NLP for Finance

Cris Chai, Dhruva Iyer, Dristi Roy, Sean Wu, Legolas Zhang, Satvik Marthi

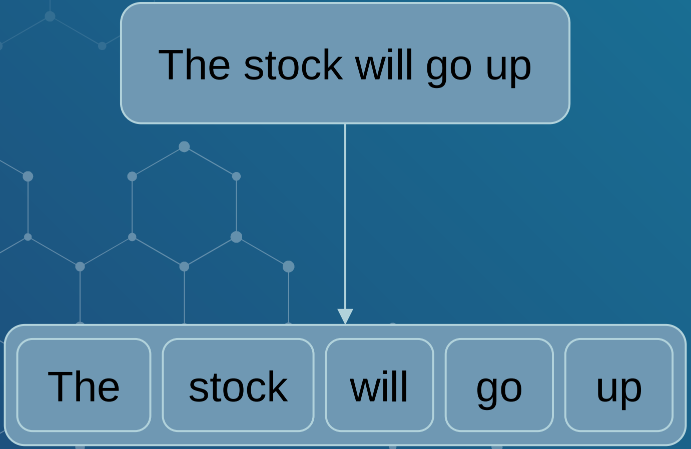
We used NLP to predict sentiments of stock market tweets, involving preprocessing and inputting them to train classification models. Sentiments can help predict changes in stock prices.

NLP pre-processing

Simplify the data

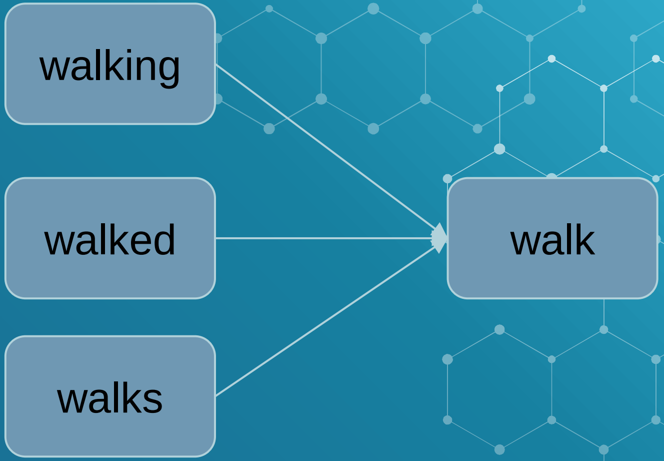
Tokenization

Split the sentence

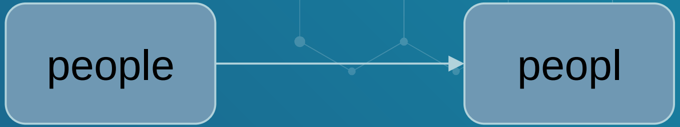


Stemming

Remove prefixes and suffixes



Some words can be incorrectly stemmed



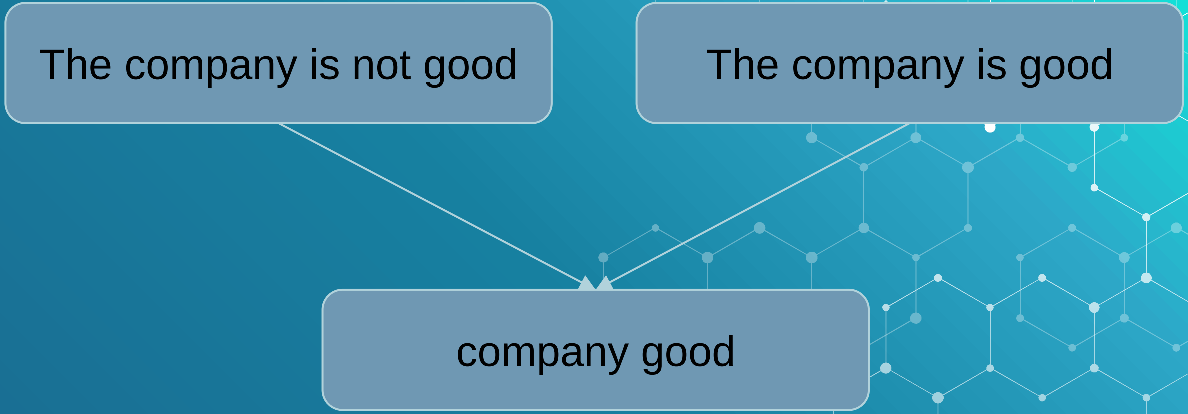
Text Cleaning

- Text cleaning is the process of simplifying a sentence
- One way is by replacing characters such as , . ; “ ‘ with spaces
- Sentences will have the same meaning

Unsimplified Sentence: I was surprised when the stock price increased.

Simplified Sentence: The stock price increased.

Stopword removal

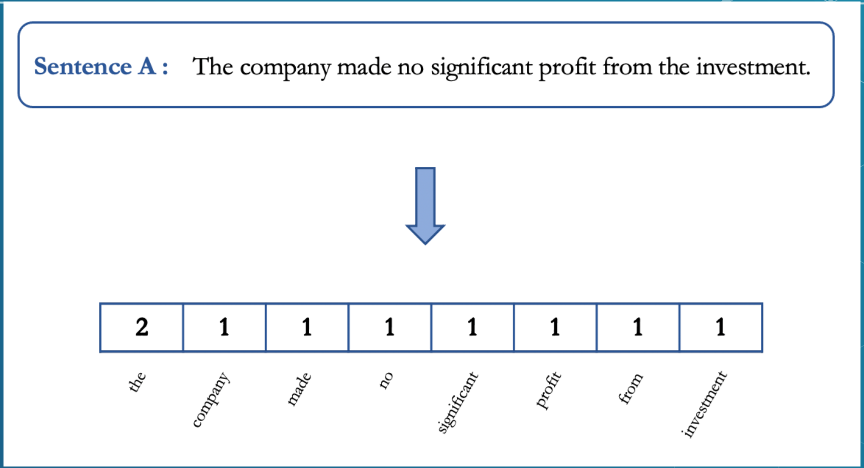


Remove stopwords like

- there
- is
- the
- and

Vectorization

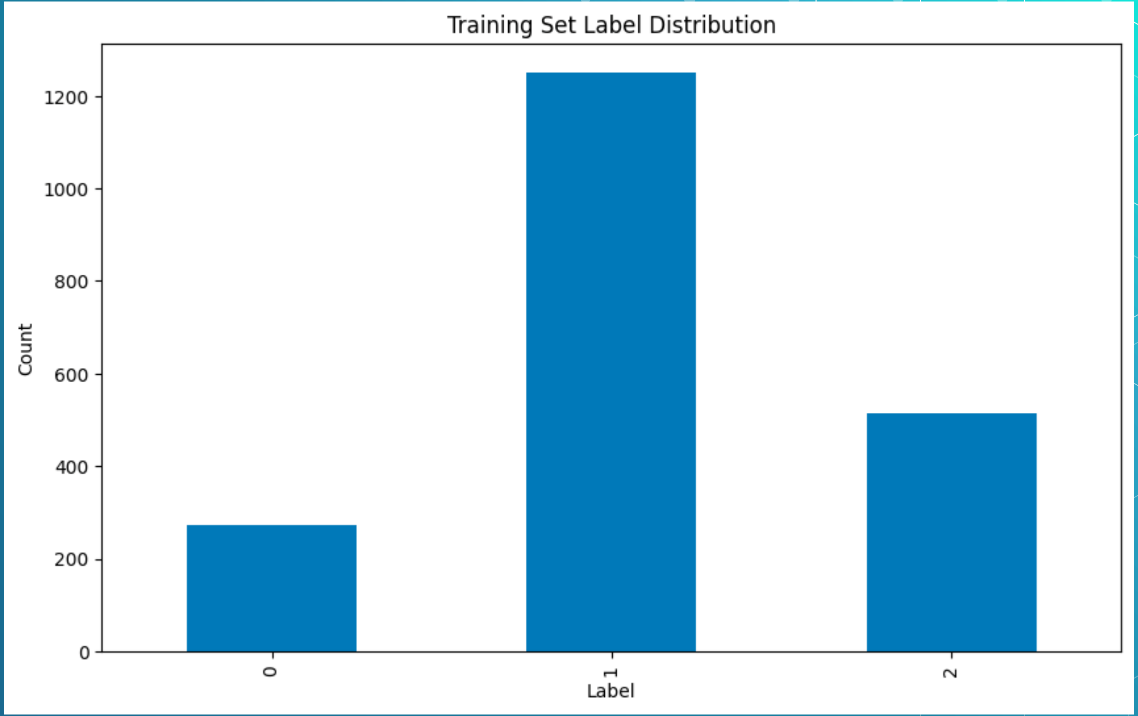
Converts a list of words into a vector of numbers


Sentence A : The company made no significant profit from the investment.
2
1
1
1
1
1
1
1
the
company
made
no
significant
profit
from
investment

Data set balancing

- Data is not balanced
- Lots of neutral data

We did not address this, but we should if we do this again.



0 is negative

1 is neutral

2 is positive

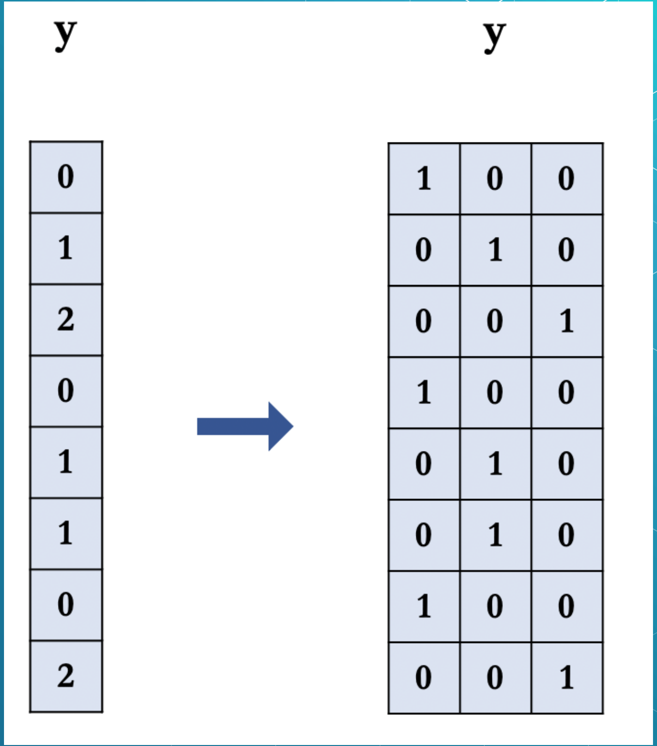
One Hot Encoding

One hot encoding is converting categorical data into numerical data

- 1 column per possible label
- Easy to interpret model's output

Before:
0: negative, 1: neutral, 2: positive

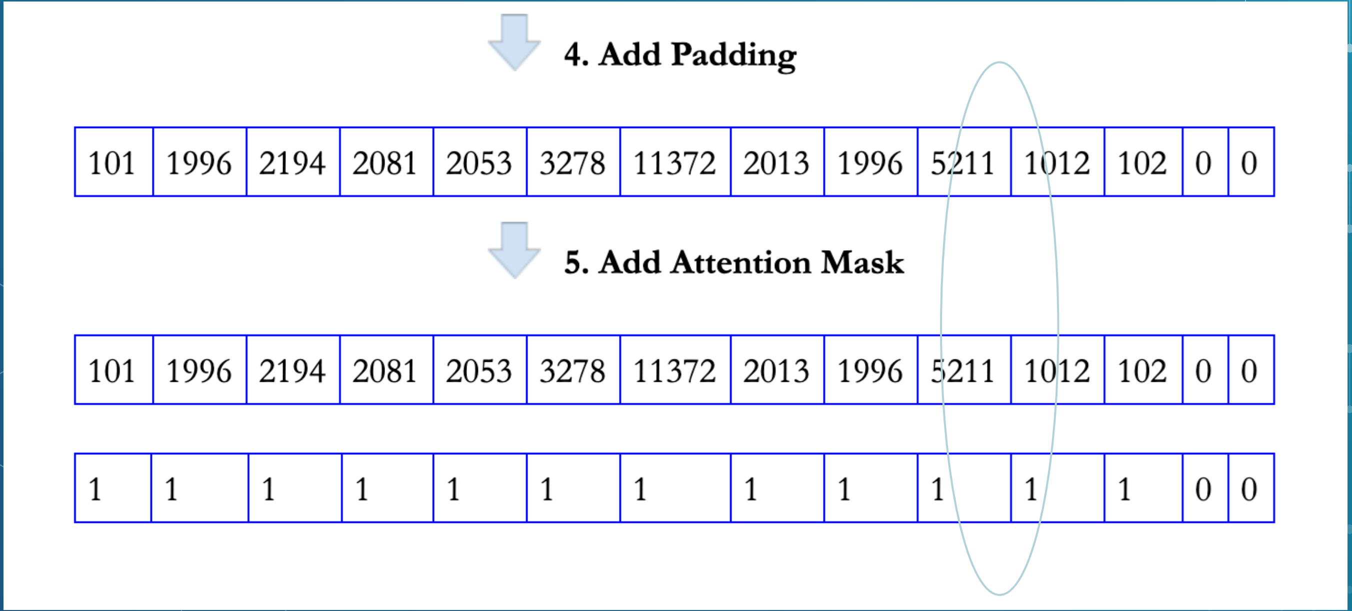
After:
Col 1: negative, Col 2: neutral, Col 3: positive



$$\begin{array}{c} \mathbf{y} \\ \hline 0 \\ 1 \\ 2 \\ 0 \\ 1 \\ 1 \\ 0 \\ 2 \end{array} \longrightarrow \begin{array}{ccc} & \mathbf{y} & \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{array}$$

Padding and Attention Masks

- Extends vectors with padding to have equal length inputs
- Attention mask draws model attention away from padding



4. Add Padding

101	1996	2194	2081	2053	3278	11372	2013	1996	5211	1012	102	0	0

5. Add Attention Mask

101	1996	2194	2081	2053	3278	11372	2013	1996	5211	1012	102	0	0
1	1	1	1	1	1	1	1	1	1	1	1	0	0

Word Order

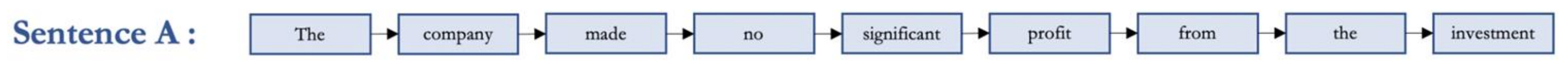
- Changes the meaning of a sentence
- Very important

Sentence A : The company made no significant profit from the investment.

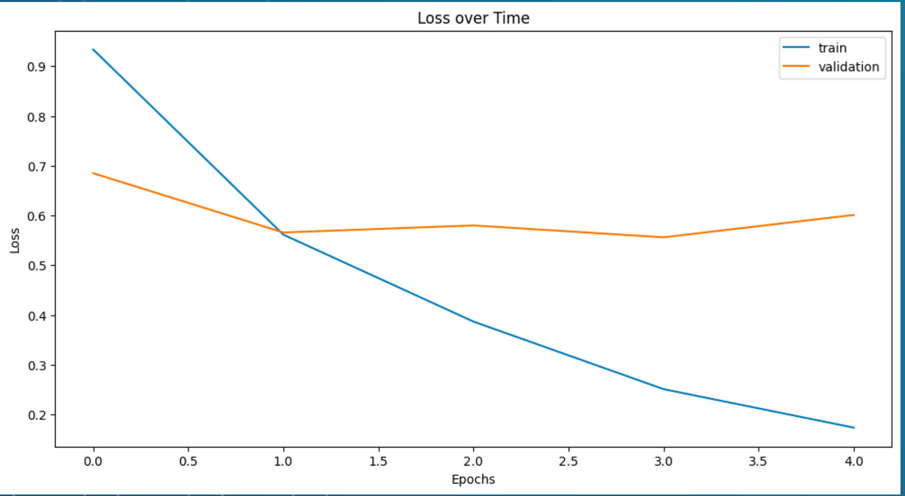
Sentence B : No, the company made significant profit from the investment.

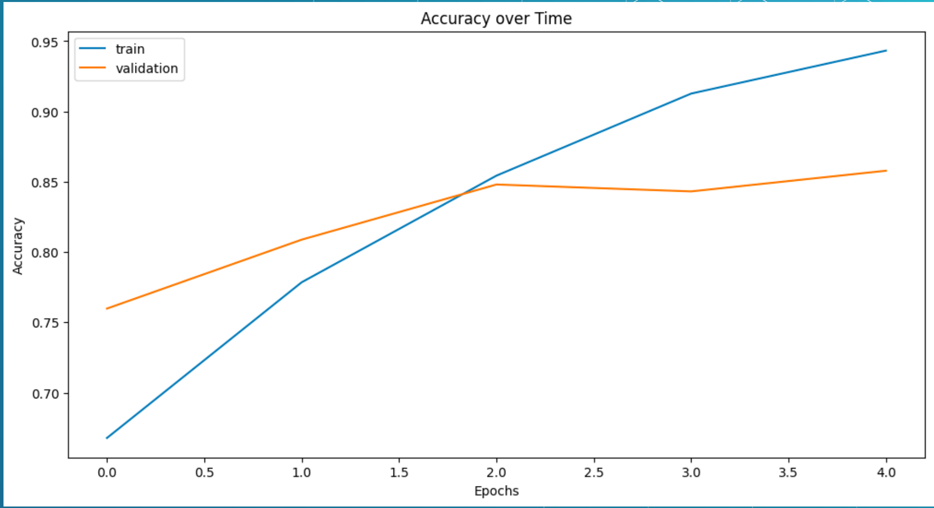
LSTM

- LSTM stands for long short-term memory
- It can help to solve word order issues
- A sequence of words rather than bag



LSTM Result





83.70% Accuracy

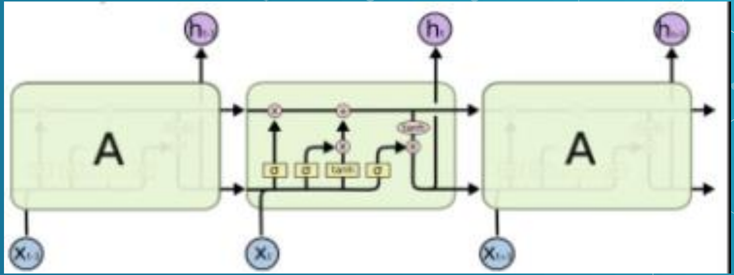
Model Architecture and Hyperparameters

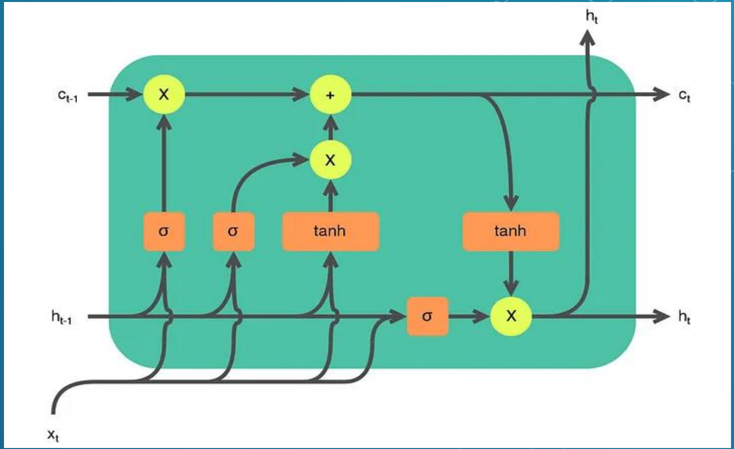
Model layers

- Embedding
- Dropout
- LSTM
- Dense

Hyperparameters

- Epochs
- Max Sequence Length
- Max NB Words





BERT Pre-trained Model

BERT is a pre-trained model from Google

- Very accurate
- Needs less effort than new model
- Speeds up training
- Needs less data

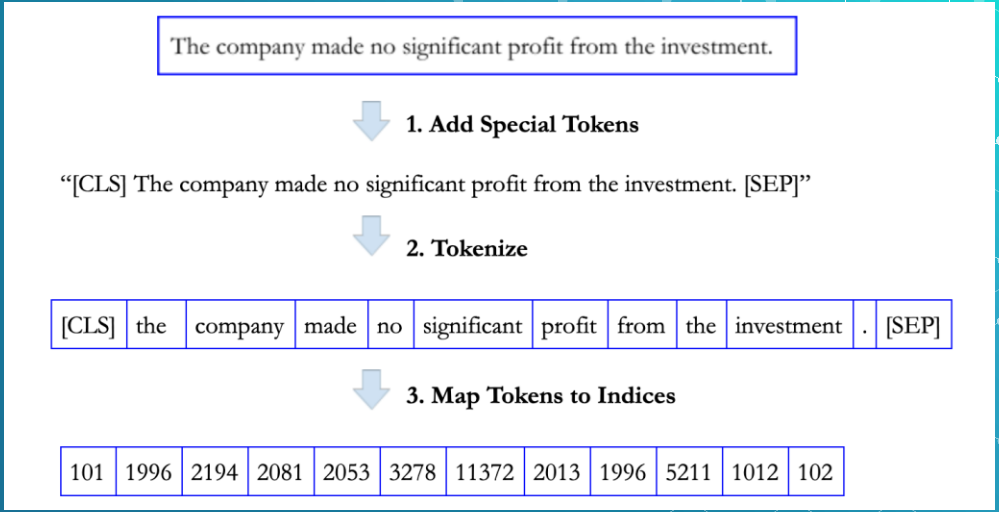
Setting Up BERT

Pre-processing steps

- Add special tokens
- Tokenize
- Convert tokens to indexes
- Add padding and attention mask

Using the model

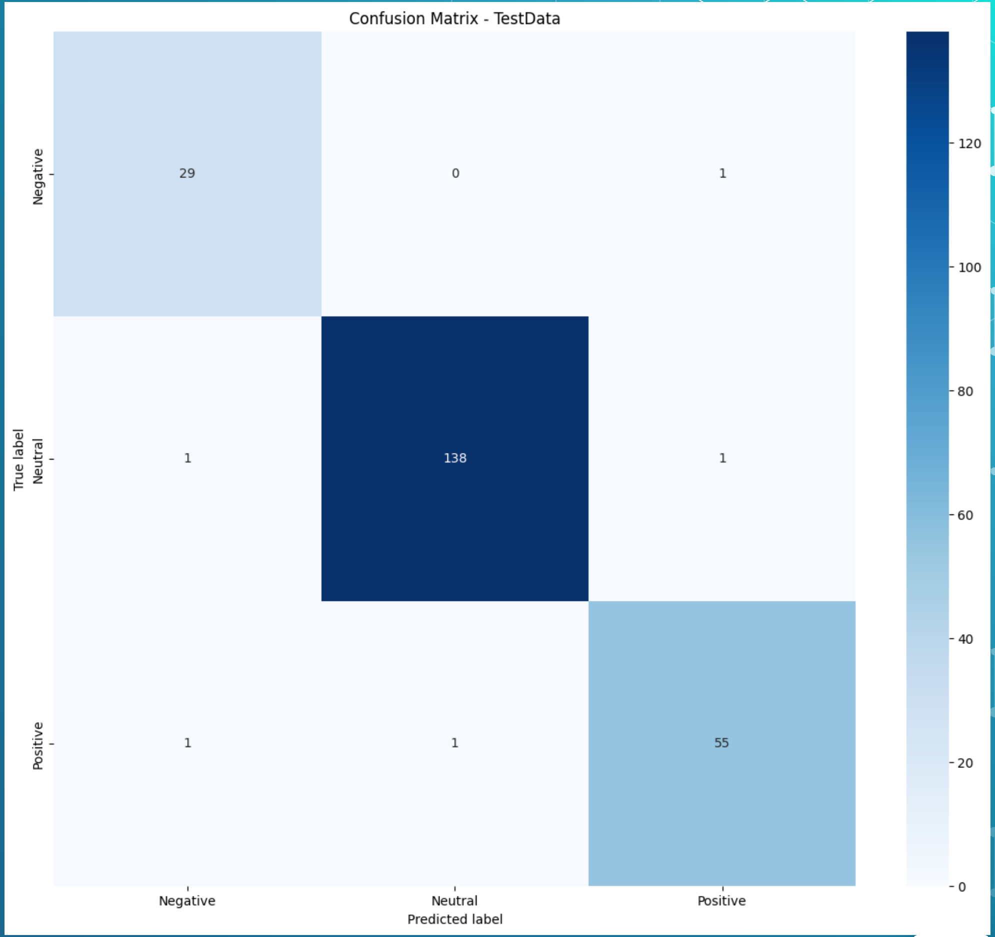
- Huggingface Pytorch has a BERT implementation
- Add an output layer to BERT
- Train like any other model


The company made no significant profit from the investment.
1. Add Special Tokens
"[CLS] The company made no significant profit from the investment. [SEP]"
2. Tokenize
[CLS] | the | company | made | no | significant | profit | from | the | investment | . | [SEP]
3. Map Tokens to Indices
101 | 1996 | 2194 | 2081 | 2053 | 3278 | 11372 | 2013 | 1996 | 5211 | 1012 | 102

BERT Results

Highly accurate in testing with no hyperparameter optimization

Only 5/227 wrong answers (98% accuracy)



Thank You