
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

predicted label	BUSINESS	ENTERTAINMENT	POLITICS	RELIGION	SPORTS	TRAVEL
	143	3	18	7	4	8
	3	172	4	5	5	9
	22	10	163	18	11	2
	7	2	4	181	1	2
	5	10	2	3	167	2
true label	BUSINESS	ENTERTAINMENT	POLITICS	RELIGION	SPORTS	TRAVEL
	14	3	1	9	7	173

Accuracy Score: 0.8325

MLP with 1 Layer, 100 Units

predicted label	BUSINESS	ENTERTAINMENT	POLITICS	RELIGION	SPORTS	TRAVEL
	139	6	17	10	5	11
	5	174	4	3	6	7
	16	5	161	17	9	2
	13	3	6	180	2	4
	2	9	3	3	169	3
true label	BUSINESS	ENTERTAINMENT	POLITICS	RELIGION	SPORTS	TRAVEL
	19	3	1	10	4	169

Accuracy Score: 0.8267

MLP with 2 Layers, 100 Units

predicted label	BUSINESS	ENTERTAINMENT	POLITICS	RELIGION	SPORTS	TRAVEL
	142	5	18	10	5	9
	5	174	4	4	5	5
	13	5	160	12	9	2
	14	3	6	185	2	4
	3	8	3	3	169	4
true label	BUSINESS	ENTERTAINMENT	POLITICS	RELIGION	SPORTS	TRAVEL
	17	5	1	9	5	172

Accuracy Score: 0.835

MLP with 3 Layers, 100 Units

predicted label	BUSINESS	ENTERTAINMENT	POLITICS	RELIGION	SPORTS	TRAVEL
	144	5	16	8	5	12
	7	169	3	3	11	5
	17	7	161	19	12	1
	12	4	8	178	7	7
	2	7	2	3	155	4
true label	BUSINESS	ENTERTAINMENT	POLITICS	RELIGION	SPORTS	TRAVEL
	12	8	2	12	5	167

Accuracy Score: 0.8117

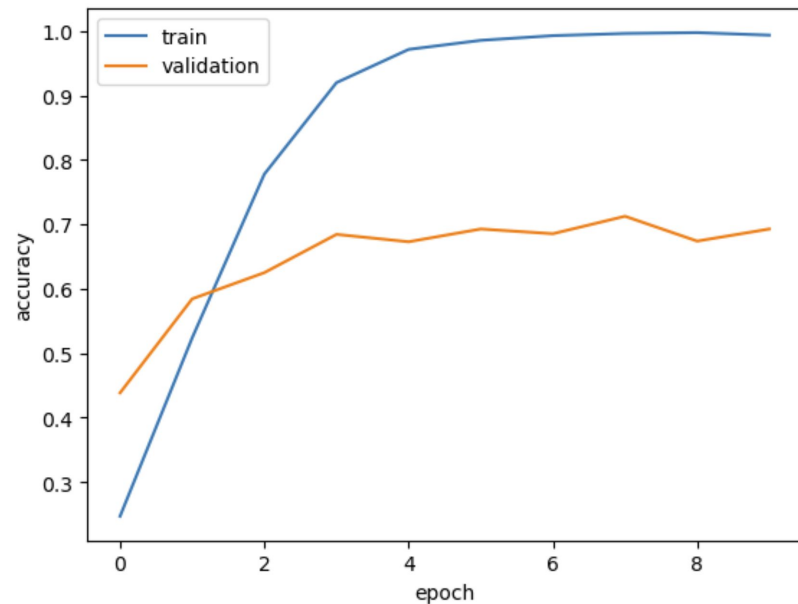
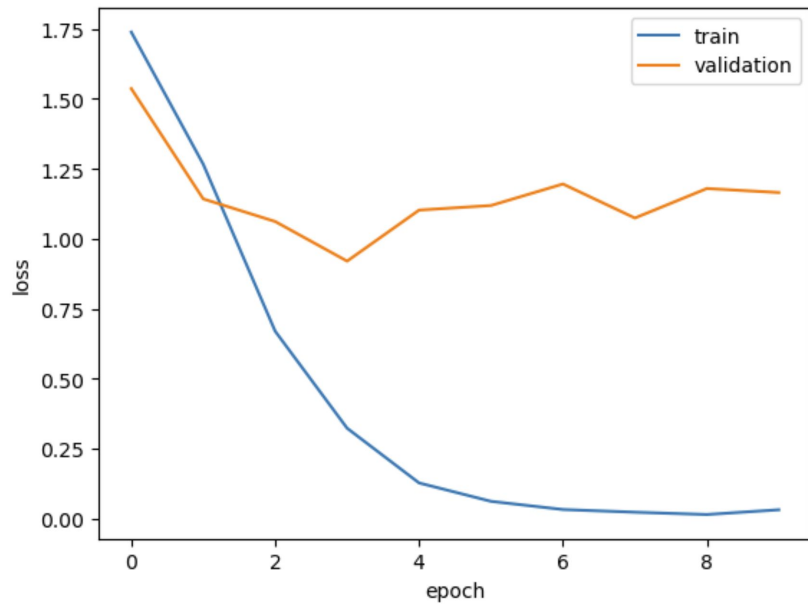
Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 81, 100)	5000000
lstm (LSTM)	(None, 100)	80400
dense (Dense)	(None, 6)	606

=====
Total params: 5,081,006

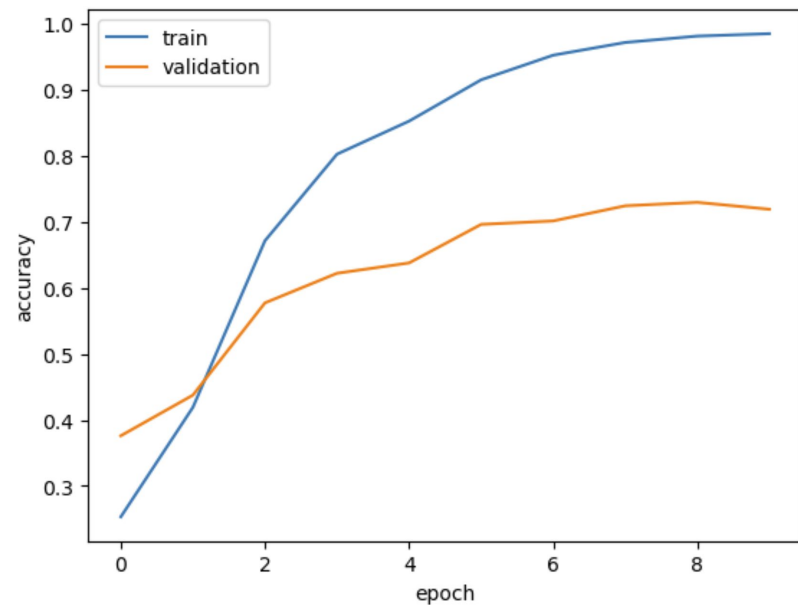
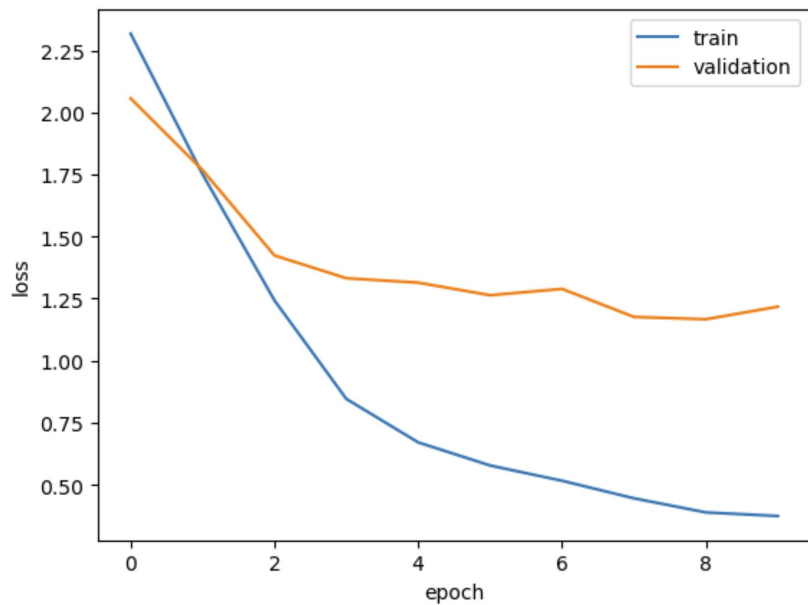
Trainable params: 5,081,006

Non-trainable params: 0
=====

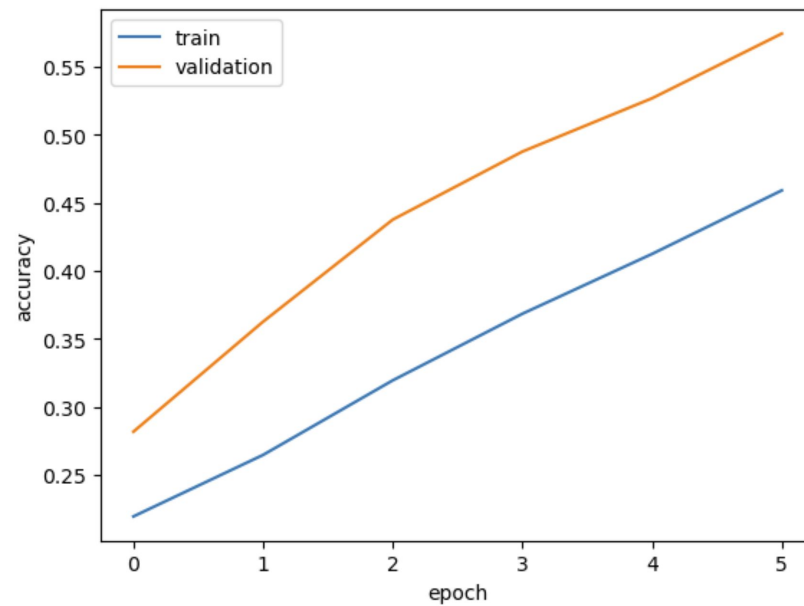
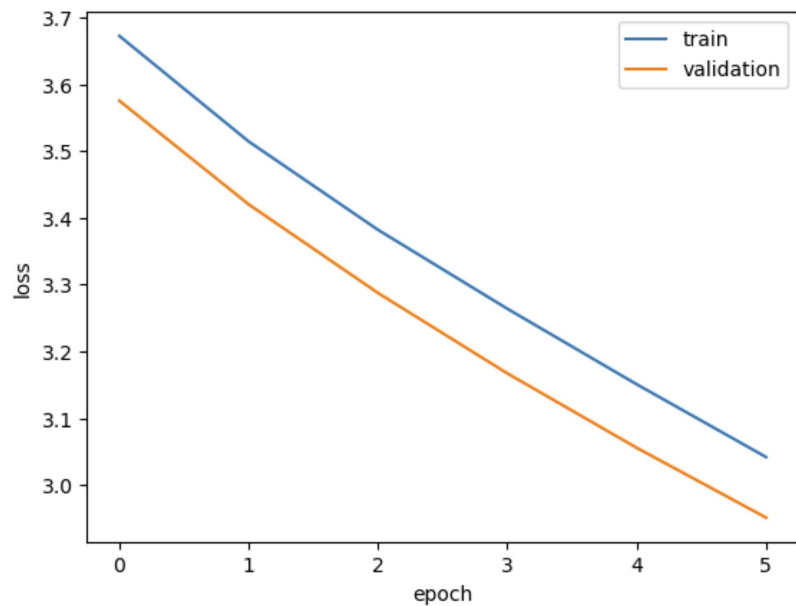
Long Short-Term Memory



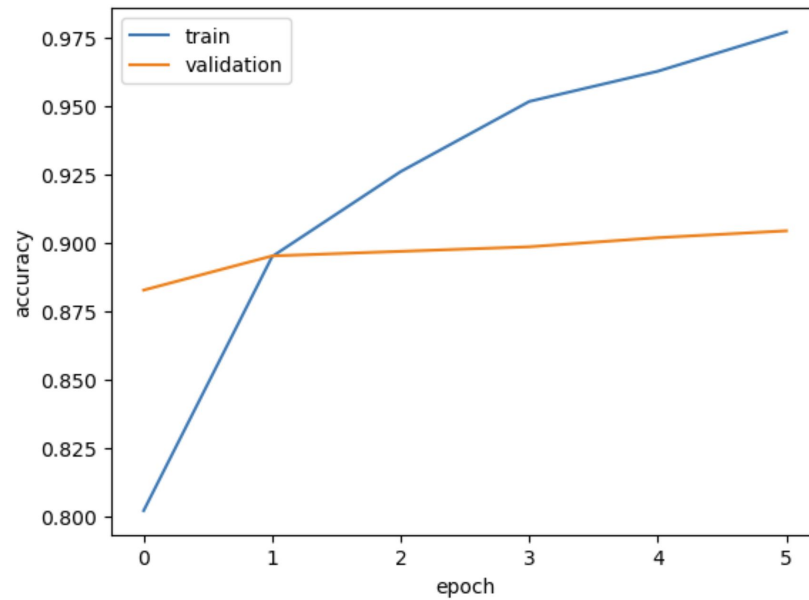
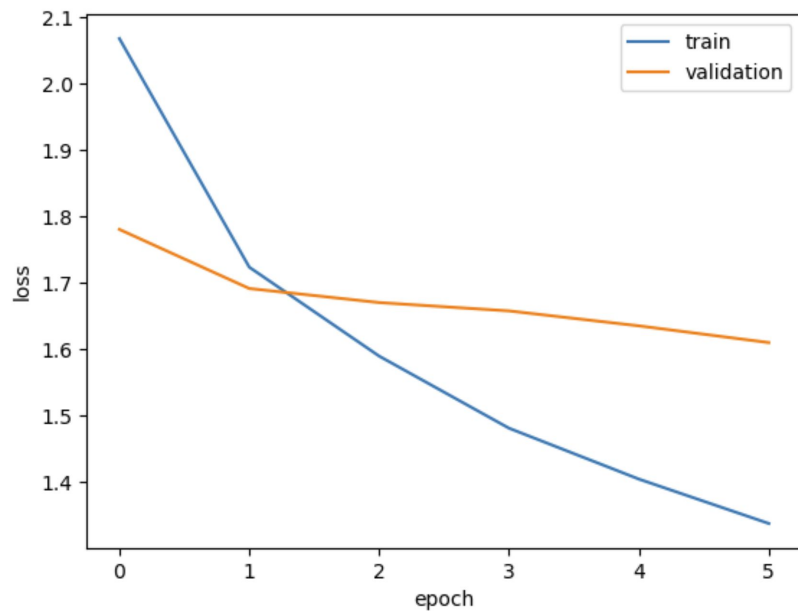
Regularized LSTM



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