DTSA 5511 Final

Deep learning for IMDB Review

Tingting Guo

Overview

IMDB dataset having 50K movie reviews for natural language processing or Text analytics.

For binary sentiment classification

Develop basic LSTM model and its variations

Practice Huggingface transformers

https://github.com/gt2onew/dtsa5511/blob/main/week6/dtsa5511final.ipynb

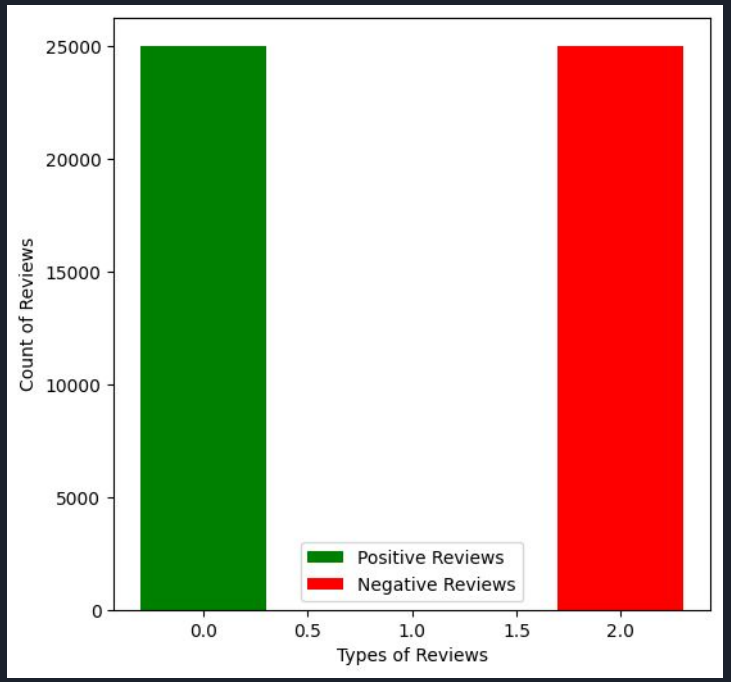
DataSet

```
df=pd.read_csv('IMDB Dataset.csv')
df.head()
```

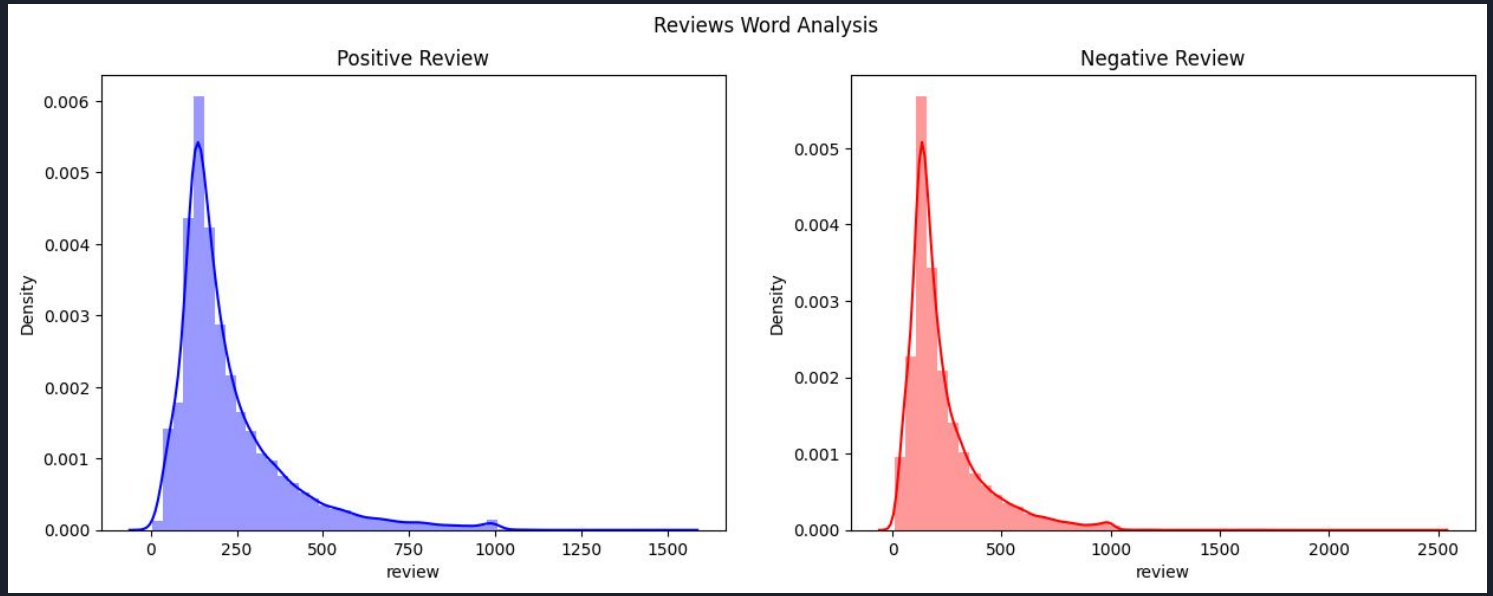
	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...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

```
df.shape
```

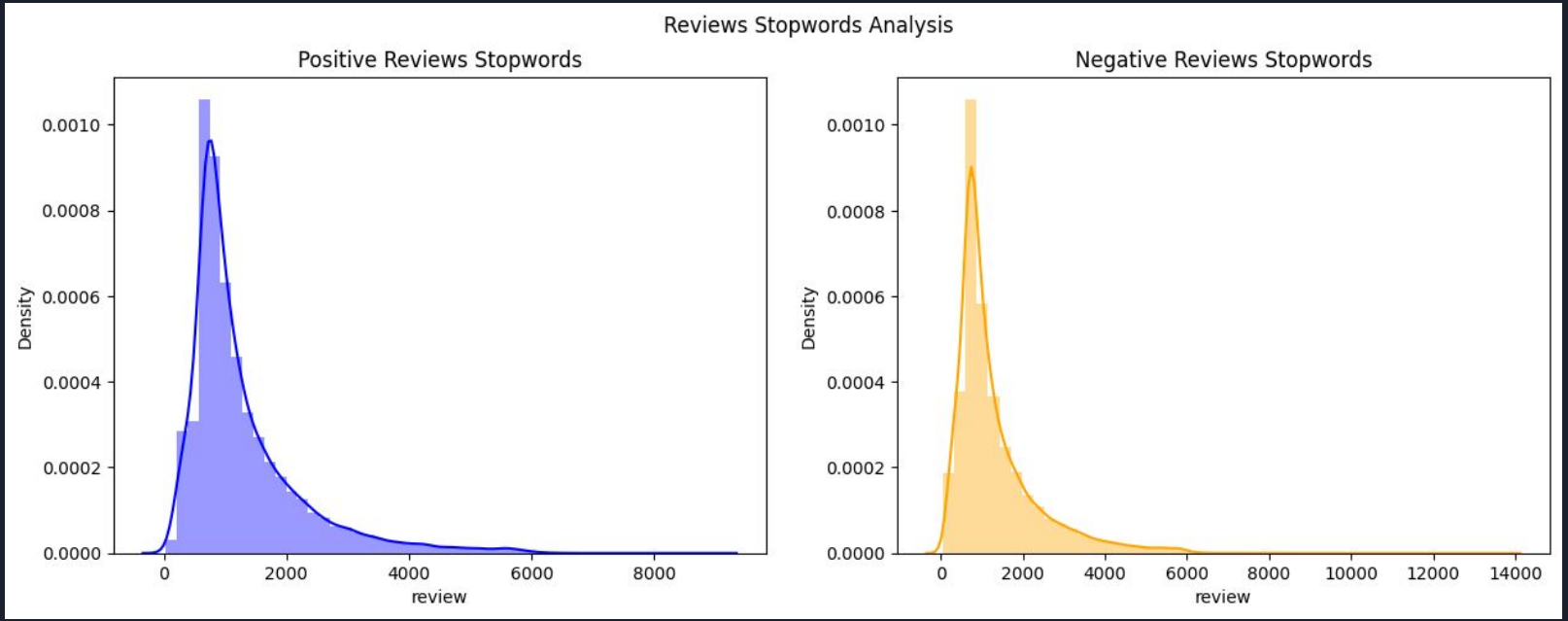
```
(50000, 2)
```



Statistics


Reviews Word Analysis
Positive Review
Density
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
250
500
750
1000
1250
1500
review
Negative Review
Density
0.005
0.004
0.003
0.002
0.001
0.000
0
500
1000
1500
2000
2500
review

Statistics


Reviews Stopwords Analysis
Positive Reviews Stopwords
Negative Reviews Stopwords
Density
review
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0
2000
4000
6000
8000
10000
12000
14000

Data Cleaning

Data cleaning

- HTML codes
- URLs
- Emojis
- Stopwords
- Punctuations
- Expanding Abbreviations

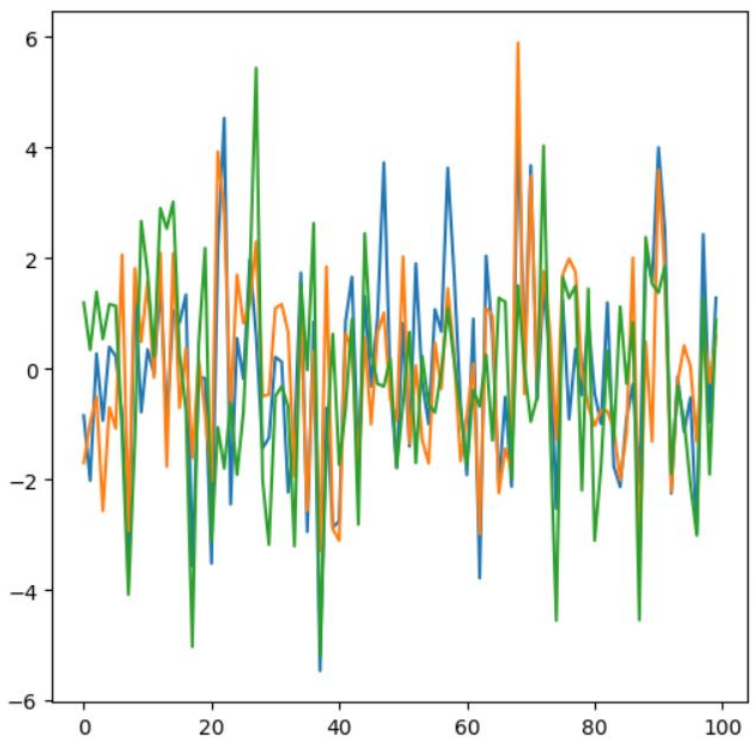
Wordcloud - positive label

Wordcloud - negative label

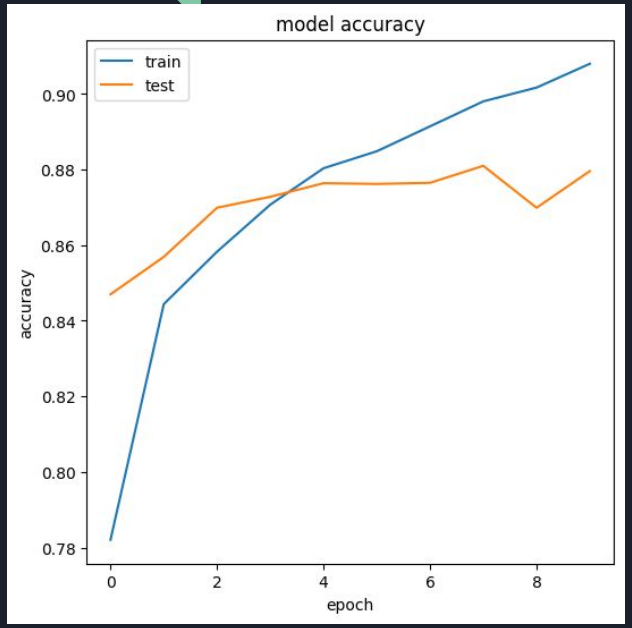
Vectorization and Embeddings

- Word2Vec
- Embedding matrix

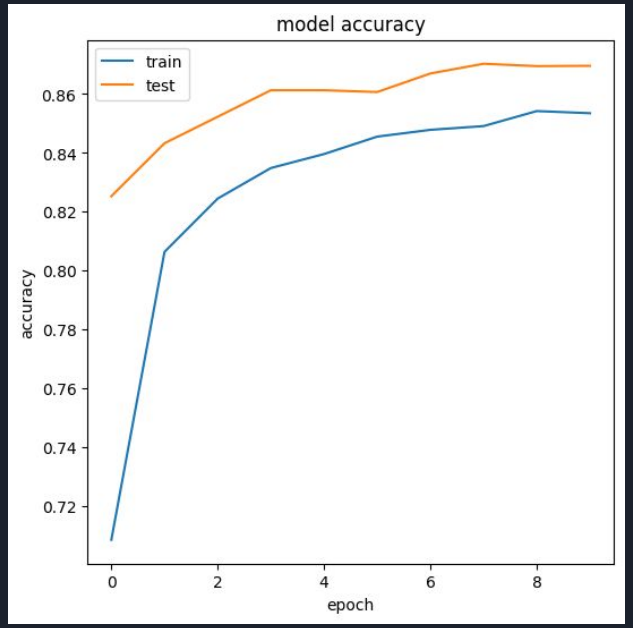
```
#Visualize the Word Vectors
plt.plot(loaded_model.wv['good'])
plt.plot(loaded_model.wv['great'])
plt.plot(loaded_model.wv['well'])

plt.show()
```



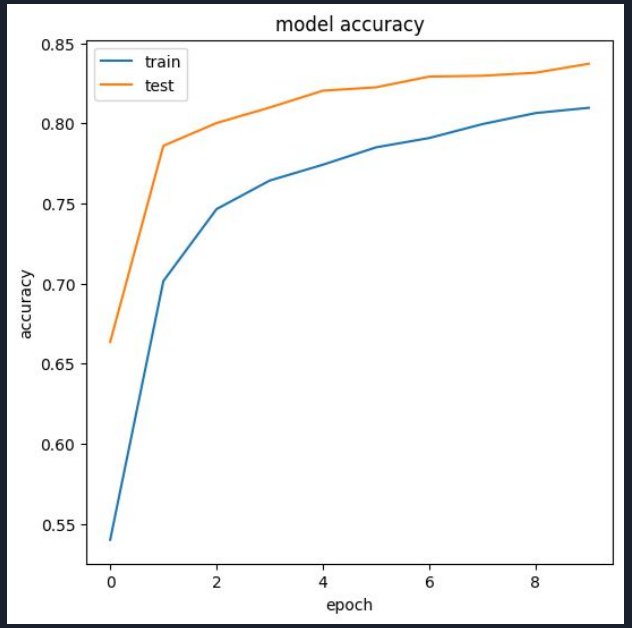
LSTM



Basic LSTM

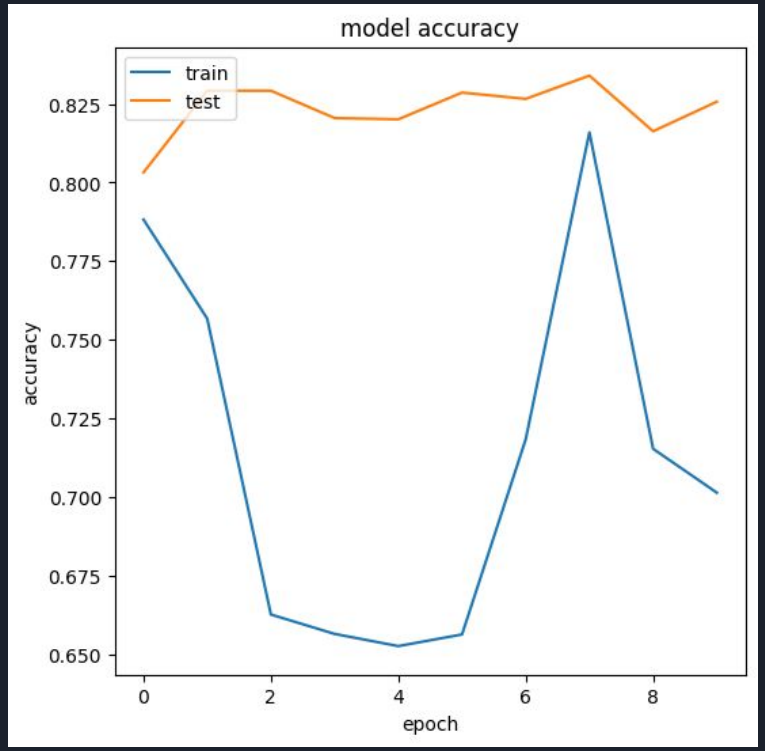


+Dense layer, dropout

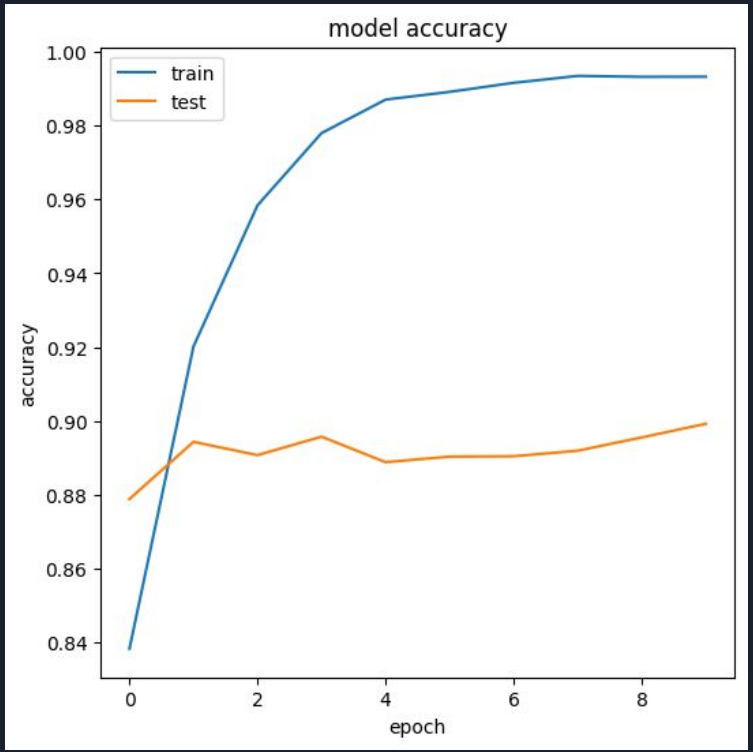


Extra LSTM layer

GPT-2 Pretrained



BERT pre-trained



Thank you!