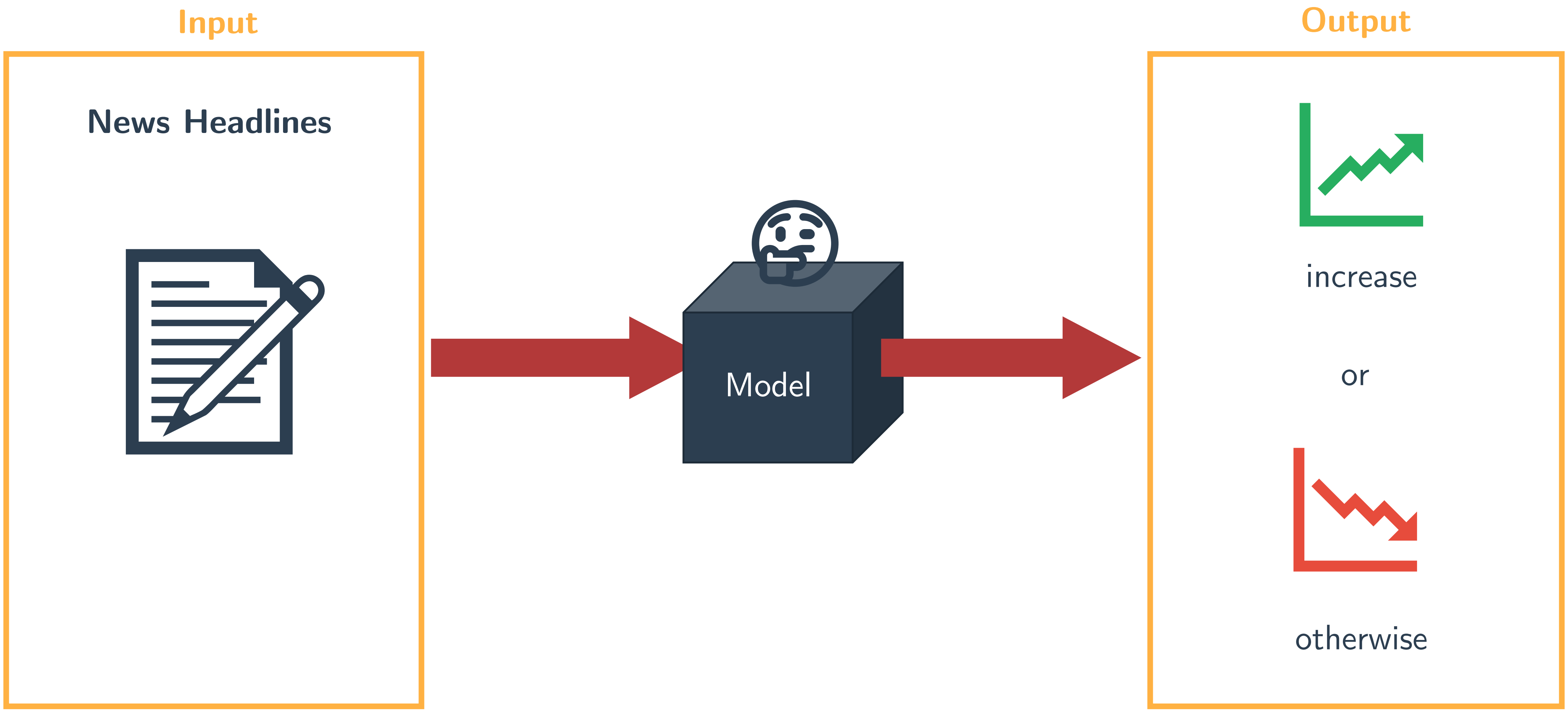
Machine Learning Project
Daily News for Stock Movement Prediction

by

Zimin Luo 417124



Input

News Headlines

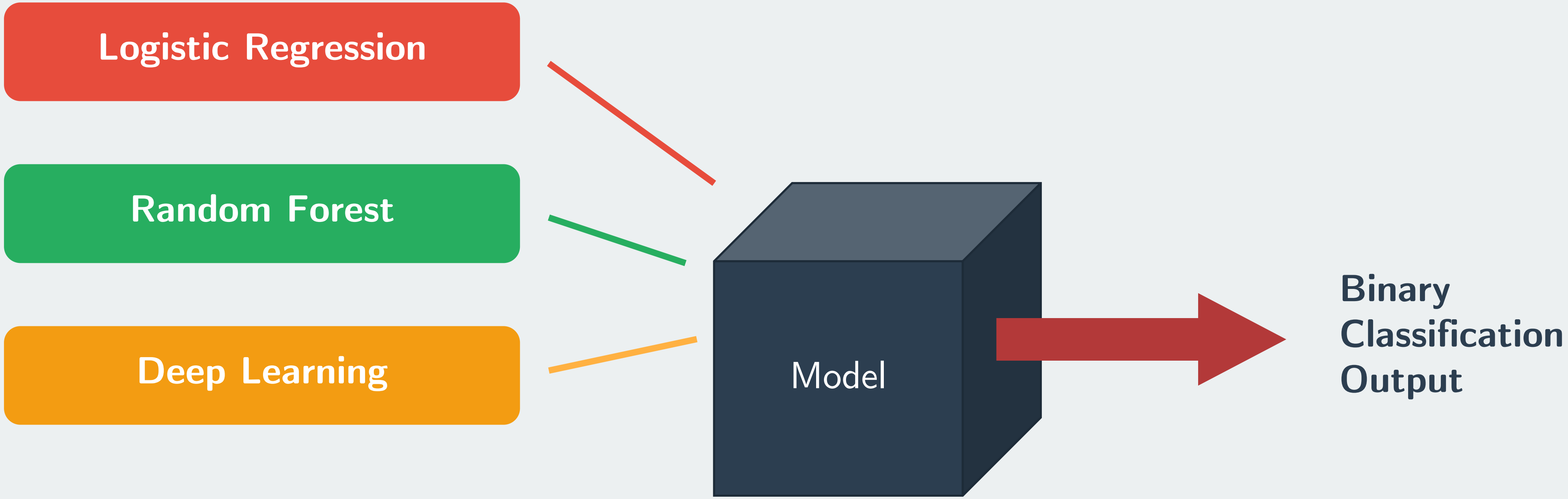
Model

Output

increase

or

otherwise



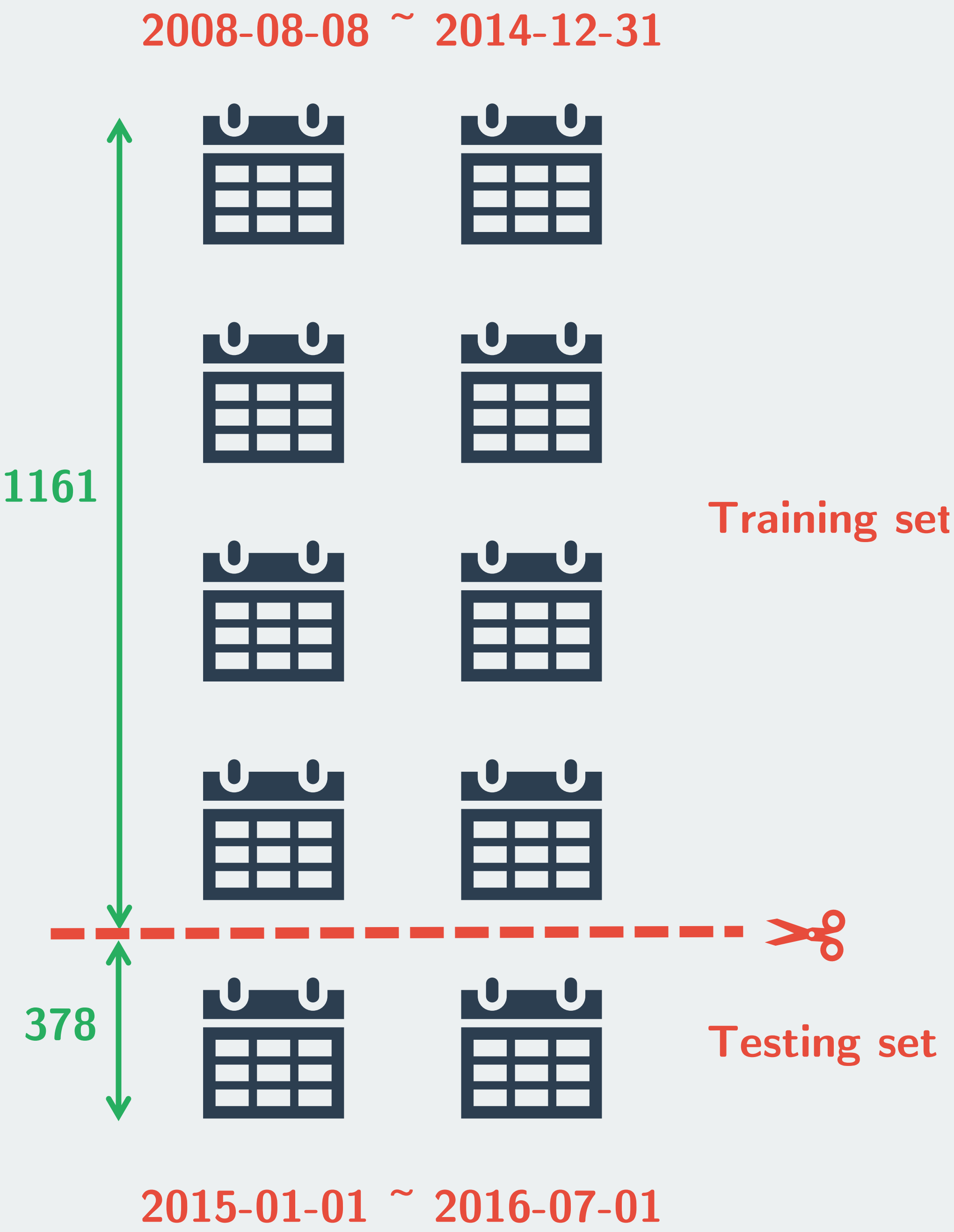
Logistic Regression

Random Forest

Deep Learning

Model

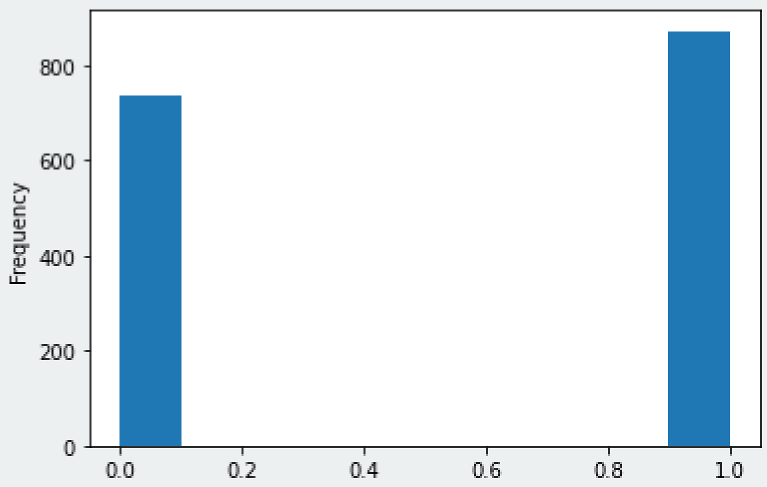
Binary
Classification
Output

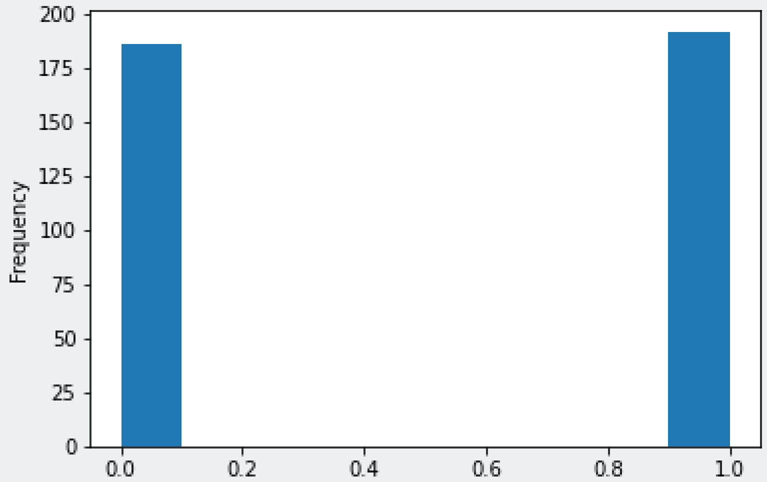


1. Combine all news headlines into one blob

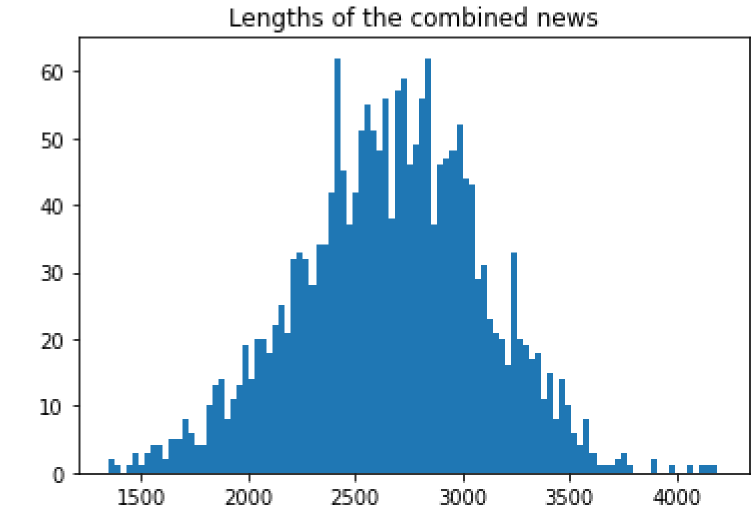
2. Clean text:
   - normalize (convert to lower cases)
   - remove HTML tags
   - remove texts in brackets []
   - remove punctuation
   - remove digits
   - fix concentration *
   - remove stop words *
   - stem words **
   - lemmatization **

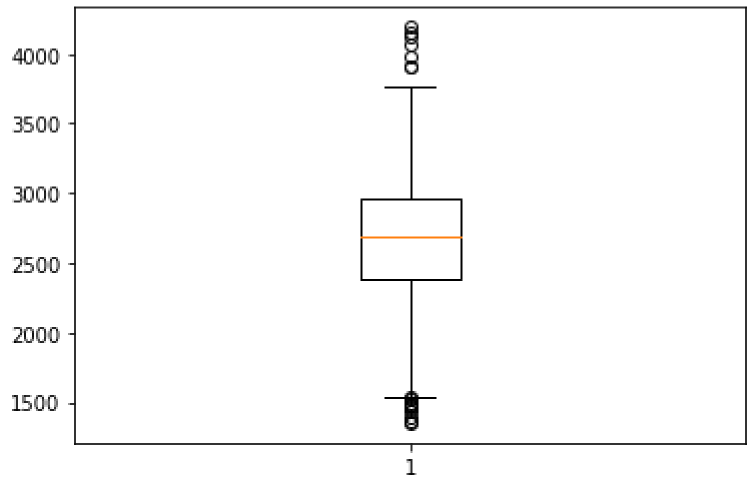
Training set



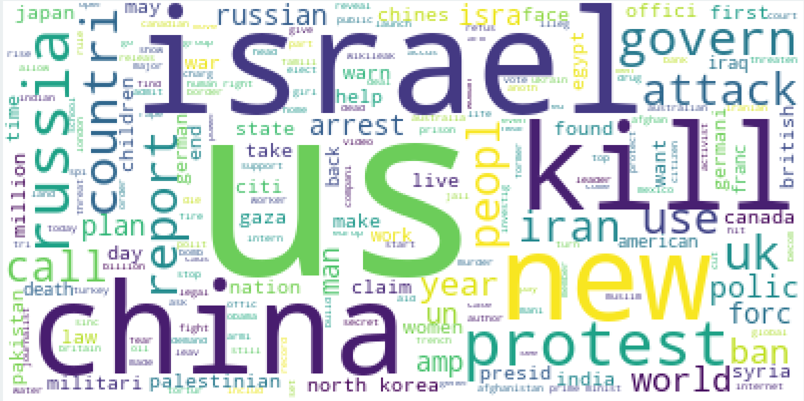


Test set

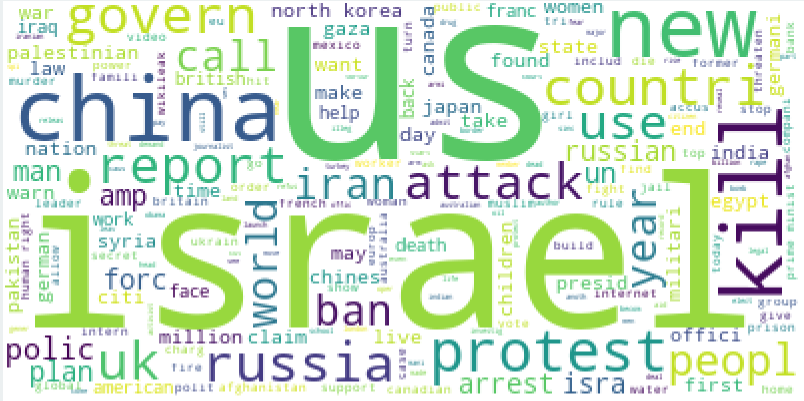

Lengths of the combined news



Word cloud of positive news ["Label" = 1]

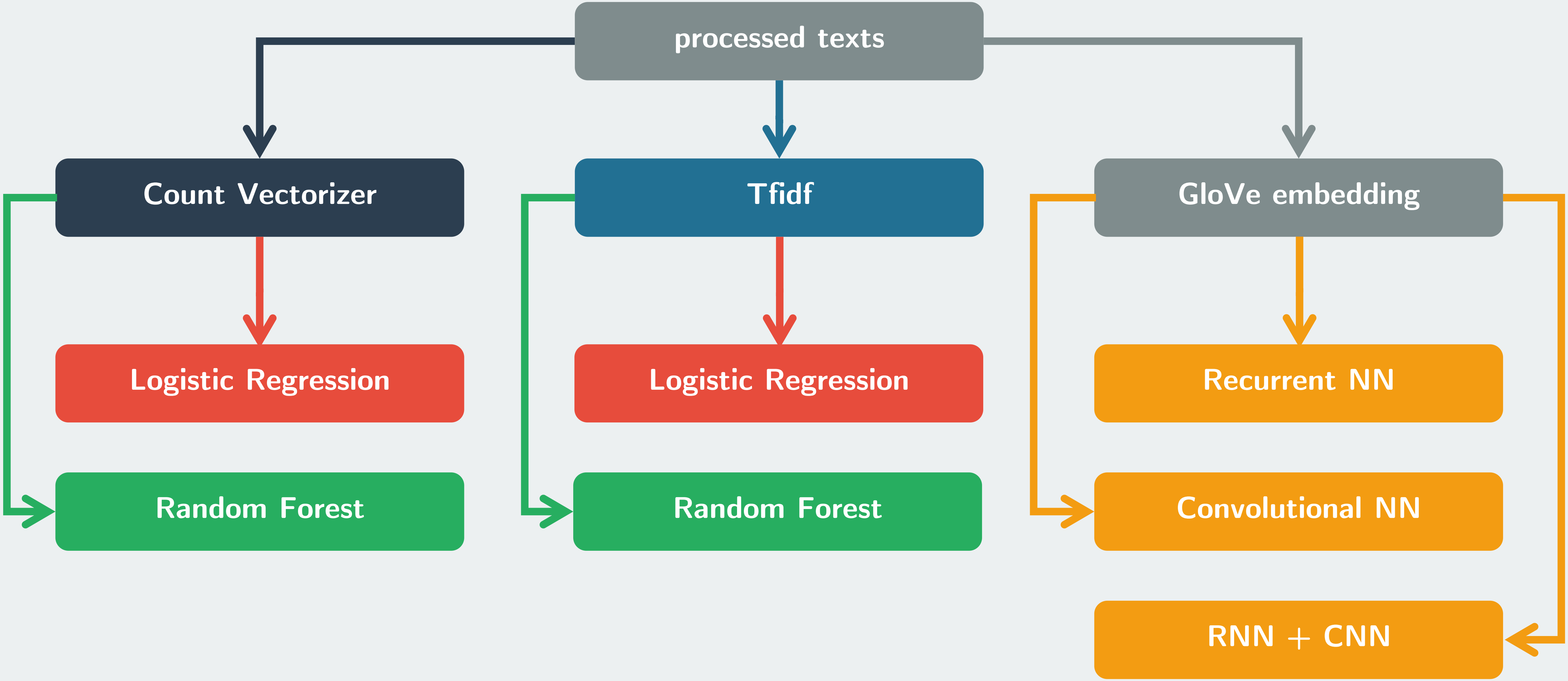


Word cloud of negative news ["Label" = 0]



Topic Modeling

```
Topic: 0
0.001*"protest" + 0.001*"egypt" + 0.001*"iran" + 0.000*"amp" + 0.000*"russia" + 0.000*"gaza" + 0.000*"russian" + 0.000*"israel" + 0.000*"canada" + 0.000*"egyptian"
Topic: 1
0.001*"israel" + 0.001*"gaza" + 0.000*"isra" + 0.000*"news" + 0.000*"korea" + 0.000*"china" + 0.000*"palestinian" + 0.000*"north" + 0.000*"nuclear" + 0.000*"bank"
Topic: 2
Words: 0.001*"gaza" + 0.001*"isra" + 0.001*"ukrain" + 0.001*"israel" + 0.000*"un" + 0.000*"amp" + 0.000*"fire" + 0.000*"nuclear" + 0.000*"protest" + 0.000*"russia"
Topic: 3
0.001*"korea" + 0.001*"north" + 0.001*"gaza" + 0.000*"palestinian" + 0.000*"iran" + 0.000*"riot" + 0.000*"isra" + 0.000*"russia" + 0.000*"amp" + 0.000*"human"
Topic: 4
0.001*"korea" + 0.000*"protest" + 0.000*"rape" + 0.000*"syria" + 0.000*"south" + 0.000*"israel" + 0.000*"iran" + 0.000*"north" + 0.000*"gaza" + 0.000*"palestinian"
Topic: 5
0.001*"wikileak" + 0.001*"isra" + 0.001*"russia" + 0.000*"israel" + 0.000*"gaza" + 0.000*"war" + 0.000*"militari" + 0.000*"drug" + 0.000*"presid" + 0.000*"nuclear"
Topic: 6
0.001*"amp" + 0.001*"korea" + 0.001*"syria" + 0.001*"north" + 0.001*"isra" + 0.000*"iran" + 0.000*"wikileak" + 0.000*"protest" + 0.000*"russia" + 0.000*"nuclear“
Topic: 7
0.001*"syria" + 0.001*"ukrain" + 0.000*"gaza" + 0.000*"oil" + 0.000*"right" + 0.000*"citi" + 0.000*"russian" + 0.000*"snowden" + 0.000*"forc" + 0.000*"amp"
Topic: 8
 0.001*"gaza" + 0.001*"protest" + 0.001*"israel" + 0.000*"isra" + 0.000*"snowden" + 0.000*"isi" + 0.000*"iran" + 0.000*"russia" + 0.000*"wikileak" + 0.000*"drug"
Topic: 9
0.001*"ukrain" + 0.001*"gaza" + 0.001*"palestinian" + 0.001*"russia" + 0.001*"wikileak" + 0.001*"israel" + 0.000*"war" + 0.000*"protest" + 0.000*"korea" + 0.000*"isra"
```



processed texts

Count Vectorizer

Tfidf

GloVe embedding

Logistic Regression

Logistic Regression

Recurrent NN

Random Forest

Random Forest

Convolutional NN

RNN + CNN

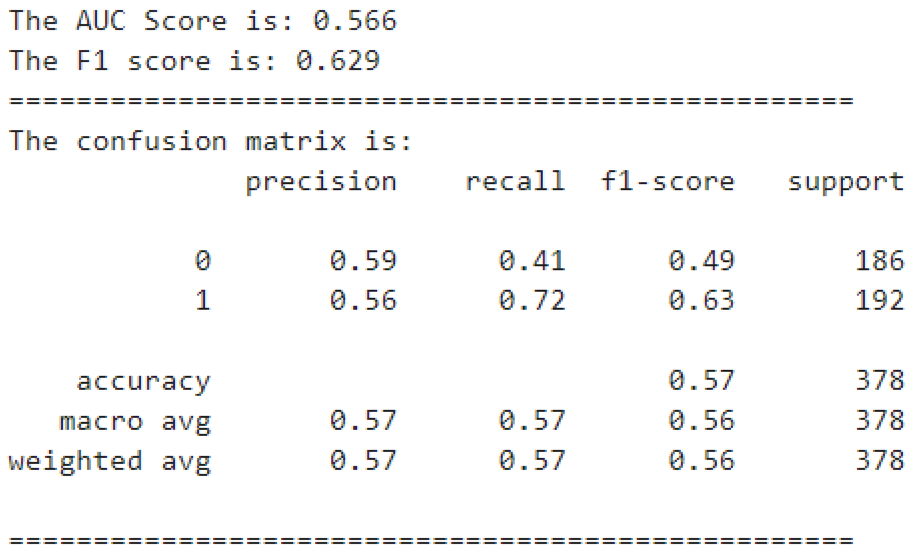
Count Vectorizer

- uni-gram
- bi-gram
- tri-gram

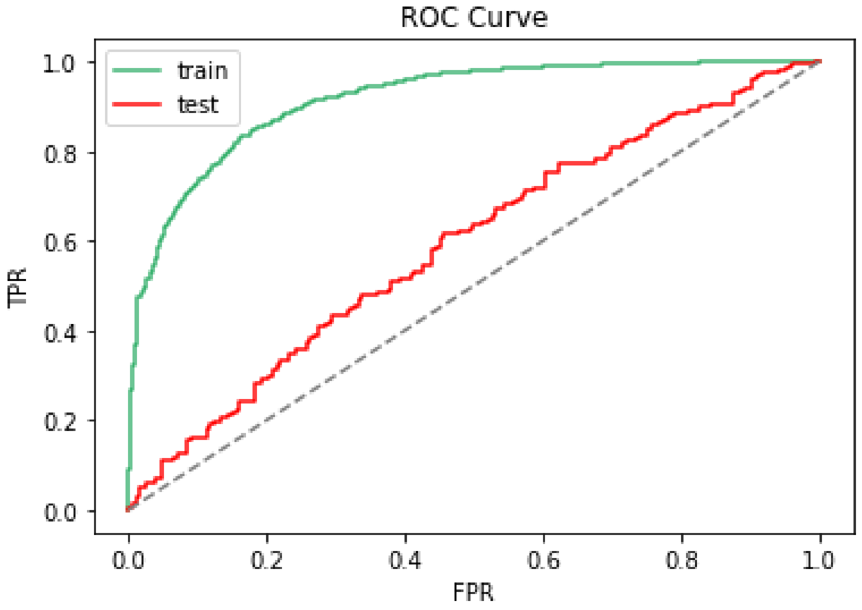
Tfidf

- uni-gram
- bi-gram
- tri-gram
- bi-gram + custom settings 🏆

Logistic Regression



```
The AUC Score is: 0.566
The F1 score is: 0.629
==================================================
The confusion matrix is:
              precision    recall  f1-score   support

           0       0.59      0.41      0.49       186
           1       0.56      0.72      0.63       192

    accuracy                           0.57       378
   macro avg       0.57      0.57      0.56       378
weighted avg       0.57      0.57      0.56       378

==================================================
```



Count Vectorizer

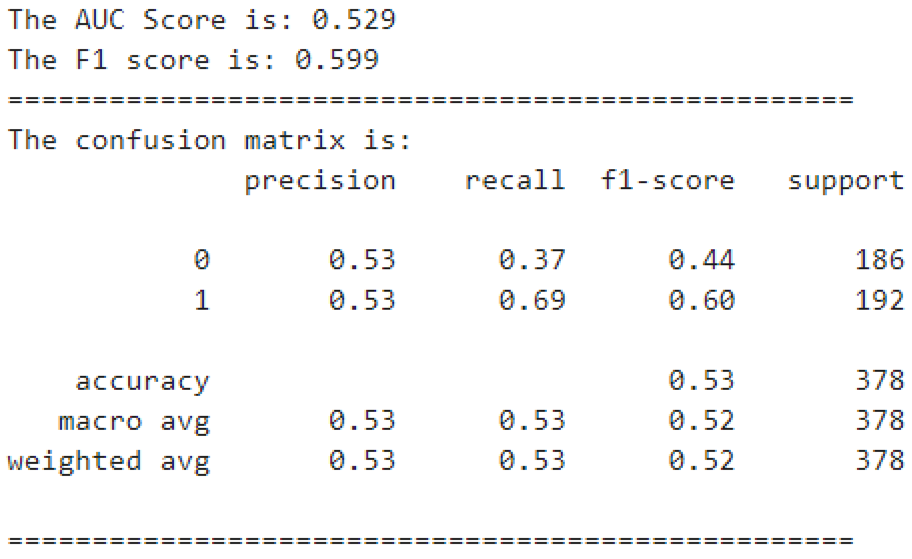
- bi-gram

Tfidf

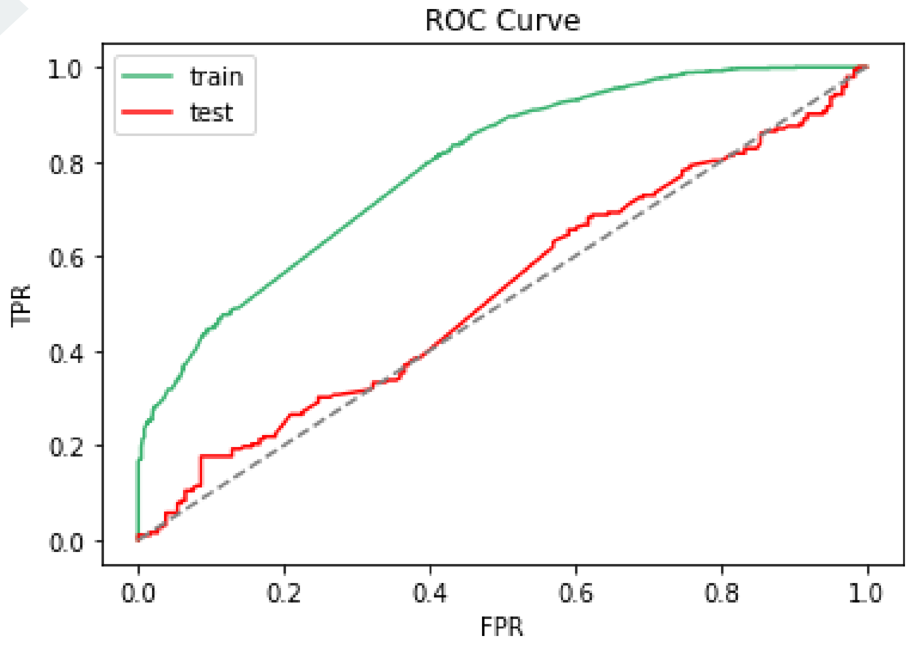
- bi-gram + custom settings

Random Forest

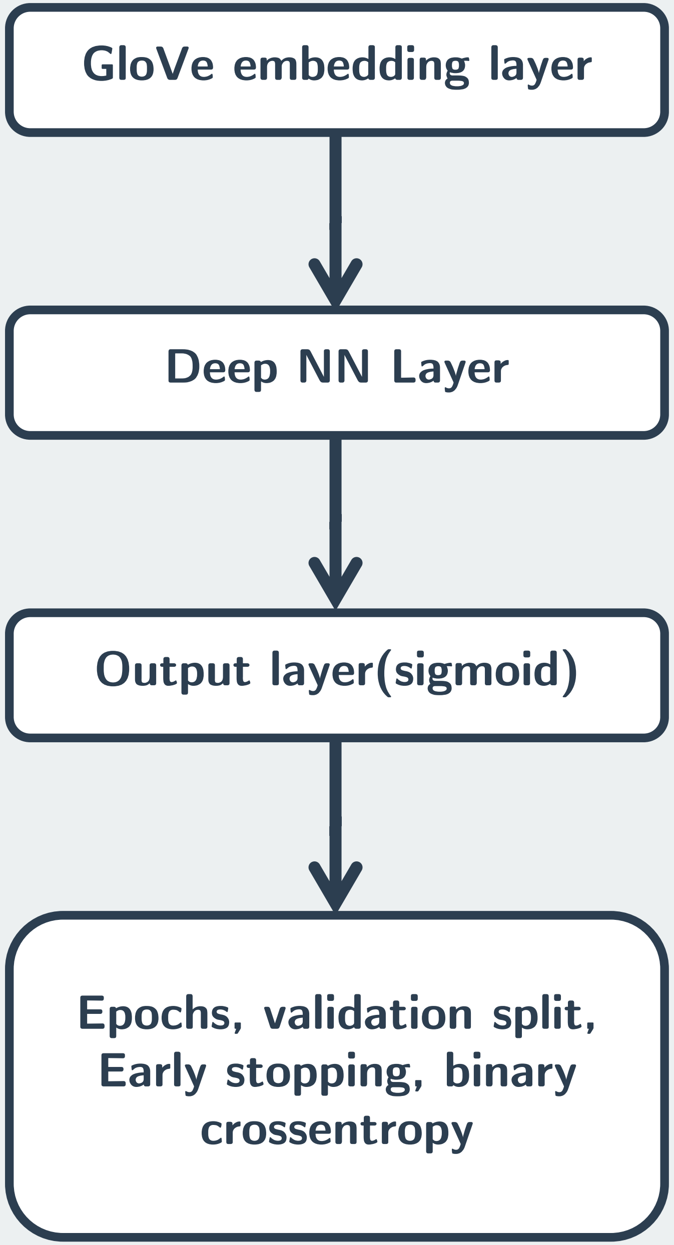
- AdaBoost 🏆
- XGBoost



```
The AUC Score is: 0.529
The F1 score is: 0.599
==================================================
The confusion matrix is:
              precision    recall  f1-score   support

           0       0.53      0.37      0.44       186
           1       0.53      0.69      0.60       192

    accuracy                           0.53       378
   macro avg       0.53      0.53      0.52       378
weighted avg       0.53      0.53      0.52       378

==================================================
```



Deep Learning



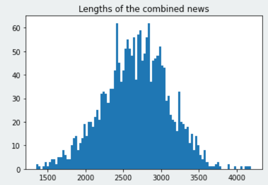
GloVe embedding layer

Deep NN Layer

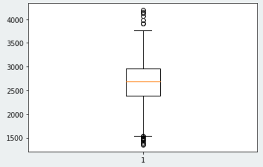
Output layer(sigmoid)

Epochs, validation split, Early stopping, binary crossentropy

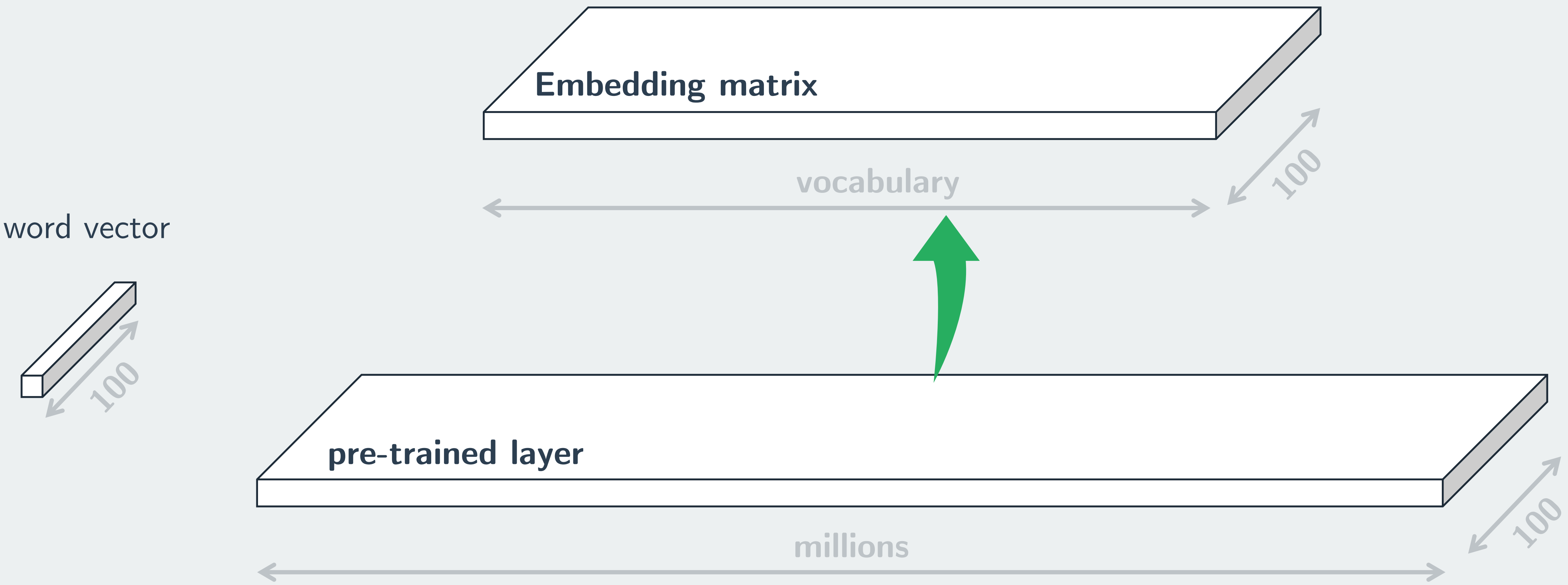
GloVe embedding layer



Lengths of the combined news



- Maximum length of each combined headline: 2500
- Dimension of word vector: 100



Embedding matrix

vocabulary

100

word vector

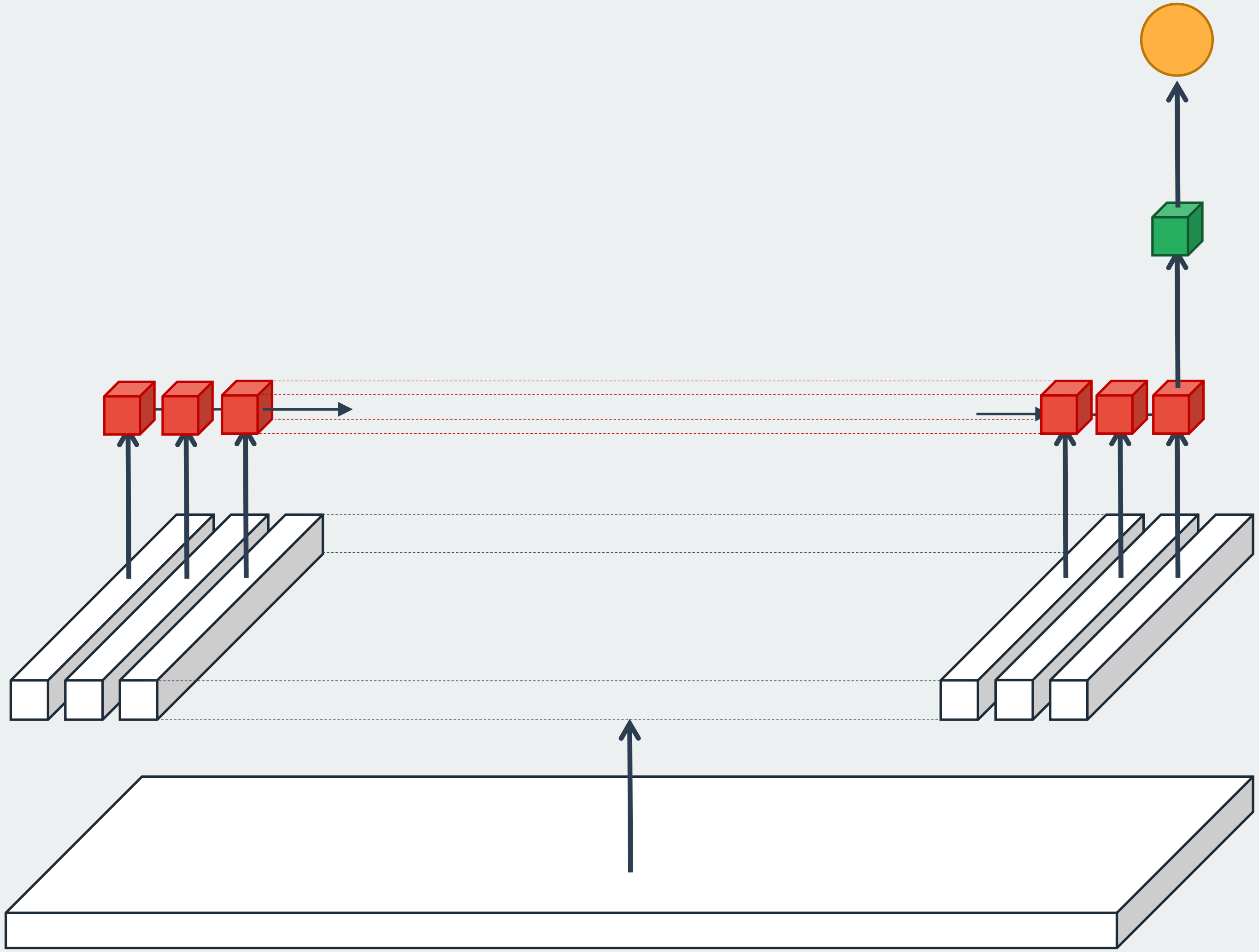
100

pre-trained layer

millions

100

RNN



Output

Dense Network

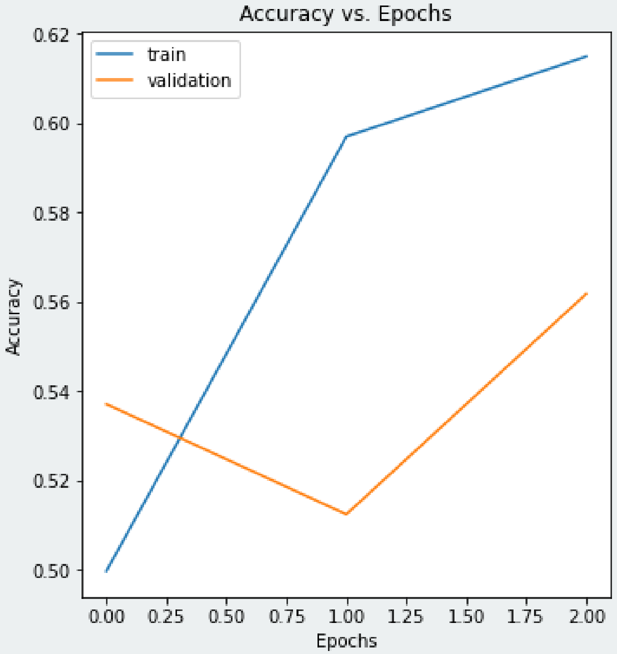
LSTM

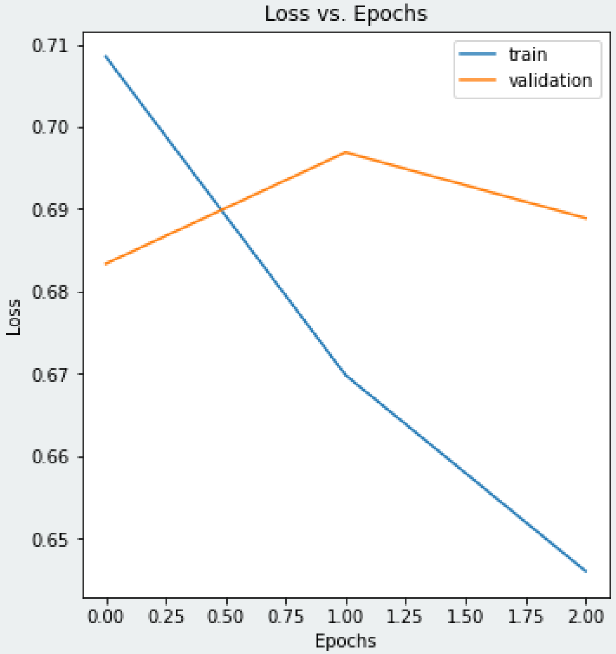
Embedded layer

Embedded matrix

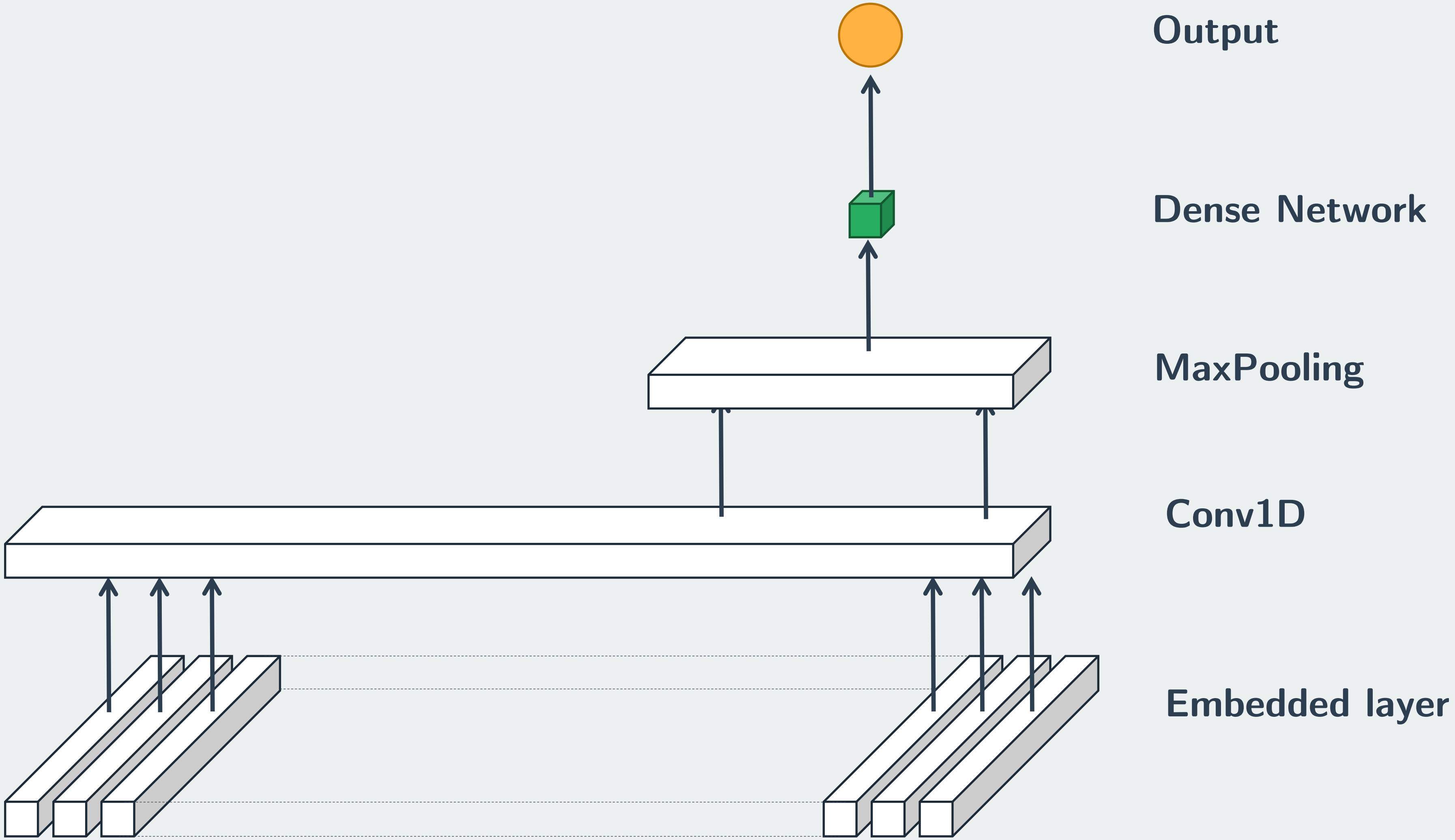
RNN

Model	Variants	Accuracy score
LSTM (RNN)	Default	0.5159
LSTM (RNN)	learning rate = 0.005	**0.5238**
LSTM (RNN)	dropout = 0.2	0.5132
LSTM (RNN)	dropout = 0.2, reccurent dropout = 0.2	0.4815
LSTM (RNN)	2 LSTM Layers	0.5132
LSTM (RNN)	optimizer = Nadam	0.4894
GRU (RNN)	Default	0.5026
GRU (RNN)	optimizer = Nadam	0.4894



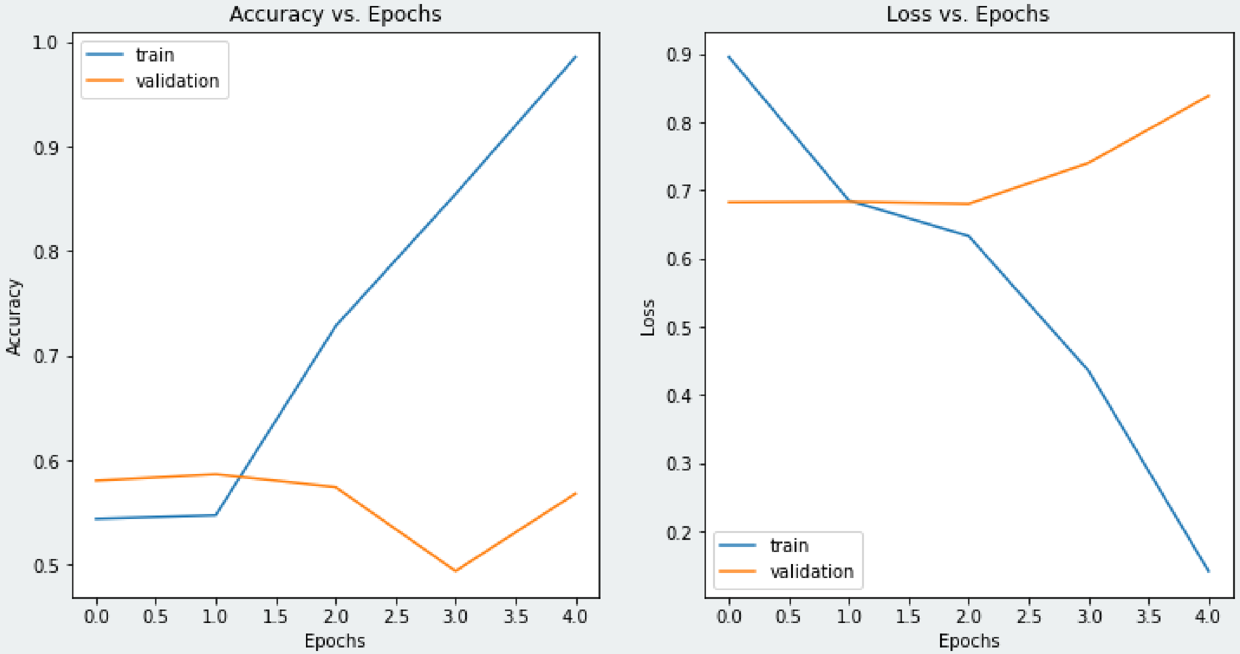


CNN

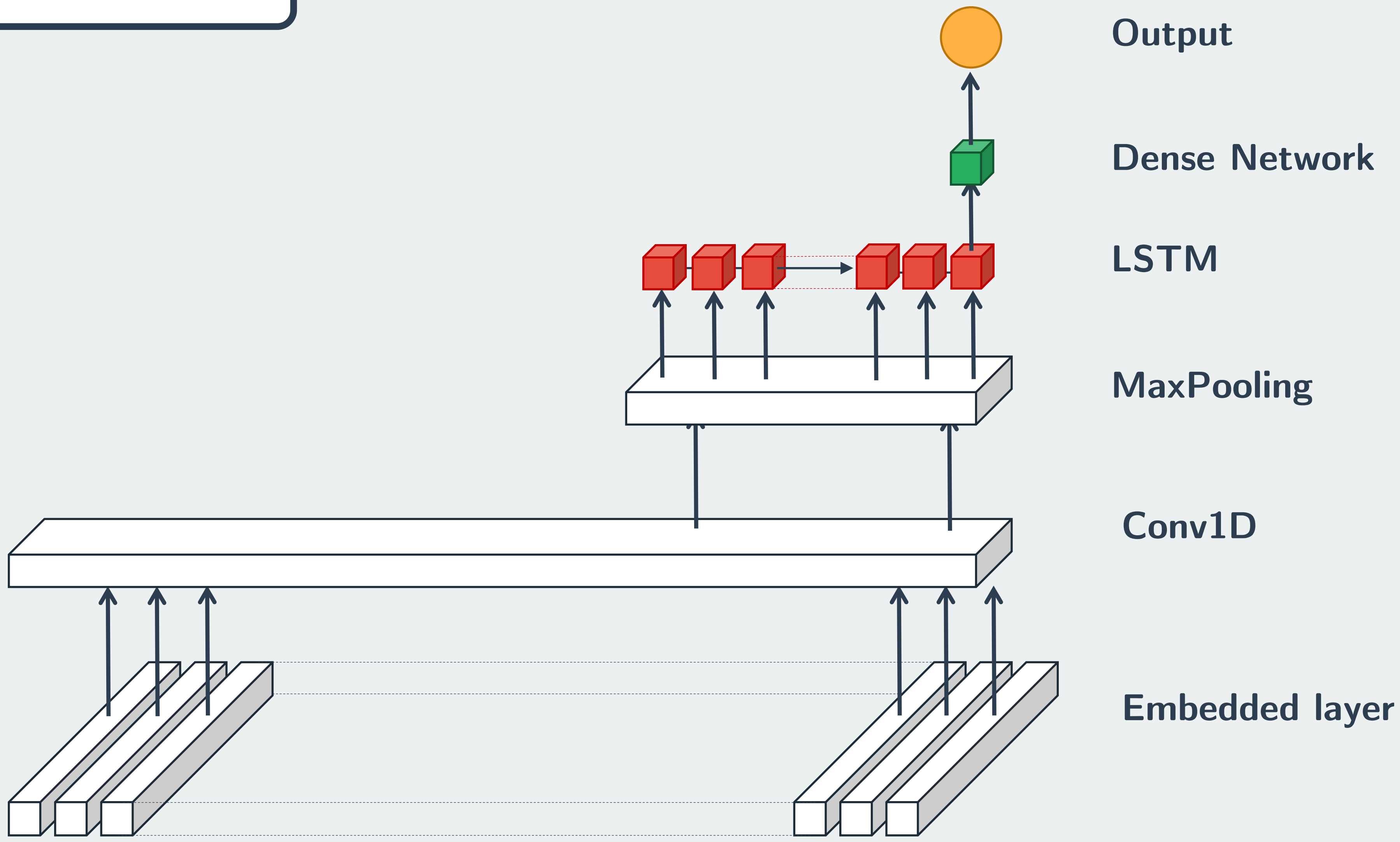


CNN

Model	Variants	Accuracy score
CNN	filters=32, kernel_size=5	**0.5079**
CNN	filters=64, kernel_size=5	0.4603
CNN	filters=100, kernel_size=5	0.5
CNN	filters=32, kernel_size=3	0.5026
CNN	filters=32, kernel_size=2	0.4841



CNN + RNN



Output

Dense Network

LSTM

MaxPooling

Conv1D

Embedded layer

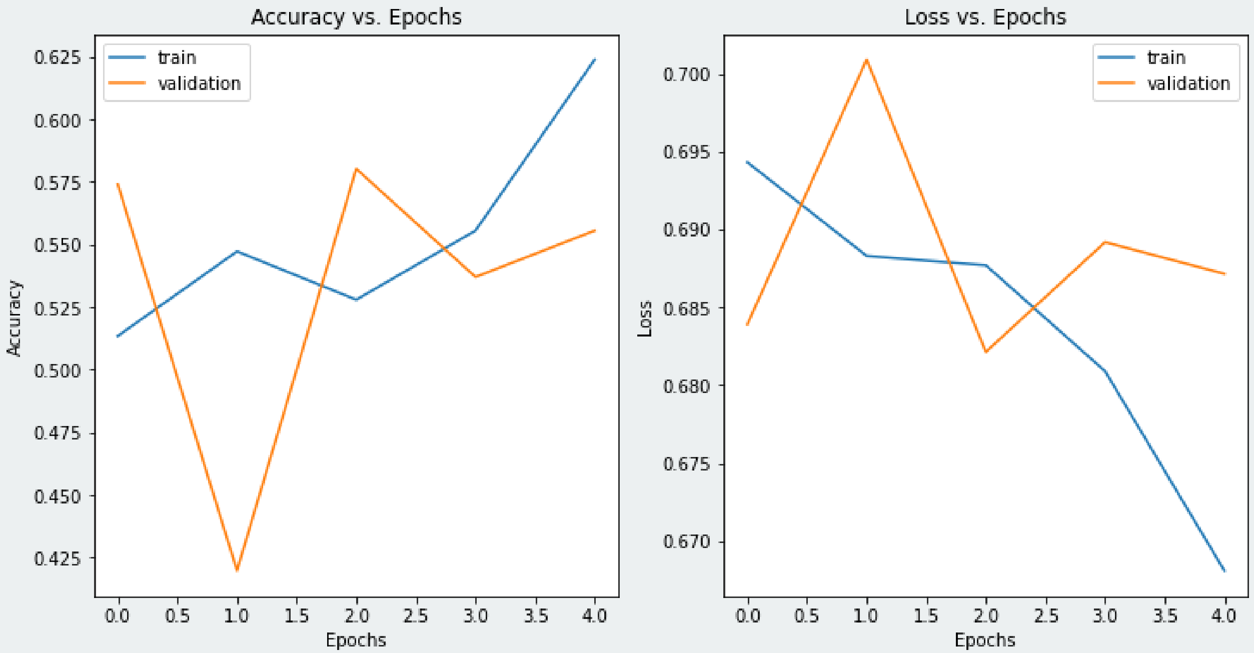
CNN + RNN

Model	Variants	Accuracy score
CNN + LSTM	filters=32, kernel_size=5	0.5079
CNN + GRU	filters=32, kernel_size=5	0.5106

CNN + GRU	filters=32, kernel_size=5,
	learning_rate = 0.005
		0.5

CNN + GRU	filters=10, kernel_size=5,
	learning_rate = 0.005
		0.5265

CNN + GRU	filters=10, kernel_size=5,
	learning_rate = 0.007
		0.4868



thank you