AMAZON SENTIMENT ANALYSIS

UEH UNIVERSITY | College of Technology and Design

Học phần: NLP

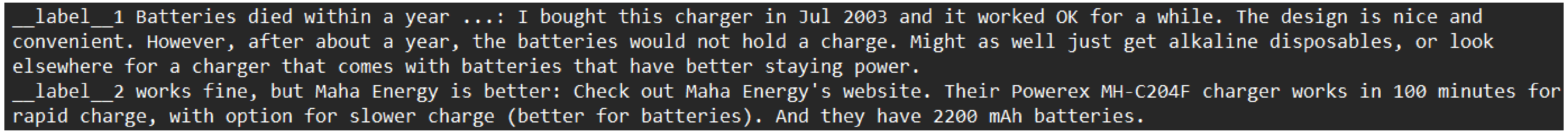
Mã lớp: 23C1INF50907601

Giảng viên: TS. Đặng Ngọc Hoàng Thành

23/12/2023

PREPROCESS & EDA

DATA FORMAT



```
__label__1 Batteries died within a year ...: I bought this charger in Jul 2003 and it worked OK for a while. The design is nice and
convenient. However, after about a year, the batteries would not hold a charge. Might as well just get alkaline disposables, or look
elsewhere for a charger that comes with batteries that have better staying power.
__label__2 works fine, but Maha Energy is better: Check out Maha Energy's website. Their Powerex MH-C204F charger works in 100 minutes for
rapid charge, with option for slower charge (better for batteries). And they have 2200 mAh batteries.
```

__label__1: negative

__label__2: postive

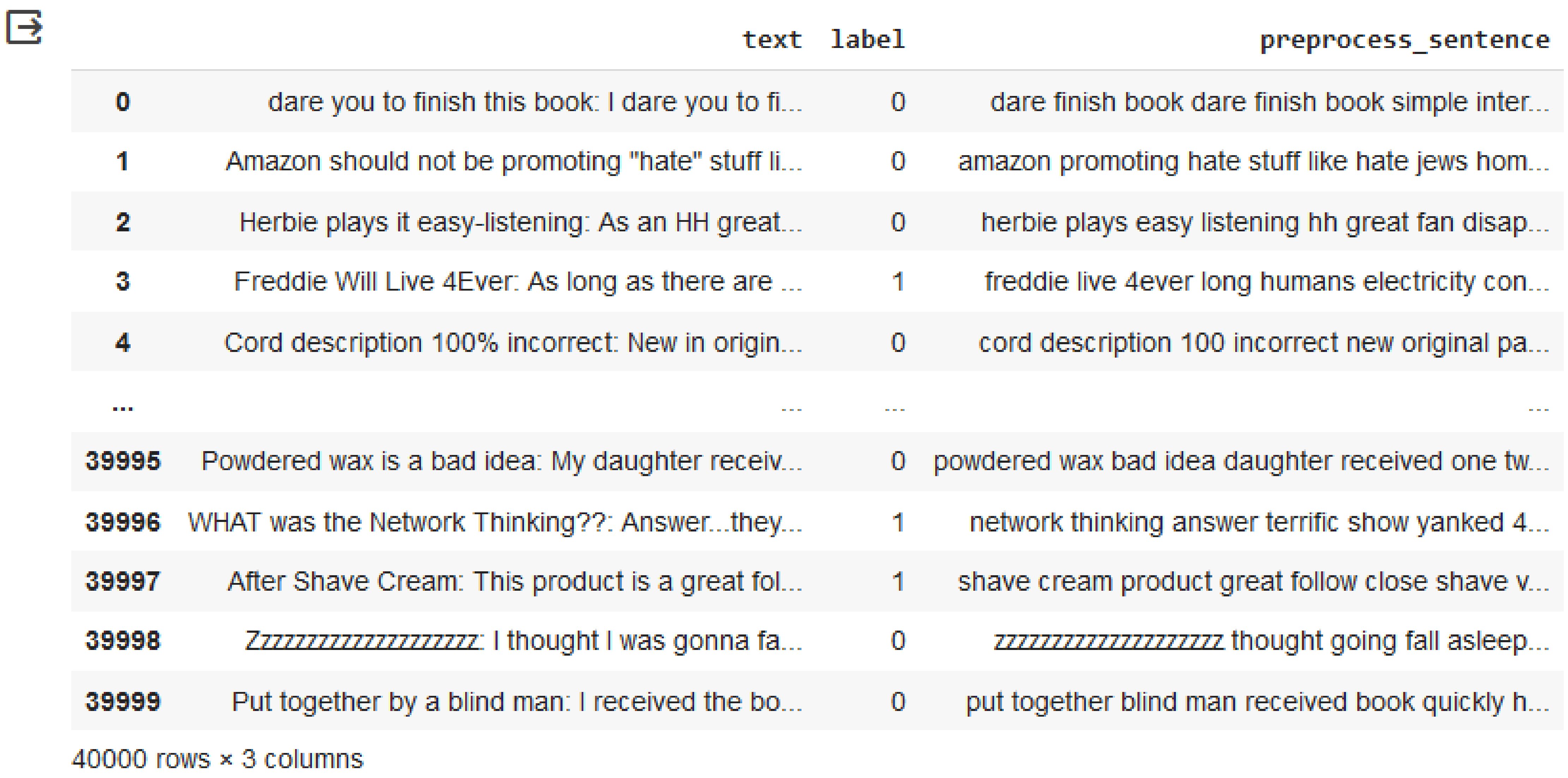
PREPROCESS

A. TIỀN XỬ LÝ

- Bước 1: Loại bỏ các đường link URL
- Bước 2: Loại bỏ các thẻ HTML
- Bước 3: Mở rộng các từ viết tắt
- Bước 4: Loại bỏ dấu chấm câu và chữ số
- Bước 5: Loại bỏ emoji
- Bước 6: Chuyển sang chữ thường
- Bước 7: Loại bỏ stopwords khỏi các câu

PREPROCESS

A. TIỀN XỬ LÝ



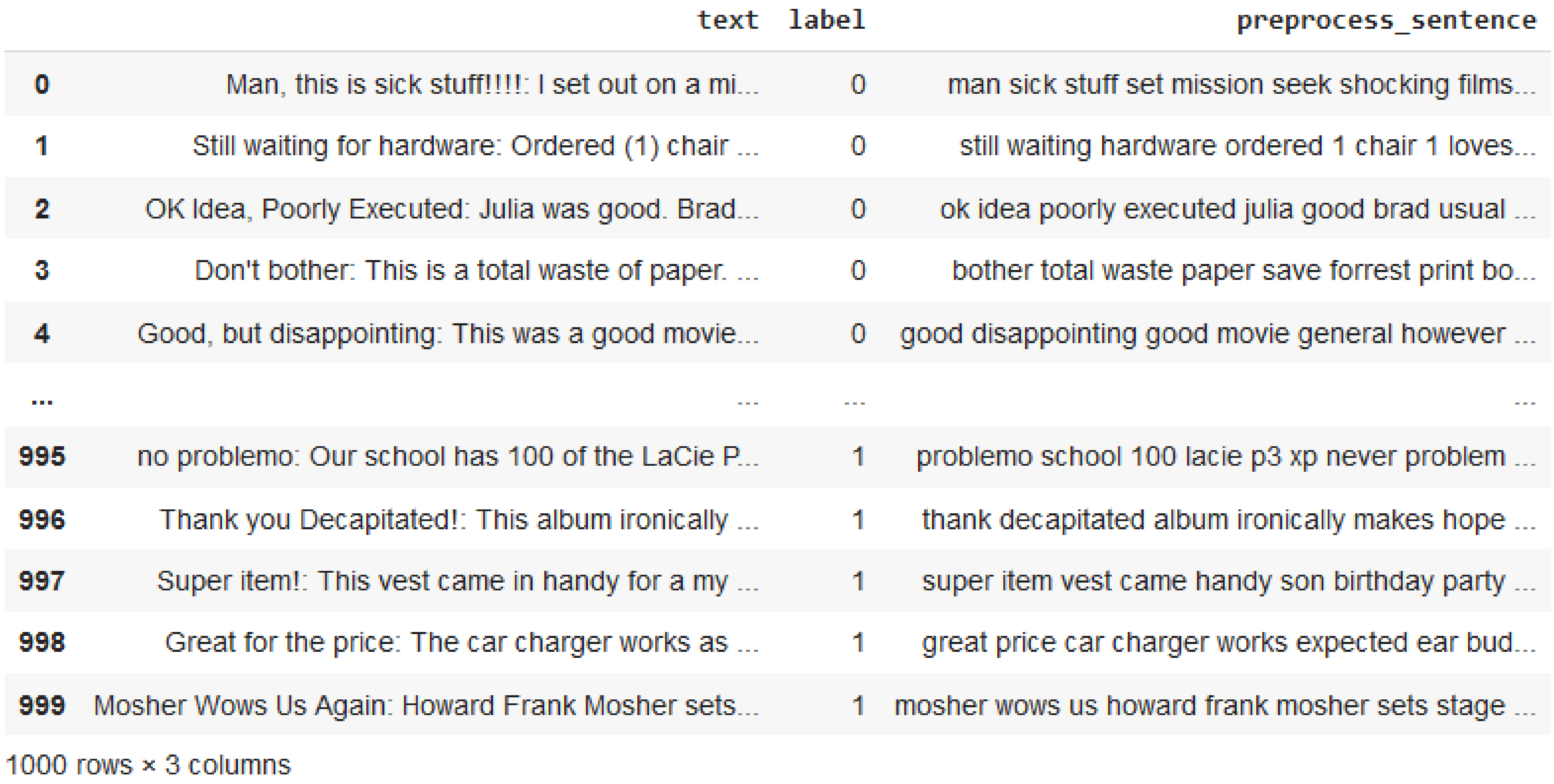
	text	label	preprocess_sentence
0	dare you to finish this book: I dare you to fi...	0	dare finish book dare finish book simple inter...
1	Amazon should not be promoting "hate" stuff li...	0	amazon promoting hate stuff like hate jews hom...
2	Herbie plays it easy-listening: As an HH great...	0	herbie plays easy listening hh great fan disap...
3	Freddie Will Live 4Ever: As long as there are ...	1	freddie live 4ever long humans electricity con...
4	Cord description 100% incorrect: New in origin...	0	cord description 100 incorrect new original pa...
...	...	...	...
39995	Powdered wax is a bad idea: My daughter receiv...	0	powdered wax bad idea daughter received one tw...
39996	WHAT was the Network Thinking??: Answer...they...	1	network thinking answer terrific show yanked 4...
39997	After Shave Cream: This product is a great fol...	1	shave cream product great follow close shave v...
39998	Zzzzzzzzzzzzzzzzzzzz: I thought I was gonna fa...	0	zzzzzzzzzzzzzzzzzzzz thought going fall asleep...
39999	Put together by a blind man: I received the bo...	0	put together blind man received book quickly h...

40000 rows × 3 columns

Train dataframe của nhóm

PREPROCESS

A. TIỀN XỬ LÝ



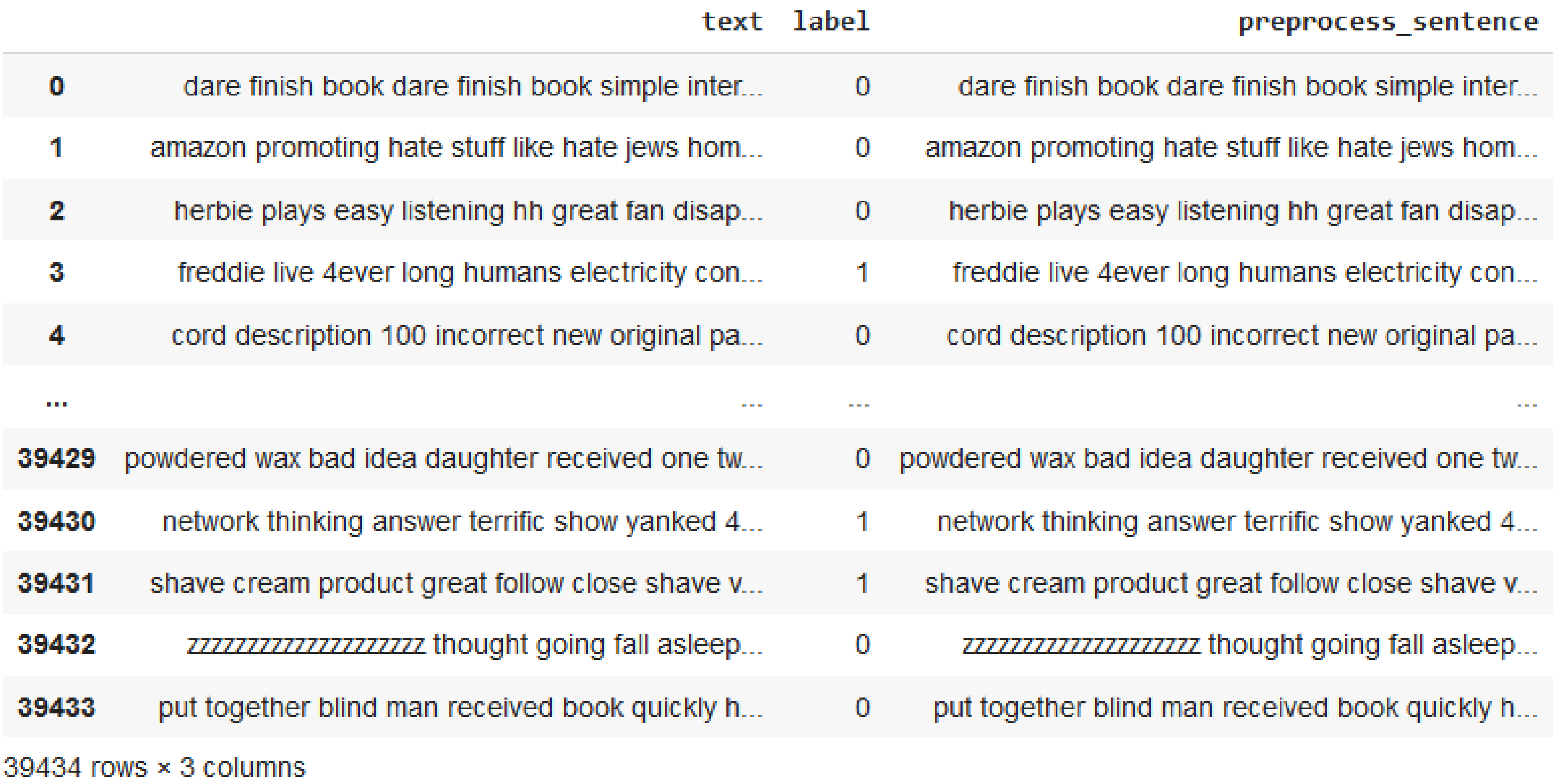
	text	label	preprocess_sentence
0	Man, this is sick stuff!!!!: I set out on a mi...	0	man sick stuff set mission seek shocking films...
1	Still waiting for hardware: Ordered (1) chair ...	0	still waiting hardware ordered 1 chair 1 loves...
2	OK Idea, Poorly Executed: Julia was good. Brad...	0	ok idea poorly executed julia good brad usual ...
3	Don't bother: This is a total waste of paper. ...	0	bother total waste paper save forrest print bo...
4	Good, but disappointing: This was a good movie...	0	good disappointing good movie general however ...
...	...	...	...
995	no problemo: Our school has 100 of the LaCie P...	1	problemo school 100 lacie p3 xp never problem ...
996	Thank you Decapitated!: This album ironically ...	1	thank decapitated album ironically makes hope ...
997	Super item!: This vest came in handy for a my ...	1	super item vest came handy son birthday party ...
998	Great for the price: The car charger works as ...	1	great price car charger works expected ear bud...
999	Mosher Wows Us Again: Howard Frank Mosher sets...	1	mosher wows us howard frank mosher sets stage ...

1000 rows × 3 columns

Test dataframe của nhóm

PREPROCESS

B. LOẠI BỎ GÁNH NẶNG

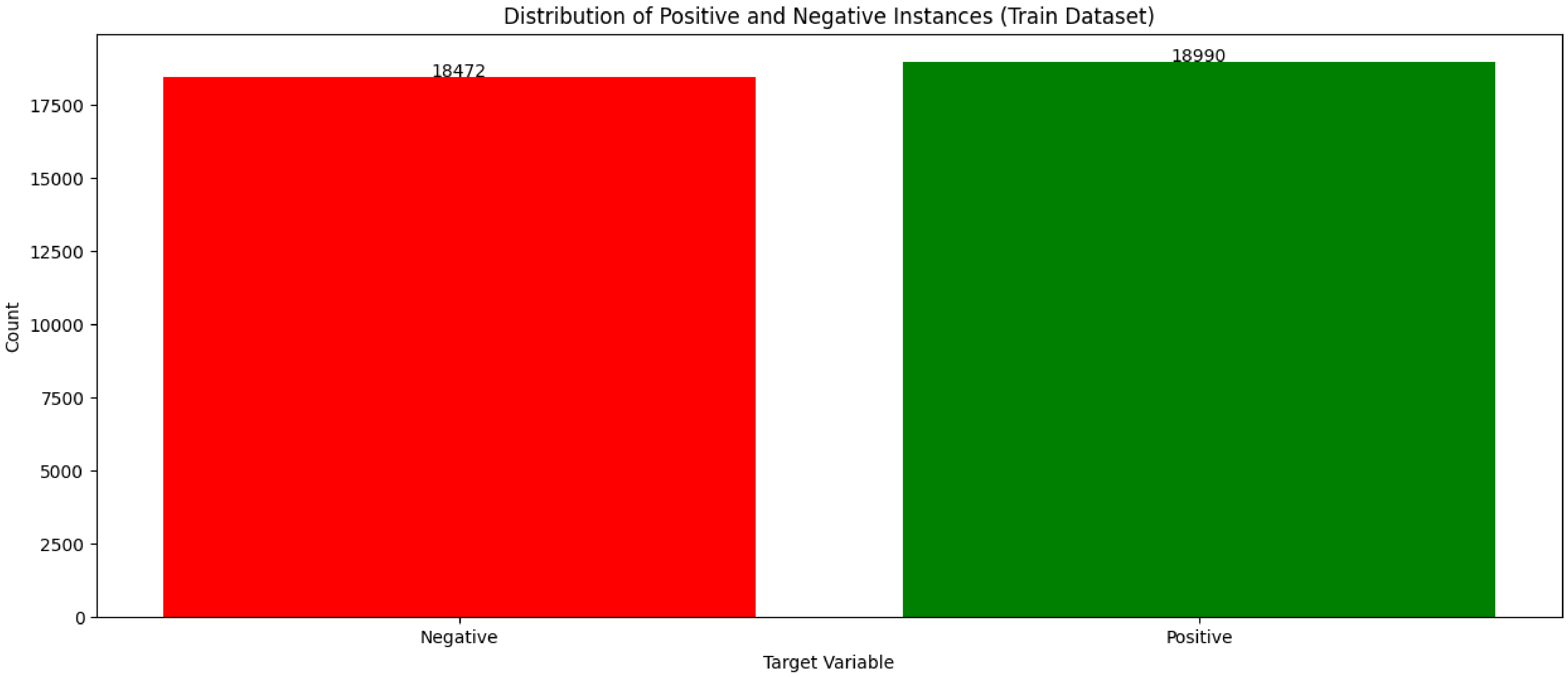


	text	label	preprocess_sentence
0	dare finish book dare finish book simple inter...	0	dare finish book dare finish book simple inter...
1	amazon promoting hate stuff like hate jews hom...	0	amazon promoting hate stuff like hate jews hom...
2	herbie plays easy listening hh great fan disap...	0	herbie plays easy listening hh great fan disap...
3	freddie live 4ever long humans electricity con...	1	freddie live 4ever long humans electricity con...
4	cord description 100 incorrect new original pa...	0	cord description 100 incorrect new original pa...
...	...	...	...
39429	powdered wax bad idea daughter received one tw...	0	powdered wax bad idea daughter received one tw...
39430	network thinking answer terrific show yanked 4...	1	network thinking answer terrific show yanked 4...
39431	shave cream product great follow close shave v...	1	shave cream product great follow close shave v...
39432	zzzzzzzzzzzzzzzzzzzz thought going fall asleep...	0	zzzzzzzzzzzzzzzzzzzz thought going fall asleep...
39433	put together blind man received book quickly h...	0	put together blind man received book quickly h...

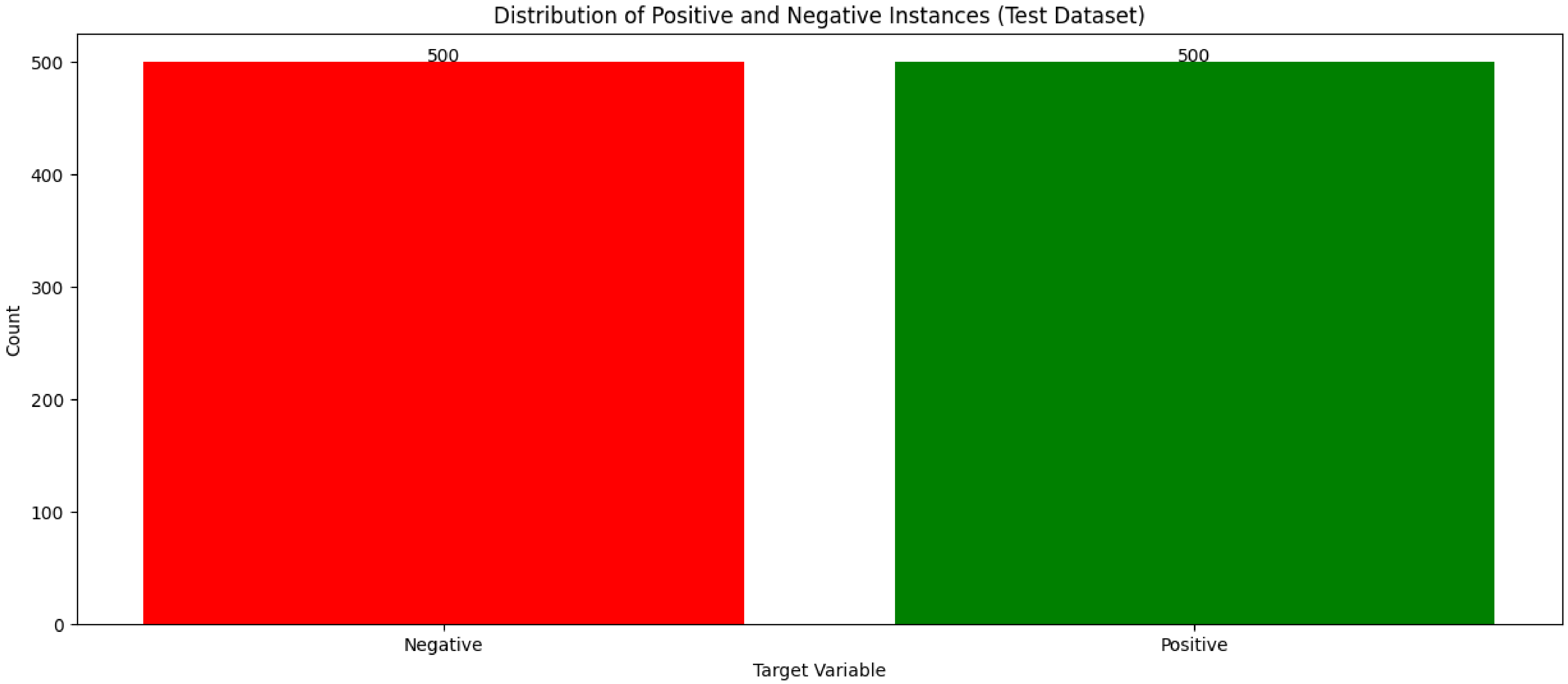
39434 rows × 3 columns

Train dataframe sau khi loại bỏ các mẫu nhãn sai

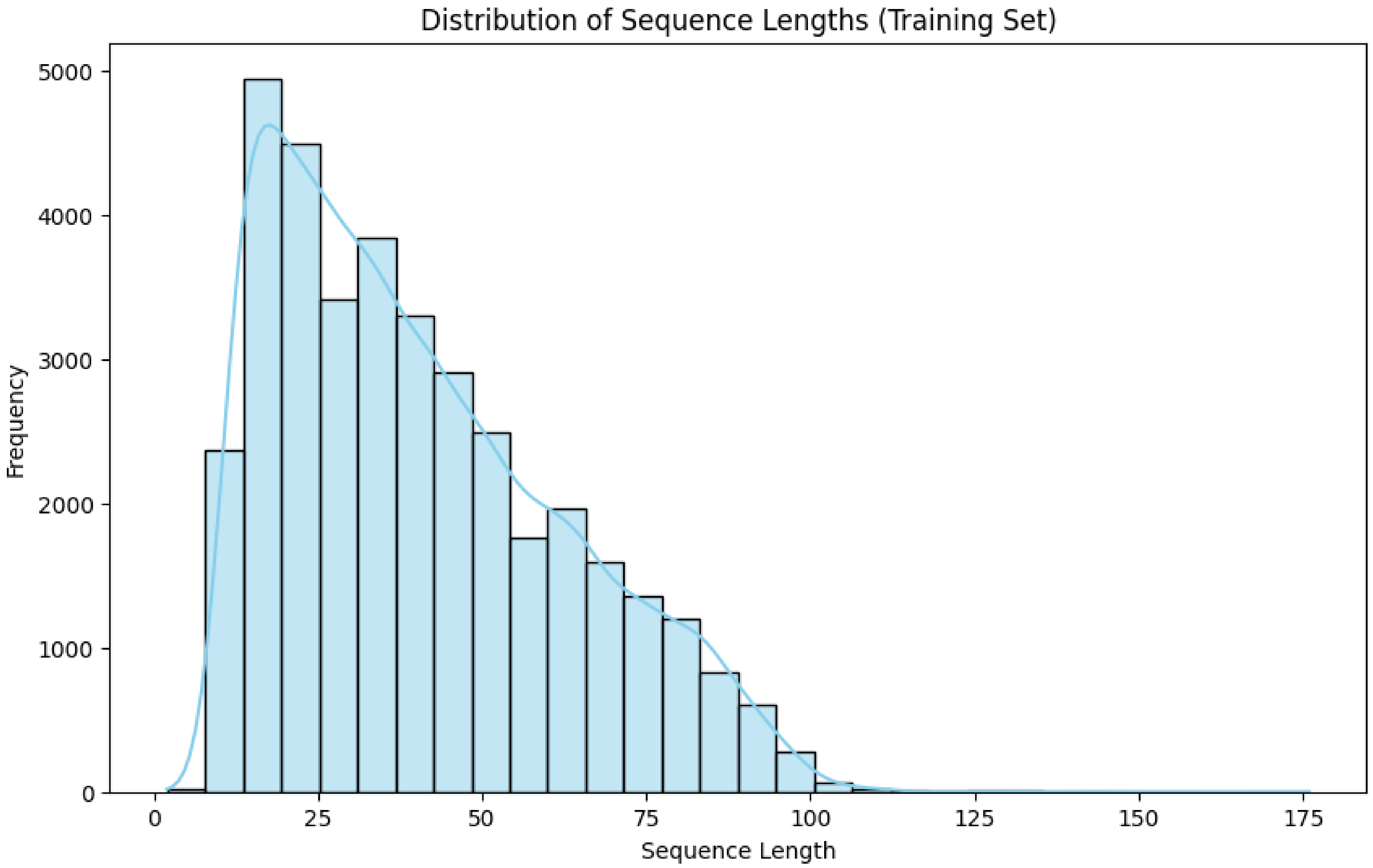
EDA


Distribution of Positive and Negative Instances (Train Dataset)
18472
18990
Count
0
2500
5000
7500
10000
12500
15000
17500
Negative
Positive
Target Variable

EDA

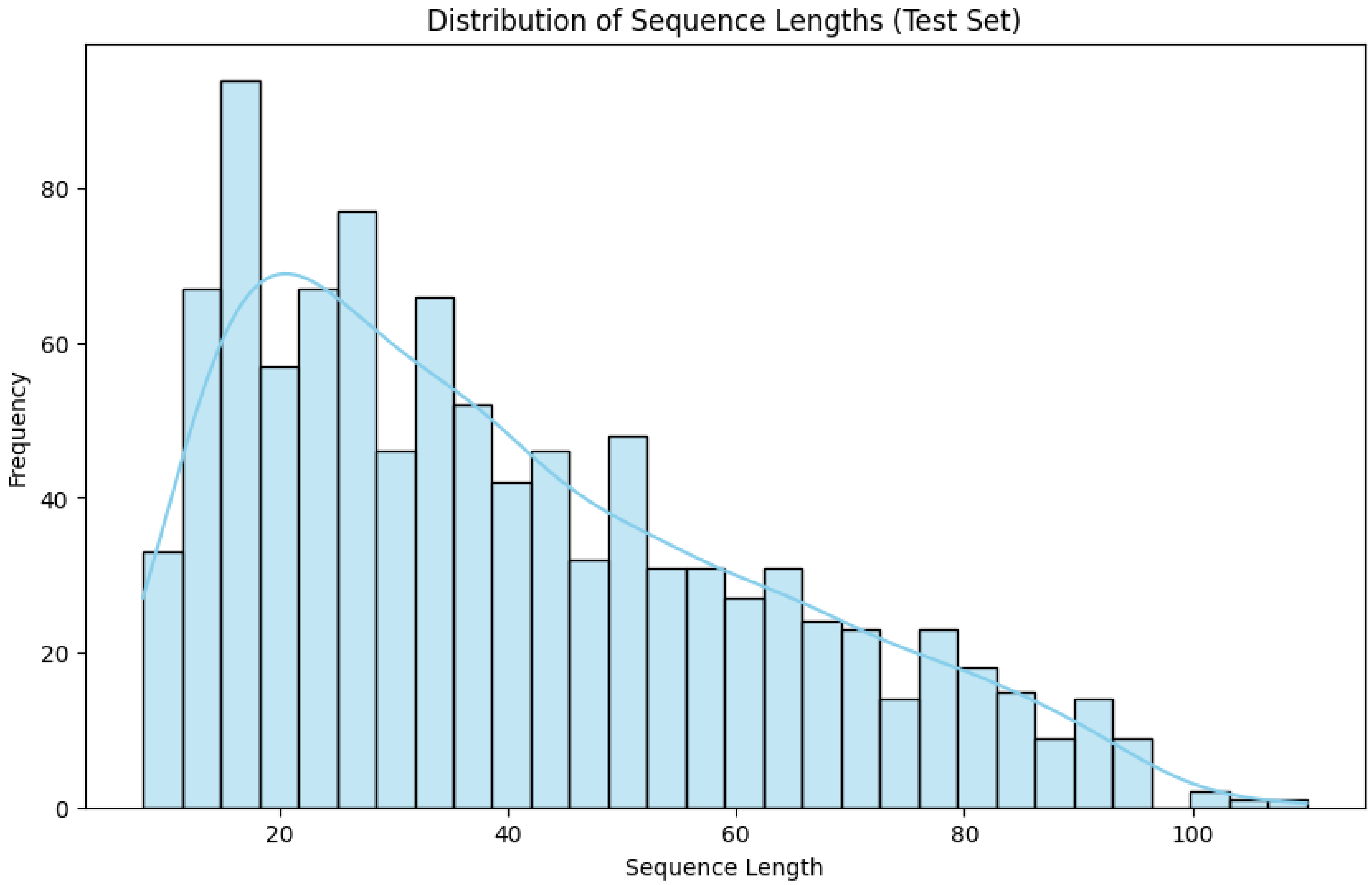

Distribution of Positive and Negative Instances (Test Dataset)
Count
500
400
300
200
100
0
500
500
Negative
Positive
Target Variable

EDA



Distribution of Sequence Lengths (Training Set)
Frequency
5000
4000
3000
2000
1000
0
0
25
50
75
100
125
150
175
Sequence Length

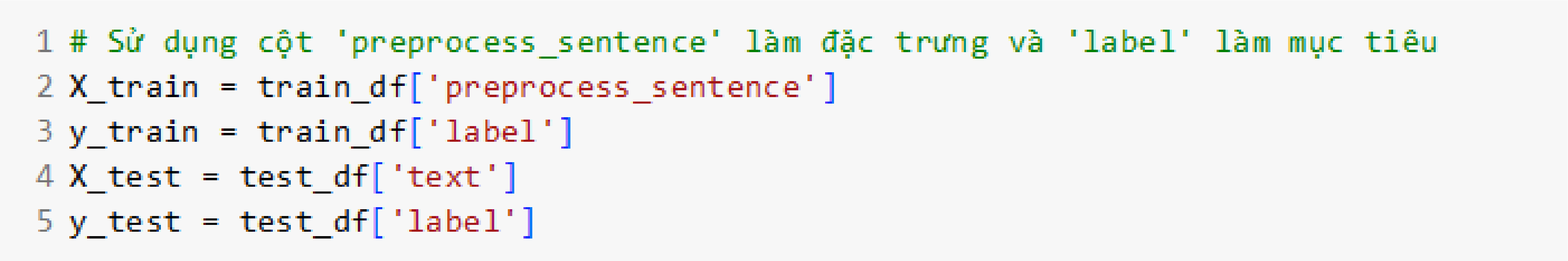
EDA



Distribution of Sequence Lengths (Test Set)
Frequency
0
20
40
60
80
20
40
60
80
100
Sequence Length

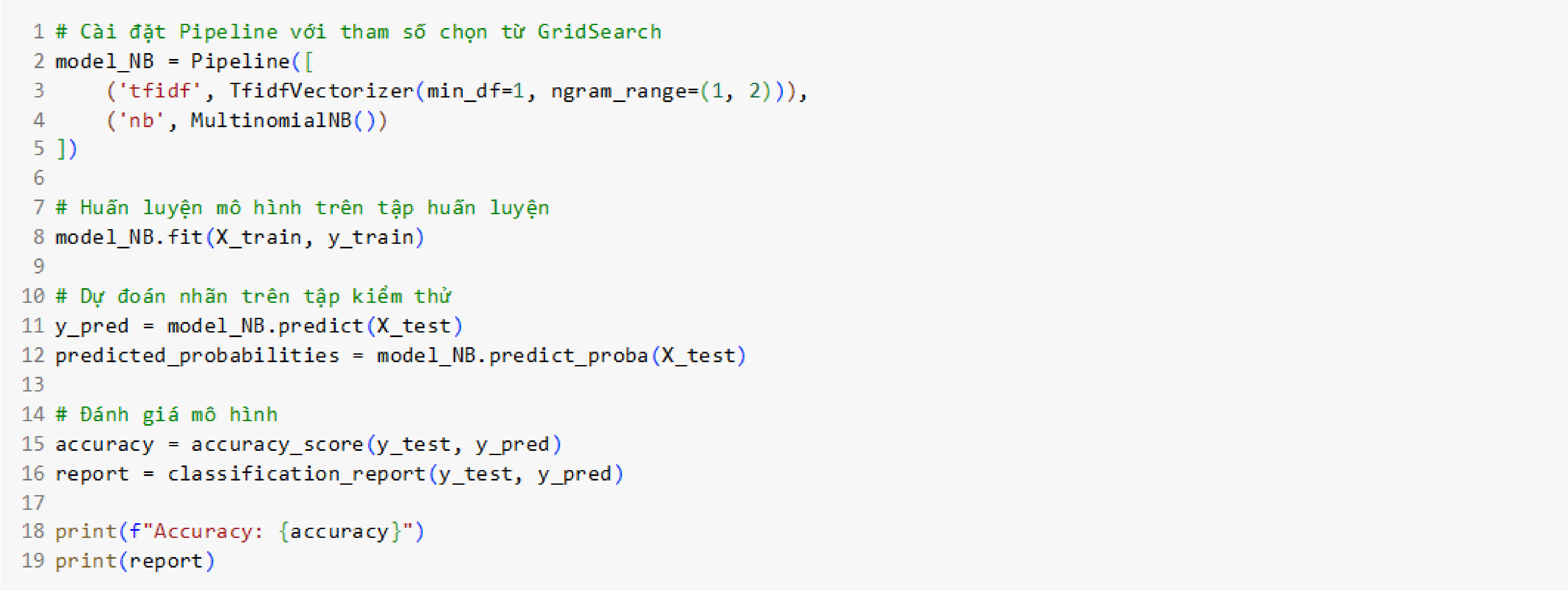
MODEL

DATA



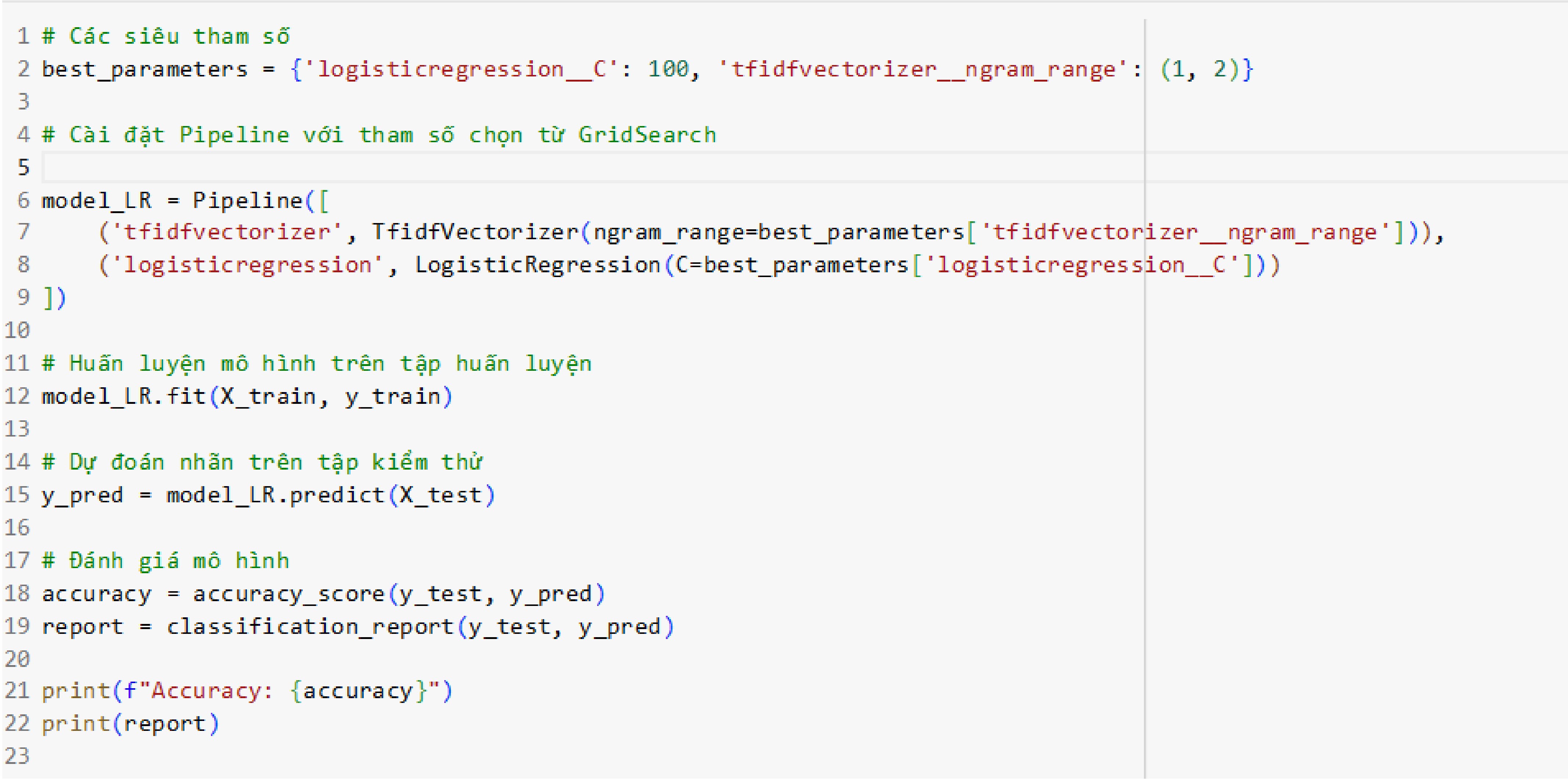
```
# Sử dụng cột 'preprocess_sentence' làm đặc trưng và 'label' làm mục tiêu
X_train = train_df['preprocess_sentence']
y_train = train_df['label']
X_test = test_df['text']
y_test = test_df['label']
```

NAIVE BAYES



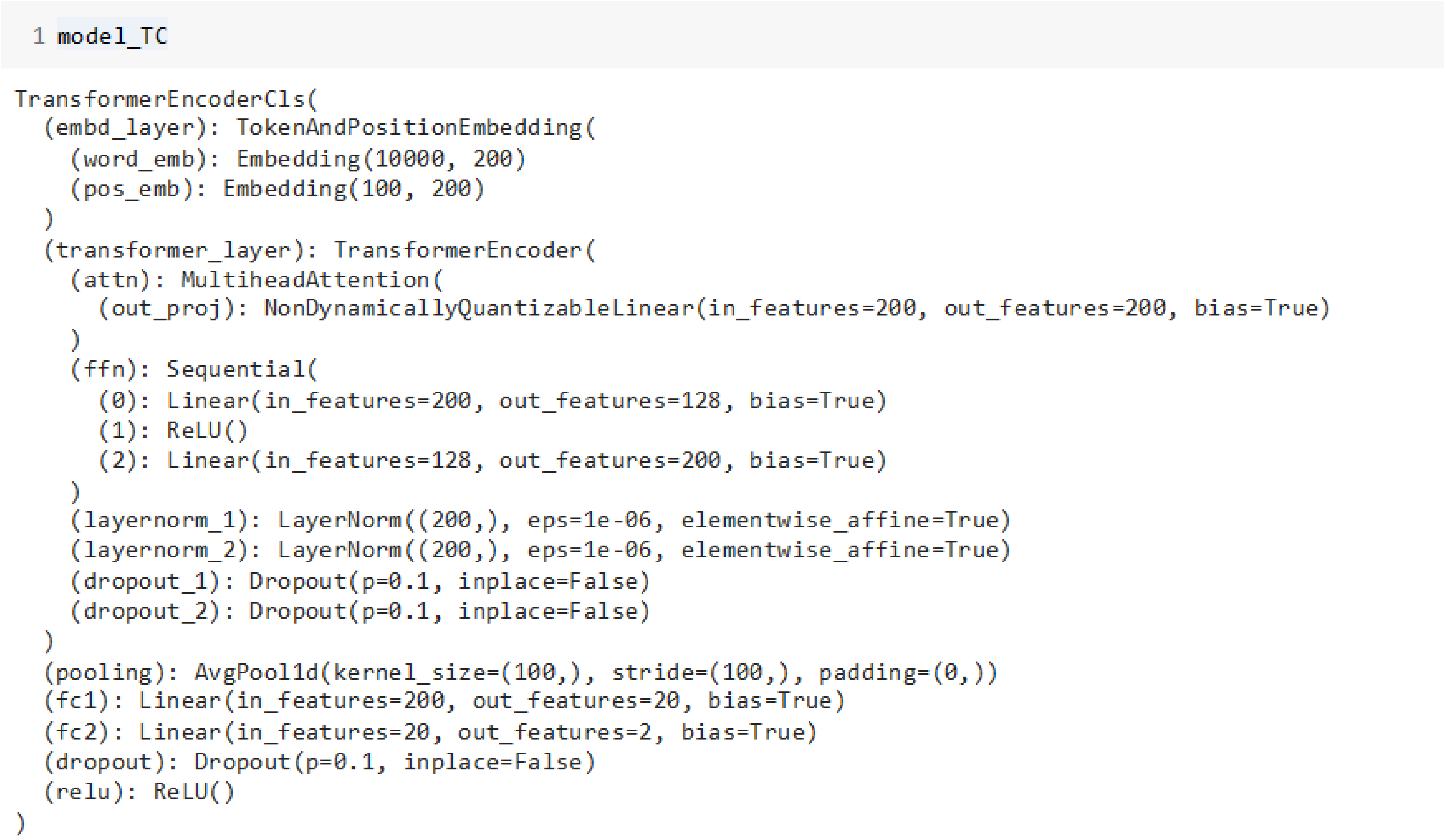
```python
# Cài đặt Pipeline với tham số chọn từ GridSearch
model_NB = Pipeline([
    ('tfidf', TfidfVectorizer(min_df=1, ngram_range=(1, 2))),
    ('nb', MultinomialNB())
])

# Huấn luyện mô hình trên tập huấn luyện
model_NB.fit(X_train, y_train)

# Dự đoán nhãn trên tập kiểm thử
y_pred = model_NB.predict(X_test)
predicted_probabilities = model_NB.predict_proba(X_test)

# Đánh giá mô hình
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)

print(f"Accuracy: {accuracy}")
print(report)
```

LOGISTIC REGRESSION



```
# Các siêu tham số
best_parameters = {'logisticregression__C': 100, 'tfidfvectorizer__ngram_range': (1, 2)}

# Cài đặt Pipeline với tham số chọn từ GridSearch

model_LR = Pipeline([
    ('tfidfvectorizer', TfidfVectorizer(ngram_range=best_parameters['tfidfvectorizer__ngram_range'])),
    ('logisticregression', LogisticRegression(C=best_parameters['logisticregression__C']))
])

# Huấn luyện mô hình trên tập huấn luyện
model_LR.fit(X_train, y_train)

# Dự đoán nhãn trên tập kiểm thử
y_pred = model_LR.predict(X_test)

# Đánh giá mô hình
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)

print(f"Accuracy: {accuracy}")
print(report)
```

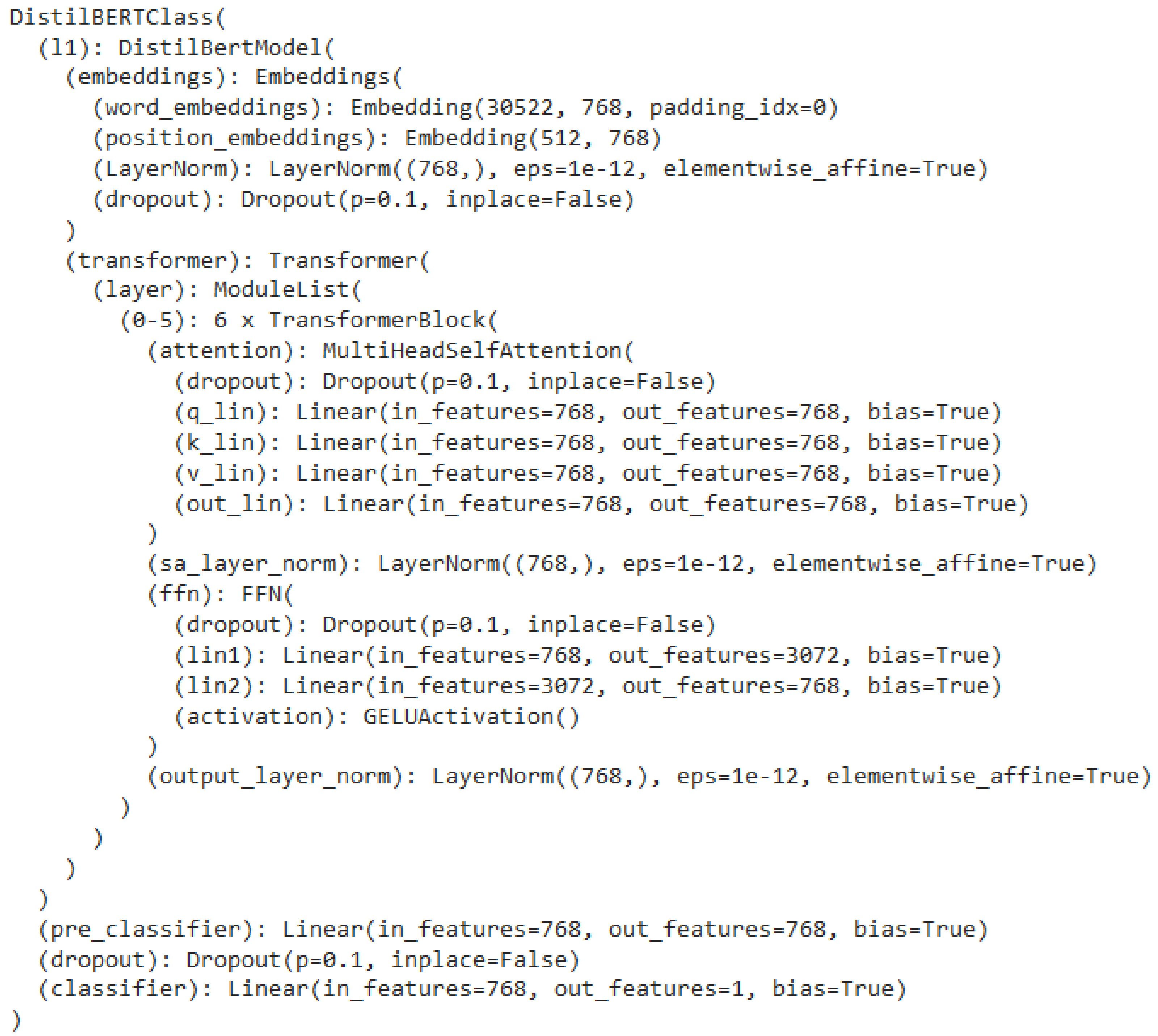
TRANSFORMER (ENCODER)



```
model_TC
```

```
TransformerEncoderCls(
  (embd_layer): TokenAndPositionEmbedding(
    (word_emb): Embedding(10000, 200)
    (pos_emb): Embedding(100, 200)
  )
  (transformer_layer): TransformerEncoder(
    (attn): MultiheadAttention(
      (out_proj): NonDynamicallyQuantizableLinear(in_features=200, out_features=200, bias=True)
    )
    (ffn): Sequential(
      (0): Linear(in_features=200, out_features=128, bias=True)
      (1): ReLU()
      (2): Linear(in_features=128, out_features=200, bias=True)
    )
    (layernorm_1): LayerNorm((200,), eps=1e-06, elementwise_affine=True)
    (layernorm_2): LayerNorm((200,), eps=1e-06, elementwise_affine=True)
    (dropout_1): Dropout(p=0.1, inplace=False)
    (dropout_2): Dropout(p=0.1, inplace=False)
  )
  (pooling): AvgPool1d(kernel_size=(100,), stride=(100,), padding=(0,))
  (fc1): Linear(in_features=200, out_features=20, bias=True)
  (fc2): Linear(in_features=20, out_features=2, bias=True)
  (dropout): Dropout(p=0.1, inplace=False)
  (relu): ReLU()
)
```

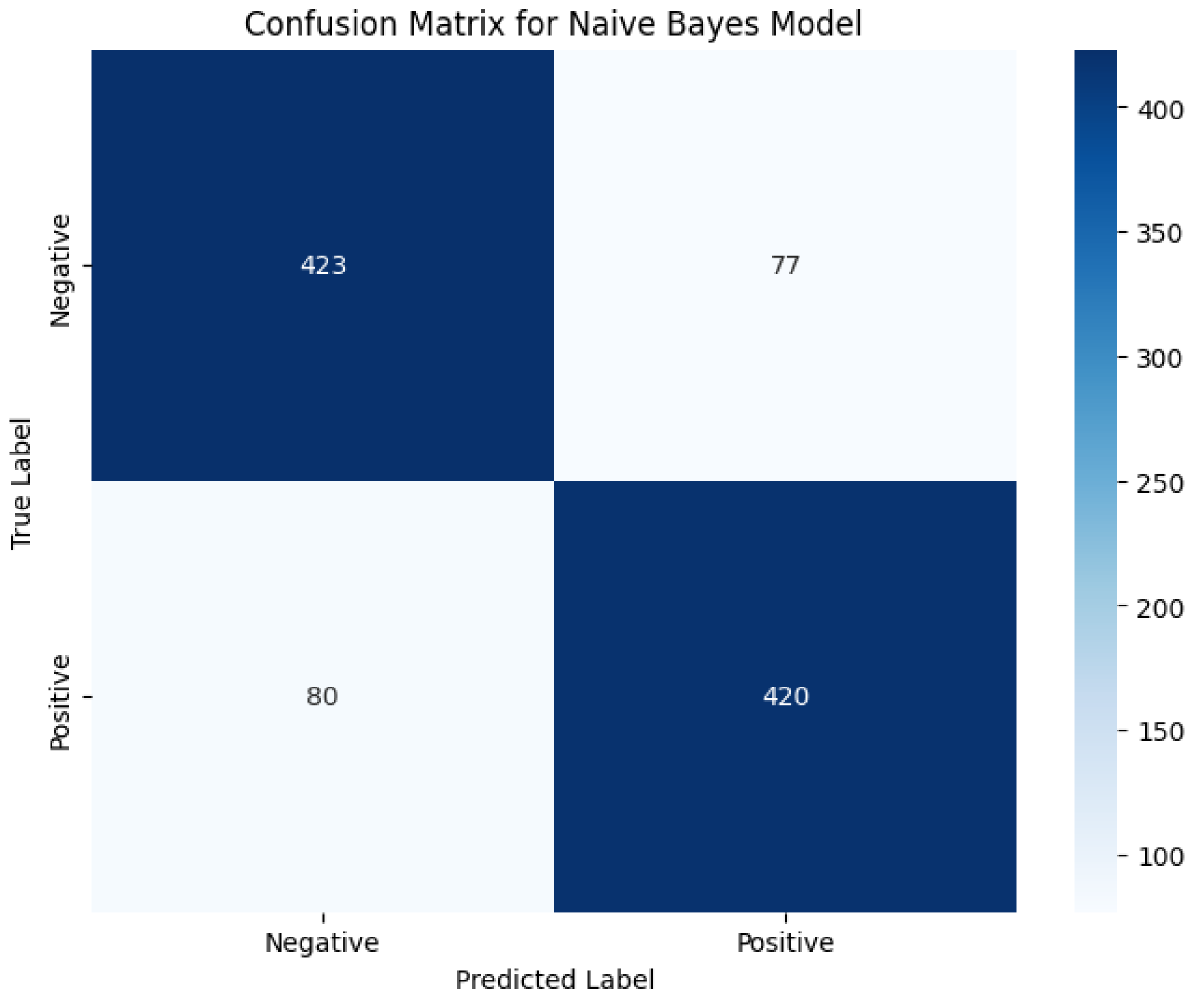
DistilBERT

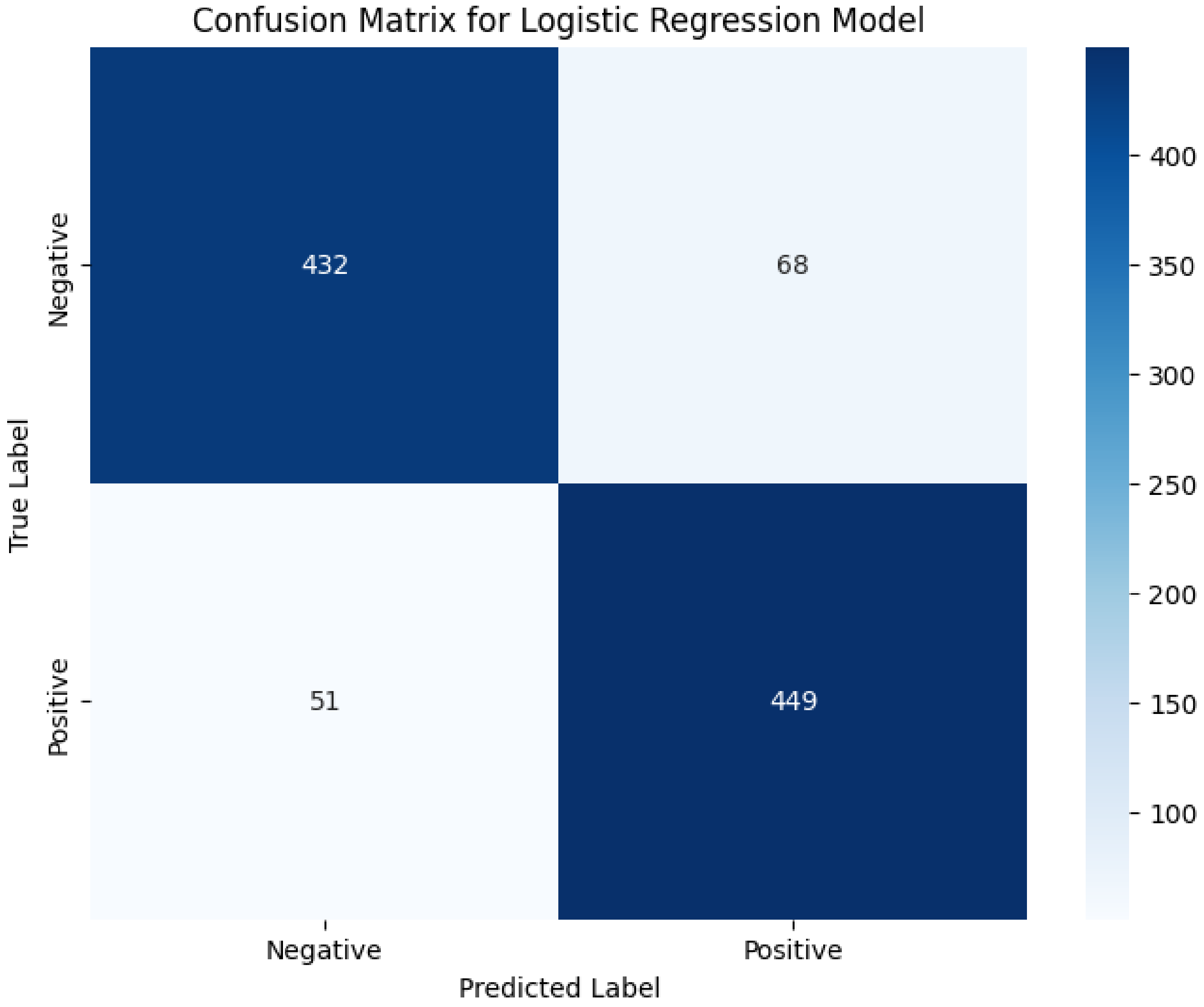


```
DistilBERTClass(
  (l1): DistilBertModel(
    (embeddings): Embeddings(
      (word_embeddings): Embedding(30522, 768, padding_idx=0)
      (position_embeddings): Embedding(512, 768)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False)
    )
    (transformer): Transformer(
      (layer): ModuleList(
        (0-5): 6 x TransformerBlock(
          (attention): MultiHeadSelfAttention(
            (dropout): Dropout(p=0.1, inplace=False)
            (q_lin): Linear(in_features=768, out_features=768, bias=True)
            (k_lin): Linear(in_features=768, out_features=768, bias=True)
            (v_lin): Linear(in_features=768, out_features=768, bias=True)
            (out_lin): Linear(in_features=768, out_features=768, bias=True)
          )
          (sa_layer_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
          (ffn): FFN(
            (dropout): Dropout(p=0.1, inplace=False)
            (lin1): Linear(in_features=768, out_features=3072, bias=True)
            (lin2): Linear(in_features=3072, out_features=768, bias=True)
            (activation): GELUActivation()
          )
          (output_layer_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
        )
      )
    )
  )
  (pre_classifier): Linear(in_features=768, out_features=768, bias=True)
  (dropout): Dropout(p=0.1, inplace=False)
  (classifier): Linear(in_features=768, out_features=1, bias=True)
)
```

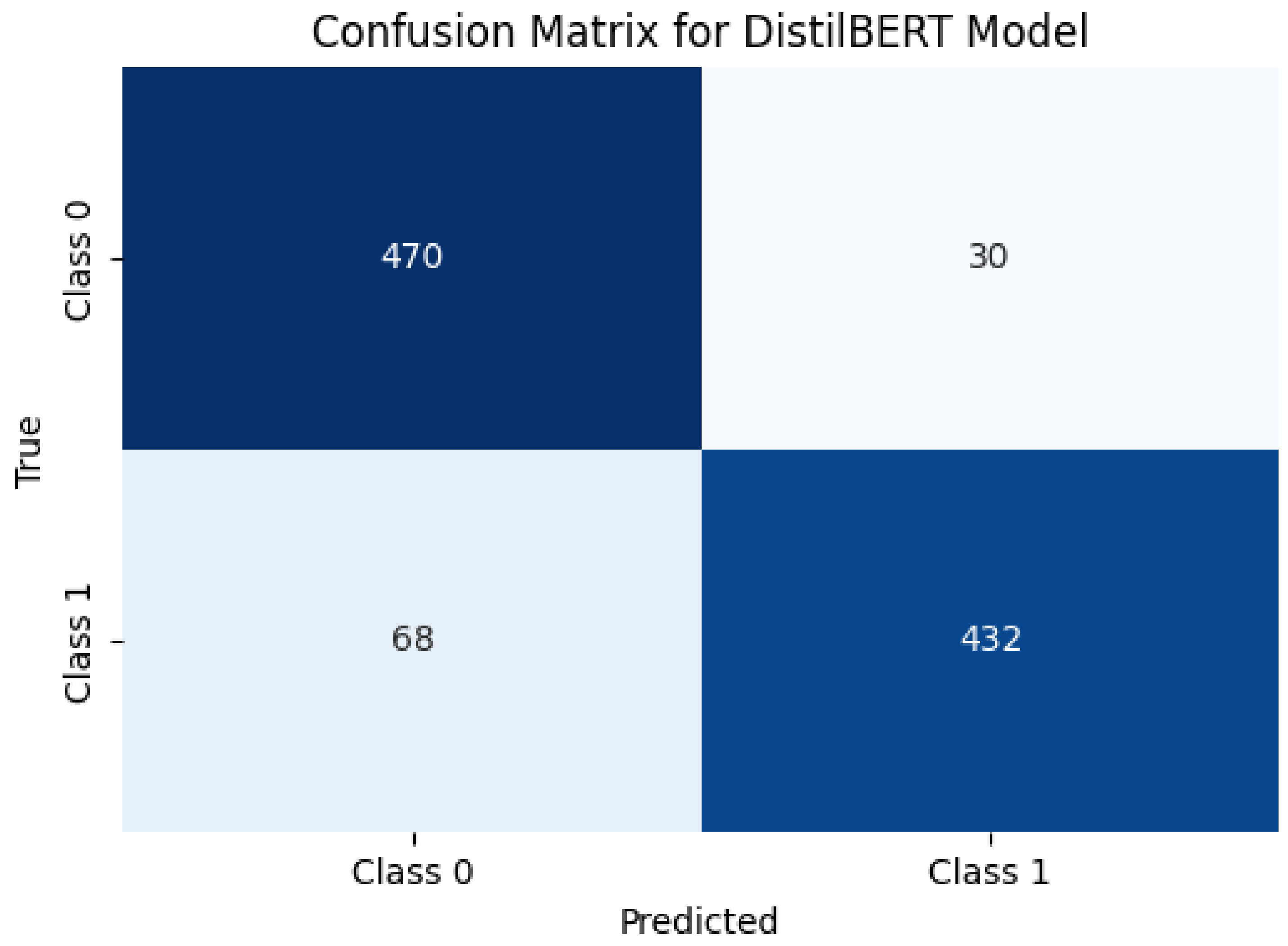
RESULT

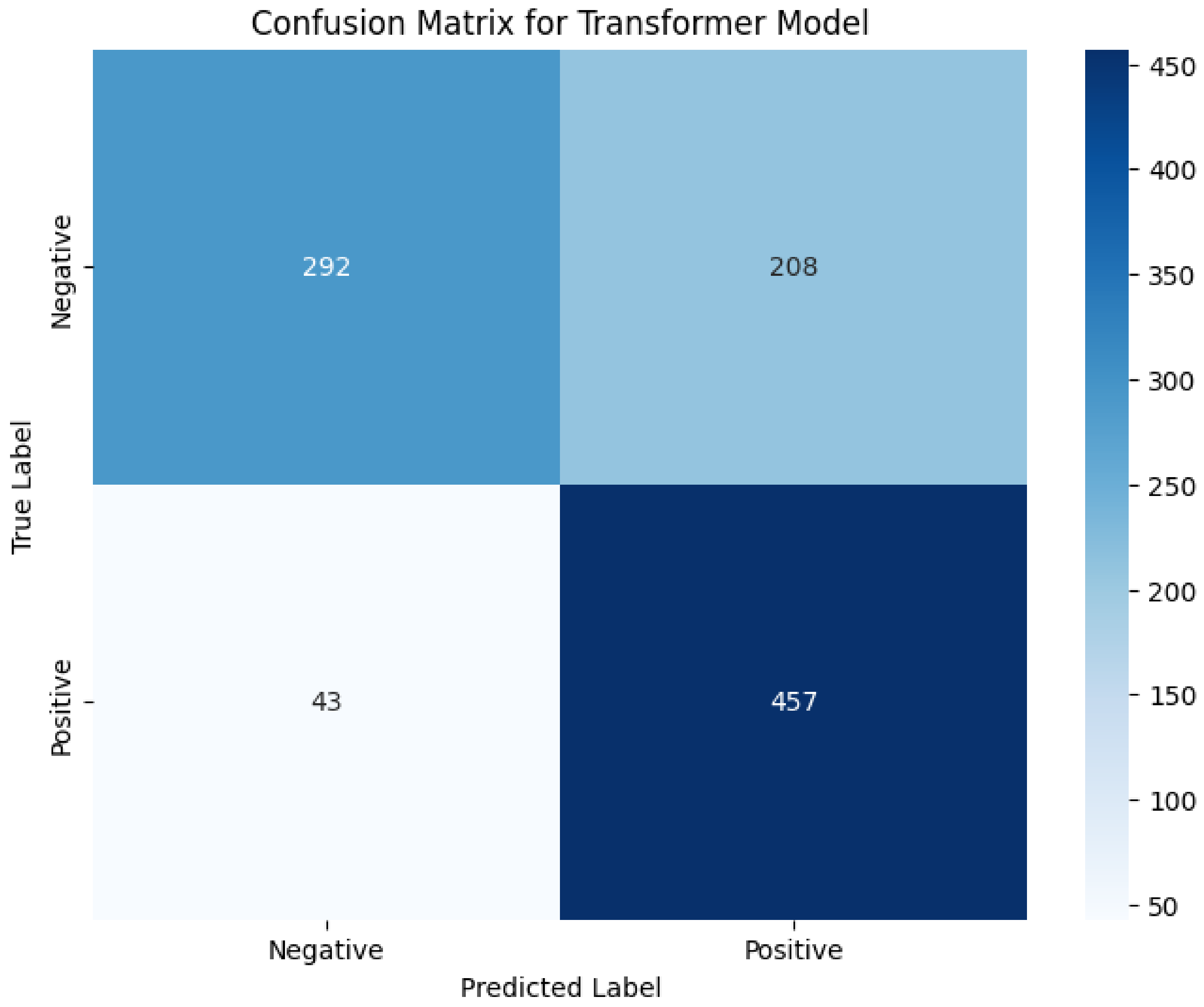
CONFUSION MATRIX





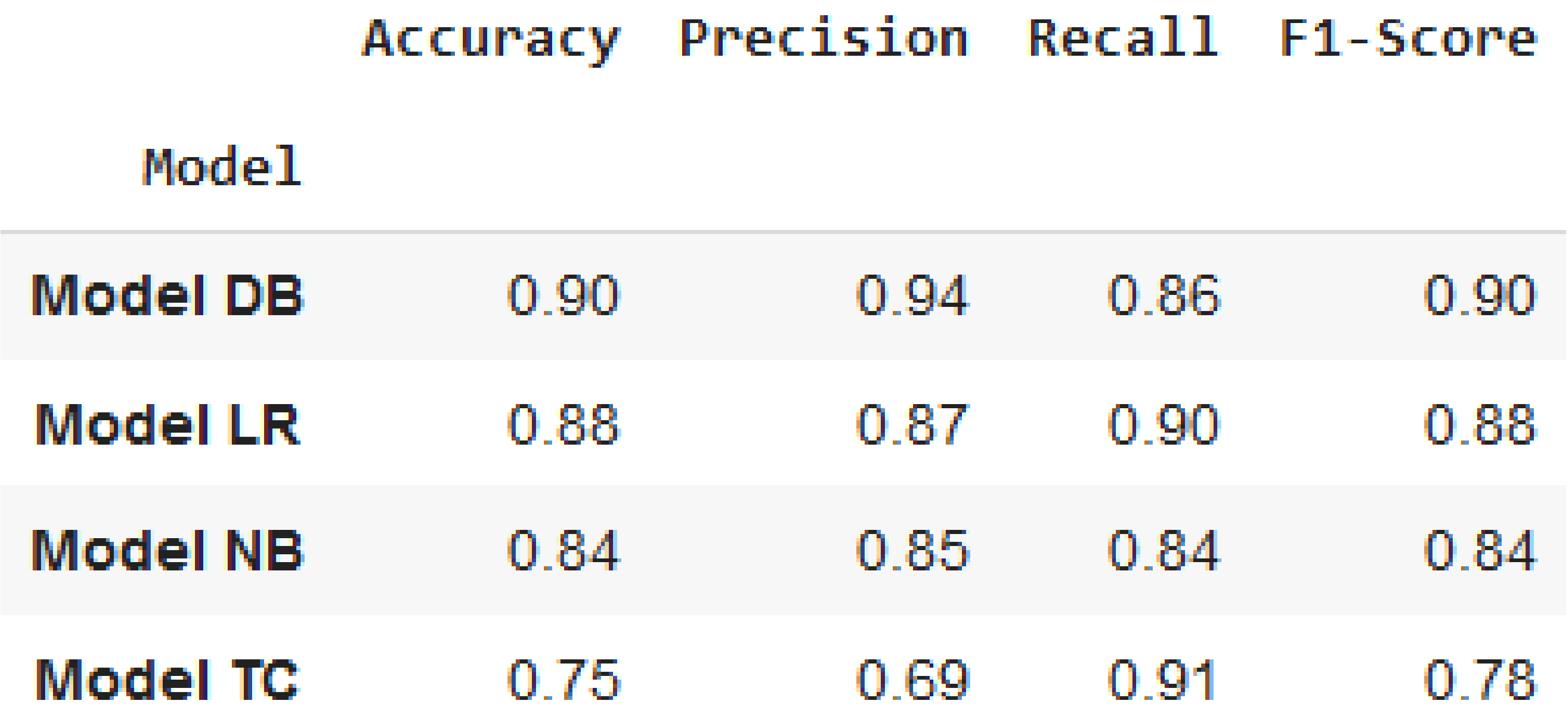
CONFUSION MATRIX





UEH UNIVERSITY College of Technology and Design

OTHER METRICS

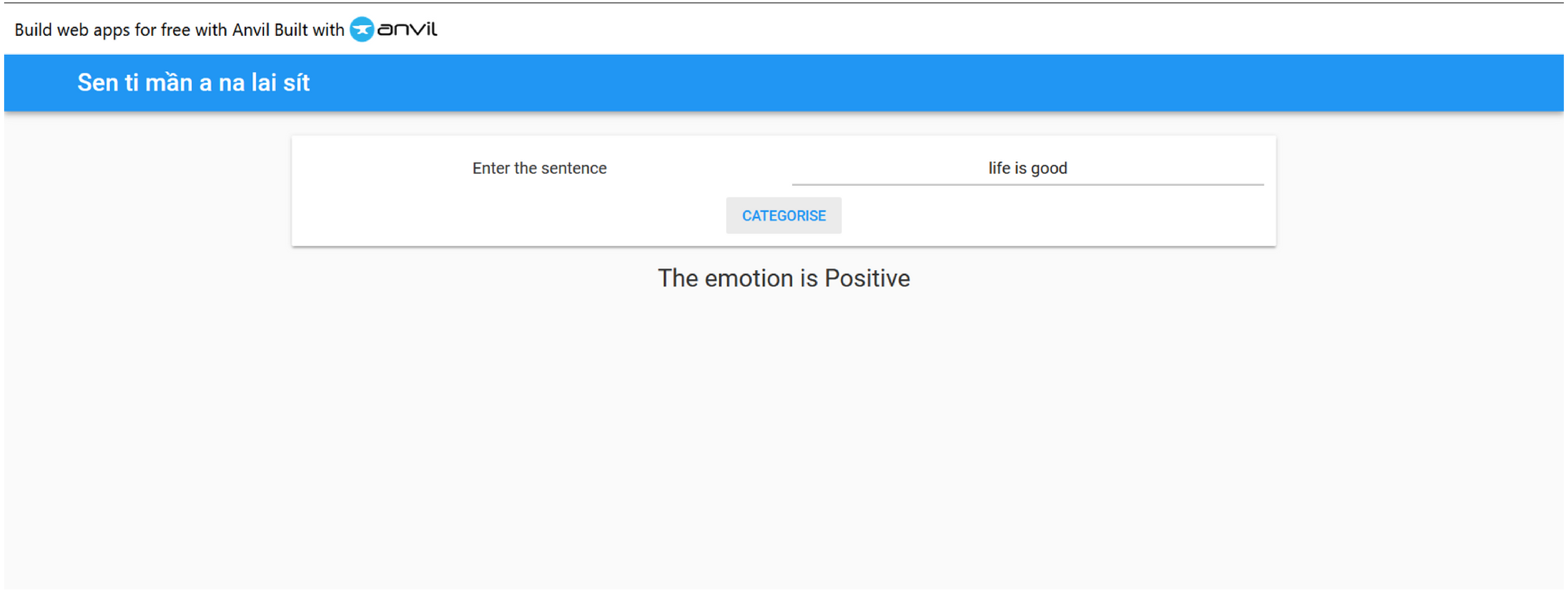


Model	Accuracy	Precision	Recall	F1-Score
Model DB	0.90	0.94	0.86	0.90
Model LR	0.88	0.87	0.90	0.88
Model NB	0.84	0.85	0.84	0.84
Model TC	0.75	0.69	0.91	0.78

CONCLUSIONS

Sau khi xem xét cả bốn chỉ số đánh giá để có cái nhìn toàn diện về hiệu suất của mỗi mô hình. Trong trường hợp này, mô hình finetune DistilBERT có vẻ là mô hình có hiệu suất tốt nhất dựa trên các chỉ số này.

DEMO (Anvil)



Build web apps for free with Anvil Built with anvil

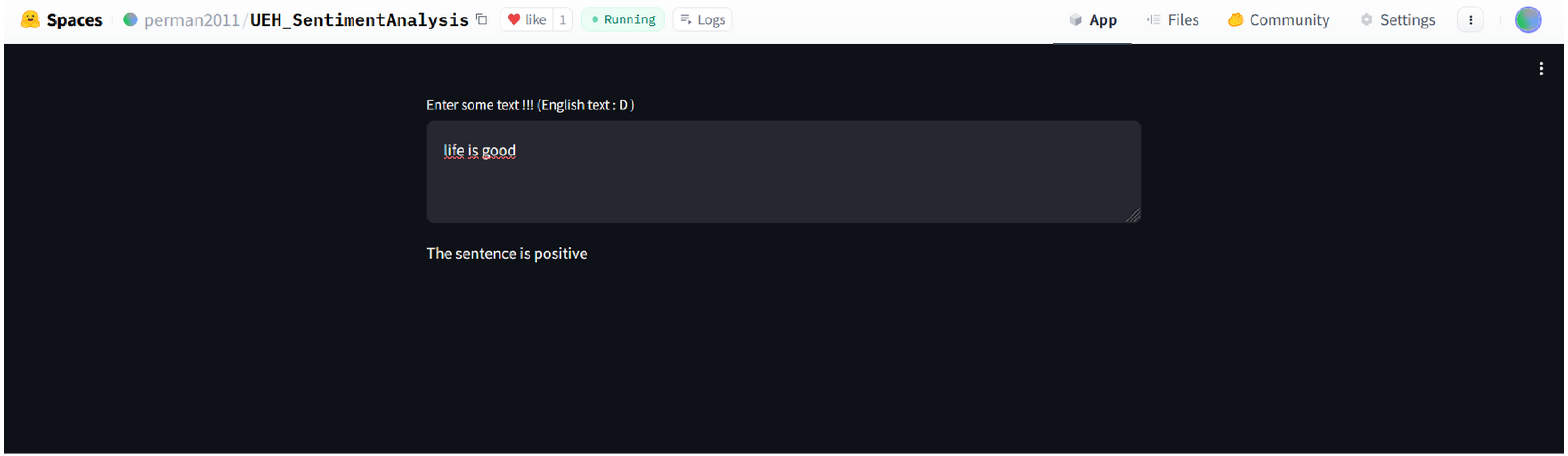
Sen ti mần a na lai sít

Enter the sentence: life is good

The emotion is Positive

SOURCE: https://understated-downright-contest.anvil.app/

DEMO (HuggingFace)



Enter some text !!! (English text : D)

life is good

The sentence is positive

SOURCE: https://huggingface.co/spaces/perman2011/UEH_SentimentAnalysis

REFERENCES

1. https://huggingface.co/spaces/perman2011/UEH_SentimentAnalysis
2. https://understated-downright-contest.anvil.app/
3. https://colab.research.google.com/drive/1yRTOP5clrG9OmAglrLpvysv5w4C2gXE0#scrollTo=M0c7nrbF3_rR

UEH UNIVERSITY | College of Technology and Design

SCHOOL OF BUSINESS INFORMATION TECHNOLOGY

CHEERS