News Classification
NLP and Deep Learning

Hoyt Lui

Steps

- Data Wrangling
- Data preprocessing
- Modeling – neural network
- 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

```
neutral                246174
corporate_action       150656
analyst_action         123136
earnings                48158
options                 17875
merger_acquisition      16471
company_guidance        14758
Name: type, dtype: int64
```

	count	%
0	246174	39.883803
3	150656	24.408484
1	123136	19.949840
2	48158	7.802303
6	17875	2.896012
4	16471	2.668544
5	14758	2.391013

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

- Training and validating – 98.6-98.7% accuracy

```
Epoch 1/6
15431/15431 [==============================] - 14s 881us/step - loss: 0.5934 - accuracy: 0.7945 - val_loss: 0.0948
- val_accuracy: 0.9748
Epoch 2/6
15431/15431 [==============================] - 14s 875us/step - loss: 0.0812 - accuracy: 0.9780 - val_loss: 0.0595
- val_accuracy: 0.9824
Epoch 3/6
15431/15431 [==============================] - 13s 862us/step - loss: 0.0544 - accuracy: 0.9841 - val_loss: 0.0480
- val_accuracy: 0.9854
Epoch 4/6
15431/15431 [==============================] - 13s 851us/step - loss: 0.0458 - accuracy: 0.9860 - val_loss: 0.0438
- val_accuracy: 0.9858
Epoch 5/6
15431/15431 [==============================] - 13s 857us/step - loss: 0.0414 - accuracy: 0.9869 - val_loss: 0.0411
- val_accuracy: 0.9867
Epoch 6/6
15431/15431 [==============================] - 13s 847us/step - loss: 0.0393 - accuracy: 0.9871 - val_loss: 0.0392
- val_accuracy: 0.9872
```

Testing

- 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',): neutral
('Elon Musk will go to the moon with his dogecoin',): neutral
('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) – 100 recent news headlines

```
0    50
2    36
4     6
3     5
5     2
6     1
Name: sent, dtype: int64
```

```
{0: 'neutral', 1: 'analyst_action', 2: 'earnings', 3: 'corporate_action', 4: 'merger_acquisition', 5: 'company_guidance', 6: 'options'}
```

Takeaway

- 5th-6th time is the optimal for modeling building
- A shy of 99% accuracy score
- It correctly classifies the 9 sample headlines entered
- It classifies 50 of 100 recent news headlines for Tesla as ‘neutral’

Further Work

- Make the dataset more balanced by adding more news in the categories other than 'neutral', which has the key of 0 in sentiment dict.
- Try other neural network models and add more layers to optimize the model.