News Classification
NLP and Deep Learning

Hoyt Lui

Steps

- Data Wrangling
- Data preprocessing
- Modeling
- Testing
- Predicting

Data Wrangling

- Classify news types in original dataset

	type	headline
0	NaN	stocks that hit 52-week highs on friday
1	NaN	stocks that hit 52-week highs on wednesday
2	NaN	71 biggest movers from friday
3	NaN	46 stocks moving in friday's mid-day session
4	NaN	b of a securities maintains neutral on agilent...
5	NaN	cfra maintains hold on agilent technologies, l...
6	NaN	ubs maintains neutral on agilent technologies,...
7	NaN	agilent technologies shares are trading higher...
8	NaN	wells fargo maintains overweight on agilent te...
9	NaN	10 biggest price target changes for friday

Rules-based

	type	count	%
6	analyst_action	126043	30.152890
2	earnings	108442	25.942255
1	corporate_action	87842	21.014179
0	redundant_meaningless	58955	14.103628
5	company_guidance	20402	4.880709
4	merger_acquisition	11674	2.792736
3	options	4655	1.113602

Manually-labelled

	type	count	%
1	redundant_meaningless	2438	24.586527
3	analyst_action	1719	17.335619
6	merger_acquisition	1571	15.843082
2	earnings	1421	14.330375
5	options	1243	12.535296
4	corporate_action	906	9.136749
0	company_guidance	618	6.232352

Data Preprocessing

- Train/test split
- Tokenizer
- Sequence padding

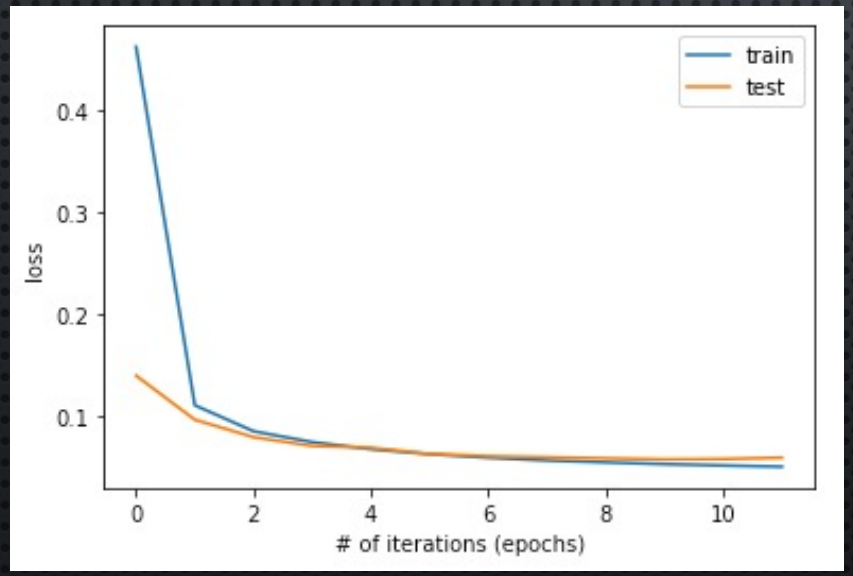
Modeling – Neural Network

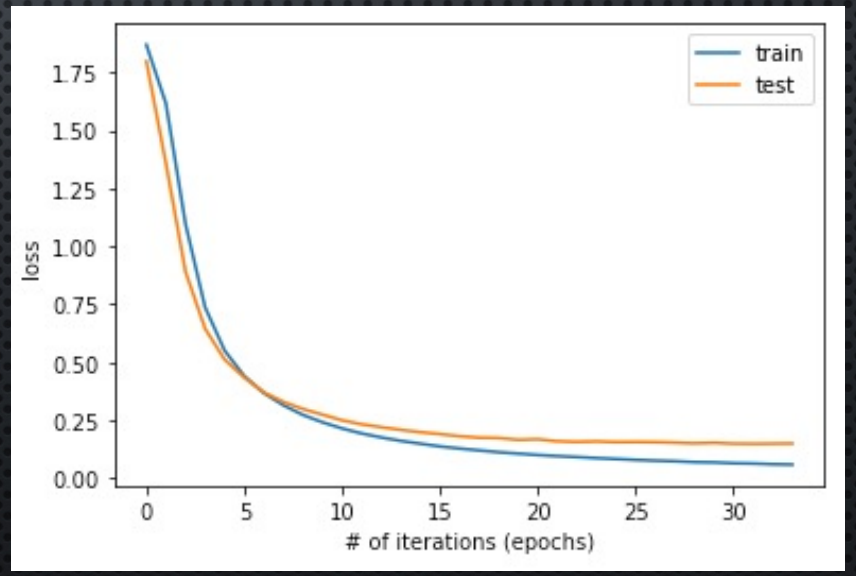
- Sequential
- Global Average Pooling 1D

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 142, 16)           16000
_________________________________________________________________
global_average_pooling1d (Gl (None, 16)                0
_________________________________________________________________
dense (Dense)                (None, 32)                544
_________________________________________________________________
dense_1 (Dense)              (None, 7)                 231
=================================================================
Total params: 16,775
Trainable params: 16,775
Non-trainable params: 0
_________________________________________________________________
```

Modeling – Neural Network

- Early stopping, 2 patience
- Rules-based: 12 epochs, 98.2% accuracy
- Manually-labelled: 34 epochs, 95.8% accuracy





Testing

- Rules-based: 9/9 correct

```
('Pfizer receives FDA approval for its COVID-19 vaccination',): corporate_action
('Tesla expects to deliver 350,000 cars in the next quarter, estimates revenue increased by 5%',): company_guidance
('Intel plans to acquire Disney to expand its consumer sector',): merger_acquisition
('Apple announces a share repurchase program for up to $25 million',): corporate_action
('Microsoft ex-CEO Bill Gates says he loves China',): redundant_meaningless
('Elon Musk will go to the moon with his dogecoin',): redundant_meaningless
('Morgan Stanley analyst upgrades Tesla to Buy, raises price target to $4,000',): analyst_action
('AMD Q4-2025 Adj. EPS $1.89 beats $0.85 estimate, sales $3.37B beat $1.33B estimate',): earnings
('Notable put options activity in Gamestop',): options
```

- Manually-labelled: 8/9 correct

```
('Pfizer receives FDA approval for its COVID-19 vaccination',): corporate_action
('Tesla expects to deliver 350,000 cars in the next quarter, estimates revenue increased by 5%',): redundant_meanin
gless
('Intel plans to acquire Disney to expand its consumer sector',): merger_acquisition
('Apple announces a share repurchase program for up to $25 million',): corporate_action
('Microsoft ex-CEO Bill Gates says he loves China',): redundant_meaningless
('Elon Musk will go to the moon with his dogecoin',): redundant_meaningless
('Morgan Stanley analyst upgrades Tesla to Buy, raises price target to $4,000',): analyst_action
('AMD Q4-2025 Adj. EPS $1.89 beats $0.85 estimate, sales $3.37B beat $1.33B estimate',): earnings
('Notable put options activity in Gamestop',): options
```

Predicting

- Tesla (TSLA) – 20 most recent news headlines

	title	Rules-based type	Manually-labelled type
0	Dow Jones Futures Fall As Tesla Earnings Beat, While Apple Leads 4 Tech Gian...	earnings	redundant_meaningless
1	PRESS DIGEST-British Business - July 27	redundant_meaningless	redundant_meaningless
2	Is Lucid Motors Stock A Buy Right Now As LCID Surges After Its Highly Antici...	redundant_meaningless	redundant_meaningless
3	Tesla Earnings Beat But Musk Warns On Chips, Semi Delayed; FSD Subscription ...	earnings	redundant_meaningless
4	Teslas Earnings Came in Strong. One Thing Is Holding the Stock Back.	earnings	redundant_meaningless
5	Tesla Beats in Q2 as Indexes Set New Closing Highs	earnings	earnings
6	Lucid Motors CEO market debut, retail investors, and what's next for EVs	redundant_meaningless	redundant_meaningless
7	Tesla (TSLA) Q2 Earnings and Revenues Top Estimates	earnings	earnings
8	Tesla posts record $1.1bn profit but warns of chip shortage	redundant_meaningless	redundant_meaningless
9	Tesla Earnings: What Happened with TSLA	earnings	redundant_meaningless
10	Tesla to come under tremendous pressure: GLJ Research Founder	redundant_meaningless	redundant_meaningless
11	Tesla tops $1 billion in profit, delays Semi launch	redundant_meaningless	redundant_meaningless
12	Is Nio Stock A Buy After Q2 Sales More Than Double?	redundant_meaningless	analyst_action
13	Tesla beats Q2 earnings estimates	earnings	earnings
14	Dow Jones Gains, Nasdaq Closes Positive; Tesla Earnings Beat As Bitcoin Soar...	earnings	redundant_meaningless
15	Amazon Job Posting Hints at Plan to Accept Cryptocurrency	redundant_meaningless	redundant_meaningless
16	Tesla beats profit and revenue estimates, delays Semi launch	earnings	earnings
17	Tesla Releases Second Quarter 2021 Financial Results	redundant_meaningless	redundant_meaningless
18	US STOCKS-S&P 500 edges up as investors eye key earnings, Fed meeting	redundant_meaningless	redundant_meaningless
19	The markets have rewarded Powell: Heritage Capital President	redundant_meaningless	redundant_meaningless

Predicting

- Tesla (TSLA) – 100 most recent news headlines classification

Rules-based

redundant_meaningless	62
earnings	33
company_guidance	3
options	1
corporate_action	1

Manually-labelled

redundant_meaningless	89
analyst_action	6
earnings	4
corporate_action	1

Takeaways

- Early stopping is useful as it reduces the chance of overfitting the model
- Rules-based model produces a higher accuracy compared to manually-labelled model (98.2% vs 95.8%), but it does not mean the result is more favorable, it depends on the objective
- Although the manually-labelled model provides a more suitable result to my needs, it shows inconsistency in the model

Further Work – Rules-based

- Make the dataset more balanced by adding more news in the Options, Merger and Acquisition, and Company Guidance categories
- Try other neural network models and select other layers and add more neuron layers to further optimize the model
- Redefine rules with more or less keywords for classification

Further Work – Manually-labelled

- Add more examples to the training set to improve consistency and prediction
- Similar to the rules-based model, try other neural network models and select other layers and add more neuron layers to further optimize the model