Event detection and time series alignment to improve stock market forecasting

Elliot Maître, Zakaria Chemli, Max Chevalier, Bernard Dousset,
Jean-Philippe Gitto, Olivier Teste

Event detection and time series alignment to improve stock market forecasting

Elliot Maître
Institut de Recherche en Informatique de Toulouse / Scalian
Toulouse, France
elliot.maitre@irit.fr

Zakaria Chemli
Scalian
Paris, France
zakaria.chemli@scalian.com

Max Chevalier
Institut de Recherche en Informatique de Toulouse
Toulouse, France
max.chevalier@irit.fr

Bernard Dousset
Institut de Recherche en Informatique de Toulouse
Toulouse, France
bernard.dousset@irit.fr

Jean-Philippe Gitto
Scalian
Blagnac, France
jean-philippe.gitto@scalian.com

Olivier Teste
Institut de Recherche en Informatique de Toulouse
Toulouse, France
olivier.teste@irit.fr

ABSTRACT

Buying commodities is a critical issue for multiple industries because the variations of stock prices are induced not only by multiple economic parameters but also by external events. Raw material buyers must keep track of information in numerous fields, which constitutes a major challenge considering the exponential growth of online data. To tackle this issue, we propose an event detection approach in order to assist them in their anticipation process. Indeed, a lot of contextual information is contained in text and exploiting it can allow one to improve its anticipation ability. Thus, we develop a framework of event detection and qualification, then we quantify the impact of these events on stock market to help buyers in their anticipation process. In this paper, we will first introduce our context, then explain the scope of our work and our goals. After detailing the related work, we will present our proposition, conclude and propose some future work possibilities.

CCS CONCEPTS

• **Information systems** → *Data management systems*; **Information retrieval**; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

Event detection, text analysis, nlp, neural networks, time series, commodities

1 INTRODUCTION

Time series play a major role in several industrial fields, such as energy [1], transport [29], economy [11] or finance [28]. Being able to accurately forecast time series is a major asset in order to anticipate the modeled phenomenon for companies. In commodities buying, the stock market is described by time series and is particularly volatile, making its forecasting both a strategical and a challenging task [7], [10]. Classic stock forecasting methods like [15] or [24] are usually based on economical data, such as currencies, indices or futures but most of them do not take into account textual data which can contain precious information. Improving time series forecasting using textual information is a challenging research issue [30].

In order to extract text data, multiple sources can be considered. An important one is micro-blogging. Several studies showed the predictive power of such media [23]. Sentiment analysis on Twitter can be helpful [2], the activity on social network can be correlated with variation of the stock [26] and Twitter data can be used to forecast polls that are then used to interpret stock variations [22]. Specialized financial website, such as Seeking Alpha, where communities of traders share their insights about the stock market, also contains meaningful information for stock market forecasting [5]. Thus, multiple sources of information like micro-blogging and specialized community websites can be combined to improve stock market forecasting.

Leveraging the expertise of several buyers via multiple interviews, we observed that they base their decisions on events happening in the real world, related by newspapers and social networks. Hence, given that the stock market reacts to news and events [8], we will particularly focus on event detection in text. Indeed, some periods are more intense than others [4] and are considered as more important. These periods, characterized by some events, are carrying more information than other periods. Being able to detect these events and quantify their impact constitute a major asset for buyers and traders. It is a difficult task, as illustrated by the impact on the stock market of the Covid-19 outbreak, which was widely discussed but largely underestimated. With adapted tools, one could have anticipated this crisis and behaved accordingly in order to mitigate the impact.

Our research aims at providing a tool leveraging information contained in text data, especially events, in order to assist people in their time series anticipation process, i.e. commodities buyers in our context. In this paper we will focus on the event detection step. We will firstly introduce our general work, then we will focus on the related work about event detection in text. Afterwards, we will develop our proposal.

2 OVERVIEW OF OUR PROPOSAL

The task of commodities price forecasting is particularly complex due to the tremendous amount of parameters that influence the

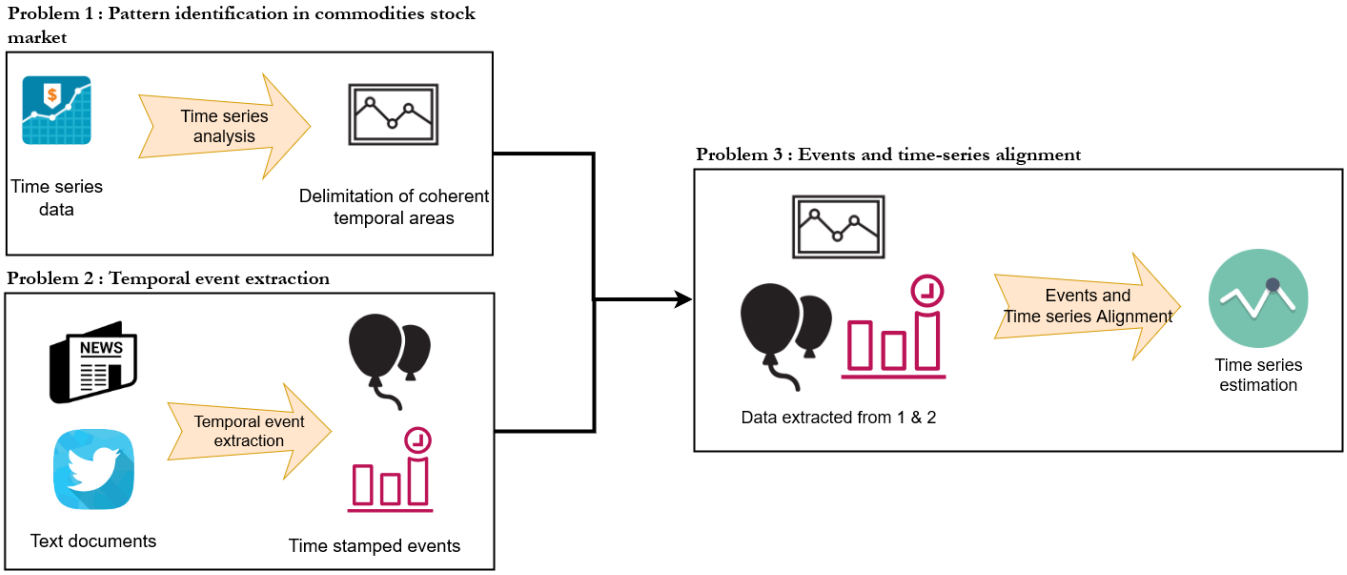

Figure 1: General approach

variations of the stock. To bring more contextual information to the buyers and to our model, we want to combine time series with text information. This is not a straightforward process and it needs to be broke down in sub-tasks. Hence, our work will be articulated around three major steps as illustrated by **Figure 1** :

(1) Time series analysis to find coherent temporal areas,
(2) Temporal event extraction,
(3) Events and time-series alignment.

While these steps are mutually dependent, it is also possible to treat them separately. Each of them constitute a scientific challenge and thus will be developed separately [20], [9], [14], [25], [24], [15]. In the rest of this paper, we will particularly focus on part (2) which is the part we are currently working on and give insights about (3) which is the next step of our work. Part (1) is currently not in the scope of this work, we plan to use existing approaches to tackle this issue.

3 RELATED WORK

There are different approaches to perform event detection in text. The two principal are topic modeling and event trigger detection. The former is a statistical approach while the latter is based on word classification.

3.1 Event trigger based approaches

The event trigger based approach is a classification method which consists in classifying words in event categories. Some words, named trigger-words, are supposed to trigger the event in the sentence and they are carrying the meaning. Detecting and classifying those words hence allow one to understand if a sentence depicts an event. ACE 2005 [12] is the reference dataset for this task and has been studied multiple times [20], [9], [14]. According to the ACE 2005 annotation guideline, in the sentence "A police officer was killed in New Jersey today", an event detection system should be able to recognize the word "killed" as a trigger for the event "Die". Currently, the state-of-the-art for this task is achieved by using neural networks and several approaches have been proposed on this base. Nguyen introduced in [20] a CNN-based approach to detect these triggers. In [9], the authors improve this work by adding a Bi-LSTM to the CNN in order to include sentence context to the detection. The authors of [14] propose a self-regulating GAN to perform the detection. In [18], the authors include even more context by a document-scale approach.

3.2 Topic modeling approaches

While the former approach is mostly based on semantic and syntactic properties, topic modeling approaches are statistical approaches. The authors of [27] propose to use Twitter users as human sensors to detect in real-time earthquake occurrences. The authors are using keywords to detect these target events and they use probabilistic models to detect the location of the events. Weng et al., in [31] analyze the wavelet signal of words in Tweets in order to filter trivial words and clusters words to detect events. In [17], the authors analyze daily topics on Twitter via Latent Dirichlet Analysis (LDA) and then determine similarity between daily topics. They detect bumps in word usage and then clusterizes topics in "eventy topics". The authors of [21] propose a sub-event detection technique using topic modeling. This technique detect sub-events linked to an event and assign a label to these sub-events. In [13], the authors propose a real-time framework to detect minor and major events on Twitter. The first module of the framework detects events and then the second module clusterizes these events.

Thus, event trigger based approaches tend to exploit the power of deep neural networks while topic modeling approaches are based on frequency of words and on what is discussed on social networks. We argue that combining the asset of each technique could be

an interesting objective. The power of representation brought by neural network is complementary to the detection approach of topic modeling.

4 OUR PROPOSAL: EVENT DETECTION COMBINING TOPIC MODELING AND NEURAL MODELS

Several constraints, such as the influence of possibly unknown parameters and the real-time nature, arise from the definition of the stock market. To predict future stock, one must exploit historical data but also real-time data. Hence, our framework must be applicable to data stream such as the Twitter stream. Moreover, some events may not be comparable to past events, so the classification must be able to handle and assign labels to unknown classes. However, we do not aim at making real-time commodities trading, we want to assist buyers in their daily buying decisions. We only want our solution to be applicable in a real-time context, i.e. with a granularity sufficient to help buyers in their daily transactions.

4.1 Motivations

Topic-modeling approaches correspond to our prerequisites, but some of them are not adapted to data-streams or does not work with unknown classes. Recent work which satisfies our constraints fails to exploit the properties of the language and are only based on a probabilistic approach linked with word apparitions.

Neural based approaches, such as the methods used in the trigger-based approaches, are powerful in order to exploit patterns discovered in past data. Moreover, they bring more information by leveraging semantics and syntactic information, with methods such as word and sentence embeddings.

Our goal is to exploit these information to improve the quality of event representation. We think that these approaches are complementary and we assert that combining them will allow us to leverage the time and frequency aspect derived from topic modeling and the representation power of neural networks, in order to optimize event classification.

4.2 Our method

To do so, we propose a novel approach based on word and sentence embeddings. The idea behind this method is to leverage the geometric power of these methods. Using the representation obtained, similar documents should have similar representations in the embedding space. By comparing the distance between documents, we will be able to create clusters of documents. Each cluster corresponds to an event. Some events may be related and clusters of similar events might be regrouped in an event cluster. This event cluster represents a class of events, such as sports events, geopolitical events... Hence, unknown events can be assimilated to events in the same event cluster. We will order documents by their apparition time, so we can adapt to the real-world context we want to apply this method to, i.e. commodities stock estimation using event detection in text data stream.

Our proposition is articulated as follows:

(1) Text data is extracted from sources previously selected by buyers, such as trusted Twitter users, in order to gather text written in regular English and focused on sharing important information. Indeed, most of the content on the internet is created by a few users.
(2) In order to have an exploitable event representation, we embed the content, using word embedding and sentence embedding.
(3) The embedded content is clusterized, leveraging the amount of information the embeddings bring. This can be done by placing the embedded content on vertices of a graph and creating an edge between each vertex, weighted by the distance between the two embeddings. If the distance is under a certain threshold, the edge is removed in order to create clusters of related contents.
(4) The clusters are labelized, by determining representative document. An example of a representative document is a document with the minimum average distance with other tweets of the cluster.

Thus, the clusters obtained are expected to be of great quality thanks to a better representation, allowing a better identification and classification of events. These detected events will have two usages : they will be used in the next steps in order to estimate the variations of the times series, and they will also be given to the buyers in order to help make their decision, alongside with our time series estimation. Since the tweets are extracted from the Twitter Stream, we will order them as their apparition order, which allows us to take time into account and adapt to the type of application we want.

4.3 Pros and cons

This methods brings more information than a regular topic modeling approach, leveraging the representation power of neural based approach. It allows us to consider the documents in a time-ordered manner which is not the case in most classification problem. This make it suitable for time-based applications such as our.

However, the efficiency of such a model for unknown events is not certain. Indeed, it is clear that neural networks sometimes fail to generalize correctly. Handling an event containing too much novelty might be misleading for some models. The time aspect may also have some impact on the efficiency of the model.

Moreover, neural based approaches require annotated data, which is not always available, especially in context such as Twitter where the amount of data is huge. This problem has been considered in recent work, notably in [19] where the authors propose a weakly-supervised approach to limit annotation time. The problem of unknown classes is not appropriately handled by these approaches. Detecting novelty without labeling it could be an insight in order to detect change in the time series, but in the mean time, we want to focus on a method allowing us to label unknown events.

Thus, this method helps us bringing more information in order to fulfill our classification objective, to adapt to our time-dependant context however it may rise several issues that we have not addressed yet.

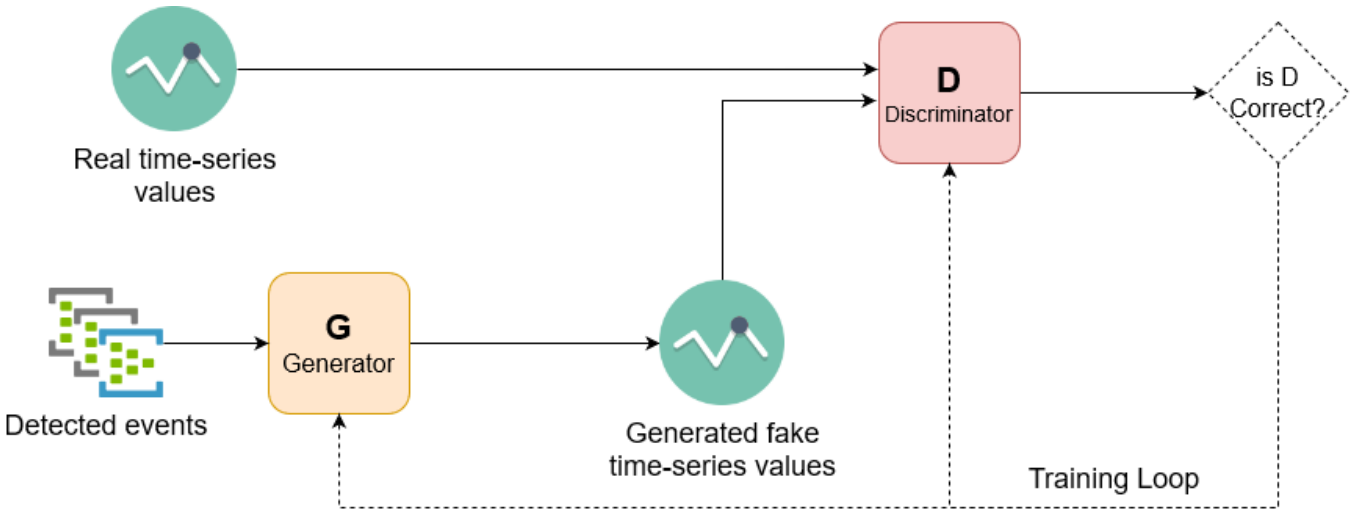

Figure 2: GAN example

5 LINKING EVENTS AND TIME SERIES VARIATIONS TO ESTIMATE FUTURE TIME SERIES VARIATIONS

Following the idea of combining time series and text, the detected events will be fed to a generative adversarial network (GAN) along with time series data, to predict expected variations of the stock prices. **Figure 2** illustrates the process we will describe. Our intuition is that the GAN will be able to link detected events and variations in the time-series. A GAN is composed of two major parts : the generator and the discriminator. The generator try to mimic the actual data and the discriminator tries to identify fake data produced by the generator. We want to produce time series estimations, so our solution is articulated as follow : the generator part of the GAN will produce time series estimations taking events as input. The discriminator will be fed with two inputs, the actual time series and the fake time-series, which is generated by the generator. The objective for the generator is to be able to produce time series estimations that are really close to reality, in order to fool the discriminator. The discriminator objective is to have a maximum accuracy in its task to differentiate fake and real input. Since the final output we want is a time series estimation, our general objective is to have a generator as optimized as possible. The discriminator is only used in the training loop, in order to give feedback to generator, to train it to produce valuable output. In order to give hints about the future time series variations, the generator will take as input the events we have previously detected, which are supposed to carry information that influences these variations. By training it properly, the generator will be able to extract information from the events and from the feedback of the discriminator. The feedback from the discriminator contains information about the time series, which are not directly available to the generator. Indeed, the final objective is to have a generator which is able to predict time series variations, by only exploiting the events we detect.

To summarize, the GAN corresponds to the event-quantifying step, and the event-time series alignment step.

Its objectives is to automatically extract information from the detected events it takes as input, and link it with the variations in the historical time series data.

6 CONCLUSION

Considering the constraints induced by our context, namely detecting possibly unknown events in order to help buyers in their daily buying decisions, we deduced that a combination of topic-modeling approaches and neural based models is a promising method to complete our task. We propose to embed content using recent models, i.e. word and sentence embeddings, in order to produce a better clusterization leveraging the representation power of these models and therefore have a better event classification.

7 FUTURE WORK

In [3], the authors temporalize word2vec to detect the mostly discussed topics during certain phases of the bitcoin time series. We would like to transpose this idea to our context, by detecting which events are activated during special phases of the commodities stock. Using time stamps of the documents, the idea is to determine which clusters of events are activated during a certain period of time and link it with stock variations. If using timestamps to order documents is not difficult, determining when an event is activated brings a lot more difficulties, such as tracking event evolution and detecting the end of an event. Another goal is to be able to directly link time series and event, in a similar method as [25]. Finally, encoder-decoder architecture are currently revolutionising the NLP domain. We would like to be able to better represent events, leveraging the power of encoder-decoder architectures such as BERT [6]. Wu et al. did something similar with news representation in [32]. Indeed, transformers are able to produce quality embeddings for both words and sentences and have proved their quality by outperforming static embedding techniques. A major drawback of transformer-based methods is their computation cost. Thus, the usage of distilled models such as TinyBERT [16] could be a solution.

REFERENCES

[1] John Asafu-Adjaye. 2000. The Relationship between Energy Consumption, Energy Prices and Economic Growth: Time Series Evidence from Asian Developing Countries. Energy Economics 22 (12 2000), 615–625. https://doi.org/10.1016/S0140-9883(00)00050-5

[2] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2010. Twitter mood predicts the stock market. CoRR abs/1010.3003 (2010). arXiv:1010.3003 http://arxiv.org/abs/1010.3003

[3] Andrew Burnie and Emine Yilmaz. 2019. An Analysis of the Change in Discussions on Social Media with Bitcoin Price. 889–892. https://doi.org/10.1145/3331184.3331304

[4] Patrick Champagne. 2000. L'événement comme enjeu. (2000). https://doi.org/10.3406/reso.2000.2231

[5] Hailiang Chen, Prabuddha De, Yu Hu, and Byoung-Hyoun Hwang. 2013. Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. Review of Financial Studies (12 2013). https://doi.org/10.2139/ssrn.1807265

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Claude B. Erb and Campbell R. Harvey. 2006. The Strategic and Tactical Value of Commodity Futures. Financial Analysts Journal 62, 2 (2006), 69–97. https://doi.org/10.2469/faj.v62.n2.4084 arXiv:https://doi.org/10.2469/faj.v62.n2.4084

[8] Eugene F. Fama. 1965. The Behavior of Stock-Market Prices. The Journal of Business 38, 1 (1965), 34–105. http://www.jstor.org/stable/2350752

[9] Xiaocheng Feng, Lifu Huang, Duyu Tang, Heng Ji, Bing Qin, and Ting Liu. 2016. A Language-Independent Neural Network for Event Detection. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Berlin, Germany, 66–71. https://doi.org/10.18653/v1/P16-2011

[10] Gary Gereffi. 1999. International trade and industrial upgrading in the apparel commodity chain. Journal of International Economics 48, 1 (June 1999), 37–70. https://ideas.repec.org/a/eee/inecon/v48y1999i1p37-70.html

[11] Clive Granger and Paul Newbold. 1986. Forecasting Economic Time Series (2 ed.). Elsevier. https://EconPapers.repec.org/RePEc:eee:monogr:9780122951831

[12] Ralph Grishman, David Westbrook, and Adam Meyers. 2005. NYU's English ACE 2005 system description. Proceedings of ACE 2005 Evaluation Workshop. Journal on Satisfiability 51 (01 2005).

[13] Mahmud Hasan, Mehmet A. Orgun, and Rolf Schwitter. 2019. Real-time event detection from the Twitter data stream using the TwitterNews+ framework. Information Processing and Management 56, 3 (5 2019), 1146–1165. https://doi.org/10.1016/j.ipm.2018.03.001

[14] Yu Hong, Wenxuan Zhou, Jingli Zhang, Guodong Zhou, and Qiaoming Zhu. 2018. Self-regulation: Employing a Generative Adversarial Network to Improve Event Detection. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, 515–526. https://doi.org/10.18653/v1/P18-1048

[15] Wei Huang, Yoshiteru Nakamori, and Shou-Yang Wang. 2005. Forecasting Stock Market Movement Direction with Support Vector Machine. Comput. Oper. Res. 32, 10 (Oct. 2005), 2513–2522. https://doi.org/10.1016/j.cor.2004.03.016

[16] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tiny{BERT}: Distilling {BERT} for Natural Language Understanding. https://openreview.net/forum?id=rJx0Q6EFPB

[17] Nathan Keane, Connie Yee, and Liang Zhou. 2015. Using Topic Modeling and Similarity Thresholds to Detect Events. In Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation. Association for Computational Linguistics, Denver, Colorado, 34–42. https://doi.org/10.3115/v1/W15-0805

[18] Dorian Kodelja, Romaric Besançon, and Olivier Ferret. 2019. Exploiting a More Global Context for Event Detection Through Bootstrapping. 763–770. https://doi.org/10.1007/978-3-030-15712-8_51

[19] Shulin Liu, Yang Li, Feng Zhang, Tao Yang, and Xinpeng Zhou. 2019. Event Detection without Triggers. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 735–744. https://doi.org/10.18653/v1/N19-1080

[20] Thien Huu Nguyen and Ralph Grishman. 2015. Event Detection and Domain Adaptation with Convolutional Neural Networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, Beijing, China, 365–371. https://doi.org/10.3115/v1/P15-2060

[21] Diogo Nolasco and Jonice Oliveira. 2019. Subevents detection through topic modeling in social media posts. Future Generation Comp. Syst. 93 (2019), 290–303.

[22] Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. International AAAI Conference on Weblogs and Social Media 11.

[23] Nuno Oliveira, Paulo Cortez, and Nelson Areal. 2016. The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Systems with Applications 73 (12 2016). https://doi.org/10.1016/j.eswa.2016.12.036

[24] Ping-Feng Pai and Chih-Sheng Lin. 2005. A hybrid ARIMA and support vector machines model in stock price forecasting. Omega 33 (12 2005), 497–505. https://doi.org/10.1016/j.omega.2004.07.024

[25] Filipe Rodrigues, Ioulia Markou, and Francisco Pereira. 2018. Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach. Information Fusion 49 (07 2018). https://doi.org/10.1016/j.inffus.2018.07.007

[26] Eduardo Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. 2012. Correlating Financial Time Series with Micro-Blogging Activity. WSDM 2012 - Proceedings of the 5th ACM International Conference on Web Search and Data Mining, 513–522. https://doi.org/10.1145/2124295.2124358

[27] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. Proceedings of the 19th International Conference on World Wide Web, WWW '10, 851–860. https://doi.org/10.1145/1772690.1772777

[28] Ruey S. Tsay. 2005. Analysis of financial time series (2. ed. ed.). Wiley-Interscience, Hoboken, NJ. http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+483463442&sourceid=fbw_bibsonomy

[29] Mascha C. van der Voort, Mark Dougherty, M.S. Dougherty, and Susan Watson. 1996. Combining Kohonen maps with Arima time series models to forecast traffic flow. Transportation research. Part C: Emerging technologies 4, 5 (1996), 307–318. https://doi.org/10.1016/S0968-090X(97)82903-8

[30] Baohua Wang, Hejiao Huang, and Xiaolong Wang. 2012. A novel text mining approach to financial time series forecasting. Neurocomputing 83 (04 2012), 136–145. https://doi.org/10.1016/j.neucom.2011.12.013

[31] Jianshu Weng and Bu-Sung Lee. 2011. Event Detection in Twitter. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2767

[32] Chuhan Wu, Fangzhao Wu, Mingxiao An, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Topic-Aware News Representation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, 1154–1159. https://doi.org/10.18653/v1/P19-1110