STAT7008 Programming for Data Science
Group Project

Examining Prediction Power of Media Sentiment on Grains Commodity Trends

Group 5C

CHAN Yik Yam Michael (UID: 3035813729)
LO Cheuk Ying (UID: 3035802873)
NG Hoi Ling Karen (UID: 3035803217)
POON Tsz Chung (UID: 3035802110)
WU Kwan Yu (UID: 3035275501)

Supervisor: Dr. Adela LAU

Fall 2020

1 Introduction

1.1 Background

Text-based analytic has been an evolving technique in machine learning that enables us to exploit hidden patterns and trends in textual data. With the increase in computational power, Natural Language Processing (NLP) has become the state-of-the-art framework to deal with statistical modelling on textual data. Recently, there are notable number of research that have proven the feasibility and capability of using NLP in providing insightful and profitable results.

A traditional approach to grain commodities trading relies heavily on fund managers' economic knowledge and sensitivity towards the market. Human intervention on trading decision is one of the key factor in managing a profitable growth fund. It is obvious that this is an aspect that currently lacks of automation in decision, and the incorporation of machine learning model is certainly beneficial in terms of portfolio management.

1.2 Project Objectives

The aim of this project is to gather and analyze publicly-available textual data from various media sources. In particular, news headlines from various news websites and public tweets from Twitter are to be obtained through web scraping techniques, followed by a construction of sentimental scores with an output of lexicon dictionaries. A data science framework is used to carry out a thorough analysis:

- Web scraping: Several web crawlers specializing in the crawling several websites, and API caller will be constructed to acquire data from the public Internet.
- Data visualisation: An exploratory data analysis will be conducted to explore the data acquired previously, so that a better understanding on the data can be gained before carrying out statistical analysis.
- Lexicon dictionary construction: Data cleaning and pre-processing will be done on the data set and methodologies will be introduced to construct a lexicon dictionary adjusted for the grains commodity space. The lexicon dictionaries contain lexicons and sentiment values that implies polarized sentiment. In addition, an additional named entity dictionary is constructed for specialized lexicons identification purpose.
- Statistical analysis with price trends: Statistical analysis will be carried out to investigate possible relationship between sentiment trends and price trends, so as to examine the ability of predicting price trends from sentiments data.

1.2 Scope of Project

Types of grains commodities to be examined are pre-chosen, namely oats, rice, corn, wheat and soybean. These five commodities are the major commodities in the market, and hence serve as an appropriate basis for a preliminary research on the subject.

2 Literature Review

2.1 Web Scraping

Web scraping is a common technique in acquiring data when no dedicated API exists for a particular website. There are previous research on utilizing publicly available news article for generating information in the financial markets (Jazbec et al., 2020)News article from the Internet can be either collected through proprietary scripting method, or with aid from third-party platform or software such as Common Crawl. The associated constituent companies can be identified during data pre-processing by matching the news headline to a list of relevant constituents.

Apart from news data, tweets on Twitter is also an important data source when opinion-related texts are to be collected. Similar to news data, dedicated tweets data provider exists as API from third-party platforms, such as Crimson Hexagon (Pellert et al., 2020) These platforms serve as a convenient method to acquire data in different time points.

2.2 Lexicon Dictionary Construction

In order to maximize the classification accuracy, it is important to consider the lexicon size and domain specificity. Pairwise Mutual Information (PMI) (Shapiro et al., 2020) had been proposed to measure the association between the word and sentiment classes, in order to reflect the possibility that the word contains sentiment. Church and Hanks, 1990 have defined PMI between word and sentiment class follow:

$$\mathrm{PMI}(w) = \mathrm{PMI}(w, \text{positive}) - \mathrm{PMI}(w, \text{negative})$$
$$\text{where } \mathrm{PMI}(w, c) = \log(P(w, c) \;/\; P(w)P(c))$$

which can reflect the degree of confidence that if the presence of that word makes it more, less or just as likely for the sentence as a whole to be positive, negative or neutral.

Shapiro, 2020 suggested that the sentiment scores will be noisy toward relatively infrequent features and words. However, taking average of the sentiment scores for the words can ease the problem of noise and improve the accuracy of classification.

2.2 Sentimental Analysis and Price Forecasting

Similar research on text analytic in the quantitative trading field has been carried out and similar statistical methods are widely adopted across the financial industry. Positive alphas (ie. outperformances) using sentiment-based portfolios were generated, which serves as a strong evidence towards the prediction power of public sentiment on lagged price changes. Several research on price prediction were made across different asset classes such as stock, cryptocurrency, and oil commodity.

Mohapatra et al., 2020 have built a price prediction platform on cryptocurrency with Twitter sentiments. Tweets are first processed with VADER (Valence Aware Dictionary and sEntiment Reasoner) by classifying texts as negative, neutral or positive, and positive correlation is found between tweets and its related underlying asset. Furthermore, two approaches were suggested for classification, namely machine learning-based approach and lexicon-based approach, where the former makes use of supervised model to differentiate text corpus characteristics, and the latter one directly assigns sentiment scores. The paper also suggested a time series-based modelling approach in price level forecasting, which is consistent with other research findings, such as Fronzetti Colladon et al., 2018.

Zhang et al., 2016 analyzed the relationship between news on stock reactions. In the paper, text data including news data and tweets are distilled through sentimental measures, and the sentiments are labelled by two general-purpose taxonomy databases (lexicon) and one specific on financial terminologies. The price attributes such as log volatility, volume and return are explained through positive and negative sentiments using panel regression. It has concluded that there are asymmetric impact on reaction for positive and negative sentiments across different level of news attention. The impacts on stocks are not evaluated only in terms of returns, but also volatility and volumes. This is similar to our research, whereas we include numerical data such as output levels in our investigation.

While Zhang et al., 2016 concluded their research using panel regression, the adequacy of the model may not be the best among all other statistical methods, such as time series modelling. Fronzetti Colladon et al., 2018 investigated prediction power of online media sources towards crude oil price, which is also in the commodity price, and compared several approaches in model fitting. Apart from regression and support vector machine (SVM) models, ARIMA and ARIMAX times series models that incorporates text mining results is also an adequate choice in statistical analysis related to sentiments.

These findings from various research have proven the possibility of extracting insightful results through the proposed framework of this project, and serve as references for the statistical methods that shall be adopted. There are more evidences on the ability of predicting price trends using sentimental analysis in more popular universes such as stock trading, and some had predicted up-or-down trends in stock indices (DOW Jones Industrial average in particular) with remarkable accuracy (Bollen et al., 2011). Yet, specific hypothetical investigations on commodity trading patterns is a domain that lacks of attention from the academia.

3 Data Science Methods

3.1 Data Sources, Data Cleaning and Pre-processing

Data Sources

- News data: News headlines data is obtained with proprietary web crawling. Three news website have been pre-chosen for web crawling, namely *Successful Farming, Financial Post* and *Ag Web.*
- Twitter data: A web crawler from the Internet is used to crawl public tweets data.
- Grains commodity data: The price data of the five grains commodity within scope is acquired by public dataset from Quandl, which contains historical daily price data.

In addition, a list of keywords (*Grains, Soybeans, Oats, Rice, Corn Wheat*) are specified.

The crawling of news websites can be achieved using the web-related packages in Python, such as *requests*, *selenium* and *beautifulsoup*.

Data Cleaning and Pre-processing

The text data are further processed by data cleaning and standard text pre-processing procedures. Data are filtered according to their respective timepoints, and unwanted texts are filtered with aid of packages such as *datetime* and *re*. Furthermore, tokenization, stemming and lemmatization were applied to the data for pre-processing purpose using packages such as *nltk* and *sklearn*. Stop words removal is to be carry out to clean texts that do not contribute any sentimental values.

3.2 Models Design

Lexicon Dictionaries Construction

The procedure of constructing the lexicon dictionaries is as follows:

1. Obtain four sentiment scores from pre-existing analyzers or dictionaries to obtain rough estimates of sentiments for each rows of headlines and an integrated score is given by a proprietary formula
2. Determine the general sentiment of news headlines and classify news headlines into three sentiment levels (positive, negative, neutral) and obtain a corpus for each sentiment levels
3. Carry out named entity recognition from news headlines to obtain a lexicon dictionary for the grains commodity space
4. Calculate Pointwise Mutual Information (PMI) scores on words from the corpus, and obtain a lexicon dictionary for each of the three sentiment levels

Four pre-existing analyzers or dictionaries, namely VADER, TestBlob, Harvard IV-4 (HIV4) Dictionary and McDonald (LM) Financial Dictionary are used to determine the rough estimate of headline sentiments. The formula for integrated score is as follows:

$$\text{Integrated score} = |\text{HIV4 score} + \text{LM score}| * \text{sign}(\text{TextBlob score}) \ / \ 2$$

However, various kinds of pros and cons can be found in each analyzer in assigning sentimental values. The summary is illustrated as follows:

	VADER analyzer	*TextBlob analyzer*	*HIV-4 dictionary*	*LM dictionary*
Attributes	Polarity	Polarity and subjectivity	Score	Score
Purpose	Detect emotions	General purpose	General news	Financial news
Works with	Tokenized texts	Overall sentences	Tokenized texts	Tokenized texts
Pros and Cons	Ignores part-of-speech	Can detect part-of-speech (eg. Negations)	Adjusted for news vocabularies	Adjusted for financial vocabularies

Table 3.1 Comparison of four analyzers or dictionaries

It has been found that while VADER, HIV-4 dictionary and LM dictionary perform well in sentiment intensity recognition, they ignores part-of-speech which is crucial in determining overall polarity, while TextBlob perform worse in sentiment strength recognition, it is generally good at detecting polarity. Hence, an integrated score is used instead of incorporating four different values obtained from the analyzers or dictionaries. On the other hand, to handle the data in twitter, VADER is adopted separately as there are literature evidences showing the outperformance of VADER than other approaches.

With the rough estimate of headlines' sentiments, they can then be classified into sentiment levels. The sentiments ranges from -1 to 1, and a threshold of -0.25 and 0.25 is manually set for classification. PMI is then carried out to calculate vocabulary-level sentiments and the three sentimental lexicon dictionaries with sentimental values are constructed according to the vocabulary-level sentiments.

Vocabularies in the sentimental lexicon dictionaries are to be constructed according to word frequencies and named entity recognition (NER) technique. Named entities are listed in the dictionary whenever the vocabulary frequency exceeds a pre-determined threshold. Named entity includes tokenized lexicons of names, geographical locations, company names or special words related to grains commodity.

<u>Sentimental Analysis and Price-Sentiment Trends Analysis</u>

The final score for each headlines are determined according to the sentimental values in the obtained from the lexicon dictionaries in the previous procedure. This allows us to examine the corresponding relationship between price trends and the sentiment scores across time. Linear regression model is applied to examine the relationship between log-return and *n*-lagged sentimental values in time periods.

Note that the sentimental values are aggregated by types of grains and various time periods. In the investigation, different time periods (*n*-daily, *n*-weekly, or *n*-monthly) are set in interpreting the sensitivity of market response. This also suppresses the adverse effect of noisy day-level sentiment data.

The linear regression model is fitted using ordinary least squares method (OLS) from package *statsmodels.api*. The sentiment values exhibits prediction power if a significant result is obtained from the regression model.

4 Result Summary

<u>4.1 Web Crawling</u>

With using the our proprietary web crawlers (see separate Python files) and a Twitter API, an approximate number of 30,000 news data and 140,000 tweets were obtained. The time period of data ranges from 2006 to December 2020. The news data contains information about date of publishing, headlines, news categories and link reference to news content, while the tweets data contains information about date of creation, tweets content, and hashtags. Below are examples of data frames of the two types of text given:

headline	url	query	date	type	hashtag	source
Wheat down 3%; soy, corn also sag on pre-holid...	/pmn/business-pmn/wheat-down-3-soy-corn-also-s...	corn	25/11/2020	PMN Business	None	news
Wheat hits near three-week high on USDA's crop...	/pmn/business-pmn/wheat-hits-near-three-week-h...	corn	25/11/2020	PMN Business	None	news

Table 4.1 News headline data

date	headline	hashtags	sources	query	url	type
27/2/2020	Pain more in the overnights for #grains, #equi...	[grains, equities, energies, coronavirus]	tweets	grains	None	tweets
4/9/2017	1878: #Iowa led the #US in #corn #production.\...	[Iowa, US, corn, production, agriculture, agri...	tweets	grains	None	tweets

Table 4.2 News headline data

4.2 Lexicon Dictionaries

According to the methodology stated in (3.2), four lexicon dictionaries (positive/neutral/negative sentiments and NER dictionary) were constructed. Each lexicons in the sentimental dictionaries are associated with an unnormalized sentimental score which is calculated with PMI. Below shows the top results from the dictionaries:

```
{'success': 4.0984485163109525,
 'better': 3.5998929851367762,
 'advanc': 3.105196743300669,
 'posit': 2.9773633717907844,
 'counti': 2.7195342624886845,
 'fungicid': 2.627160942357669,
 'clear': 2.627160942357669,
 'complet': 2.577564001218297,
 'target': 2.560469567858997,
 'approv': 2.560469567858997,
 'treatment': 2.525378248047727,
 'conserv': 2.489010603876852,
 'win': 2.451270275894005,
 'profit': 2.4357660893580393,
 'program': 2.391846855423204,
 'lead': 2.350174159022636,
 'gener': 2.3286679538016726,
 'bargain': 2.3286679538016726,
 'minist': 2.3286679538016726,
 ...
```

```
{'crude': -4.208023643789632,
 'bad': -3.897868715485793,
 'fail': -3.8978687154857927,
 'due': -3.8978687154857927,
 'troubl': -3.8025585356814684,
 'daili': -3.350325689639035,
 'lose': -3.125940657072733,
 'deficit': -3.109411355121523,
 'small': -3.004050839463696,
 'choppi': -3.004050839463696,
 'sharp': -3.004050839463696,
 'oppos': -3.004050839463696,
 'advers': -2.8862678038073133,
 'heavi': -2.8862678038073133,
 'jitter': -2.886267803807313,
 'weak': -2.8668497179502115,
 'doubt': -2.808306262337601,
 'weaken': -2.7527364111827906,
 'ring': -2.7527364111827906,
 'alleg': -2.7527364111827906,
 ...
```

Figure 4.1 Top positive lexicons (left) and negative lexicons (right)

```
['uschina trade news',
 'kim',
 'qa bryan',
 'brazilian',
 'northwest il',
 'soil food web analysis',
 'syrian',
 'alltech',
 'naftas',
 'nigerian',
 'oregon',
 'west african',
 'toban dyck',
 'refine higholeic soybean',
 'sichuan province',
 'wyoming',
 'dallas',
 'pepsico',
 'primer nitrogen',
 'mexican',
 'france',
 'spring',
 ...
```

headline	lexicon_score
Positive ROI Drives Bidders to Farmland Auctions	4.597282
Converting CRP land to grain crop production t...	4.593807
Brazil deems truck stops 'essential' to keep g...	4.592483
May Successful Marketing Newsletter Is Out	4.590827
Farm Management Software: No easy way out	4.589187
Farm Bureau president urges nutrient management	4.587587
Money Managers Increase Net-Long Bets in Soybe...	4.586822
A farm land value bubble?	4.586429
BASF introduces new soybean seed treatments	4.586216
Dow AgroSciences Announces Enlist Weed Control...	4.584143
Soy extends rally as China keeps buying; USDA ...	4.583459

Figure 4.2 Named entities (left) and fitted news headlines sentimental score (right)

4.3 Price-Sentiment Trend Analysis

The investigation of relationship between price trends and sentiments are set at the base assumption of using OLS linear regression. Yet, there are flexibility in choice of time period aggregation (such as daily, weekly, bi-weekly and monthly) when exploring and determining the independent variables, according to the belief of market sensitivity. A shorter-term time period aggregation focuses on higher-frequency movements while longer-term time period aggregation focuses on the slow-moving trends. On the other hand, filters on headlines can be either applied or not applied according to the grains types (soybeans, oats, wheat, rice, corn) depending on whether the overall grains commodity market or specific grains commodity market is needed for investigation.

It has been found that the soybean sentiment values are predictive in forecasting price trends of the soybeans commodity market. Below shows the result of the fitted regression:

	lag-1 score	*lag-2 score*	*lag-3 score*	*lag-4 score*	*lag-5 score*	*lag-6 score*
coefficient	-0.0626	-0.0679	0.0002	0.0186	0.0251	0.0393
p-value	0.025**	0.003**	0.991	0.342	0.165	0.101
R^2 value	0.659					

Table 4.3 Regression result of soybean commodity price trends on soybean sentiments

The above regression is based on a 2-month aggregation of sentiment values. A significant coefficient on lag-1 score and lag-2 scores implies a predictive power of lag-1 sentiment (2-month) and lag-2 sentiment (4-month) in forecasting the price trend. With a relatively long period of aggregation time period, this reflects the sentiments prediction in the slower-moving trend of the soybean commodity market.

On the other hand, in examination of linear regression models with higher-frequency periods, it has been found that the 3-time-period-lagged overall grains commodity sentiment is predictive in price trends of the corn commodity. Below shows the result of the fitted regression:

	lag-1 score	*lag-2 score*	*lag-3 score*	*lag-4 score*	*lag-5 score*	*lag-6 score*
coefficient	0.0011	-0.0009	0.0032	-0.0019	0.0003	-0.0010
p-value	0.379	0.457	0.013**	0.148	0.832	0.445
R^2 value	0.034					

Table 4.4 Regression result of corn commodity price trends on grains commodity market sentiments

The above regression is based on a 4-day aggregation of sentiment values. The significant p-value on lag-3 sores implies a predictive power of lag-3 sentiment (12-day) in forecasting the price trend. These are the evidences that the market sentiments is somewhat predictive in forecasting price trends in both low-frequency and high-frequency basis.

5 Discussions

5.1 Lexicon Dictionaries

From Section (4.2), the lexicon dictionaries are successfully generated with adjusted scores and adjusted context for the scope of grains commodity market. It can be seen that the sentiment values assigned to the words are consistent with human sentiment. For instance, 'success' and 'better' are some positive words in nearly all news headlines or tweets that might contain them, and the lexicon dictionary generated is able to capture these strongly positive sentiments. On the other hand, vocabularies such as 'bad' and 'fail' are the vocabularies contained in the negative sentiment dictionary with top magnitude of polarity. These are the words that very likely to represent a negative sentiment in an actual news headline. Hence, the data science methodology designed for this project's purpose provides us a certain degree of confident of accuracy.

5.2 Price-Sentiment Trend Analysis

To examine whether the sentiment scores are able to predict future price trends, several linear regression models with a variety of number of lags, independent variables (market or specific types of grains) and period of aggregation (n-daily, n-weekly, or n-monthly) were carried through. It has been found that not all combinations of the above mentioned parameters produces significant results. Yet, several settings of the models provide insights for further investigation.

In Section (4.3) we have fitted two regression models with different parameter settings, where each of them emphasizes in the examination of predictive power under different commodity market sensitivity scenario as an assumption. The two models have shown a satisfactory prediction power on specific types of grains commodity, especially on soybean and corn with overall market sentiment. The former model related to soybean shows a remarkable R^2 value of 0.659 and significant coefficients with p-value smaller than 0.05 on first two lagged sentiment variables, hinting a pattern somewhat similar to autoregressive models, which shows a result consistent to other literatures such as Zhang et al., 2016, mentioned in Chapter 2. The latter model related to corn aimed to examine the quicker moving time period of aggregation. Significance correlation was found between log return of corn price with the lag-3 overall grains commodity market sentiment, with a p-value smaller than 0.05. It can be interpreted that the trend of sentiment score related to overall market also has a predictive power on corn price trend in shorter time period. However, this result suffered from R^2 value, which reflect that the information contained in sentiment score data is not solely enough to explain market price trends. This is consistent with our belief that models can be more accurate if more logical factors are included, such as lagged price data and volume trends. For the purpose in examining predictive power of sentiment scores, it is reasonable for a lack of adequate regression fit.

6 Conclusions

This project aims to create lexicon databases through web crawling on online media websites and examine the predictive power of sentiments score using our proprietary lexicon dictionaries. The project objectives are achieved through the proposed data science workflow explained in Chapter 3. Web crawling of historical headlines and tweets on media and news websites is first done. Lexicon dictionaries using analyzers, dictionaries, with proposed adjustment formula and Pointwise Mutual Information method are constructed as in Section (3.2). The lexicon dictionaries with assigned sentiment scores in vocabulary-level are used to determine final score for crawled headlines. Lastly, statistical analysis on price-sentiment data are conducted as an evidence towards the usability of the lexicon dictionaries and accuracy of sentiments, while demonstrations of prediction power of sentiments are examined under linear regression statistical framework in Section (4.3). A fairly good prediction result towards soybean and corn commodity space is found under a certain parameter settings as discussed in Chapter 5.

Nevertheless, it can been observed that the results quality of this project are restricted and limited due to several reasons:

- Data Availability: In can be observed that the crawled data suffered from imbalance in availability in different time periods. There are readily more data available in time points nearer to present than the others. A more thoughtful and thorough method should be considered in order to obtain more data in historical time points.
- Data Quality: This project considered only news headlines as a representation of sentiment of news. The sentiment scores can be found more accurate and representable if the news content can also be crawled. Due to computational capacity reasons, the news content were not included. Yet, URLs of news headlines were recorded for this purpose.
- Data Source Inadequacy: With a budget, time and labour constraint, only three news website were included as news data sources. This may result in bias appearing in the news. Financial news collection providers can be a better option in collecting news data, instead of manually choosing websites.
- Subjectivity in Statistical Model: The choice of OLS linear regression results in an easy-to-interpret result. However, the choice of linear regression as fitted model is not thoroughly supported by any statistical evidences. This have to be carefully discussed in order to robustly support any evidences of insightful results. Yet, the discussion is beyond the scope of this project.

7 References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on language in social media (LSM 2011) (pp. 30-38).

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2 (1), 1-8. https://doi.org/10.1016/j.jocs.2010.12.007

Chowdhary, K. R. (2020). Natural language processing. In Fundamentals of Artificial Intelligence (pp. 603-649). Springer, New Delhi.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., ... & Štajdohar, M. (2013). Orange: data mining toolbox in Python. the Journal of machine Learning research, 14(1), 2349-2353.

Dunis, C. L., & Nathani, A. (2007). Quantitative trading of gold and silver using nonlinear models. Neural Network World, 17(2), 93.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.

Fronzetti Colladon, A., Elshendy, M., Battistoni, E., & Gloor, P. (2018). Using four different online media sources to forecast the crude oil price. Journal of Information Science, 44, 408-421. https://doi.org/10.1177/0165551517698298

Jazbec, M., Pasztor, B., Faltings, F., Antulov-Fantulin, N., & Kolm, P. N. (2020). On the impact of publicly available news and information transfer to financial markets.

Lawson, R. (2015). Web scraping with Python. Packt Publishing Ltd.

Li, Q., Wang, T., Gong, Q., Chen, Y., Lin, Z., & Song, S. K. (2014). Media-aware quantitative trading based on public Web information. Decision support systems, 61, 93-105.
Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of natural language processing, 2(2010), 627-666.

Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Information Processing & Management, 102212.

Mohapatra, S., Ahmed, N., & Alencar, P. (2020). Kryptooracle: A real-time cryptocurrency price prediction platform using twitter sentiments.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), 544-551.

Pellert, M., Lasser, J., Metzler, H., & Garcia, D. (2020). Dashboard of sentiment in austrian social media during covid-19.

Richardson, L. (2007). Beautiful soup documentation. Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018].

Salunke, S. S. (2014). Selenium Webdriver in Python: Learn with Examples. CreateSpace Independent Publishing Platform.

Seabold, S., & Perktold, J. (2010, June). Statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference (Vol. 57, p. 61).

Shah, D., Isah, H., & Zulkernine, F. (2018, December). Predicting the effects of news sentiments on the stock market. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 4705-4708). IEEE.

Shapiro, A. H., Sudhof, M., & Wilson, D. (2020). Measuring News Sentiment (pp. 1-49, Working paper). Federal Reserve Bank of San Francisco.

Van Rossum, G. (2007, June). Python programming language. In USENIX annual technical conference (Vol. 41, p. 36).

Zhang, J. L., Hardle, W. K., Chen, C. Y., & Bommes, E. (2016). Distillation of news flow into analysis of stock reactions. Journal of Business Economic Statistics, 34 (4), 547-563. https://doi.org/10.1080/07350015.2015.1110525

8 Appendix

Please proceed to the separate Python files and Jupyter notebooks.