STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

BIA 667 Final Project Report

Daily News for Stock Market Prediction

Siqi Jiang

Yanan Liu

Chao Tian

Executive summary

Market prediction offers great profit avenues and is a fundamental stimulus for most researchers in this area. The goal of this project is to predict stock market movement using stock price information and daily news headline. The news data is collected from Reddit news and top 25 headlines, ranked based on reddit user votes, are taken on each day. The stock market data, DJIA (Dow Jones Industrial Average) of each day is collected from Yahoo finance. Combined both datasets to process and apply modeling techniques further to get desired results. Machine learning models are used to address the problem defined. We computed accuracies of various models to know the best performing model among the models we applied. The accuracy of four machine learning models is about 50%. To get higher accuracy, we use two deep learning models: CNN and LSTM. The accuracy of two deep learning models is about 52%. Therefore, the deep learning model performs better in this scenario. We got the best accuracy for sentiment analyzing with CNN algorithm, able to improve the accuracy to 52.3%.

Project Introduction

Background

Stock market prediction aims to determine the future movement of the stock value of a financial exchange. The accurate prediction of share price movement will lead to more profit investors can make. Predicting how the stock market will move is one of the most challenging issues due to many factors that involved in the stock prediction, such as interest rates, politics, and economic growth that make the stock market volatile and very hard to predict accurately.

Since stock investment is a major financial market activity, a lack of accurate knowledge and detailed information would lead to an inevitable loss of investment. Stock market prediction methods are divided into two main categories: technical and fundamental analysis. Technical analysis focuses on analyzing historical stock prices to predict future stock values. On the other hand, fundamental analysis relies mostly on analyzing textual information like financial news and earning reports. These days more and more critical information about the stock market has become available on the Web, such as BBC, Bloomberg, and Yahoo Finance. This provides a good chance for us to use online textual financial news to predict the stock market.

Motivation & Research question

It has always been an alluring dream for the quants in Wall street to predict the stock market by looking at the historical records. Active research has been going in this field for many years. Many researchers have pointed out that apart from historical stock price data, there are some unquantifiable forces play a major role in the trend of the stock market. The un-quantifiable factors could be a political development, policy changes or some time related event. The news tabloids are an interesting collection of records of such events which are also ordered chronologically and indexed to a great extent. Our project is an attempt to analyze these news top headlines to find if there is a correlation between the contents of such top headlines to the performance of the stock. We will analyze word patterns and sentiment of the top headlines to predict the performance of the stock market.

Literature Review

Predicting the movement of stock market indices is of great importance to entire industries. The investors determine stock prices by using publicly available information to predict how the stock market will react, where "publicly available information" means mostly (financial) news. Nowadays, news come almost exclusively via web sources in the form of text. This is the reason why many researchers have proposed methods that use text information for analyzing the stock market. Many papers study the relationship between stock price movements and the market sentiments.

The works of Bollen, Mao, and Zeng (2001) and Chowdhury, Routh, and Chakrabarti (2014) fall into the category of papers that describe the use of sentiment in text to predict the stock market. While Bollen, Mao and Zeng (2001) use sentiment analysis on tweets, Chowdhury, Routh and Chakrabarti, (2014) try to extract sentiment from news.

Mudinas A, Zhang D, Levene M (2019) examined the ability to use sentiment polarity (positive and negative) and sentiment emotions selected from financial news or tweets to predict the market movements. For sentiment analysis, they have collected a large dataset of the top 25 historical financial news headlines in addition to a large set of financial tweets collected from Twitter. Furthermore, they collected stock historical price data for many S&P 500 companies and used the close price as an indicator of the stock movements. The authors examined two machine learning methods SVM and LSTM. The experiment result illustrated that for some stocks, adding sentiment emotions to the machine learning market prediction model will increase the prediction accuracy.

Checkley MS, Higón DA, Alles H (2017) examined the efficiency of using sentiment analysis of microblogging sites to forecast the stock price returns, volatility, and trading volume. The extracted intraday data from the two sources of information, Twitter and StockTwits, were collected for 2 years. For the evaluation, the authors used five famous stocks, namely, Amazon, Apple, Goldman Sachs, Google, and IBM. Prices were represented every 2 min, and the sentiment data were collected for the same period span of each trading day. The experiments indicate that there is a causal link between Twitter sentiments and stock market returns, volatility, and volume. Among all five stocks, market volatility and volume seem to be more predictable than market direction or return.

Bujari A, Furini M, Laina N. (2017) aims to analyze whether tweet messages could be used to predict future trends of stocks for particular companies listed on the Dow Jones stock market, focusing on 12 companies related to 3 distinct and crucial economic branches in technology, services, and health care. The authors gathered the company's market data and Twitter posts for a 70-day period for analysis. The companies of each category were chosen based on the volume of messages that mention the company names on the StockTwits website. The study illustrates that some of the proposed ad hoc forecasting models well predict the next day direction of the stock movements for some companies with 82% of success and there is no unified method to be used with all cases. The results also indicate that more volume of a tweet will yield better prediction results. Moreover, the study proved the robust correlation between tweet's posts and the trend movements for some companies.

Overall, past studies indicate that there is a strong relationship between market movements and information published in news and social media. The information on social media contributes to enhancing the prediction models with all of the discussed papers. The evaluation of event sentiment may affect the market returns further and boost the outcome of forecasting.

In contrast to the other current review articles that concentrate on discussing many methods used for forecasting the stock market, this study aims to compare many machine learning (ML) and deep learning (DL) methods used for sentiment analysis to find which method could be more effective in prediction and for which types and amount of data.

Data Preprocessing

Dataset

We used data from two independent sources:

• News data: historical news headlines from Reddit World News Channel. They are ranked by Reddit users' votes, and only the top 25 headlines are considered for a single date.

	Date	News
2	2016/7/1	A 117-year-old woman in Mexico City finally received her birth certificate, and died a few hours later. Trini
3	2016/7/1	IMF chief backs Athens as permanent Olympic host
4	2016/7/1	The president of France says if Brexit won, so can Donald Trump
5	2016/7/1	British Man Who Must Give Police 24 Hours' Notice of Sex Threatens Hunger Strike: The man is the subject
6	2016/7/1	100+ Nobel laureates urge Greenpeace to stop opposing GMOs
7	2016/7/1	Brazil: Huge spike in number of police killings in Rio ahead of Olympics
8	2016/7/1	Austria's highest court annuls presidential election narrowly lost by right-wing candidate.
9	2016/7/1	Facebook wins privacy case, can track any Belgian it wants: Doesn't matter if Internet users are logged intc
10	2016/7/1	Switzerland denies Muslim girls citizenship after they refuse to swim with boys at school: The 12- and 14-ye
11	2016/7/1	China kills millions of innocent meditators for their organs, report finds
12	2016/7/1	France Cracks Down on Factory Farms - A viral video campaign has moved the govt to act. In footage share
13	2016/7/1	Abbas PLO Faction Calls Killer of 13-Year-Old American-Israeli Girl a Martyr
14	2016/7/1	Taiwanese warship accidentally fires missile towards China
15	2016/7/1	Iran celebrates American Human Rights Week, mocks U.S. rights record
16	2016/7/1	U.N. panel moves to curb bias against L.G.B.T. people despite fierce resistance from Muslim and African cc
17	2016/7/1	The United States has placed Myanmar, Uzbekistan, Sudan and Haiti on its list of worst human trafficking o
18	2016/7/1	S&P revises European Union credit rating to 'AA' from 'AA+'

• Stock data: Dow Jones Industrial Average (DJIA) daily index values were used. The dataset contains 1989 records with dates ranging from 08-08-2008 to 07-01-2016. On each date, the "open", "high", "low", "close", "volume", and "adjusted close" values are recorded.

	Date	Open	High	Low	Close	Volume	Adj Close
2	2016/7/1	17924.24	18002.38	17916.91	17949.37	82160000	17949.37
3	2016/6/30	17712.76	17930.61	17711.8	17929.99	1.33E+08	17929.99
4	2016/6/29	17456.02	17704.51	17456.02	17694.68	1.06E+08	17694.68
5	2016/6/28	17190.51	17409.72	17190.51	17409.72	1.12E+08	17409.72
6	2016/6/27	17355.21	17355.21	17063.08	17140.24	1.39E+08	17140.24
7	2016/6/24	17946.63	17946.63	17356.34	17400.75	2.39E+08	17400.75
8	2016/6/23	17844.11	18011.07	17844.11	18011.07	98070000	18011.07
9	2016/6/22	17832.67	17920.16	17770.36	17780.83	89440000	17780.83
10	2016/6/21	17827.33	17877.84	17799.8	17829.73	85130000	17829.73
11	2016/6/20	17736.87	17946.36	17736.87	17804.87	99380000	17804.87
12	2016/6/17	17733.44	17733.44	17602.78	17675.16	2.49E+08	17675.16
13	2016/6/16	17602.23	17754.91	17471.29	17733.1	91950000	17733.1
14	2016/6/15	17703.65	17762.96	17629.01	17640.17	94130000	17640.17
15	2016/6/14	17710.77	17733.92	17595.79	17674.82	93740000	17674.82
16	2016/6/13	17830.5	17893.28	17731.35	17732.48	1.02E+08	17732.48
17	2016/6/10	17938.82	17938.82	17812.34	17865.34	90540000	17865.34
18	2016/6/9	17969.98	18005.22	17915.88	17985.19	69690000	17985.19

Data processing

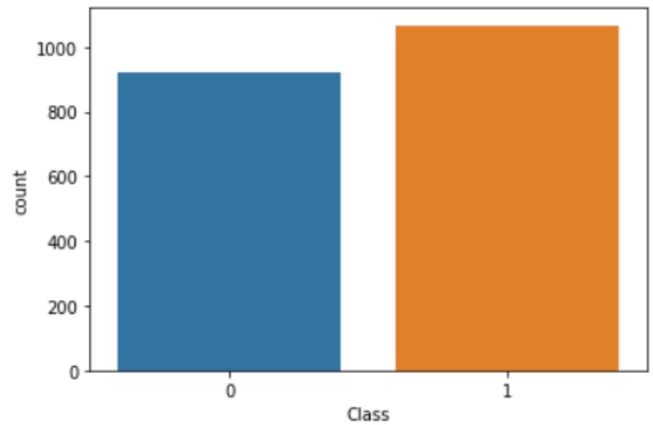
News and stock data were merged into a single dataset by aligning the news headlines with the trading days of the stock data – for each of the 1989 trading days all the DJIA index values for that day together with the most voted 25 news headlines for the previous day were recorded.

Since the goal of this project is to predict stock market movement using stock price information and daily news headline， we will use the change of "close" value to define the trend of the stock market. To be specific, we replace all the stock data by a "label" feature in the dataset with value "0" if the closing value of DJIA has decreased from previous day's close, and as "1" if the closing value of DJIA has increased or stayed same as previous day's close.

The entire processed dataset consists of 1989 rows, representing the trading days, and 27 columns that represent the features of which 25 are the news headlines (as text), and 1 is the date of the trading day. The feature "label" reflects the rising or falling of the DJIA index.

	Date	Label	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	
0	2008-08-08	0	b"Georgia 'downs two Russian warplanes' as cou...	b'BREAKING: Musharraf to be impeached.'	b'Russia Today: Columns of troops roll into So...	b'Russian tanks are moving towards the capital...	b"Afghan children raped with 'impunity,' U.N. ...	b'150 Russian tanks have entered South Ossetia...	b"Breaking: Georgia invades South Ossetia, Rus...	b"The 'enemy combatent' trials are nothing but...	tr Os
1	2008-08-11	1	b'Why wont America and Nato help us? If they w...	b'Bush puts foot down on Georgian conflict'	b"Jewish Georgian minister: Thanks to Israeli ...	b'Georgian army flees in disarray as Russians ...	b"Olympic opening ceremony fireworks 'faked'"	b'What were the Mossad with fraudulent New Zea...	b'Russia angered by Israeli military sale to G...	b'An American citizen living in S.Ossetia blam...	b'\ W I
2	2008-08-12	0	b'Remember that adorable 9-year-old who sang a...	b"Russia 'ends Georgia operation'"	b'"If we had no sexual harassment we would hav...	b"Al-Qa'eda is losing support in Iraq because ...	b'Ceasefire in Georgia: Putin Outmaneuvers the...	b'Why Microsoft and Intel tried to kill the XO...	b'Stratfor: The Russo-Georgian War and the Bal...	b"I'm Trying to Get a Sense of This Whole Geor...	s t
3	2008-08-13	0	b' U.S. refuses Israel weapons to attack Iran:...	b"When the president ordered to attack Tskhinv...	b' Israel clears troops who killed Reuters cam...	b'Britain\'s policy of being tough on drugs is...	b'Body of 14 year old found in trunk; Latest (...	b'China has moved 10 *million* quake survivors...	b"Bush announces Operation Get All Up In Russi...	b'Russian forces sink Georgian ships '	co recc
4	2008-08-14	1	b'All the experts admit that we should legalis...	b'War in South Osetia - 89 pictures made by a ...	b'Swedish wrestler Ara Abrahamian throws away ...	b'Russia exaggerated the death toll in South O...	b'Missile That Killed 9 Inside Pakistan May Ha...	b"Rushdie Condemns Random House's Refusal to P...	b'Poland and US agree to missle defense deal. ...	b'Will the Russians conquer Tblisi? Bet on it,...	e Sou

The data is balanced with a data split of 47/53, with 925 records labeled as "0" and 1064 records labeled as "1" in the total 1989 records present in the data.



The news headline contains some extra information which we do not need or not desirable for text processing. For example, the top1 news in 2008-08-08, b"Georgia 'downs two Russian warplanes' as countries move to brink of war", contains letter b, double quotes, single quote, upper case letters. All of them should be removed in order to further process the text and apply our text classification models. There are many such headlines present in the data, which have similar issues as the headline show.

To address these issues in the data, we use the function stopwords and tokenize from nltk package to process the data. We removed extra characters and meaningless words, such as quotes, extra letters coming between (‘b). Then we converted the words into lower case letters. At the same time, we divided the sentences into words and created a table with count of those words.

	Date	Label	Top1_	Top2_	Top3_	Top4_	Top5_	Top6_	Top7_	Top8_	Top9_
0	2008-08-08	0	georgia downs two russian warplanes countries ...	breaking musharraf impeached	russia today columns troops roll south ossetia...	russian tanks moving towards capital south oss...	afghan children raped impunity ,' official say...	150 russian tanks entered south ossetia whilst...	breaking georgia invades south ossetia russia ...	the enemy combatent trials nothing sham salim ...	georgian troops retreat osettain capital presu...
1	2008-08-11	1	why wont america nato help us if wont help us ...	bush puts foot georgian conflict	jewish georgian minister thanks israeli traini...	georgian army flees disarray russians advance ...	olympic opening ceremony fireworks faked '"	what mossad fraudulent new zealand passports i...	russia angered israeli military sale georgia	an american citizen living ossetia blames geor...	welcome to world war iv now in high definition !'
2	2008-08-12	0	remember adorable year old sang opening ceremo...	russia ends georgia operation '"	'" if sexual harassment would children ...'"	al qa eda losing support iraq brutal crackdown...	ceasefire georgia putin outmaneuvers west	why microsoft intel tried kill xo 100 laptop	stratfor the russo georgian war balance power	trying get sense this whole georgia russia war...	the us military surprised timing swiftness rus...
3	2008-08-13	0	refuses israel weapons attack iran report	when president ordered attack tskhinvali capit...	israel clears troops killed reuters cameraman	britain \' policy tough drugs pointless ", say...	body 14 year old found trunk latest ransom pai...	china moved 10 million quake survivors prefab ...	bush announces operation get all up in russia ...	russian forces sink georgian ships	the commander navy air reconnaissance squadron...

Finally, we combined all the sentences or headlines in a row, and we use TFIDF to transform text into matrix.

Methodology & Result

After we got the data of daily news which we want, the next goal was to find the relation between the up or down of Dow Jones Industrial Average and daily news. In other words, this is a two-class problem. Compared with yesterday, if the DJIA (Dow Jones Industrial Average) remains unchanged or rises, we will mark it as 1, and if it falls, it will be marked as 0. This part mainly contains four sections: split data to train set and test set, use four classifiers to predict, use two neural network model to solve the problem, compare and analyze all methods 'results.

Training set and Test set

To verify the effectiveness of methods, we treated 80 percent of the dataset as the training set and the rest 20 percent dataset as the test set. The total row number of datasets was 1989.

Algorithm

In this part, we use six different methods to predict the stock market. Four of them came from machine learning classifiers, and the other two are neural network models.

Logistic Regression Classifier

Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature. Logistic regression is most commonly used when the data in question has binary output, so when it belongs to one class or another, or is either a 0 or 1.

Random Forest Classifier

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction (see figure below). The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science-speak, the reason that the random forest model works so well is A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

Multinomial Naive Bayesian classifier

Naive Bayes classifier is a general term that refers to conditional independence of each of the features in the model, while Multinomial Naive Bayes classifier is a specific instance of a Naive Bayes classifier that uses a multinomial distribution for each of the features. And Multinomial Naive Bayes classifier is one of the two classic naive Bayes variants used in text classification.

XGBoost classifier

XGBoost is short for "Extreme Gradient Boosting". The algorithm applied by XGBoost is a gradient boosting decision tree, which can be used for both classification and regression problems. Gradient boosting is to try to correct the residuals of all previous weak learners by adding new weak learners. In the end, if multiple learners are added together to make the final prediction, the accuracy will be higher than a single one. It is called Gradient because the gradient descent algorithm is used to minimize the loss when adding a new model.

LSTM Network

Long Short Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter & Schmidhuber (1997) and were refined and popularized by many people in the following work. They work tremendously well on a large variety of problems and are now widely used. LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods is practically their default behavior, not something they struggle to learn.

CNN Network

Convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analyzing visual imagery. In this case, we use CNN to do a text task. Technically, deep learning CNN models to train and test, each input will pass it through a series of convolution layers with filters (Kernals), Pooling, fully connected layers (FC) and apply Softmax function to classify an object with probabilistic values between 0 and 1.

Experiments & Results

Machine Learning

For four classifiers, we did not set some special parameter. After training these models using training set, we use testing set to evaluate the performance of classifiers. The result as following:

```
a,b,c,d = all_classifier()
```

```
Logistic Classifier 0.507537688442211
Logistic ROC Curve 0.5506668013740149
RandomForest Classifier 0.5025125628140703
MultinomialNB Classifier 0.507537688442211
XGBClassifier Accuracy: 50.00%
```

Then, I use tf-vectorizer to see the positive and negative of the words:

```
50000 50000
Positive                       Words  Coefficient
29143           new zealand     0.063290
15216            first time     0.063036
42435              tear gas     0.054554
22085         islamic state     0.053448
39855          social media     0.053029
21122  intelligence agencies     0.051963
39044       sexual violence     0.051112
38336      security council     0.050514
1958          air pollution     0.049268
29542      nobel peace prize    0.049127
Negative                   Words  Coefficient
18800  haiti earthquake    -0.049158
47581       wall street    -0.050373
40176      south korean    -0.051054
21603      iran nuclear    -0.052222
39021      sexual abuse    -0.052383
32720     phone hacking    -0.053675
30067   nuclear weapons    -0.058092
3474       around world    -0.064438
5571          bin laden    -0.065974
139              10 000    -0.066501
```

Looks like positive and negative words are making some sense.

Deep Learning

I. LSTM

In this model, I decide the max_len = 400, emb_output_dims = 100. And I also add a dropout laryer to prevent overfitting. The following pictures are summary of LSTM model and the result.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 400, 100)          3470000
_________________________________________________________________
lstm (LSTM)                  (None, 32)                17024
_________________________________________________________________
dropout (Dropout)            (None, 32)                0
_________________________________________________________________
dense (Dense)                (None, 1)                 33
=================================================================
Total params: 3,487,057
Trainable params: 3,487,057
Non-trainable params: 0
_________________________________________________________________
```

```
results = model.evaluate(X_test, y_test)
print('test loss, test acc:', results)
```

```
13/13 [==============================] - 0s 25ms/step - loss:
test loss, test acc: [2.093090534210205, 0.520100474357605]
```

II. CNN

The max_len and emb_output_dim is same with LSTM. Because this is binary classification problem, the activate function of dense layer is sigmoid. The number of epochs was decided to 20.

```
Model: "functional_3"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
main_input (InputLayer)         [(None, 400)]        0
__________________________________________________________________________________________________
embedding (Embedding)           (None, 400, 100)     3470000     main_input[0][0]
__________________________________________________________________________________________________
conv_unigram (Conv1D)           (None, 400, 64)      6464        embedding[0][0]
__________________________________________________________________________________________________
conv_bigram (Conv1D)            (None, 399, 64)      12864       embedding[0][0]
__________________________________________________________________________________________________
conv_trigram (Conv1D)           (None, 398, 64)      19264       embedding[0][0]
__________________________________________________________________________________________________
pool_unigram (MaxPooling1D)     (None, 1, 64)        0           conv_unigram[0][0]
__________________________________________________________________________________________________
pool_bigram (MaxPooling1D)      (None, 1, 64)        0           conv_bigram[0][0]
__________________________________________________________________________________________________
pool_trigram (MaxPooling1D)     (None, 1, 64)        0           conv_trigram[0][0]
__________________________________________________________________________________________________
flat_unigram (Flatten)          (None, 64)           0           pool_unigram[0][0]
__________________________________________________________________________________________________
flat_bigram (Flatten)           (None, 64)           0           pool_bigram[0][0]
__________________________________________________________________________________________________
flat_trigram (Flatten)          (None, 64)           0           pool_trigram[0][0]
__________________________________________________________________________________________________
concate (Concatenate)           (None, 192)          0           flat_unigram[0][0]
                                                                 flat_bigram[0][0]
                                                                 flat_trigram[0][0]
__________________________________________________________________________________________________
dropout (Dropout)               (None, 192)          0           concate[0][0]
__________________________________________________________________________________________________
output (Dense)                  (None, 1)            193         dropout[0][0]
==================================================================================================
Total params: 3,508,785
Trainable params: 3,508,785
Non-trainable params: 0
__________________________________________________________________________________________________
```

```
score = model_2.evaluate(X_test, y_test, batch_size=32,verbose=1)
print('Test score:', score[0])
print('Test accuracy:', score[1])
```

```
13/13 [==============================] - 0s 24ms/step - loss: 0.913
Test score: 0.9132059216499329
Test accuracy: 0.5226130485534668
```

Compare & Analysis

Now, we put all of the results coming from different methods in the same table. Compare and analyze their performance.

	Accuracy
Logistic Regression Classifier	0.507537
Random Forest Classifier	0.502513
Multinomial Naive Bayesian classifier	0.507538
XGBoost classifier	0.50
LSTM	0.5201
CNN	0.522613

According to the table, we can find that overall the accuracy rate is around 50%. The performance of the first 4 classifiers is not very different. The accuracy of LSTM and CNN is slightly better than baseline models such as machine learning. But the advantage is not very obvious. It may be due to the small amount of data, different text processing methods, and the adjustment of the neural network model structure, which will bring certain errors. As a result, the overall accuracy rate is only about 50%.

Conclusion

After dealing and analyzing the top daily news from 08-08-2008 to 07-01-2016, we can easily find these existing relations between media news and stock market movement. Using machine learning and deep learning prediction model, we predict the 'adj close price' of stock will up or down in the future. In these models, we verify that positive and negative words are making some sense by using tf-vectorizer. Finally, we compare and analyze the performance of these prediction model. Meanwhile, we listed some reasons that may influence the results.

From a practical perspective, this research topic is of practical significance. Stock market prediction aims to determine the future movement of the stock value of a financial exchange. The accurate prediction of share price movement will lead to more profit investors can make. Moreover, the application of machine learning and deep learning technology in the financial field is a major development in the computer field, and it is also a hot research direction now.

Future work

All of the accuracy of these prediction models is nearly about 50%. This is not high enough, and there is room for further improvement. This may be due to two reasons. One is data. We only used the 1989 rows data to build models. Perhaps the model accuracy is not high enough due to the small amount of data. On the other hand, it is the method. Probably the structure of model we use is not reasonable enough, and there are algorithms more suitable for this model.

In the future, we plan to improve our model. Because this is a text classification problem, maybe BERT model will have better performance for this task. Besides, it is possible obtaining a more accurate results to utilize another one word to vector method.

Reference

[1] Mudinas A, Zhang D, Levene M. Market trend prediction using sentiment analysis: lessons learned and paths forward[J]. arXiv preprint arXiv:1903.05440, 2019.

[2] Checkley M S, Higón D A, Alles H. The hasty wisdom of the mob: How market sentiment predicts stock market behavior[J]. Expert Systems with applications, 2017, 77: 256-263.

[3] Souza T T P, Aste T. Predicting future stock market structure by combining social and financial network information[J]. Physica A: Statistical Mechanics and its Applications, 2019, 535: 122343.

[4] Wu G G R, Hou T C T, Lin J L. Can economic news predict Taiwan stock market returns?[J]. Asia Pacific management review, 2019, 24(1): 54-59.

[5] Bujari A, Furini M, Laina N. On using cashtags to predict companies stock trends[C]//2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC). IEEE, 2017: 25-28.

[6] Uysal A K, Murphey Y L. Sentiment classification: Feature selection based approaches versus deep learning[C]//2017 IEEE International Conference on Computer and Information Technology (CIT). IEEE, 2017: 23-30.

[7] Sohangir S, Wang D, Pomeranets A, et al. Big Data: Deep Learning for financial sentiment analysis[J]. Journal of Big Data, 2018, 5(1): 3.

[8] Singhal P, Bhattacharyya P. Sentiment analysis and deep learning: a survey[J]. Center for Indian Language Technology, Indian Institute of Technology, Bombay, 2016.

[9] Ding X, Zhang Y, Liu T, et al. Using structured events to predict stock price movement: An empirical investigation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1415-1425.