

... ['tiffanylue',
     'i',
     'know',
     'i',
     'was',
     'listenin',
     'to',
     'bad',
     'habit',
     'earlier',
     'and',
     'i',
     'started',
     'freakin',
     'at',
     'his',
     'part']
```

Filtering :

Filtering

```
#filter kata tidak penting dalam bahasa inggris

nltk.download('stopwords')

w_list = stopwords.words("english")
def stopwords_removal(words):
    return [word for word in words if word not in w_list]

data.apply(stopwords_removal)
print(data.head())
```

[118] ✓ 5.8s Python

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

```
0    tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part
1                                ayin n bed with a headache ughhhh waitin on your call
2                                uneral ceremony gloomy friday
3                                wants to hang out with friends
4    dannycastillo e want to trade with someone who has ouston tickets but no one will
Name: content, dtype: object
```

Stemming :

```
#stemming index ke-10

#defining the object for stemming
porter_stemmer = PorterStemmer()

#defining a function for stemming
def stemming(text):
    stem_text = [porter_stemmer.stem(word) for word in text]
    return stem_text

data
```

[127] ✓ 0.4s Python

... 0 tiffanylue i know i was listenin to bad habit earlier and i
started freakin at his part
1 ayin n bed with a headache ughhhh
waitin on your call
2 uneral
ceremony gloomy friday
3 wants to
hang out with friends
4 dannycastillo e want to trade with someone who has ouston
tickets but no one will
...
39995 ohn loyd aylor
39996 appy
others ay ll my love
39997 appy other s ay to all the mommies out there be you woman or man as long as you re momma to
someone this is your day

CLUSTERING

encoding kedalam numerik

```
Clustering

# Menentukan jumlah cluster
stdScaler = StandardScaler()
stdScaler.fit(data)
data_scaled = stdScaler.transform(data)

data_scaled = pd.DataFrame(data_scaled, columns=['sentiment', 'encode_result_table'])

kmeans = KMeans(n_clusters=3, random_state=0)
y_predict = kmeans.fit_predict(data_scaled)
y_predict

df['komentar'] = y_predict
df
```

	tweet_id	sentiment	content	clean_punct	encode_result_table	komentar
0	1956967341	2	@tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part =[tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part	17653	2
1	1956967666	10	Layin n bed with a headache ughhhh...waitin on your call...	ayin n bed with a headache ughhhh waitin on your call	26038	0
2	1956967696	10	Funeral ceremony...gloomy friday...	uneral ceremony gloomy friday	21048	0
3	1956967789	3	wants to hang out with friends SOON!	wants to hang out with friends	39077	2
4	1956968416	8	@dannycastillo We want to trade with someone who has Houston tickets, but no one will.	dannycastillo e want to trade with someone who has ouston tickets but no one will	9591	1
...
39995	1753918954	8	@JohnLloydTaylor	ohn loyd aylor	3597	1
39996	1753919001	7	Happy Mothers Day All my love	appy others ay ll my love	22134	2

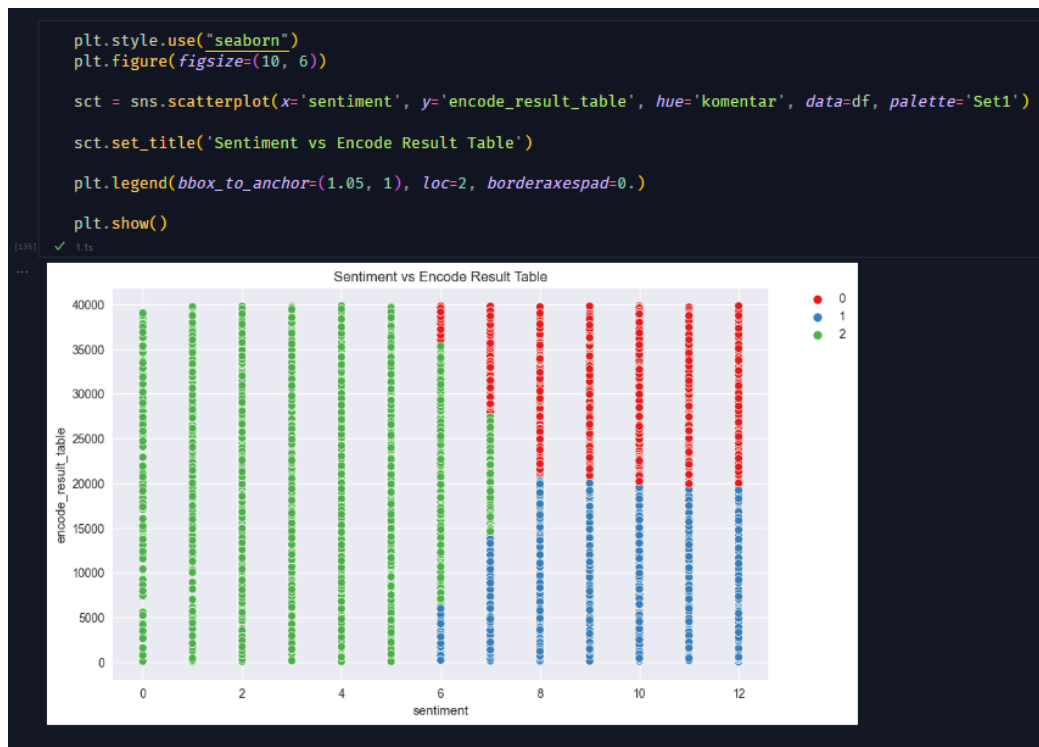
Data Scaling

Tujuan dari data scaling ini adalah membuat data berada dalam rentang (range) yang tidak terlalu jauh. Terkadang di dalam data muncul outlier atau pencilan, yang nanti bisa mengganggu dalam proses clustering. Untuk data scaling kita akan menggunakan StandardScaler, yang akan diimplementasikan terhadap data yang ada pada kolom 'sentiment' dan 'content(encode)'

K-Means Clusterring

pada bagian y_predicted , kita melakukan clustering dari persebaran data penghasilan terhadap sentiment, lalu menyimpan hasilnya di variabel y_predicted . Lalu pada bagian akhir yaitu df_columns['jenis_komentar'] = y_predicted , kita memasukkan hasil clustering yang berupa one-dimensional array ke dataframe di kolom 'jenis_komentar'.

Note: nilai hasil cluster (nilai 0, 1, 2) tiap data bisa jadi berbeda dengan hasil di atas, namun setiap data masih berada dalam kelompok/kategori yang sama dengan data lainnya.



Dari cluster di atas, coba perhatikan pada sentiment yang minimal terdapat 8 orang. Pada daerah tersebut, kita bisa melihat bahwa warna merah melambangkan kelompok yang contentnya negative lebih tinggi dibanding warna hitam yang contentnya netral. Artinya tingkat content negative pada sentiment 6 lebih tinggi disbanding dengan content netral

Labeling

```

new_cols = {
    'sentiment' : 'label_sentiment',
    'komentar' : 'label_komentar'
}

df = df.rename(columns=new_cols)
df

```

	tweet_id	label_sentiment	content	clean_punct	encode_result_table	label_komentar
0	1956967341	2	@tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part =[tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part	17653	2
1	1956967666	10	Layin n bed with a headache ughhhh...waitin on your call...	ayin n bed with a headache ughhhh waitin on your call	26038	0
2	1956967696	10	Funeral ceremony...gloomy friday...	uneral ceremony gloomy friday	21048	0
3	1956967789	3	wants to hang out with friends SOON!	wants to hang out with friends	39077	2
4	1956968416	8	@dannycastillo We want to trade with someone who has Houston tickets, but no one will.	dannycastillo e want to trade with someone who has ouston tickets but no one will	9591	1
...
39995	1753918954	8	@JohnLloydTaylor	ohn loyd aylor	3597	1
39996	1753919001	7	Happy Mothers Day All my love	appy others ay ll my love	22134	2
39997	1753919005	7	Happy Mother's Day to all the mommies out there, be you woman or man as long as you're 'momma' to someone this is your day!	appy other s ay to all the mommies out there be you woman or man as long as you re momma to someone this is your day	22067	2
39998	1753919043	5	@niallley WASUP BEAUTIFUL!!! FOLLOW ME!! PEEP OUT MY NEW HIT SINGLES WWW.MYSPACE.COM/IPSOHOT I DEF. WAT U IN THE VIDEO!!	niallley	14943	2
39999	1753919049	7	@mopedronin bullet train from tokyo the gf and i have been visiting japan since thursday vacation/sightseeing gaijin godzilla	mopedronin bullet train from tokyo the gf and i have been visiting japan since thursday vacationsightseeing gaijin godzilla	14603	2

Hasil dari labelling

```

df = pd.DataFrame({
    'label' : df['label_sentiment'],
    'clean' : df['clean_punct'],
    'encode_result_table' : df['encode_result_table'],
    'komentar' : df['label_komentar']
})
df

```

	label	clean	encode_result_table	komentar
0	2	tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part	17653	2
1	10	ayin n bed with a headache ughhhh waitin on your call	26038	0
2	10	uneral ceremony gloomy friday	21048	0
3	3	wants to hang out with friends	39077	2
4	8	dannycastillo e want to trade with someone who has ouston tickets but no one will	9591	1
...
39995	8	ohn loyd aylor	3597	1
39996	7	appy others ay ll my love	22134	2
39997	7	appy other s ay to all the mommies out there be you woman or man as long as you re momma to someone this is your day	22067	2
39998	5	niallley	14943	2
39999	7	mopedronin bullet train from tokyo the gf and i have been visiting japan since thursday vacationsightseeing gaijin godzilla	14603	2

40000 rows x 4 columns

Memisahkan fitur dengan label

Terdapat variabel x (data source) dan y (data target) yang masing-masing menyimpan value dari kolom content dan kolom sentiment.

Klasifikasi dan Prediksi

```

# #Memisahkan Fitur dengan Label
X = df['hasil_tweet'].values
y = df['Labels_Komentar'].values

```

Klasifikasi dan Prediksi

```
# Memisahkan Fitur dengan Label
X = df['hasil_tweet'].values
y = df['Labels_komentar'].values
```

Python

```
# EKSTRAKSI FITUR
# Ekstraksi Fitur yang saya gunakan adalah konsep Bag of Words dengan menggunakan fungsi TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer

# Split data training dan testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=50)

# Inisiasi TfidfVectorizer
bow = TfidfVectorizer(stop_words='english')

# Fitting dan transform X_train dengan TfidfVectorizer
X_train = bow.fit_transform(X_train)

# Transform X_test
X_test = bow.transform(X_test)
```

Python

```
# Klasifikasi yang digunakan adalah algoritma Random Forest
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

# Fit ke model
rdf_model = RandomForestClassifier().fit(X_train, y_train)

# Prediksi dengan data training
y_pred_train = rdf_model.predict(X_train)
```

```
# Klasifikasi yang digunakan adalah algoritma Random Forest
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

# Fit ke model
rdf_model = RandomForestClassifier().fit(X_train, y_train)

# Prediksi dengan data training
y_pred_train = rdf_model.predict(X_train)

# Evaluasi akurasi data training
acc_train = accuracy_score(y_train, y_pred_train)

# Prediksi dengan data training
y_pred_test = rdf_model.predict(X_test)

# Evaluasi akurasi data training
acc_test = accuracy_score(y_test, y_pred_test)

# Print hasil evaluasi
print(f'Hasil akurasi data train: {acc_train}')
print(f'Hasil akurasi data test: {acc_test}')
```

Python

```
Hasil akurasi data train: 0.994875
Hasil akurasi data test: 0.444875
```

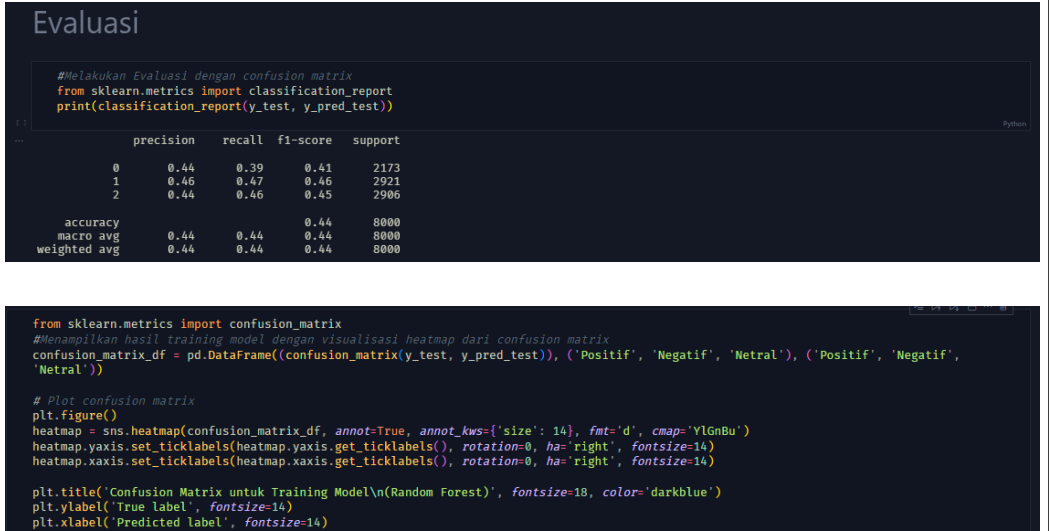
Classification menggunakan model algoritma Decision Tree. Pertama kita harus mengimport library yang dibutuhkan seperti, numpy (untuk melakukan perhitungan array dimensi), pandas (untuk analisis data dalam bentuk dataframe), dan DecisionTreeClassifier (untuk membuat model klasifikasi dengan Decision Tree)

DecisionTree dan juga terdapat kode program mendefinisikan variabel classifier untuk proses decision tree classifier. Pada baris kode `dt.fit(X_train, y_train)` berfungsi membuat model DecisionTreeClassifier untuk training set.

Kemudian mendefinisikan `y_pred_dt` untuk memprediksi hasil model DecisionTreeClassifier ke test set. Setelah memprediksi hasil model DecisionTreeClassifier ke test set, selanjutnya yaitu mengevaluasi kinerja pada set

pengujian dengan menghitung akurasi antara `y_test` dengan `y_pred_dt`.

Dengan menggunakan function `predict()` kita bisa melakukan prediksi pada data test. Hasil prediksi menunjukkan, model kita memiliki hasil accuracy yang kurang baik

	sehingga algoritma Decision Tree ini kurang cocok untuk kebutuhan klasifikasi jenis komentar pengguna twitter berdasarkan teks.
8	<p>Evaluasi</p>  <pre> #Melakukan Evaluasi dengan confusion matrix from sklearn.metrics import classification_report print(classification_report(y_test, y_pred_test)) precision recall f1-score support 0 0.44 0.39 0.41 2173 1 0.46 0.47 0.46 2921 2 0.44 0.46 0.45 2906 accuracy 0.44 0.44 0.44 8000 macro avg 0.44 0.44 0.44 8000 weighted avg 0.44 0.44 0.44 8000 from sklearn.metrics import confusion_matrix #Menampilkan hasil training model dengan visualisasi heatmap dari confusion matrix confusion_matrix_df = pd.DataFrame((confusion_matrix(y_test, y_pred_test)), ('Positif', 'Negatif', 'Netral'), ('Positif', 'Negatif', 'Netral')) # Plot confusion matrix plt.figure() heatmap = sns.heatmap(confusion_matrix_df, annot=True, annot_kws={'size': 14}, fmt='d', cmap='YlGnBu') heatmap.yaxis.set_ticklabels(heatmap.yaxis.get_ticklabels(), rotation=0, ha='right', fontsize=14) heatmap.xaxis.set_ticklabels(heatmap.xaxis.get_ticklabels(), rotation=0, ha='right', fontsize=14) plt.title('Confusion Matrix untuk Training Model\n(Random Forest)', fontsize=18, color='darkblue') plt.ylabel('True label', fontsize=14) plt.xlabel('Predicted label', fontsize=14) </pre>

Link Google Colabs :

https://colab.research.google.com/drive/1wNj29Fr3Aj5k6OuxF73Ph_A5xoK1ylef?usp=sharing