**Stock Movement Prediction with Event Aggregation on web-based social data**

Names and Uni: Shangzi Xie sx2257 Chenyu Huang ch3470

Topic: Market Intelligence — Event Linkage and Impact Prediction

Date: 02/14

**Abstract:**

In recent years, many scholars are using methods based on machine learning or deep learning to predict stock price movement using web-based social data. However, the growing volume of opinionated text and complexity of the market caused by chaotic event interactions, makes it almost impossible to come up with a precise strategy for decision making in the stock market. So as to fix this problem, we proposed an event aggregation model from the event linkage perspective to acquire a better feature before predicting the stock movement. By eliminating the redundancy of features and the necessity of iterative computation, our model is evaluated to perform better than several traditional models.

1. **Background (Review of Related Literature):**

We divide the literature research into the following 3 parts: Event Modeling, Event Linkage and Impact Prediction.

* Event Modeling

First of all, we have to model our events so that we can apply a certain metric to find the similarity of events or evaluate the correlation between an event and the market. Graph, vector representation and dimension reduction methods are widely used in the related literatures.

Graph representation is one of the most popular and classical methods to model an event. Disjoint sets (Ruiz, Hristidis, Castillo, Gionis, & Jaimes, 2012) have been used to predict the stock price and traded volume time series. N-gram graph is also used along with metrics including value similarity or value ratio (Pittaras, Giannakopoulos, Tsekouras, & Varlamis, 2018) to facilitate later text clustering or classification.

A great number of vector-based representations are springing out in the recent times. High-dimensional word embeddings (Kusner, Matt, et al. 2015) are used to compute the distances between documents so as to represent the similarity. Pre-trained vector model of BERT (Devlin, Chang, Lee, & Toutanova, 2018) are proved to be extraordinarily effective in recent studies.

Dimension reduction methods are also widely used to convert web-based social events into a set of labels (Dickinson & Hu, 2015) (Pagolu, Reddy, Panda, & Majhi, 2017), which represents the classification of the events. Some statistics methods like logarithmic differentiation (Gilbert & Karahalios, 2010) are also used to map the original event to lower dimensional representation.

* Event Linkage

From the econometric perspective, some researchers are using Granger causality (Gilbert & Karahalios, 2010) to draw a relationship between web-based social data and the S&P 500 Index. However, this can only be used to naively accept or reject a relationship. In order to come up with a stronger correlation, K-means clustering (Yu, Hye-Yeon, et al., 2019) can be applied on some graph based representations because of their intrinsic distance property. Cosine similarity (Fauzi, Utomo, Pramukantoro, & Setiawan, 2017) is extremely widely-used in vector representations to describe the similarity between events. Semi-supervised learning techniques can even be applied to automatically generate labels according to the cosine similarity metric.

* Impact Prediction

With a great number of previous studies, impact prediction is considered as a classification problem in the market intelligence field. We are to provide a predictor such that it can tell whether the stock price of the next day will increase or decrease. Classifiers such as SMO (Reddy, Shiva, Abhilash, & Yoganandam, 2018) and XGBoost (Chen & Guestrin, 2016) are popular in this task because of their outstanding performance. Of course, a more challenging perspective is to consider impact prediction as a regression problem, which requires a much higher accuracy and computation frequency. Support Vector Regression (SVR) are coming into our eyes in the recent times. Among them, a variant of SVR with brain storm optimization (Wang, Hou, Wang, & Shen, 2016) outperformed 3 other traditional modals in all evaluation criteria. Other variants of SVR (Qiu, Zhu, Suganthan, & Amaratunga, 2017; Zhang, Teng, & Chen, 2019) are also providing a feasible and effective option to forecast stock price “continuously”. What’s more, some scholars are thinking out of the time domain and come up with a minute-level price forecasting approach with spectrum analysis (Lahmiri, 2018).

1. **Introduction to the Project:**

Our goal of this research is to draw a correlation between web-based events and the movement of the stock price.

As there has been a huge number of literatures studying on how to predict the stock price with either classification or regression, our main focus would be on the event linkage. In other words, how to correctly aggregate similar events so that we can compute the impact of this aggregated group of events altogether instead of computing the impact iteratively for each event. Because the latter method may consider the impact of event linear, however, in most practical scenarios, it turns out to be nonlinear. We shall come up with an advanced representation of event linkage either with modified vector addition or dimensionality reduction. Features can be extracted from the linked events and will be used for the successive model.

Since we don’t want to predict the movement of the stock price solely on the web-based features, we also have to fuse these features and the historical stock price to achieve the best possible prediction. One possible solution is to naively apply linear regression, but as we have discussed above, it will not work very well in most time. However, we can still use traditional methods like linear regression as our benchmarks, and evaluate our model. Maybe we can apply principle component analysis to rule out the redundant information contained in historical data and our features. Because the historical stock price is pretty likely to be affected by the features as well. In this case, I think the vector representation might achieve a good performance.

1. **Introduction to the Dataset:**

We use API provided by Twitter and Stocktwits to collect tweets. In twitter, the Tweepy API can be used to search and filter tweets. Stock price dataset can be acquired from Yahoo Finance. These 2 datasets will be aligned to make sure that they are within the same time slot.

For the tweets, we will do a preprocessing so that our dataset will be a clean column of word lists with number of likes and re-tweets. And for the stock price, we will only consider the stock price of Alphabet and compute the increment of price for each day.

A big challenge here is that if we want to do classification in the dimensionality reduction of events, we have to label the original dataset.

One possible solution is to use a model-based opinion mining technique provided by a part-of-speech graphical model to extract user’s opinions and test it from stock market social network. Another possible solution is to use semi-supervised learning, generate 768 dimensional base vectors and manually label 5~10% of our dataset, use this labeled data to infect the other unlabeled data so that we only have to do a small part of the whole labeling job.

1. **Plan:**

Milestone 1 + Progress 1:

* Data Gathering
  + Tweets Dataset
  + Stock Price Dataset
* Data Cleaning
  + Special Character Removal
  + Tokenization
  + Stop Word Removal
* Data Preprocessing

We will make sure that the dataset is ready for us to use to predict the stock price. We might have to apply some benchmarking models on our dataset to validate the usability. Also, for the semi-supervised learning approach in the dimensionality reduction part, we shall try to use the method I discussed above in the 3rd part to generate labels.

Milestone 2 + Progress 2:

* Event Modeling

Since we cannot confirm the effect of our event modeling before we finish our whole model and get the evaluation result, so we shall start the preparation of several candidates of event modeling representations in this period for our future use.

* Event Linkage

How do we draw the correlation of events so that they can be appropriately aggregated to become features for the successive model is the most important part of this project. Here we are about to come up with own method to aggregate similar events either with vector operation or dimensionality reduction and extract feature from the aggregated clusters. Also, we would like to use traditional feature extractors as our benchmark.

Milestone 3 + Progress 3:

* Fusion of Feature and Historical Price:

Try to fuse the features and historical price with principle component analysis or other methods to generate the input training data. If this step turns out to give a bad performance, regard it.

* Stock Price Prediction (Classification):

With available classifiers including LR, SMO and XGBoost, we shall train our predictor with the fused input data and our complete stock price output. A partition of training set and testing set is absolutely a must.

Also, we can go back to the previous steps according to the performance here to modify our modeling.

**Reference:**

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13*-*17*-*Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. (Mlm). Retrieved from http://arxiv.org/abs/1810.04805

Dickinson, B., & Hu, W. (2015). Sentiment Analysis of Investor Opinions on Twitter. *Social Networking*, *04*(03), 62–71. https://doi.org/10.4236/sn.2015.43008

Fauzi, M. A., Utomo, D. C., Pramukantoro, E. S., & Setiawan, B. D. (2017). Automatic essay scoring system using N-GRAM and cosine similarity for gamification based elearning. *ACM International Conference Proceeding Series*, *Part F131200*, 151–155. https://doi.org/10.1145/3133264.3133303

Gilbert, E., & Karahalios, K. (2010). Widespread worry and the stock market. *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 58–65.

Lahmiri, S. (2018). Minute-ahead stock price forecasting based on singular spectrum analysis and support vector regression. *Applied Mathematics and Computation*, *320*, 444–451. https://doi.org/10.1016/j.amc.2017.09.049

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2017). Sentiment analysis of Twitter data for predicting stock market movements. *International Conference on Signal Processing, Communication, Power and Embedded System, SCOPES 2016 - Proceedings*, 1345–1350. https://doi.org/10.1109/SCOPES.2016.7955659

Pittaras, N., Giannakopoulos, G., Tsekouras, L., & Varlamis, I. (2018). Document clustering as a record linkage problem. *Proceedings of the ACM Symposium on Document Engineering 2018, DocEng 2018*. https://doi.org/10.1145/3209280.3229109

Qiu, X., Zhu, H., Suganthan, P. N., & Amaratunga, G. A. J. (2017). Stock price forecasting with empirical mode decomposition based ensemble ν-Support Vector regression model. *Communications in Computer and Information Science*, *775*, 22–34. https://doi.org/10.1007/978-981-10-6427-2\_2

Reddy, K. P. K., Shiva, K. L., Abhilash, K., & Yoganandam, Y. (2018). *Database Assisted Automatic Modulation Classification Using Sequential Minimal Optimization*. Retrieved from http://arxiv.org/abs/1806.07566

Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., & Jaimes, A. (2012). Correlating financial time series with micro-blogging activity. *WSDM 2012 - Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, 513–521. https://doi.org/10.1145/2124295.2124358

Wang, J., Hou, R., Wang, C., & Shen, L. (2016). Improved v -Support vector regression model based on variable selection and brain storm optimization for stock price forecasting. *Applied Soft Computing Journal*, *49*, 164–178. https://doi.org/10.1016/j.asoc.2016.07.024

Zhang, J., Teng, Y. F., & Chen, W. (2019). Support vector regression with modified firefly algorithm for stock price forecasting. *Applied Intelligence*, *49*(5), 1658–1674. https://doi.org/10.1007/s10489-018-1351-7

Yu, H. Y., Park, S., Cheong, Y. G., Kim, M. H., & Bae, B. C. (2019, November). Emotion-Based Story Event Clustering. In *International Conference on Interactive Digital Storytelling,* 348-353. Springer, Cham.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, June). From word embeddings to document distances. In *International conference on machine learning*, 957-966.