KOREA UNIVERSITY

Movie Review Classifier

Project Proposal Presentation

WOO-CHAN KIM

I. Project title

Building the Movie review classifier (Positive or Negative review)

II. Project introduction

1) Objective

: The project objective is to build a classifier that evaluates the positive and negative sentiments of movie reviews. This model has high applicability as it can be used in various NLP fields such as document classification.

2) Motivation

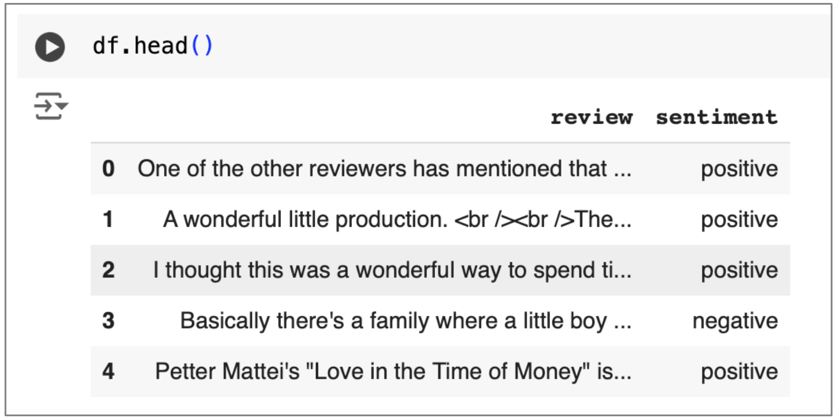
: As a Large Language Model researcher, I want to choose an NLP task.
So, I picked the Movie Review Classifier task, which is a famous NLP task.
I want to compare the performances of ML and DL in NLP fields by making two models.

KOREA UNIVERSITY

III. Dataset description

1) What is IMDB Dataset?

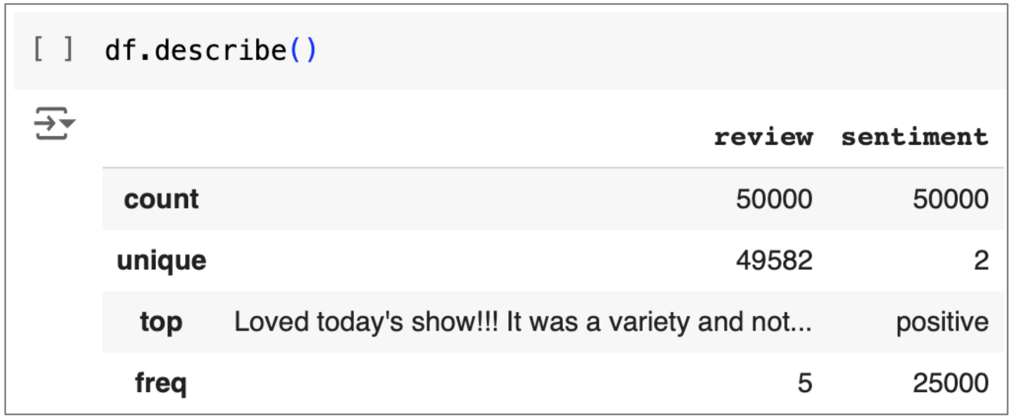
: This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. They provide a set of 25,000 highly polar movie reviews for training, and 25,000 for testing.



```
df.head()
```

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. <br /><br />The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

5 heads of IMDB Dataset



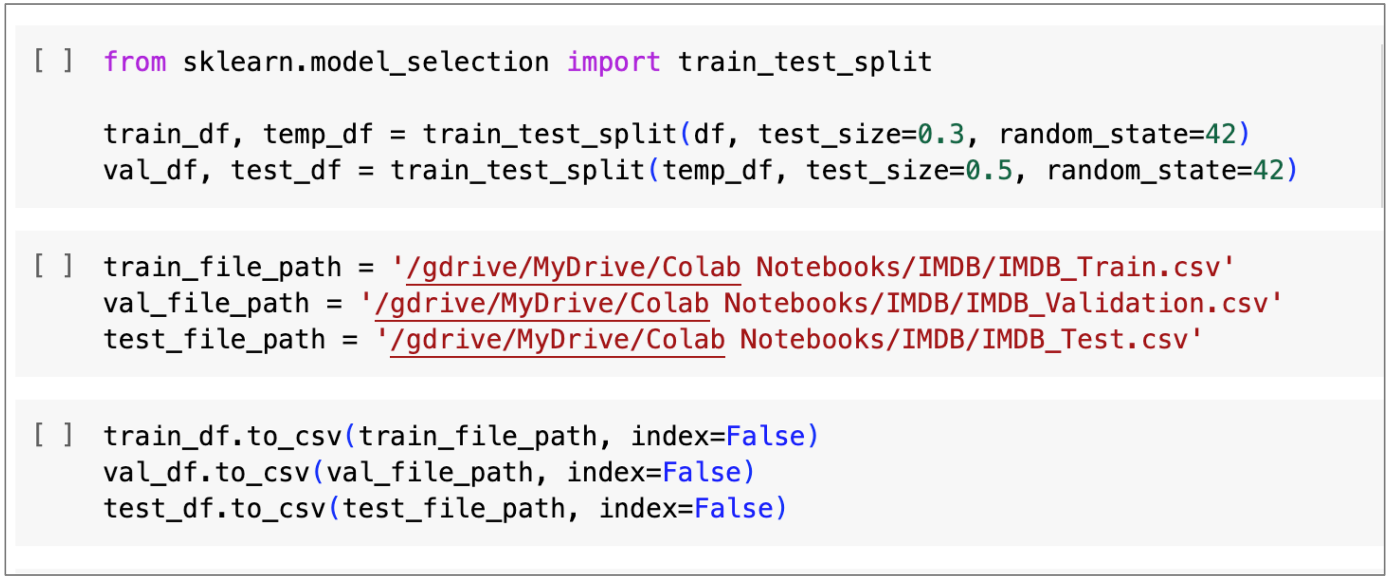
```
[ ] df.describe()
```

	review	sentiment
count	50000	50000
unique	49582	2
top	Loved today's show!!! It was a variety and not...	positive
freq	5	25000

Describe of IMDB Dataset

III. Dataset description

2) Train / Validation / Test dataset



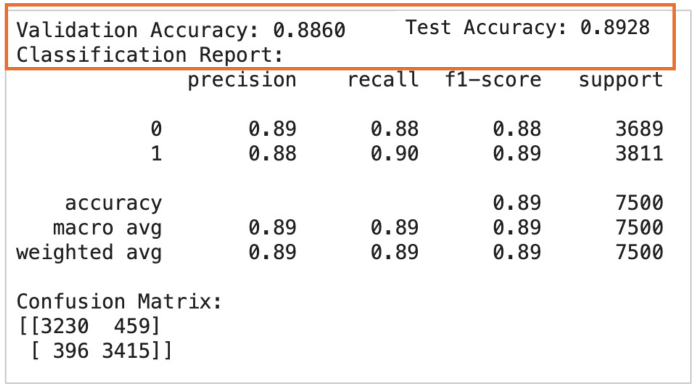
```
from sklearn.model_selection import train_test_split

train_df, temp_df = train_test_split(df, test_size=0.3, random_state=42)
val_df, test_df = train_test_split(temp_df, test_size=0.5, random_state=42)
```

```
train_file_path = '/gdrive/MyDrive/Colab Notebooks/IMDB/IMDB_Train.csv'
val_file_path = '/gdrive/MyDrive/Colab Notebooks/IMDB/IMDB_Validation.csv'
test_file_path = '/gdrive/MyDrive/Colab Notebooks/IMDB/IMDB_Test.csv'
```

```
train_df.to_csv(train_file_path, index=False)
val_df.to_csv(val_file_path, index=False)
test_df.to_csv(test_file_path, index=False)
```

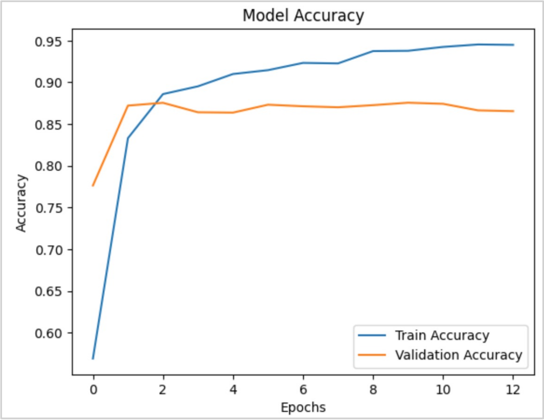
IV. Conclusion

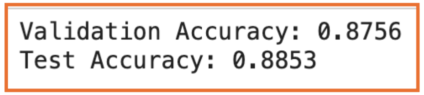
I developed 2 kinds of movie review classifiers using both logistic regression and LSTM models, both of which demonstrated high performance. While I initially encountered overfitting with the LSTM model, I resolved this by simplifying the model's structure. Typically, deep learning outperforms machine learning, but due to the relatively small size of the IMDB dataset, both models delivered similar results.



```
Validation Accuracy: 0.8860      Test Accuracy: 0.8928
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.88      0.88      3689
           1       0.88      0.90      0.89      3811

    accuracy                           0.89      7500
   macro avg       0.89      0.89      0.89      7500
weighted avg       0.89      0.89      0.89      7500

Confusion Matrix:
[[3230  459]
 [ 396 3415]]
```

Logistic Regression





```
Validation Accuracy: 0.8756
Test Accuracy: 0.8853
```

LSTM

KOREA
UNIVERSITY

Q & A

E-mail : kimwc620@korea.ac.kr

GitHub address :

https://github.com/SkyDreamer14/IMDB_Dataset_Movie_Reviews