News Category Classification

Benjamin Barnett and Yaniv Bronshtein

Introduction

How to categorize news articles using only their titles and short descriptions?

Our plan to solve this problem: **Multinomial Classification**

Used Naive Bayes classifier, 1-3 hidden layer of size 100 MLP nets as baseline

Trained LSTM and DistilBERT on 80% of data with 20% validation split

Regularized tuned BERT had best performance with over 90% testing accuracy

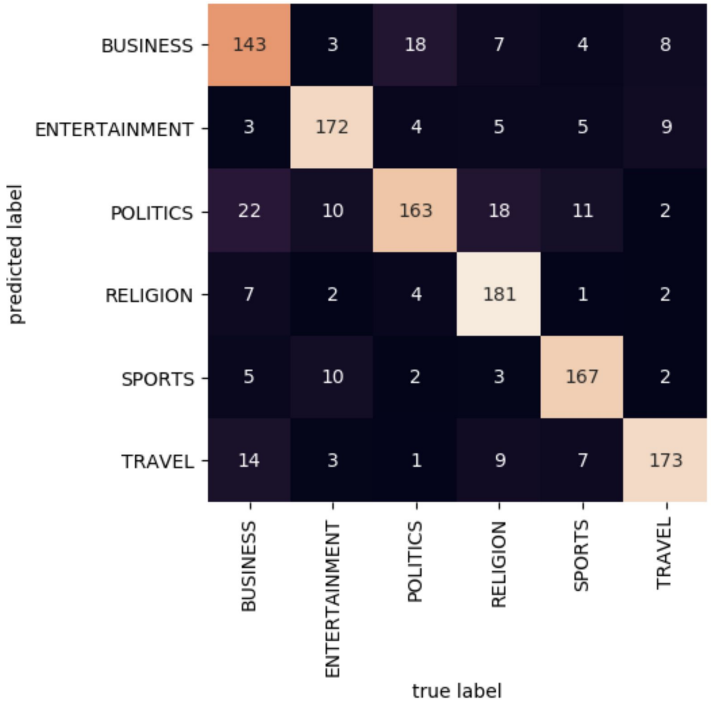
Data Collection

Collected 200k HuffPost news articles published from 2012-2018 from Kaggle, but filtered data to select 6k articles from 6 distinct categories, taking 1k each

Article Categories: Politics, Entertainment, Travel, Business, Sports, & Religion

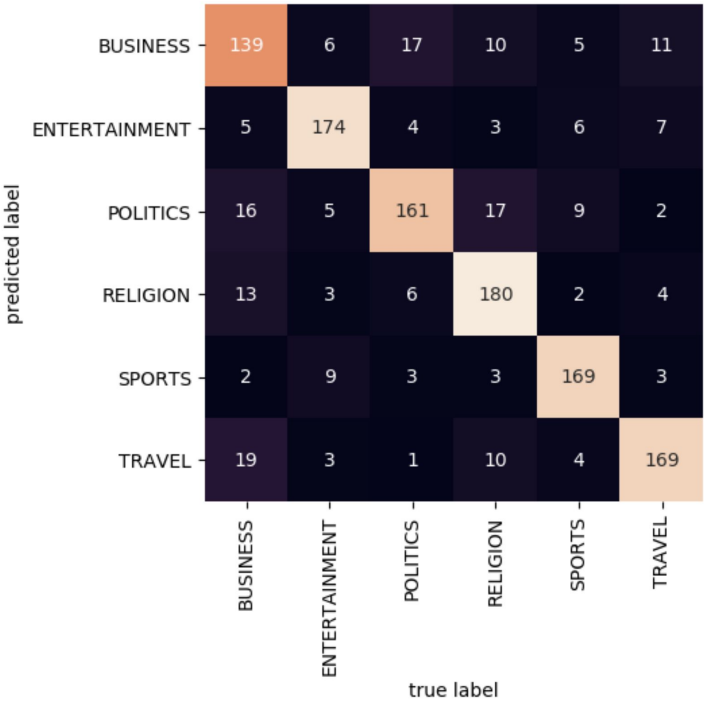
Replaced unknown words with <UNK> tokens

Multinomial Bayes



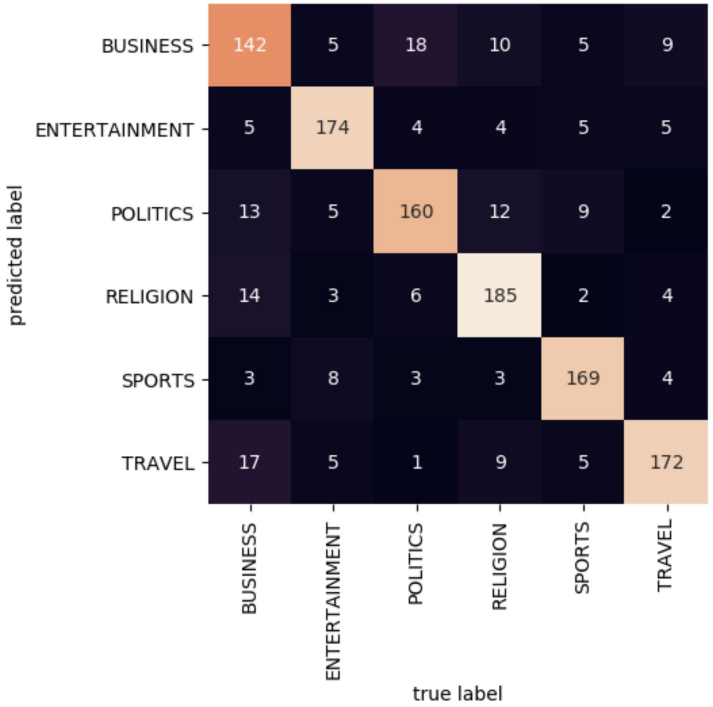
Accuracy Score: 0.8325

MLP with 1 Layer, 100 Units



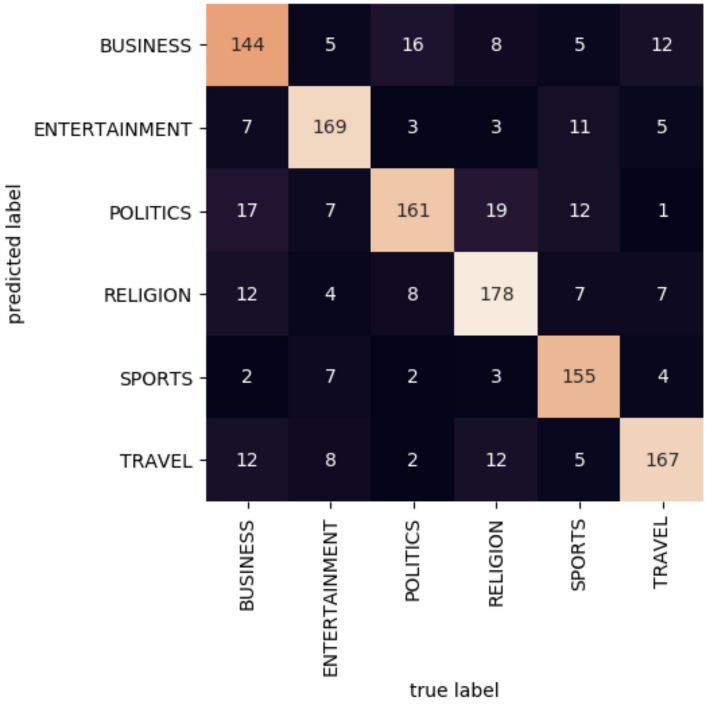
Accuracy Score: 0.8267

MLP with 2 Layers, 100 Units



Accuracy Score: 0.835

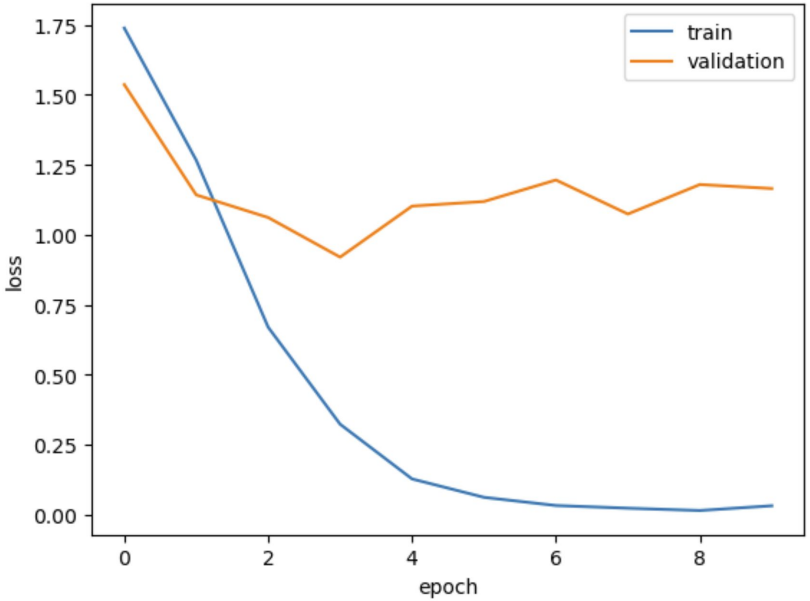
MLP with 3 Layers, 100 Units

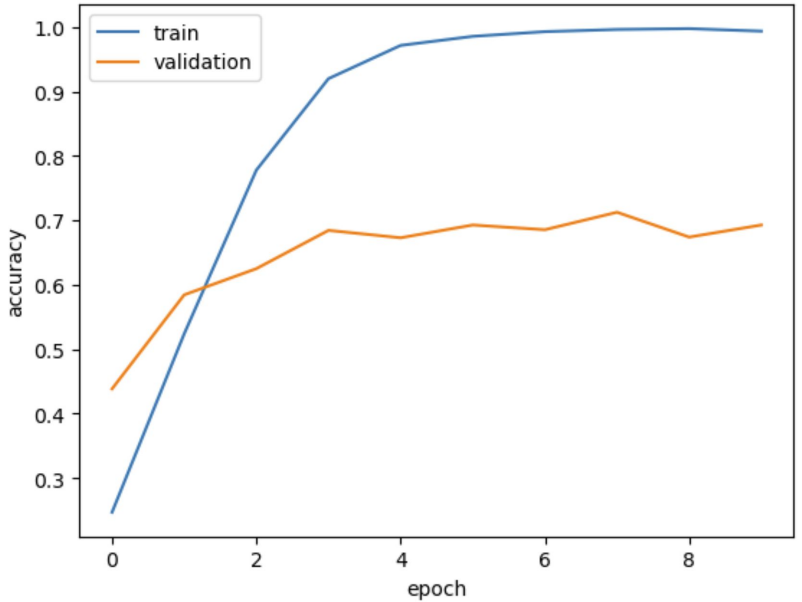


Accuracy Score: 0.8117

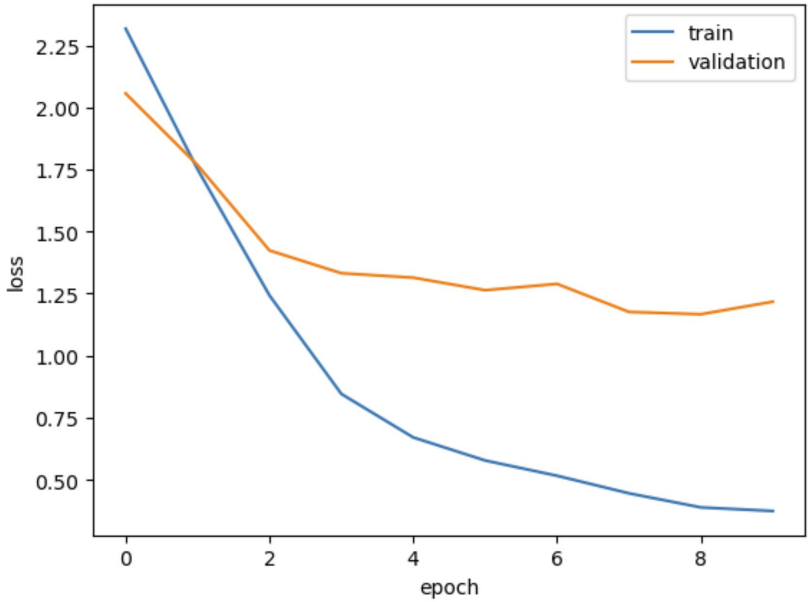
```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 embedding (Embedding)        (None, 81, 100)           5000000

 lstm (LSTM)                  (None, 100)               80400

 dense (Dense)                (None, 6)                 606

=================================================================
Total params: 5,081,006
Trainable params: 5,081,006
Non-trainable params: 0
_________________________________________________________________
```

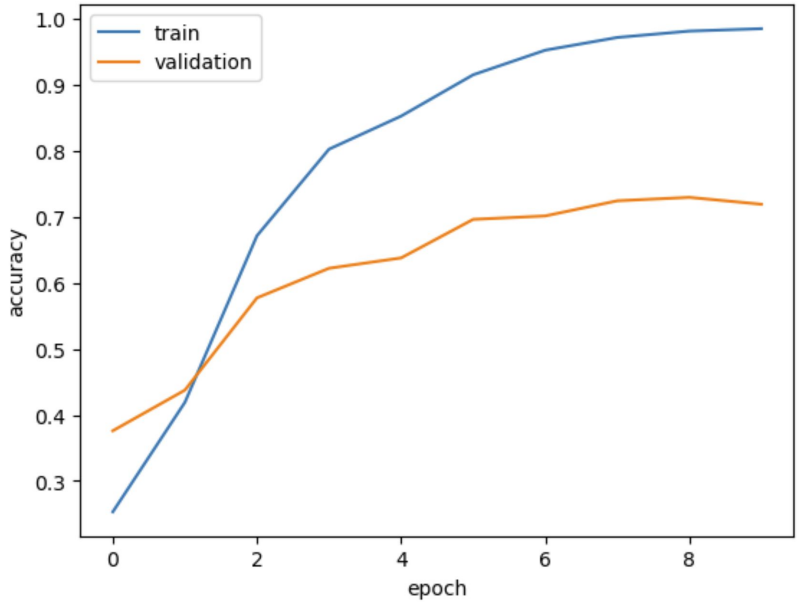
Long Short-Term Memory



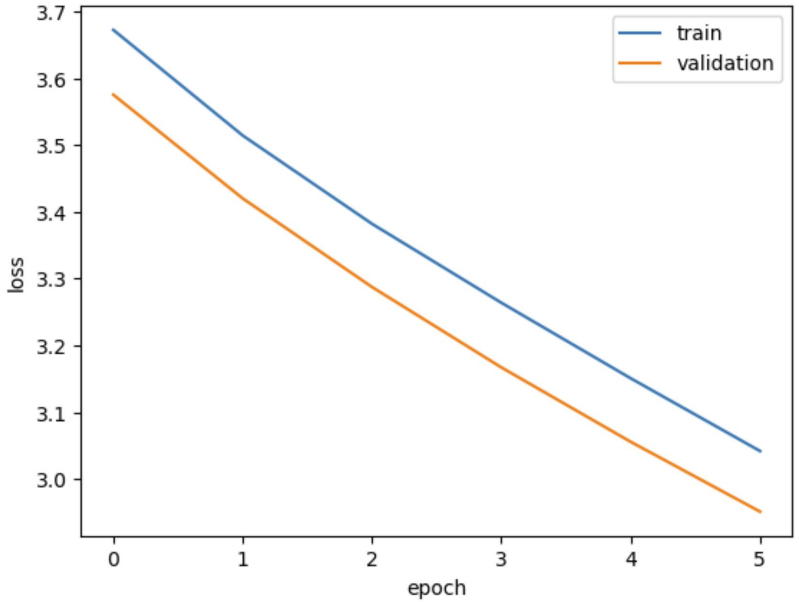


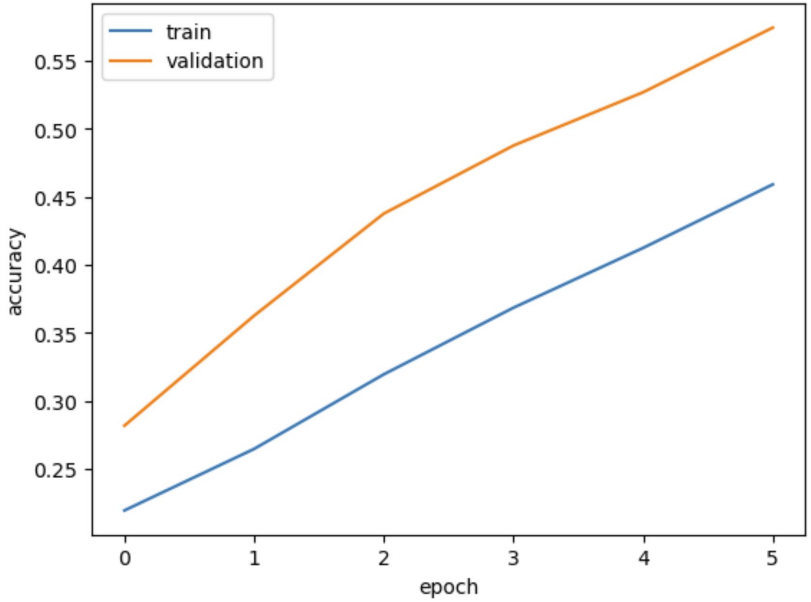
Regularized LSTM



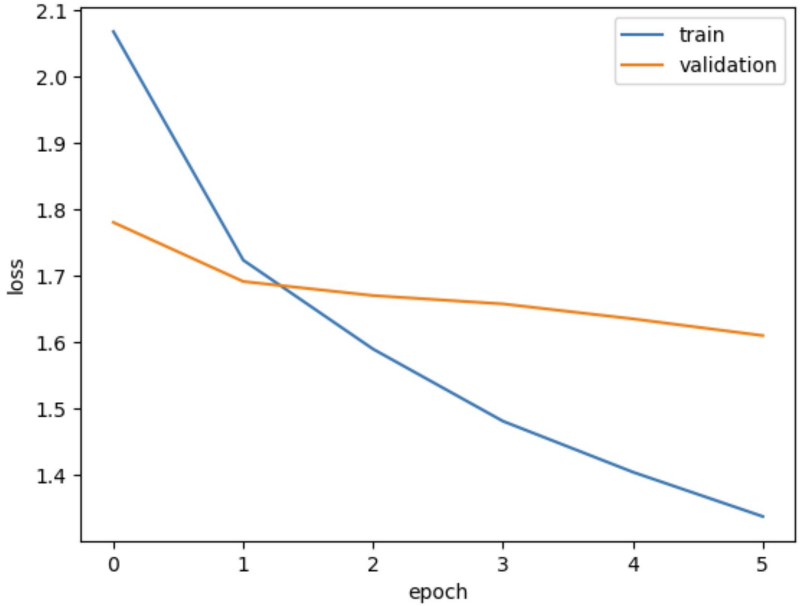


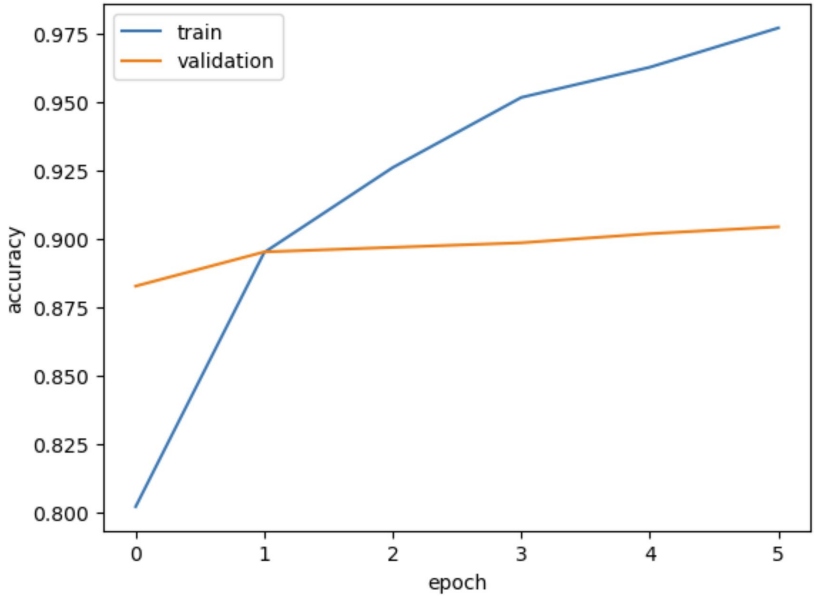
Regularized Untuned DistilBERT





Regularized Tuned DistilBERT





Conclusion

Accuracy Scores on validation data:

Base Models

Naive Bayes	Default MLP	2 Hidden MLP	3 Hidden MLP
0.8325	0.8267	0.8350	0.8117

Tensorflow/ Transformer Models

LSTM	Regularized LSTM	Regularized Untuned BERT	Regularized Tuned BERT
0.6927	0.7188	0.5742	0.9042