The Financial Attention Index to Measure Impact of Crisis from Microblog

Zhongsheng Wang, Kiyoaki Shirai

School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology

wzswan1.0@jaist.ac.jp, kshirai@jaist.ac.jp

Abstract

This paper proposes a new financial index called Financial Attention Index (FAI) to measure an extent of a financial risk. A stock market drastically changes when many novice investors participate in it. At the same time, they often posts messages on Social Networking Service. Therefore, the FAI is calculated as ratio of financial related comments on microblog to detect a financial crisis. Furthermore, we train a model to predict future stock prices from a history of stock values and the FAI. Results of experiments show effectiveness of our proposed index to capture an impact of a financial crisis.

Keywords: Financial Index, Financial Crisis, Social Networking Service, Machine Learning

1. Introduction

A financial crisis, such as a drastic drop of stock prices, may cause considerable losses to many investors. At the same time, due to globalization, a financial crisis in one country may influence stock markets across the whole world. Because a stock market is an innate complex, dynamic and chaotic, the management of financial risk has been proved to be a very difficult task. For many decades, researchers have tried to analyze historical stock prices or a company's financial statements to measure financial risk (Fama et al., 1969; Fama, 1991; Cootner, 1964). However, the results were still not quite helpful for risk management.

In the Socio-economic Theory of Finance (Prechter Jr et al., 2012; Prechter Jr and Parker, 2007), irrational speculation behaviors play an import role in a financial crisis. Social Networking Service (SNS), such as Twitter or Weibo, are now widely used by large numbers of people. These social networks provide us with a considerable amount of information that can be used to monitor the financial market and which can also be used to predict a financial crisis.

The goal of this present research is to propose a new financial index that measures the extent of a financial crisis. Microblogs are used as a source of our new index. The hypotheses behind our financial crisis index are as follows.

- Not many investors focus on financial markets daily. When a bull market begins and stock prices keep going up, more beginners come into the markets.

- These people are intense and they make irrational speculation decisions. This may cause a panic and this can lead to a bear market. The behaviors of beginners may make a financial risk more serious.

- When many novice investors come into the market, they post messages about financial topics on SNSs. Therefore, the intensity of attention towards finance on microblogs can have a positive correlation with the level of a financial bubble. In particular, the numbers of financial messages posted on SNSs can be used as a financial crisis index.

- This index can be used as an observable financial market singularity for risk management.

In this paper, we will also attempt to use our proposed financial crisis index to predict the financial market trend for risk management. Specifically, a model to predict future stock prices is trained with past stock prices and our index that is derived from texts posted on a microblog. Given that historical prices are sequential data, Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) have been used to predict a stock's price (Chen et al., 2015; Heaton et al., 2017). We train the LSTM and measure its predictive ability to evaluate the effectiveness of our proposed financial crisis index.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 explains our proposed financial index for risk management. Section 4 presents the training of the LSTM for prediction of stock values. Section 5 reports the results our experiments and evaluates our proposed method. Finally, Section 6 concludes this paper.

2. Related work

2.1. Financial analysis theory

Financial market analysis has been one of the most attractive areas of previous research. Many researchers have tried to interpret the current financial situation from several different academic perspectives. The econometrics theory model is the most famous and influential of these methods. In addition, many financial market studies have been based on Fama's Efficient Market hypothesis (Fama et al., 1969; Fama, 1991). Although this hypothesis considers that the current price of an asset always reflects all of the previous information available for it instantly, it is impossible for us to collect all of the necessary information to make a prediction of the future price.

The other famous economic theory is the Random-walk hypothesis (Cootner, 1964; Malkiel, 1973), which claimed that a stock price changed independently of its history. However, information other than historical prices can also be used for stock price prediction, such as the financial

news that is released every day. At the same time, this theory considers that it is impossible to predict a financial market. In contrast, many previous studies have already proven that the stock market only followed those theories during specific periods (Glantz and Kissell, 2013).

2.2. Use of textual data for financial market analysis

The previous studies have only worked on history data and past stock prices. However, many other factors can be taken into consideration when analyzing the market. For example, thanks to their widespread use, textual information from Web forums and SNSs can be used for market analysis.

Nguyen et al. proposed a method based on sentiment analysis on social media to predict the movement of stock prices (Nguyen and Shirai, 2015; Nguyen et al., 2015). A new topic model, called Topic Sentiment Latent Dirichlet Allocation (TSLDA), infers topics and their sentiments simultaneously and has been incorporated into the prediction model.

Jaramillo et al. proposed a method to predict a stock price using a history of prices, and also the polarity of company reports and news (Jaramillo et al., 2017). In this study, the polarity of the texts is identified by Support Vector Machine (SVM).

Jianhong et al. applied a deep learning method on sentiment-aware stock market prediction (Li et al., 2017). They tried to analyze sentiment of documents in a stock forum with a Naive Bayes model. They then trained an LSTM neural network to a predict stock value using the results of the sentiment analysis as an input.

Similar to these previous studies, we also use textual information for a stock market analysis. However, our main focus is not to predict a stock's value but instead to avoid financial risk.

2.3. Prediction of financial market and risk

Machine learning is often applied for financial market forecasting with historical data. These methods are important related work because this paper also applies machine learning for stock movement prediction with our proposed index. For example, Nelson et al. used LSTM neural networks to predict stock market price movement (Nelson et al., 2017). In addition, several methods have proposed to apply deep learning for multivariate financial series (Batres-estrada, 2015; Heaton et al., 2017). Chen et al. used an LSTM-based method for China stock market return prediction (Chen et al., 2015).

Although most researchers and investors care about a good return in the financial market, managing risk can be more important because it can help to avoid massive loss. Some of the previous work on the financial risk management was based on historical prices only; however, several studies have also tried to use textual analysis.

For example, Niemira and Saaty proposed a method to build Analytic Network Process model for financial crisis forecasting (Niemira and Saaty, 2004). This paper trained a turning point model to forecast a financial crisis likelihood based on an Analytic Network Process framework.

Meanwhile, Oh et al. proposed a method to use neural networks to support early warning system for financial crisis forecasting (Oh et al., 2005). Using nonlinear programming, the procedure of DFCI (daily financial condition indicator) construction is calculated by integrating three sub-DFCIs, which are based on different financial variables.

Trusov et al. used company financial reports in a multi-representation approach to text regression of financial risks (Trusov et al., 2015). Finally, Kogan et al. also used financial report regression for financial risk prediction (Kogan et al., 2009). In particular, they used Support Vector Regression (SVR) as a prediction model.

Although these previous studies have tried to use textual information for financial risk management, texts on SNS were not given attention. In this paper, comments on Weibo are used as a source of textual information for risk management.

3.   Financial Attention Index

3.1.   Definition

We propose to use the Financial Attention Index (FAI) to measure the extent of a financial risk. The FAI is defined as in Equation (1).

$$FAI \stackrel{def}{=} \frac{\text{number of financial related comments on SNS}}{\text{total number of comments on SNS}} \quad (1)$$

As discussed in Section 1, we suppose that financial topics are to more mentioned and discussed on SNSs when many novice investors participate in the market, which may cause market instability. Therefore, the FAI is assumed to be positively correlated to a financial risk. Although a financial crisis can be measured from various points of view, the FAI can only be used a financial crisis index from one perspective.

Figure 1 shows the procedure to calculate FAI. The comments on SNS are classified to determine whether or not they are related to financial topics. In this study, we will use the Weibo. The number of financial comments and total comments are then counted to get the FAI. However, a classifier of financial related comments is not trained from Weibo comments and, instead, the labeled data of news articles is used for training.
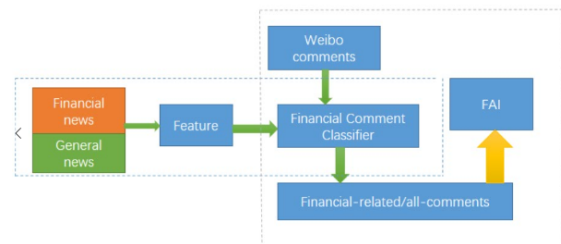

Figure 1: Procedure to calculate FAI

In our research, FAI is calculated for every week; the number of comments posted in a period of one week is counted to get FAI.

3.2. Calculating FAI

3.2.1. Type of classifier

Two kinds of financial classifiers are trained.

- Two-way classifier

  This classifies a comment on Weibo as a financial or non-financial related comment.

- Three-way classifier

  This classifies a comment on Weibo into three classes: (1) Stock-related, which is a comment related to the stock market, (2) Financial-related, which is a comment related to financial markets such as future market, bond market and so on, but not related to the stock market, and (3) Other, which is a comment not related to financial topics. Because we mainly focus on analysis of stock prices to detect a financial crisis, we distinguish topics about stocks with other financial related topics. When the three-way classifier is used for calculation of FAI, both stock-related and financial-related comments are treated as financial comments.

3.2.2. Data collection

Two kinds of data are collected to calculate FAI.

Weibo dataset

Comments on Weibo posted from 2013-7-1 to 2016-12-5 are crawled. This period contains stable, bull, and bear markets, as reported later. The number of collected comments is 2,104,746. They are collected with their posting time to calculate the FAI for each period of one week.

News dataset

News articles are collected from two sources: the Tencent news website and the text collection of THUC Project. Tencent includes the websites of general news[1] and financial news[2]. In the latter, news stories are categorized into several topics. The news in the topic category of "The New Third Board"[3] are used as stock-related documents, while the news stories in the other topic categories are financial-related documents. News stories in the general news website are crawled as other (non-financial) documents. The THUC text collection is developed by Tsinghua University (Sun et al., 2006). In this dataset, each document is annotated with its topic (stock related, financial related and other). Table 1 shows the number of documents in the news dataset.

The news dataset is used to train both the two-way and three-way financial classifiers. When the two-way classifier is trained, news articles of the stock and financial classes are treated as the financial related documents.

_____

[1]http://news.qq.com/

[2]http://finance.qq.com/

[3]A name of stock market in China

Table 1: News dataset

Class	Website	THUC Project
Financial related	800,000	5,000,000
Stock related	5,000	200,000,000
Other	165,000	500,000,000

3.2.3. Training the classifier

SVM is used to train the financial classifier. SVM has been widely applied in the classification of documents, such as sentiment analysis. It is considered as the most appropriate learning algorithm for unbalanced datasets with a large number of features. A linear function is chosen as the kernel function of SVM. The gensim (Řehůřek and Sojka, 2010), sklearn (Pedregosa et al., 2011), scipy (Jones et al., 2001) and jieba (Rossum and Guido, 1995) [4] tools are used to train SVM.

Bag-of-words are used as features for training SVM. The bag-of-words model is a simplified representation of a document, which is widely used in natural language processing and information retrieval (McTear et al., 2016). In information retrieval, a weight of an index term is often determined by the TF-IDF (term frequency-inverse document frequency), which reflects how important the term is in a text collection (Jaramillo et al., 2017). In this study, the function words are removed by preprocessing. All content words in a document are extracted as the features. The weight of each feature is set as the TF-IDF score.

We found that the number of the features was high; that is, nearly 200,000. Therefore, we apply Latent Semantic Analysis (LSA) to reduce the feature space. LSA is used to reduce the size of a matrix of words by documents using Singular Value Decomposition (SVD). In this study, the number of the features is reduced to 50.

4. Stock price prediction with FAI

To evaluate its risk management ability, the FAI is used to predict a stock index. LSTM (Xiong et al., 2015) is chosen to train the prediction model because it is well used in the prediction of time series. In our model, the input of LSTM is a time sequence of either the stock index, a difference of the stock index, or FAI. The difference of the stock index is defined as a change of the stock index between the current and previous periods. We also train a model where these three kinds of values are concatenated as a vector and passed to the input layer of LSTM. The output of LSTM is a stock index of the next period. We define a period of LSTM as one transaction day. Note that FAI is calculated for each week. If the FAI is used for LSTM, then the same value is entered during days in a week. Our LSTM structure consists of one input layer, two LSTM layers and one output layer. The input and output layers consist on one neuron node. The first and second LSTM layers contain 5 and 100 nodes, respectively.

LSTM is learned through training by the Python deep learning library Keras (Chollet, 2015). The activate function in

_____

[4]It is used for word segmentation of Chinese texts.

LSTM units is 'linear'. The model is trained by the rmsprop method with 1 example in a batch, with categorical cross entropy as the objective loss function. The validation fraction is set as 0.1%. The learning rate is set as 0.001. All of the initial weights are set to be small positive constant values. To prevent overfitting, a dropout is set at 20% and an L2 regularization constraint is set as 0.01.

5. Evaluation

5.1. Classification of financial comment

The classifier to judge whether a comment is related to financial topics takes an import role in FAI. First, the financial classifier is empirically evaluated by a 10-fold cross validation on our news dataset. The performance of the classifier is measured by the accuracy, which is defined as a ratio of the number of correctly classified comments to the total number of comments.

Table 2 shows the accuracy of the two-way and three-way classifiers. Recall that the comments are classified as either "financial-related" or "other", even when the three-way classifier is used, while both stock-related and financial-related comments are regarded as financial-related comments.

Table 2: Result of classification of financial comments

Model	Accuracy
Two-way classifier	83.35%
Three-way classifier	86.18%

The performance of the financial classifiers is satisfying. We found that the three-way classifier outperformed the two-way classifier and, therefore, the three-way classifier is used in our next experiments.

5.2. Correlation between FAI and the real stock index

To evaluate how well the FAI can work when used to predict a financial crisis, the correlation between FAI and a real stock index is measured. Given that our FAI is derived from comments of Weibo, which is a Chinese microblogging service, the SSE Composite Index (SCI) is chosen as the stock index in this experiment. The SCI is computed from the stock prices of Chinese companies. It is a tool that is widely used by investors to describe the market.

The SCI values are obtained from the Finance Sina website[5]. The dataset contains the SCI values of 679 trading days during 2013-10-28 to 2016-8-2, which is almost the same period where the comments on Weibo are downloaded.

The F-test, which is a common method for statistical test, is applied to measure the correlation between FAI values and the real SCI values for 679 trading days. We also apply the F-test between FAI and the difference of the stock index, in addition to the stock index and the difference.

Table 3 shows the results of the F-test. It can be seen that FAI strongly correlates with the stock index, but not with the difference of the stock index. It seems difficult to measure the change of the stock index using only FAI.

---

[5] http://finance.sina.com.cn/

Table 3: Result of F-test

Two variables	F-test
FAI vs. Stock Index	0.000111
FAI vs. Difference	0.0155
Stock Index vs. Difference	0.00718

5.3. Prediction of stock index movement

5.3.1. Experimental setting

Our LSTM-based stock index prediction model is evaluated for its ability to predict not the stock index but, instead, a movement of the stock index. More precisely, we consider a task to classify a movement of the stock index in one week into one of the following three classes.

Up : This is a case where the stock index goes up by T or more, as shown in (2),

$$P_e - P_s > T. \tag{2}$$

, where P_s and P_e are the stock index at the start and end of the period, respectively.

Keep : This is a case where the stock index does not change drastically as shown in (3),

$$|P_e - P_s| \leq T. \tag{3}$$

Down : This is a case where the stock index goes down by T or more, as shown in (4),

$$P_e - P_s < -T. \tag{4}$$

We set T as 0.02 in this experiment.

We have chosen 120 continuous weeks in our stock index dataset to be used as test periods. For each test period, all of the past values are used as the training data as shown in Figure 2. Note that more training data is available when the stock movement of the later week is predicted. For the first test period, data of previous 80 transaction days is prepared for training. We chose this experimental setting so that more data can be used for training the LSTM.

5.3.2. Evaluation criteria

Two evaluation criteria are used in this experiment:

Accuracy

This is a proportion of the number of the test periods for which the predicted movement class agrees with the true class.

No Lost Accuracy

The main goal of this research is not to help investors get a good return but to help them avoid financial risks. Consequently, it is more important to predict the movement "Down" to avoid making a loss. Therefore, we introduce another criterion, which is called "No Lost Accuracy". This is the accuracy of the stock movement prediction task where the test periods are classified as either "Down" or not. That is, "Up" and "Keep" classes are merged into one class: "Not-Down".

Figure 2: Test and training periods of prediction model

5.3.3.   Result and discussion

Table 4 shows the Accuracy (A) and No Lost Accuracy (NLA) of four different prediction models. Here, "SI", "DI", "FAI" stand for the model using only the stock index, the difference of the stock index, and FAI, respectively. "All" stands for the model using three values.

Table 4: Results of the stock movement prediction

Model	SI	DI	FAI	All
A	31.1%	42.9%	41.2%	35.3%
NLA	47.9%	62.2%	54.6%	57.1%

From these results, it can be seen that "FAI" is better than "SI" in terms of Accuracy and No Lost Accuracy. This indicates that FAI is effective and is able to predict the movement of the stock index. When combining FAI with other indexes, the No Lost Accuracy is also improved ; however, "FAI" does not outperform "DI". Therefore, it is found that the difference of the stock index is a strong indicator of the movement of the stock market.

Figure 3 shows a change of SSE Composite Index in our dataset. Both a bull market and a bear market are found in this graph. To evaluate the performance of the prediction models in different situations (bull market, bear market etc.), we divide the test periods into 12 terms, where each term consists of 10 test periods (10 weeks), and we then measure an average of Accuracy and No Lost Accuracy on each term. The results are shown in Table 5. The tables include the situation, which is graphically shown in Figure 3, for each term.

The results of our experiment show a tendency for the "FAI" model to achieve better performance than the other models in both a bull and a bear market. In particular, it is good on a front bull situation (T5). In terms of a bear market (T8, T9, and T10), the "FAI" model is better than or comparable to the others. These results indicate that FAI is effective and it is able to predict a drastic movement of the market. In contrast, the model using the difference of the stock index (DI) works well for stable situations. The model using the stock index (SI) has the worst results in our experiment. Unexpectedly, the "All" model is not al-

Table 5: Prediction for different situations

(a) Accuracy

Term	Situation	SI	DI	FAI	All
T1	stable	50%	60%	40%	40%
T2	stable	10%	60%	50%	40%
T3	stable	30%	40%	40%	20%
T4	prim bull	30%	70%	10%	50%
T5	front bull	40%	30%	90%	50%
T6	mid bull	40%	30%	40%	40%
T7	late bull	20%	50%	40%	30%
T8	front bear	20%	30%	40%	0%
T9	mid bear	10%	40%	30%	50%
T10	late bear	40%	20%	50%	10%
T11	stable	30%	40%	40%	50%
T12	stable	44%	56%	22%	44%

(b) No Lost Accuracy

Term	Situation	SI	DI	FAI	All
T1	stable	70%	60%	50%	70%
T2	stable	10%	80%	60%	70%
T3	stable	70%	60%	50%	20%
T4	prim bull	50%	90%	40%	50%
T5	front bull	60%	50%	100%	80%
T6	mid bull	50%	50%	50%	70%
T7	late bull	30%	60%	40%	50%
T8	front bear	40%	50%	50%	50%
T9	mid bear	30%	60%	70%	60%
T10	late bear	50%	50%	50%	40%
T11	stable	50%	70%	50%	70%
T12	stable	67%	67%	44%	56%

ways the best choice. Although three values are simply combined as one input vector in our model, this might be a bit too naive.

6.   Conclusion

This paper investigates whether the financial attention of investors as measured from a large collection of comments on Weibo could predict a down occasion on the SSE Composite Index (SCI). The FAI is defined as a proportion of the financial related comments and estimated by the financial classifier trained from the news articles. The FAI, in addition to the stock index and the difference of the index, were used as the input of the LSTM model to predict the future stock index.

The results of our experiment showed that the accuracy of the LSTM with FAI was no better than the model with the difference of the stock index on average but it was better or comparable in bear markets. Therefore, the FAI can be an effective index to predict a financial risk and this can help investors to avoid making a massive loss. In addition, the FAI can also effectively predict the beginning of a bull market because the prediction model with FAI worked well in a front bull term. That is, the FAI can be used to detect turning points in a stock market, no matter if prices move down or up.

Although the results of the experiments have proven the effectiveness of our proposed method, there is still room for

Figure 3: Change of SSE Composite Index

improvement. Given that the number of SNS comments might be insufficient, we plan to get more comments from Weibo to improve the quality of the FAI. Currently, three indexes (the stock index, the difference of the index, and FAI) are simply concatenated in our LSTM-based prediction model; however, a means to combine them should be explored in more detail. Consequently, in our future work, we will investigate the design of a structure of LSTM to accept multiple inputs.

7. Bibliographical References

Batres-estrada, G. (2015). Deep learning for multivariate financial time series.

Chen, K., Zhou, Y., and Dai, F., (2015). *A LSTM-based method for stock returns prediction: A case study of China stock market*, pages 2823–2824. Institute of Electrical and Electronics Engineers Inc.

Chollet, F. (2015). Keras. *Git-hub*.

Cootner, P. H. (1964). *The random character of stock market prices*. M.I.T. Press.

Fama, E. F., Fisher, L., Jensen, M. C., and Roll, R. (1969). The adjustment of stock prices to new information. *International Economic Review*, 10(1):1–21.

Fama, E. (1991). Efficient capital markets: II. *Journal of Finance*, 46(5):1575–617.

Glantz, M. and Kissell, R. (2013). *Multi-asset Risk Modeling, 1st ed.* Academic Press.

Heaton, J. B., Polson, N. G., and Witte, J. H. (2017). Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1):3–12, 10.

Jaramillo, J., Velasquez, J. D., and Franco, C. J. (2017). Research in financial time series forecasting with SVM: Contributions from literature. *IEEE Latin America Transactions*, 15(1):145–153, Jan.

Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009). Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL ,09, pages 272–280.

Li, J., Bu, H., and Wu, J. (2017). Sentiment-aware stock market prediction: A deep learning method. In *2017 International Conference on Service Systems and Service Management*, pages 1–6, June.

Malkiel, B. G. (1973). *A Random Walk Down Wall Street*. Norton, New York.

McTear, M., Callejas, Z., and Griol, D., (2016). *The Dawn of the Conversational Interface*, pages 11–24. Cham.

Nelson, D. M. Q., Pereira, A. C. M., and Oliveira, R. A. d. (2017). Stock market's price movement prediction with LSTM neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1419–1426.

Nguyen, T. H. and Shirai, K. (2015). Topic modeling based sentiment analysis on social media for stock market prediction. In *ACL*.

Nguyen, T. H., Shirai, K., and Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Syst. Appl.*, 42(24):9603–9611, December.

Niemira, M. and Saaty, T. (2004). An analytic network process model for financial-crisis forecasting. *International Journal of Forecasting*, 20:573–587.

Oh, K. J., Kim, T. Y., Lee, H. Y., and Lee, H., (2005). *Using Neural Networks to Support Early Warning System for Financial Crisis Forecasting*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Prechter Jr, R. R. and Parker, W. D. (2007). The financial/economic dichotomy in social behavioral dynamics: The socionomic perspective. *Journal of Behavioral Finance*, 8(2):84–108.

Prechter Jr, R. R., Goel, D., Parker, W. D., and Lampert, M. (2012). Social mood, stock market performance, and U.S. presidential elections: A socionomic perspective on voting results. *SAGE Open*, 2(4):2158244012459194.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings*

of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 45–50.

Rossum and Guido. (1995). Python reference manual. *CWI (Centre for Mathematics and Computer Science)*.

Sun, M., Guo, Z., Zhao, Y., Zheng, Y., Si, X., and Liu, Z. (2006). Thu Chinese text classification.

Trusov, R., Natekin, A., Kalaidin, P., Ovcharenko, S., Knoll, A., and Fazylova, A. (2015). Multi-representation approach to text regression of financial risks. In *2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT)*, pages 110–117.

Xiong, R., Nichols, E., and Shen, Y. (2015). Deep learning stock volatilities with google domestic trends.