Institute of Data

2022

Data Science and AI

Sentiment analysis on financial news headlines

Capstone Project

Suchada (Amee) Sudlert

Agenda

- Bio
- Project Context
- Define
- Design
- Deliver
- Summary, conclusions and next steps
- Appendix: list of supporting documents

Bio

2 years' experience in software development with in-depth of software designing and testing for the railway industry. Bachelor's degree of Control system engineering and currently studying a RMIT industry accredited graduate certificate in Data Science and AI. Joined an internship at Curtin University, Australia, to make an elderly’s assistant robot. Creating algorithms for trading based on Machine Learning.

Project context

- **Business aspects**
  - Industry or domain: Finance / trading sectors
  - Problem area: sentiment analysis for financial news
  - Why is this area interesting: fundamental analysis, helps classify good or bad news

Define

- **Business aspects**
  - Stakeholders: AI researchers and investors
  - Business question: how well we can classify positive and negative sentiments on the dataset that contains news headlines
  - Business value: Positive news will normally cause individuals to buy stocks.
- **Technical perspective**
  - Techniques used: text pre-processing and Count Vectorization
  - Pipeline: NLP, Machine Learning and Deep Learning
  - Model validation results: Accuracy score and Confusion Metrix

Dataset

1. ['negative', 'The international electronic industry company Elcoteq has laid off tens of employees from its Tallinn facility ; contrary to earlier layoffs the company contracted the ranks of its office workers , the daily Postimees reported .'],
2. ['neutral', 'Technopolis plans to develop in stages an area of no less than 100,000 square meters in order to host companies working in computer technologies and telecommunications , the statement said .']
3. ['positive', 'With the new production plant the company would increase its capacity to meet the expected increase in demand and would improve the use of raw materials and therefore increase the production profitability .'],

Pre-processing text

1. **Lower casing:** Converting a word to lower case (**NLP -> nlp)**.
   Words like ***Book*** and ***book*** mean the same but when not converted to the lower case those two are represented as two different words in the vector space model (resulting in more dimensions).
2. **Tokenization:** Break the raw text into small chunks. Tokenization breaks the raw text into words.
3. **Stop words removal:** Stop words are very commonly used words (a, an, the, etc.) in the documents. These words do not really signify any importance as they do not help in distinguishing two documents.
4. **Lemmatization**: Unlike stemming, lemmatization reduces the words to a word existing in the language.

Pre-processing text

- **Original text/news headline**
  - 'compared with the ftse index which rose points or on the day this was a relative price change of '

- **After done some text cleaning steps**
  - compare ftse index rise point day relative price change

Count Vectorizer

- The text is transformed to a sparse matrix

```
text = ['Hello my name is james, this is my python notebook']
```

The text is transformed to a sparse matrix as shown below.

	hello	is	james	my	name	notebook	python	this
0	1	2	1	2	1	1	1	1

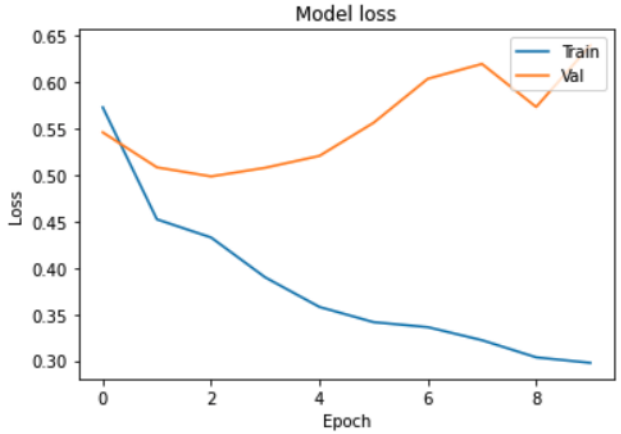
Source : https://towardsdatascience.com/basics-of-countvectorizer-e26677900f9c

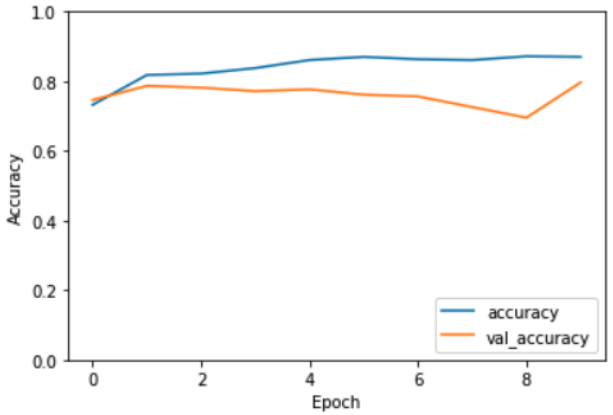
Feature Engineering

- **SelectKbest**
  - The premise with **SelectKBest is combining the univariate statistical test with selecting the K-number of features based on the statistical result** between the X and y.

Feature Engineering

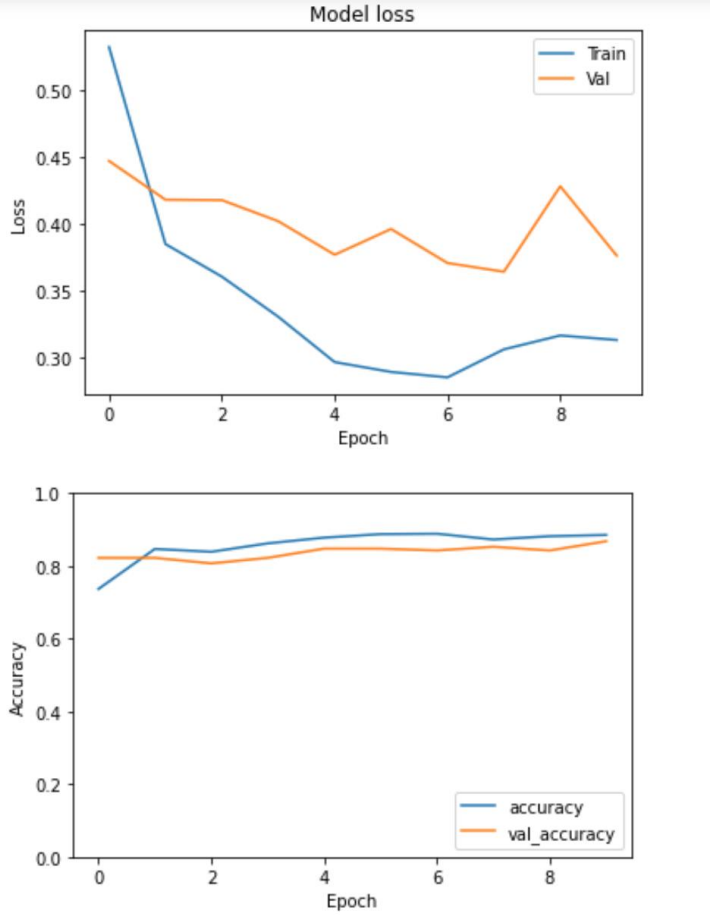
- Neural Network **without** feature selection





Feature Engineering

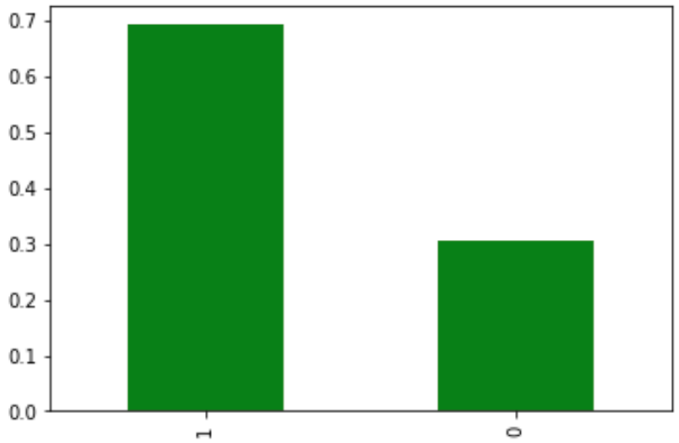
- Neural Network **wi**th feature selection



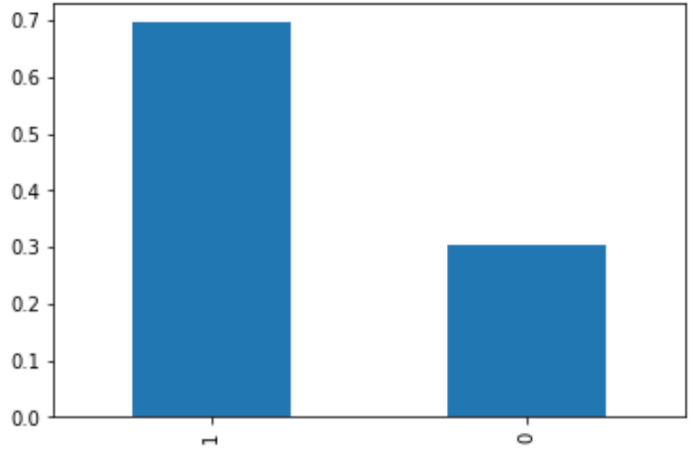
Split Data

- Split data into training and testing set

```
1    1226
0     544
Name: Sentiment, dtype: int64

<AxesSubplot:>
```

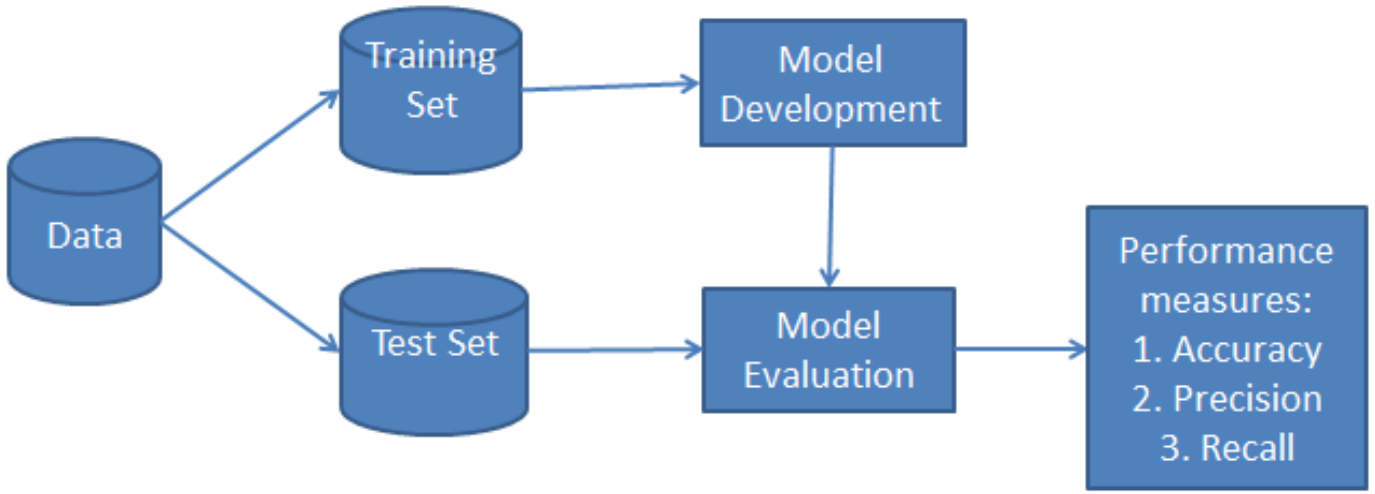



```
1    137
0     60
Name: Sentiment, dtype: int64

: <AxesSubplot:>
```



Model

- **1st Model: Multinomial Naïve Bayes (**single model)

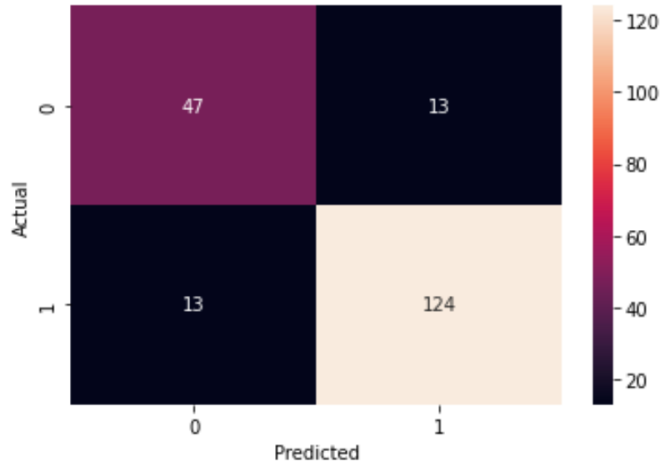


Source: https://www.datacamp.com/tutorial/naive-bayes-scikit-learn

Model Evaluation

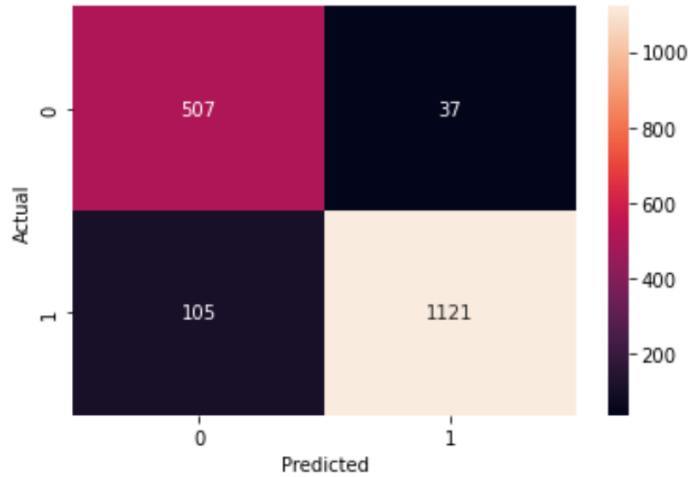
- **1st Model: Multinomial Naïve Bayes (**single model)

```
Report :
                precision    recall  f1-score   support

           0       0.78      0.78      0.78        60
           1       0.91      0.91      0.91       137

    accuracy                           0.87       197
   macro avg       0.84      0.84      0.84       197
weighted avg       0.87      0.87      0.87       197

Accuracy score  :  0.868020304568528
```



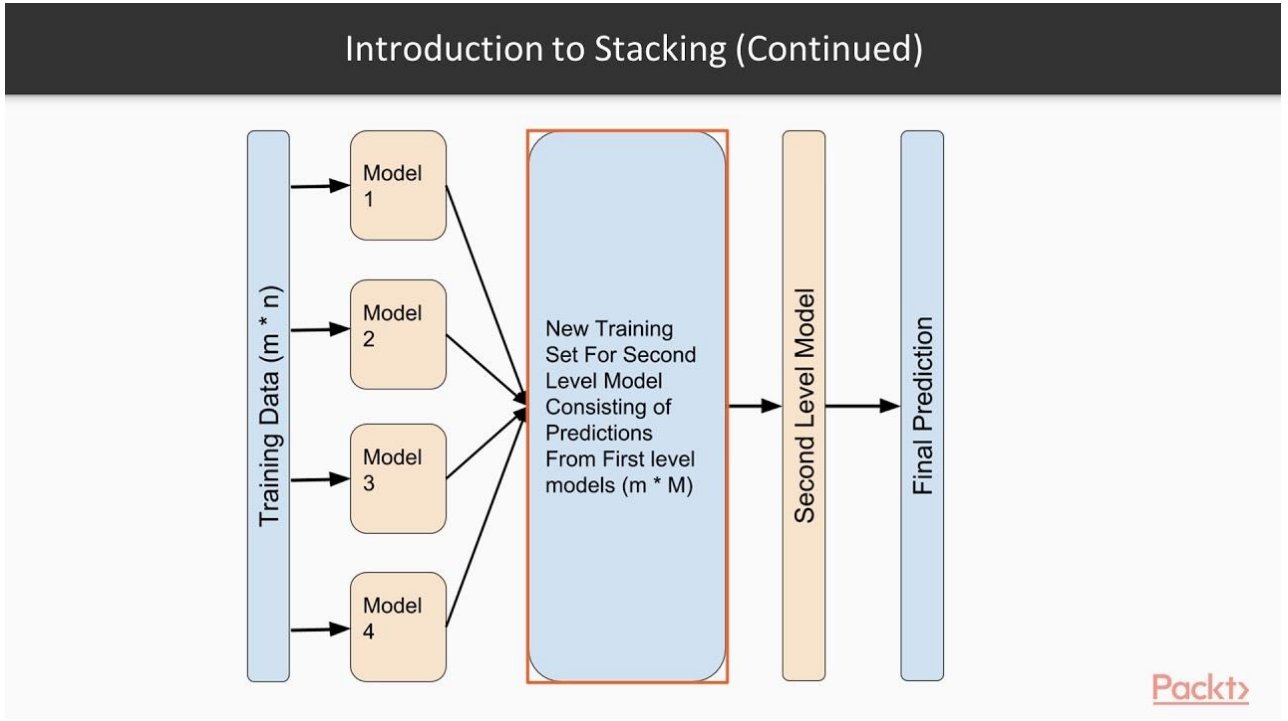
Model Evaluation

```
mean_score of cross validation =  0.7814339778613038
Report :
                precision    recall  f1-score   support

           0       0.83      0.93      0.88       544
           1       0.97      0.91      0.94      1226

    accuracy                           0.92      1770
   macro avg       0.90      0.92      0.91      1770
weighted avg       0.93      0.92      0.92      1770

Accuracy score  :  0.919774011299435
```



Model

- **2nd Model: Stacking Model (Ensemble)**



Source: https://www.youtube.com/watch?v=DCrcoh7cMHU

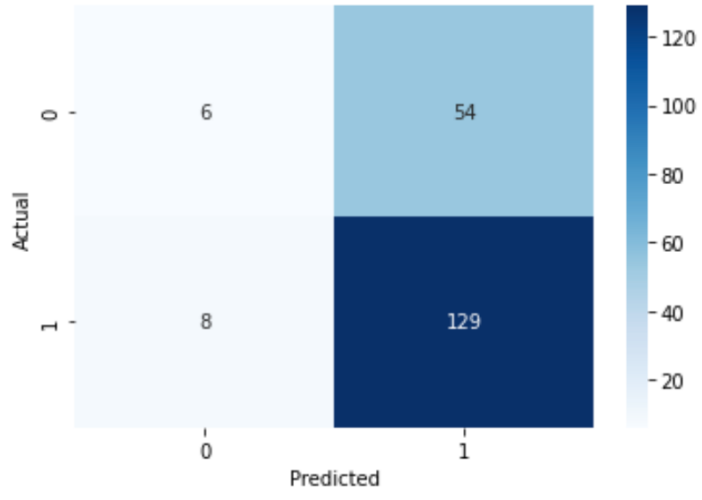
Model Evaluation

- **2nd Model: Stacking Model (Ensemble)**

Report :

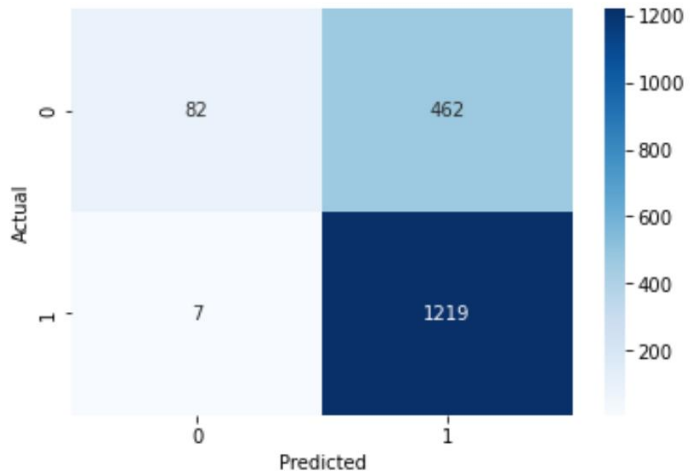
	precision	recall	f1-score	support
0	0.43	0.10	0.16	60
1	0.70	0.94	0.81	137
accuracy			0.69	197
macro avg	0.57	0.52	0.48	197
weighted avg	0.62	0.69	0.61	197

Accuracy score : 0.6852791878172588



Model Evaluation

```
Report :
              precision    recall  f1-score   support

           0       0.92      0.15      0.26       544
           1       0.73      0.99      0.84      1226

    accuracy                           0.74      1770
   macro avg       0.82      0.57      0.55      1770
weighted avg       0.79      0.74      0.66      1770

Accuracy score  :  0.7350282485875707
```



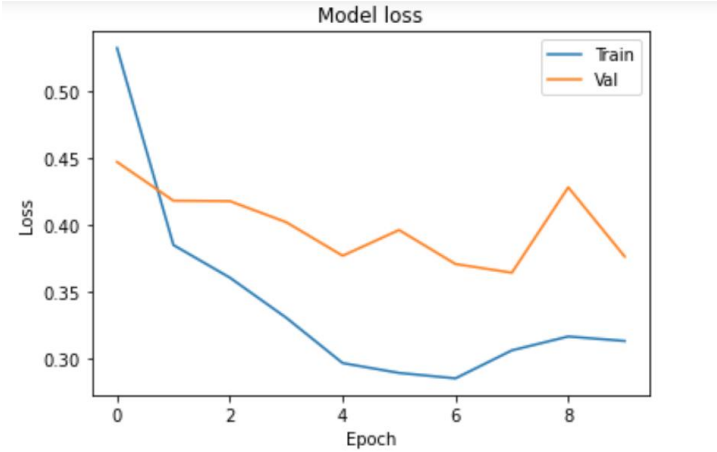
Model

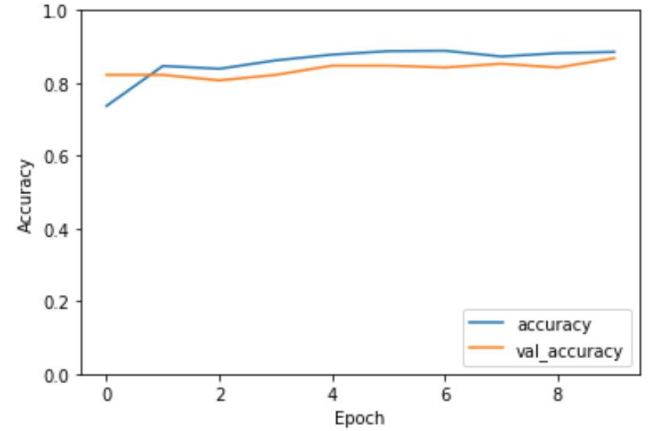
- **3rd Model : Neural Network model (Deep learning)**

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 300)	900300
dense_2 (Dense)	(None, 30)	9030
dense_3 (Dense)	(None, 4)	124
dense_output (Dense)	(None, 1)	5

Model Evaluation

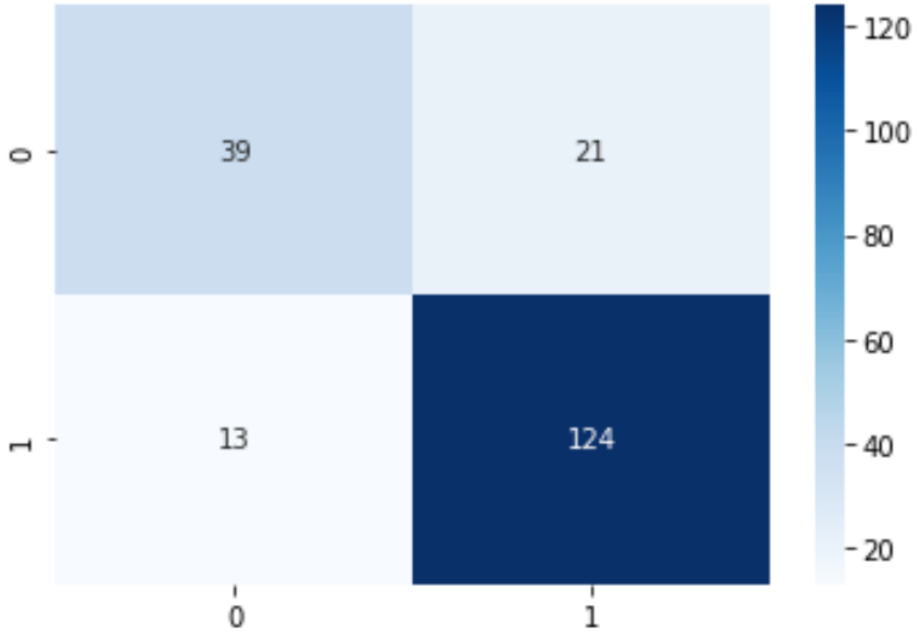
- **3rd Model : Neural Network model (Deep learning)**





Model Evaluation

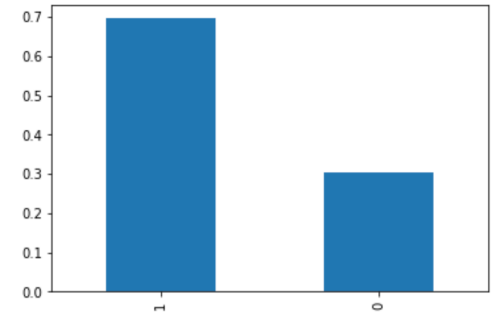
- **3rd Model : Neural Network model (Deep learning)**

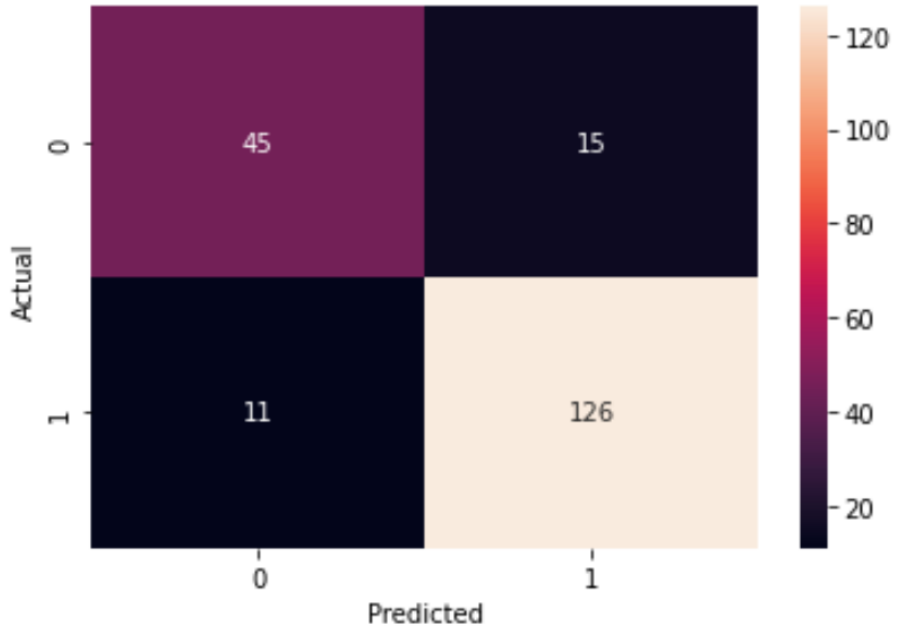


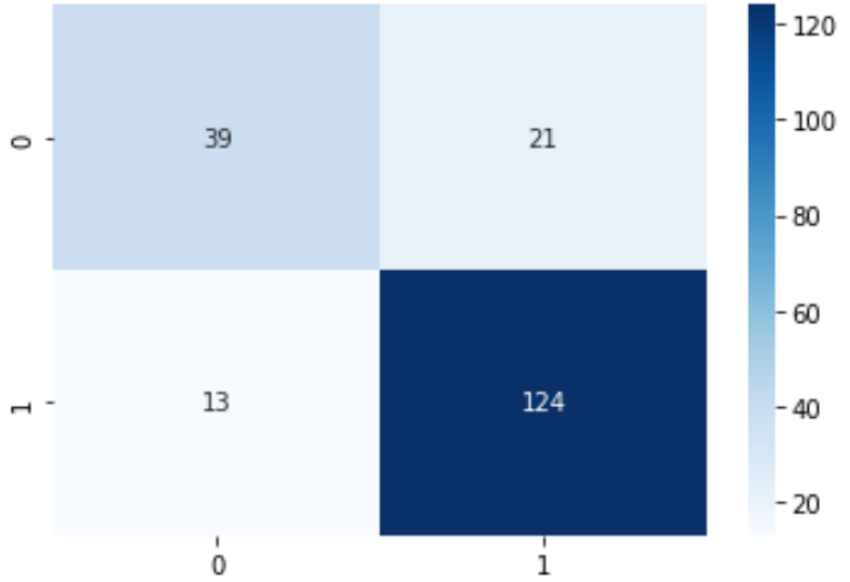
Conclusions

- MultinomialNB VS Neural Network

```
1    137
0     60
Name: Sentiment, dtype: int64

: <AxesSubplot:>
```

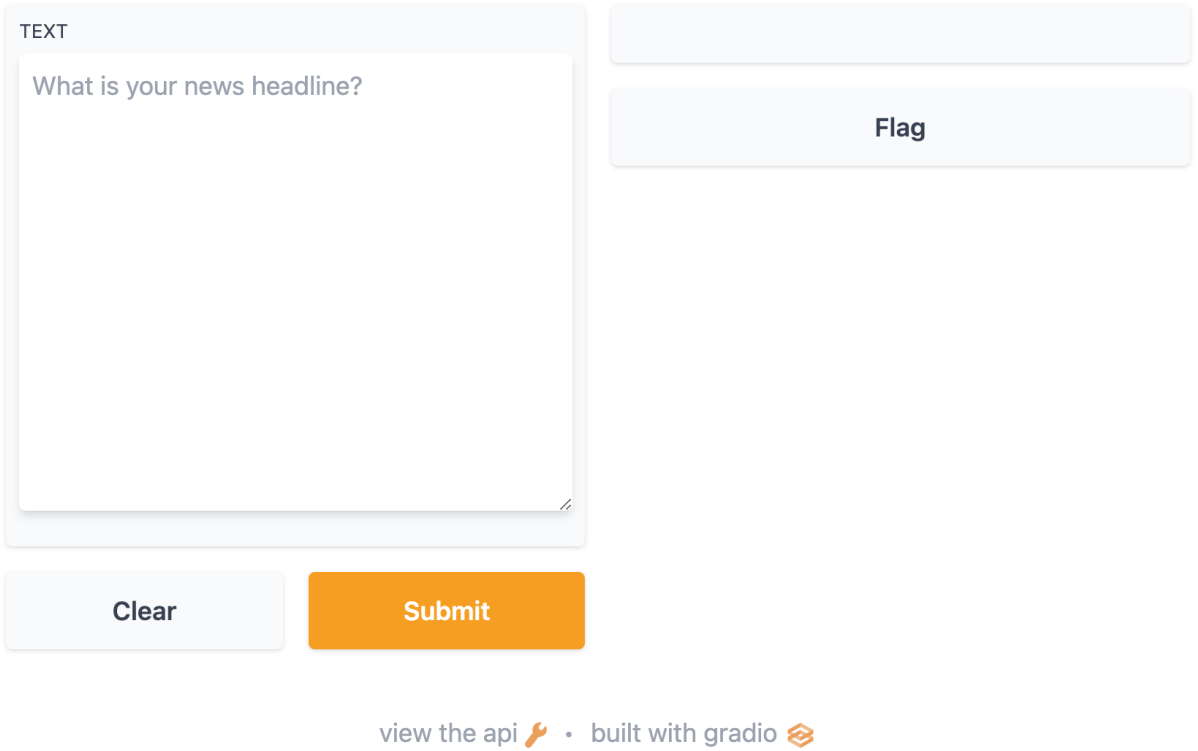






Deployment

- Gradio



Gradio

- News (sample)
  - Negative - "Nokia 's share price fell less than one percent to 18.70 euros ($ 25.41) in Helsinki , while Siemens shares fell 1.02 percent to 90.19 euros ($ 122.57) in Frankfurt ."
  - Positive - 'With the new production plant the company would increase its capacity to meet the expected increase in demand and would improve the use of raw materials and therefore increase the production profitability.
  - Unseen (from headstocks.au) - Pureplay AI stock Appen gets takeover offer, share price jumps almost 30pc

Next Steps

- Train with larger data or in many areas
- Explore more techniques in text pre-processing in NLP

Questions?

Appendices

Look at the most likely words for positive news

```
946         increase
2142            rise
2297         service
1562             new
2335            sign
42         agreement
1677           order
788             grow
194         business
1373          market
932          improve
2405        solution
363         customer
548           expand
112            award
1219            lead
2947             win
2556          supply
1660       operation
1927         product
326         contract
1473          mobile
2506     strengthen
2636     technology
1556        network
1941        project
757             good
14       acquisition
789           growth
2513          strong
Name: feature, dtype: object
```

Look at the most likely words for negative news

```
385         decrease
1934          profit
577             fall
1303            loss
1659       operating
285          compare
1216             lay
804             half
1985         quarter
2120          result
1785          period
439             drop
1307             low
366              cut
469         employee
2456           staff
1546     negotiation
446            early
2241         scanfil
384          decline
1792       personnel
2092          report
1043            item
2660     temporarily
2661       temporary
1308           lower
2048       reduction
857          helsinki
542          exclude
2042           recur
Name: feature, dtype: object
```

End of Presentation!