News Classification

Natural Language Processing and Deep Learning

Hoyt Lui

Content

Problem Statement…….. 2

Workflow…… 3

Data Wrangling…….. 4

Data Preprocessing……..6

Modeling……8

Testing……..11

Predicting…… 12

Takeaways…… 15

Further Work……15

Business Recommendations……16

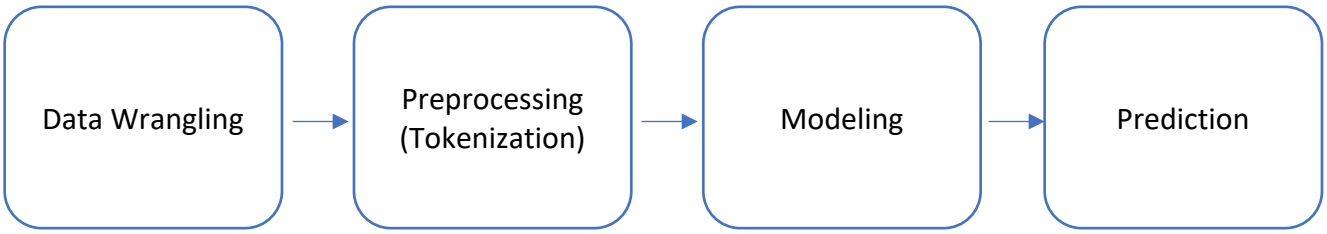
Problem Statement

Natural Language Processing (NLP) can be applied in many areas. In finance, particularly for investing and trading, it is widely used for sentiment analysis. While this approach is applicable to the investing world, we also need to understand that it will have a better result in long-term investing than in short-term trading, mainly because 90% of the traders lose money, and the top-tier traders who make profit consistently are those who set a correct trading mentality for the markets instead of relying on sentiment analysis. In other words, the reason that sentiment analysis does not work well in trading is because many tweets and discussions posted are from the traders or someone who shows interest in the stocks. And since majority of the traders lose money long-term, most textual data, which is also their opinions, is not too relevant to predict the stock price.

If we agree on NLP driving a better result in long-term investing, then we are also aware that sentiment analysis is not too useful for investing purposes. Part of the reason is that long-term investing focuses on company's performance, from products, revenue and cost, to debt, capital expenditure, and profit. It would be better to analyze the company's financial statements than their speeches on either the MD&A (Management Discussion and Analysis) session or earnings call presentation. A simple positive, negative or neutral sentiment will not provide much benefit to stock price movement.

So, what is the use of NLP for investing and trading? This is how I apply NLP to my trading strategies. News acts as a trigger to drive a stock price either up or down, depending on the type of news. Some news creates a more longer-term effect, which may push the stock price upward for a few days up to a few weeks. It is also true for the opposite. Some news can plummet the stock price to close to nothing, creating ripple effects in consecutive days of play. Whereas some news has little or no effect. So, instead of classifying the news as "positive", "negative", or "neutral" for sentiment analysis, which most people on the Internet do, I classify news using a different methodology – news types. For examples, an earnings beater for a company may most likely have a much more positive effect than the company giving out their positive guidance for their next fiscal year. Or, a price target decrease announced by couple street analysts may have a more negative effect than a failed product launch. The whole model usually would combine NLP with other statistical modeling to yield the best prediction, but NLP is the first step of finding the right catalyst.

Workflow



In this project, we will look at both rules-based and manually-labelled models so that we can compare which method works better than the other. For rules-based, I classified the news using keywords in 7 categories; while for the other model, I manually labelled close to 10,000 headlines, also in 7 categories.

The models are tokenized in the same way, with the same parameters. The difference will be in the modeling part, where I allow the neural network to train both models before they get overfit. Due to the stochastic nature of neural networks, in the 8 trials tested for both models, the rules-based model has the number of iterations between 4 and 15 with 2 patience added. whereas the manually-labelled model has 25-37 iterations, also with 2 patience added. We will further discuss the process and the results below.

Data Wrangling

I downloaded the financial news data from one of the Kaggle datasets. In fact, any news headline datasets will do the work after being categorized.

There are only two columns needed for this analysis – headline and type. At first, the type column contains only empty values because it is newly created. Our job is to categorize as much news in the headline as possible.

First, we will have to convert all text into lowercase, as shown below

	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

Then we will determine which categories are needed for classification. In this project, we classify news in 7 types.

1. Redundant/meaningless. Any news that has little or no impact on stock price movement.
2. Analyst action. Initial coverage, ratings or price target adjustment by analysts.
3. Earnings. Earnings performance or related news.
4. Corporate action. Dividend or share repurchase program announcement, new contract or partnership, settlement of litigation, etc.
5. Merger and acquisition. M&A with other companies.
6. Company guidance. Company providing estimates guidance to their future performance.
7. Options. Any news related to options or derivatives activities of the company.

Rules-based model

Below is the distribution of the news in each category. 30% comes from Analyst Action, 26% from Earnings, 21% Corporate Action, 14% Redundant/Meaningless. And the minority includes Company Guidance, Merger and Acquisition, and Options. There is a total of 418,013 news headlines in the dataset. The range from 1% to 30% shows imbalance in data. But let's proceed for now and see the results first.

	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 model

Below shows that 25% from Redundant/Meaningless, 17% from Analyst Action, 16% Merger and Acquisition, 14% Earnings, 13% Options, and the rest includes Corporate Action and Company Guidance. This is a more balanced dataset, which contains 9,916 news headlines. However, it is just 2.4% of the quantity compared to the rules-based model above.

	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

We split dataset into 80/20 training and test sets. Then we process the data using tokenization.

What is tokenization? And why is it important?
Tokenization in lexical analysis means that a text string is broken down into chunks of words, each word is called a token. These tokens will then be passed on to some other form of processing for further analysis.

Computers are different from human. They do not know the meaning of words. However, we can assign meaning of the words to the computers. Then they will categorize the assigned words into that meaning the next time they encounter the same word, which is the essence of tokenization – breaking down the words for the computer to understand.

We declare the variables for tokenization purposes as followed:

- Input dimension. We assign computer to memorize 1000 vocabulary words as the input dimension.
- Output dimension, also dense embedding dimension. Here, we set 16 as the shape of the output dimension.
- Input length. It is the length of input sequences. Without it, the shape of dense outputs cannot be computed. We set the max input length to be 142.
- OOV token. We assign "OOV" (out-of-vocabulary) so that it will be added to the word index as a complement to the existing vocabulary words.

What is turning texts into sequences? Why is the concept important?
After creating a tokenizer object, we will convert words into sequences of integers that is indexed with the vocabulary words we set earlier, i.e., 0-999 in this case. This process will have the program returned a list of words (or tokens).

For example, we have 4 words:
'elon'
'musk'
'loves'
'dogecoin'

After we turn text into sequence, we now have ['elon', 'musk', 'loves', 'dogecoin'], or in computer vision, e.g., [586, 587, 23, 748].

What is sequence padding? And why is it important?

Next, we pad sequences into all same-length sequences. The process is mandatory because all neural networks require inputs to have the same shape and size, as it is reading an array as an input. Therefore, we ensure all sequences having the same length by truncating and/or filling 0 in them.

- Max length. We set the max length to be the same input length we set for tokenizer.
- Truncating type. Since I think key ideas are usually presented early in the sentence, I set "post" as the truncating type, so any sequence longer than the max length will be truncated on the end. You can set "pre" alternatively to have it cut on the beginning otherwise.
- Padding type. I set "post" as the padding type, which means adding 0's on the end of any shortened sequences. Again, you can set "pre" alternative to add 0's on the beginning of sequences if you prefer to.

For example,
[['amd', 'will', 'supersede', intel', 'in', 'sales'],
['in', 'the', 'hunt', 'for', 'life', 'outside', 'earth']]

These 6-word and 7-word sequences will look like below in computer vision:
[[96, 34, 452, 194, 22, 48],
[22, 13, 56, 25, 65, 235, 385]]

If we set the max length at "8" words, and pad the sequences with "post". Then the computer will receive:
[[96, 34, 452, 194, 22, 48, 0, 0],
[22, 13, 56, 25, 65, 235, 385, 0]]

If we set the max length at "7" words, and pad the sequences with "pre". Then the computer will receive:
[[0, 96, 34, 452, 194, 22, 48],
[22, 13, 56, 25, 65, 235, 385]]

If we set the max length at "6" words, and truncate the sequences with "post". Then the computer will receive:
[[96, 34, 452, 194, 22, 48],
[22, 13, 56, 25, 65, 235]]

If we set the max length at "5" words, and truncate the sequences with "pre". Then the computer will receive:
[[34, 452, 194, 22, 48],
[56, 25, 65, 235, 385]]

Modeling

What is Sequential model? How does it differ from Functional model?
In short, sequential model is a linear stack of layers that creates layer-by-layer for problems. It is straightforward and easy to set up compared to functional model, which is more flexible and capable of creating complex networks that specifically connect different input and output layers. For this project, we can build a simple model using sequential model to group a linear stack of layers into it.

Layer selection – why pooling layers?
Pooling layers use pooling operation to calculate the value in each patch of each feature map. (Max pooling will calculate the maximum value, while average pooling will calculate the average value.) The problem with feature maps is that they record precise position of the input features, meaning small movements in the position of the input feature will result in a different feature map. By using pooling layers, the results are down sampled (or pooled) feature maps that summarize the presence of features in the patches of the feature map.

This can be the most activated or the average presence of a feature in the patch, depending if you use max pooling or average pooling methods.

I used average pooling layers because I wanted to find the average, or more common, value in the news headlines. However, you can try max pooling layers to return the maximum value. You can also try the other layers such as recurrent layers (especially LSTM) or activation layers, which are also the popular alternatives.

Parameters are set below for building the neural network:

```
# build neural network
model = Sequential()
model.add(Embedding(vocab_size, embedding_dim, input_length=max_length))
model.add(GlobalAveragePooling1D())
model.add(Dense(32, activation='relu'))
model.add(Dense(7, activation='softmax'))

model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
model.summary()
```

First off, note that this model consists of one 1000-neuron input layer, one average pooling layer, one 32-neuron hidden layer, and one 7-neuron output layer.

- Activation. I set “relu” (or rectified linear activation function) so that it returns all positive values in the hidden layers. In the output layer, I set “softmax” as the function that normalizes the output to a probability distribution over 7 predicted classes in this case.

- Loss. “Categorical cross-entropy” is used because the output can only belong to 1 out of 7 possible categories. The lower the loss score, the lower the variance, the more accurate the model.
- Optimizer. Adam optimization is used, which applies stochastic gradient descent (SGD) to allow the function to jump to better local minima.
- Metrics. Accuracy is used to calculate how often predictions equal labels.

```
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
_________________________________________________________________
```

```
# add early stopping with patience
es = EarlyStopping(monitor='val_loss', mode='min', verbose=1, patience=2)
```

Early stopping – why adding it?

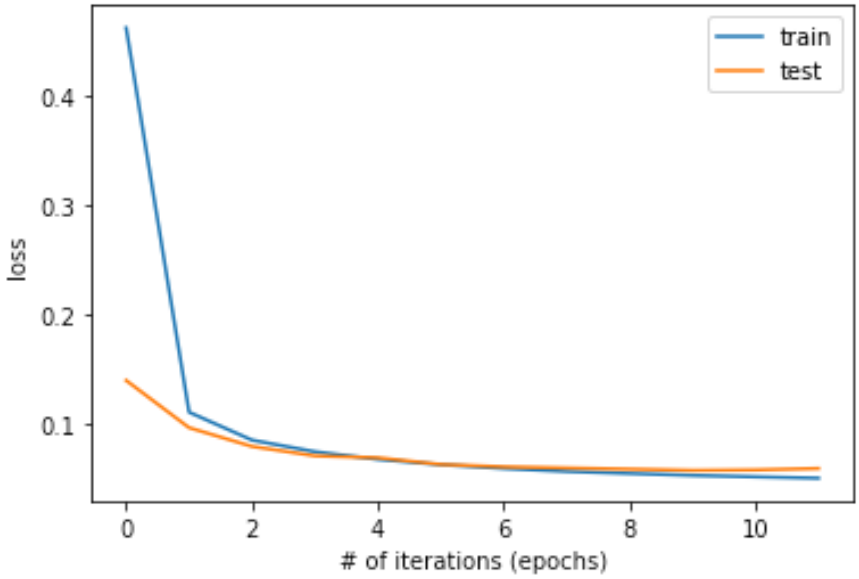
Early stopping is a mechanism that automatically stops training the model once the model performance stops improving on the validation dataset in loss. It will reduce overfitting the model, which happens casually in neural networks.

Adding patience

Patience is used to delay a trigger in stoppage when we see no improvement in training the model. In this project, we set patience to be 2, which means that the training will iterate 2 more times after early stopping kicks in. This prevents the training from stopping too early, causing underfitting the model. As we can see in the plots below, the cost of underfitting is much greater than that of overfitting by training the model 2 more times.

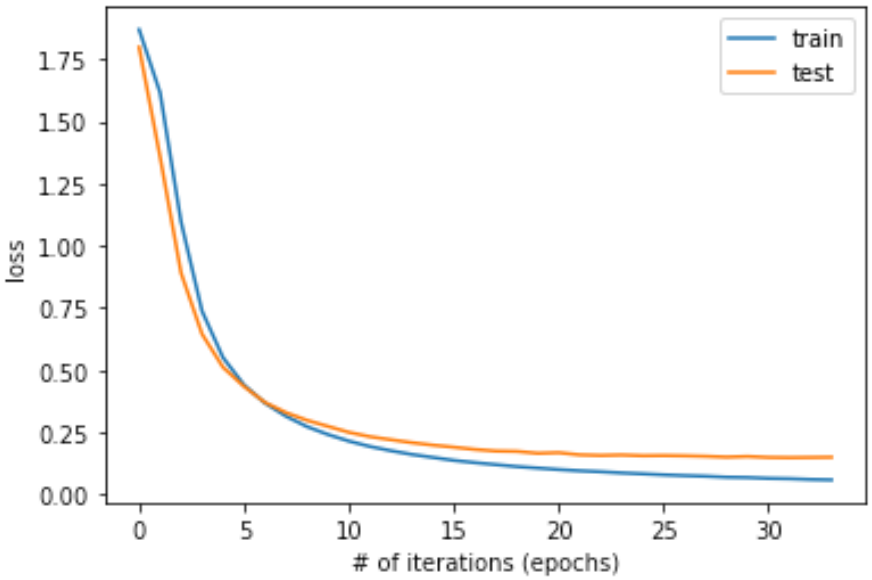
In the rules-based model, epoch stops at 12 with an accuracy of 98.2% on the validation dataset. In the 8 trials we tested, the model stopped in the range of 4 to 15 times, with an accuracy between 97.0 and 98.6%.

```
Epoch 00012: early stopping
Training accuracy: 0.984, Test accuracy: 0.982
```



In the manually-labelled model, epoch stops at 34 with an accuracy of 95.8% on the validation dataset. In the 8 trials we tested, the model stopped in the range of 25 to 37 times, with an accuracy between 94.2 and 97.0%.

```
Epoch 00034: early stopping
Training accuracy: 0.987, Test accuracy: 0.958
```



Testing

The simple made-up phrases below are used to test the validity of the model. It will also spot the weakness of the model in case of wrong classification.

The rules-based model got 9 out of 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
```

The manually-labelled model got 8 out of 9 correct, which shows some inconsistency in the model.

```
('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

We further used the model to predict the 100 most recent news of Tesla. Below is how I would classify 20 of them with my rationale. 6/20 is earnings and 14/20 is Redundant/Meaningless. Please be noted that Tesla just held their Q2-2021 earnings call; hence, more news in earnings than usual is expected.

Title	Type	Rationale
Dow Jones Futures Fall As Tesla Earnings Beat, While Apple Leads 4 Tech Giants Set To Report	Redundant/meaningless	News includes other stocks
PRESS DIGEST-British Business - July 27	Redundant/meaningless	Meaningless title
Is Lucid Motors Stock A Buy Right Now As LCID Surges After Its Highly Anticipated IPO?	Redundant/meaningless	Irrelevant to TSLA
Tesla Earnings Beat But Musk Warns On Chips, Semi Delayed; FSD Subscription 'Debatable'	Redundant/meaningless	View rather than earnings news
Teslas Earnings Came in Strong. One Thing Is Holding the Stock Back.	Redundant/meaningless	View rather than earnings news
Tesla Beats in Q2 as Indexes Set New Closing Highs	Earnings	
Lucid Motors CEO market debut, retail investors, and what's next for EVs	Redundant/meaningless	Irrelevant to TSLA
Tesla (TSLA) Q2 Earnings and Revenues Top Estimates	Earnings	
Tesla posts record $1.1bn profit but warns of chip shortage	Earnings	
Tesla Earnings: What Happened with TSLA	Redundant/meaningless	Meaningless title
Tesla to come under tremendous pressure: GLJ Research Founder	Redundant/meaningless	View
Tesla tops $1 billion in profit, delays Semi launch	Earnings	
Is Nio Stock A Buy After Q2 Sales More Than Double?	Redundant/meaningless	Irrelevant to TSLA
Tesla beats Q2 earnings estimates	Earnings	
Dow Jones Gains, Nasdaq Closes Positive; Tesla Earnings Beat As Bitcoin Soars; IPO Passes Buy	Redundant/meaningless	General market news
Amazon Job Posting Hints at Plan to Accept Cryptocurrency	Redundant/meaningless	Irrelevant to TSLA
Tesla beats profit and revenue estimates, delays Semi launch	Earnings	
Tesla Releases Second Quarter 2021 Financial Results	Redundant/meaningless	Meaningless title
US STOCKS-S&P 500 edges up as investors eye key earnings, Fed meeting	Redundant/meaningless	General market news
The markets have rewarded Powell: Heritage Capital President	Redundant/meaningless	General market news

Below is the result from the rules-based model. Out of the 20 news headlines, it classified 9 in earnings and 11 in Redundant/Meaningless. It is more sensitive than I would like when compared to my way of classification.

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

Out of the 100 most recent news, 62 are classified as redundant/meaningless, 33 are earnings, the rest includes company guidance, options and corporate action.

redundant_meaningless	62
earnings	33
company_guidance	3
options	1
corporate_action	1

Below is the result from the manually-labelled model. We can tell 2 characteristics right off the bat.

1. It classifies a lot of headlines as Redundant/Meaningless
2. There is inconsistency of the method, e.g., the one headline classified an Analyst Action

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

Out of the 100 recent Tesla news, a whopping 89% are Redundant/Meaningless. I am surprised that news in Analyst Action is more than that in Earnings, which is hugely different from the rules-based model. And out of the 4 Earnings headlines, all of them existed in the 20 most recent news. It holds true because Tesla just announced their earnings in the last couple hours by the time I was running the model. All the earnings news before their earnings call should be neglected because they do not provide any meaningful value in the absence of the true results.

```
redundant_meaningless    89
analyst_action            6
earnings                  4
corporate_action          1
```

There are definitely pros and cons within the manually-labelled model.
Pros:

- It is more conservative. By identifying the majority of the headlines as meaningless news, we only focus on the few ones that may cause actions in the stock price movement

Cons:

- As shown above, it misclassified 1 of the 9 made-up headline as Redundant/Meaningless, and 1 of the 20 most recent news headlines of Tesla as Analyst Action where it should be Redundant/Meaningless. This shows inconsistency in the model and needs more sample in the training set to improve its prediction.

Now, keep in mind that the way of classification is subjective. Some people may prefer the rules-based model as it captures all the news headlines as Earnings using the keywords. However, if I only focus on the absolutely essential news without other unnecessary distraction, the manually-labelled model fits my needs better. For examples, if other stock tickers are mentioned along in the headlines, it will be classified as a general news because the words used to describe the other companies may not be relevant to the target company. And to reinforce that, the problem with the current manual model is that it needs more examples in the training set to perform better such that it will be less likely to classify an obvious news headline into some other categories.

Takeaways

This project tells us few key 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.
- Although the manually-labelled model provides a more suitable result to my needs, it shows inconsistency in the model. It implies that an approximation of 1,000 samples in each category in the training set is not enough, and more samples are needed to improve its predictive power.

Further work

Approaches can be considered for further improvement of the rules-based model:

- 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.

Approaches can be considered for further improvement of the manually-labelled model:

- 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.

Business Recommendation

So, we created this news classification model through NLP and deep learning. The baseline of the manually-labelled model results in a 95.8% accuracy. And it will improve once we accumulate more data in the training set.

Why is the model valuable? How is it applied? Who will benefit from using it?
It goes back to our original intention of creating this model – why we need to classify news. We know that different news will lead to different impact to the stock price movement. By combing news catalyst and the past stock movement using statistical approach, we will have an idea of how certain news type will drive the stock price in which direction, and by how much approximately. And the first step will always go back to correctly classifying the news type. The goal of this project is to create insights to retail and/or institutional investors and traders who make investment decisions based on event-driven strategies.