Student Information

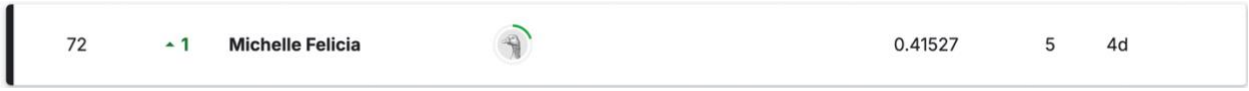
Name: 黃明霞 Michelle Grace

Student ID: 110006311

GitHub ID: chellegrace855

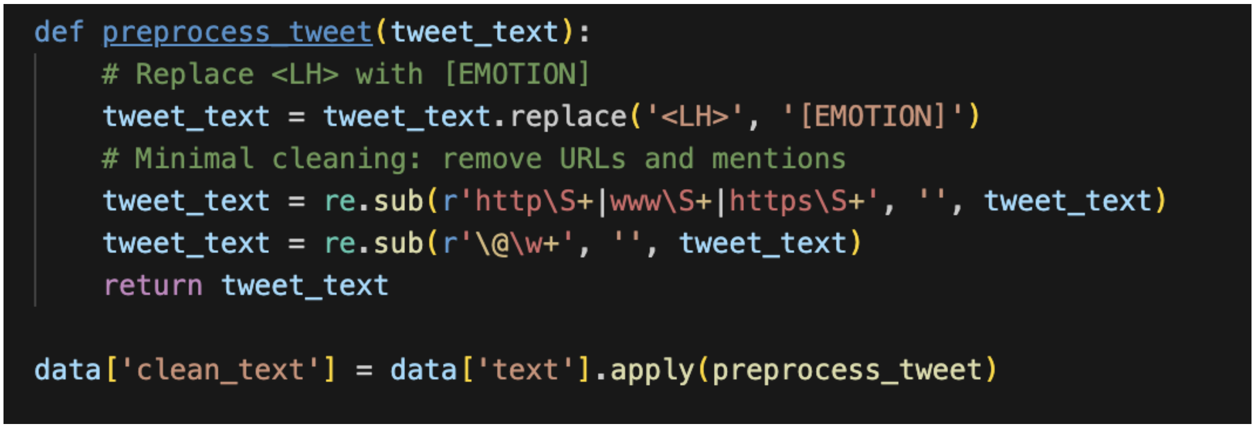
Kaggle name: michellefelicia

Kaggle private scoreboard snapshot:



72	▲ 1	Michelle Felicia	0.41527	5	4d

Insights: From seeing the dataset, most emotions have some specifically highly correlated words. For example joy is tightly correlated with excited, etc. Therefore, simple algorithm such as TF-IDF or BOW might work better or comparable with other highly complex algorithm. Moreover, I notice that there is almost always "<LH>" in each tweet, which might be the concealed emotion mentioned.

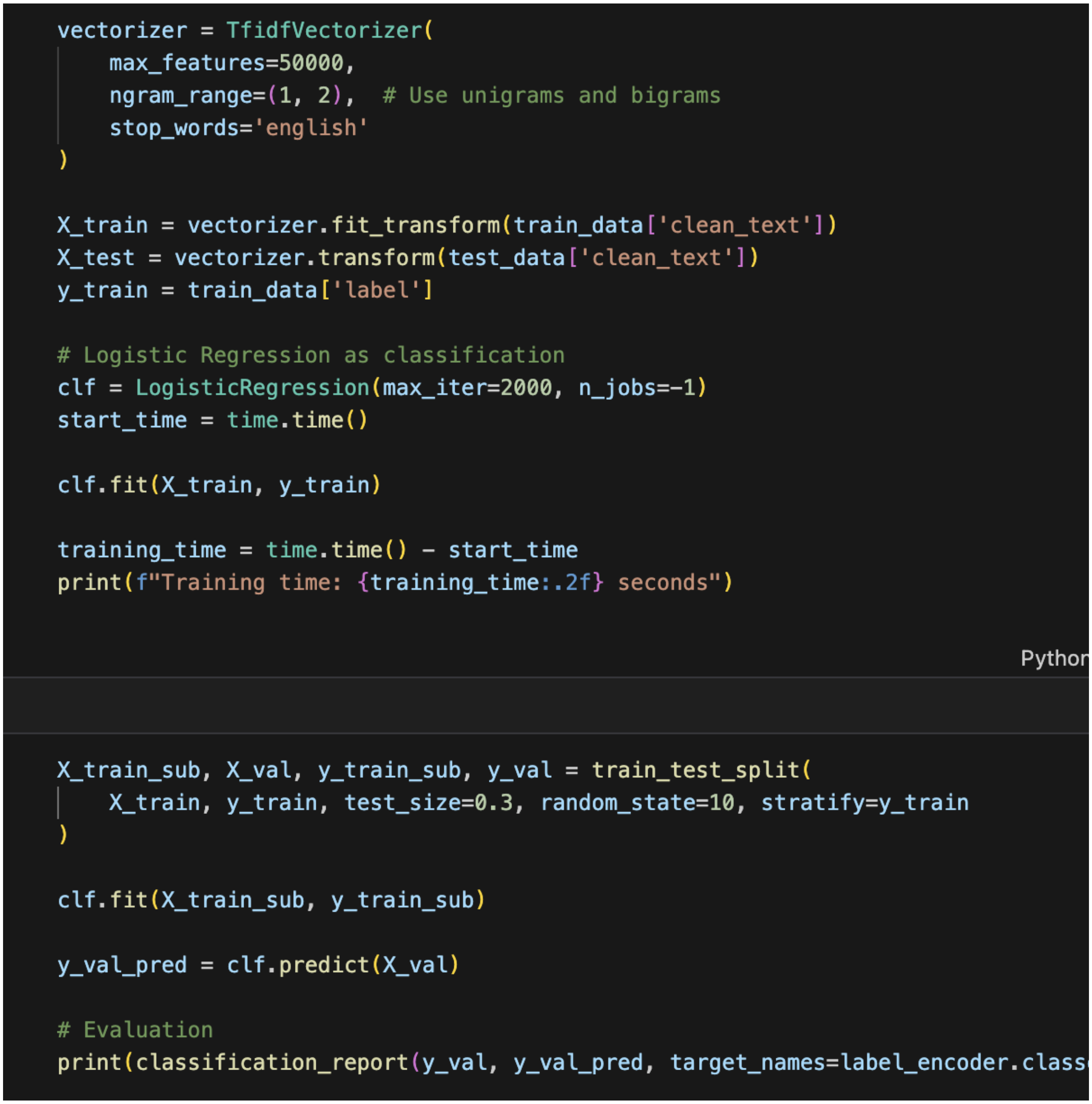


```python
def preprocess_tweet(tweet_text):
    # Replace <LH> with [EMOTION]
    tweet_text = tweet_text.replace('<LH>', '[EMOTION]')
    # Minimal cleaning: remove URLs and mentions
    tweet_text = re.sub(r'http\S+|www\S+|https\S+', '', tweet_text)
    tweet_text = re.sub(r'\@\w+', '', tweet_text)
    return tweet_text

data['clean_text'] = data['text'].apply(preprocess_tweet)
```

Preprocessing:

- Replace <LH> with [EMOTION] for the model to notice this placeholder
- Remove URLs and mentions
- Split Training and Validation to 70/30



```python
vectorizer = TfidfVectorizer(
    max_features=50000,
    ngram_range=(1, 2),  # Use unigrams and bigrams
    stop_words='english'
)

X_train = vectorizer.fit_transform(train_data['clean_text'])
X_test = vectorizer.transform(test_data['clean_text'])
y_train = train_data['label']

# Logistic Regression as classification
clf = LogisticRegression(max_iter=2000, n_jobs=-1)
start_time = time.time()

clf.fit(X_train, y_train)

training_time = time.time() - start_time
print(f"Training time: {training_time:.2f} seconds")
```

```python
X_train_sub, X_val, y_train_sub, y_val = train_test_split(
    X_train, y_train, test_size=0.3, random_state=10, stratify=y_train
)

clf.fit(X_train_sub, y_train_sub)

y_val_pred = clf.predict(X_val)

# Evaluation
print(classification_report(y_val, y_val_pred, target_names=label_encoder.class
```

Model:

- TF-IDF = convert text into numerical features based on the importance of words in the dataset
- Logistic Regression = use for classifying tweets of the 8 emotions