nlp_bert

May 31, 2024

[1]:
```
# Gerekli Kütüphaneleri Yükleme
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re
import nltk
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, classification_report
# Hugging Face Transformers kütüphanesini yükleme
from transformers import BertTokenizer, BertModel
from sklearn.preprocessing import LabelEncoder
```

[2]:
```
# NLTK verilerini indir
nltk.download('stopwords')
from nltk.corpus import stopwords

# Stop words ve diğer ön işlemler
stop_words = set(stopwords.words('english'))
```

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

[3]:
```
# Veri setlerini yükle
train_data = pd.read_csv('train.csv')
test_data = pd.read_csv('valid.csv')
```

[4]:
```
# Etiketleri sayısallaştırma
label_encoder = LabelEncoder()
train_data['Y'] = label_encoder.fit_transform(train_data['Y'])
test_data['Y'] = label_encoder.transform(test_data['Y'])
train_data.head()
```

[4]:

	Id	Title	\
0	34552656	Java: Repeat Task Every Random Seconds	

```
1  34553034                  Why are Java Optionals immutable?
2  34553174  Text Overlay Image with Darkened Opacity React…
3  34553318         Why ternary operator in swift is so picky?
4  34553755                 hide/show fab with scale animation

                                                 Body  \
0  <p>I'm already familiar with repeating tasks e…
1  <p>I'd like to understand why Java 8 Optionals…
2  <p>I am attempting to overlay a title over an …
3  <p>The question is very simple, but I just cou…
4  <p>I'm using custom floatingactionmenu. I need…

                                                 Tags         CreationDate  Y
0                                   <java><repeat>  2016-01-01 00:21:59  1
1                                 <java><optional>  2016-01-01 02:03:20  0
2  <javascript><image><overlay><react-native><opa…  2016-01-01 02:48:24  0
3  <swift><operators><whitespace><ternary-operato…  2016-01-01 03:30:17  0
4  <android><material-design><floating-action-but…  2016-01-01 05:21:48  0
```

[5]:
```
# Metin ön işleme fonksiyonu
def preprocess_text(text):
    text = text.lower()  # Büyük harfleri küçültme
    text = re.sub(r'[^\w\s]', '', text)  # Noktalama işaretlerini kaldırma
    text = re.sub(r'\d+', '', text)  # Sayıları kaldırma
    text = ' '.join([word for word in text.split() if word not in stop_words]) ␣
 ↪# Durma kelimelerini kaldırma
    return text
```

[6]:
```
# Metinleri ön işleme
train_data['cleaned_text'] = (train_data['Title'] + " " + train_data['Body']).
 ↪apply(preprocess_text)
test_data['cleaned_text'] = (test_data['Title'] + " " + test_data['Body']).
 ↪apply(preprocess_text)
train_data.head()
```

[6]:
```
         Id                                           Title  \
0  34552656             Java: Repeat Task Every Random Seconds
1  34553034                  Why are Java Optionals immutable?
2  34553174  Text Overlay Image with Darkened Opacity React…
3  34553318         Why ternary operator in swift is so picky?
4  34553755                 hide/show fab with scale animation

                                                 Body  \
0  <p>I'm already familiar with repeating tasks e…
1  <p>I'd like to understand why Java 8 Optionals…
2  <p>I am attempting to overlay a title over an …
3  <p>The question is very simple, but I just cou…
```

```
4  <p>I'm using custom floatingactionmenu. I need…

                                                 Tags         CreationDate  Y  \
0                                    <java><repeat>  2016-01-01 00:21:59  1
1                                  <java><optional>  2016-01-01 02:03:20  0
2  <javascript><image><overlay><react-native><opa…  2016-01-01 02:48:24  0
3  <swift><operators><whitespace><ternary-operato…  2016-01-01 03:30:17  0
4  <android><material-design><floating-action-but…  2016-01-01 05:21:48  0

                                        cleaned_text
0  java repeat task every random seconds pim alre…
1  java optionals immutable pid like understand j…
2  text overlay image darkened opacity react nati…
3  ternary operator swift picky pthe question sim…
4  hideshow fab scale animation pim using custom …
```

[7]:
```
# BERT modelini ve tokenizer'ı yükleme
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
model = BertModel.from_pretrained('bert-base-uncased')
```

[8]:
```
# Metinleri BERT ile dönüştürme
def get_bert_embeddings(text, tokenizer, model):
    inputs = tokenizer(text, return_tensors='pt', truncation=True,␣
 ↪padding=True, max_length=512)
    outputs = model(**inputs)
    return outputs.last_hidden_state[:, 0, :].detach().numpy()  # [CLS]␣
 ↪token'ın gömülmelerini döndür
```

[9]:
```
# Eğitim ve test setleri için BERT gömülmelerini elde etme
train_embeddings = np.vstack([get_bert_embeddings(text, tokenizer, model) for␣
 ↪text in train_data['cleaned_text']])
test_embeddings = np.vstack([get_bert_embeddings(text, tokenizer, model) for␣
 ↪text in test_data['cleaned_text']])
```

[10]:
```
# Model Eğitimi ve Değerlendirme
models = {
    'Logistic Regression': LogisticRegression(max_iter=1000),
    'Random Forest': RandomForestClassifier(n_estimators=100, random_state=42),
    'SVM': SVC(kernel='linear', random_state=42)
}

for model_name, model in models.items():
    print(f'\n{model_name} Modeli Eğitiliyor...')
    model.fit(train_embeddings.squeeze(), train_data['Y'])

    # Test verisi üzerinde tahminler yapma
    y_pred = model.predict(test_embeddings.squeeze())
```

```
    # Modelin performansını değerlendirme
    accuracy = accuracy_score(test_data['Y'], y_pred)
    report = classification_report(test_data['Y'], y_pred)

    print(f'\n{model_name}')
    print(f'Accuracy: {accuracy}')
    print('Classification Report:')
    print(report)
```

```
Logistic Regression Modeli Eğitiliyor…

C:\Users\OMENPC\anaconda3\Lib\site-
packages\sklearn\linear_model\_logistic.py:458: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(


Logistic Regression
Accuracy: 0.8478666666666667
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.81      0.81      5000
           1       0.80      0.77      0.79      5000
           2       0.94      0.96      0.95      5000

    accuracy                           0.85     15000
   macro avg       0.85      0.85      0.85     15000
weighted avg       0.85      0.85      0.85     15000


Random Forest Modeli Eğitiliyor…

Random Forest
Accuracy: 0.7387333333333334
Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.73      0.72      5000
           1       0.68      0.64      0.66      5000
```

```
              2       0.81      0.85      0.83      5000

    accuracy                           0.74     15000
   macro avg       0.74      0.74      0.74     15000
weighted avg       0.74      0.74      0.74     15000


SVM Modeli Eğitiliyor…

SVM
Accuracy: 0.8477333333333333
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.81      0.80      5000
           1       0.80      0.77      0.79      5000
           2       0.95      0.96      0.95      5000

    accuracy                           0.85     15000
   macro avg       0.85      0.85      0.85     15000
weighted avg       0.85      0.85      0.85     15000
```

[ ]: