Unsupervised Word Influencer Networks from News Streams

Ananth Balashankar
New York University
ananth@nyu.edu

Sunandan Chakraborty
Indiana University
sunchak@iu.edu

Lakshminarayanan Subramanian
New York University
lakshmi@cs.nyu.edu

Abstract

In this paper, we propose a new unsupervised learning framework to use news events for predicting trends in stock prices. We present *Word Influencer Networks (WIN)*, a graph framework to extract longitudinal temporal relationships between any pair of informative words from news streams. Using the temporal occurrence of words, WIN measures how the appearance of one word in a news stream *influences* the emergence of another set of words in the future. The latent word-word influencer relationships in WIN are the building blocks for causal reasoning and predictive modeling. We demonstrate the efficacy of WIN by using it for unsupervised extraction of latent features for stock price prediction and obtain 2 orders lower prediction error compared to a similar causal graph based method. WIN discovered influencer links from seemingly unrelated words from topics like politics to finance. WIN also validated 67% of the causal evidence found manually in the text through a direct edge and the rest 33% through a path of length 2.

1 Introduction

Stock price prediction using financial news events and social media sentiments have been studied extensively in literature. Most of these works rely on extracting rich features from relevant financial news of companies (Falinouss, 2007; Kalyani et al., 2016; Hagenau et al., 2013; Shynkevich et al., 2015), Twitter sentiments of financial terms (Mao et al., 2011; Rao and Srivastava, 2012; Bernardo et al., 2018) and market volatility measures (Balcilar et al., 2017; Sun et al., 2014) as features to predict trends in their stock prices. However, none of these approaches tried to exploit *unknown* or *little known* relationships between news events and stock prices. Previous works used “known” factors and used them as features to predict stock prices by extracting them from news stories. There might be other unknown (and non-finance related) factors potentially influencing stock prices that cannot be discovered using these methods.

This paper aims to understand unknown and latent relationships between words that describe events in news streams to potentially uncover *hidden* links between news events and apply those *new* relationships to build a news-driven predictive model for stock prices. The appearance of these relationship entities in news, may be well separated over time. For example, market volatility is known to be triggered by recessions; this hidden relationship may manifest in new streams with a frequency spike in the word ”recession” followed by a frequency spike in the word ”volatility”, a *few weeks later*. Thus, mining large news datasets can potentially reveal influencing factors behind the surge of a particular word in news. This notion can be generalized to discover the influence of one event to another, where the events are manifested by specific words appearing in news.

In this paper, we propose a new framework – *Word Influencer Networks (WIN)* that aims at detecting the latent relationships between words, where such relationships are not directly observed. WIN differs from existing relationship extraction and representational frameworks across two dimensions – (1) unsupervised causal relationships instead of associative ones that can be used to understand a path of *influence* among news items, (2) finding inter-topic influence relationships outside the “context” or the confines of a single document. Construction of WIN can be used to build

Proceedings of the First Workshop on Economics and Natural Language Processing, pages 62–68
Melbourne, Australia, July 20, 2018. ©2018 Association for Computational Linguistics

predictive models for numerous *news-dependent* variables, including stock prices.

We constructed WIN from a news corpus of around $700,000$ articles and evaluated it to extract features for stock price prediction and obtained two orders lower prediction error compared to a similar causal graph based method. WIN also validated 67% of the causal evidence found manually in the text through a direct edge in the network and the rest through a path of length 2. We also evaluated the network qualitatively for sparsity and its capacity to generate "out of context" inter-topic word relationships on the entire vocabulary.

2 Related Work

Online news articles are a popular source for mining real-world events, including extraction of causal relationships. Radinsky and Horvitz (Radinsky and Horvitz, 2013) proposed a framework to find causal relationships between events to predict future events from News but caters to a small number of events. Causal relationships extracted from news using Granger causality have also been used for predicting variables, such as stock prices (Kang et al., 2017; Verma et al., 2017; Darrat et al., 2007). A similar causal relationship generation model has been proposed by Hashimoto et al. (2015) to extract causal relationships from natural language text. A similar approach can be observed in (Kozareva, 2012; Do et al., 2011), whereas CATENA system (Mirza and Tonelli, 2016) used a hybrid approach consisting of a rule-based component and a supervised classifier. WIN differs from these approaches as it explores latent inter-topic causal relationships in an unsupervised manner from the entire vocabulary of words and collocated N-grams.

Apart from using causality, there are many other methods explored to extract information from news and are used in time series based forecasting. Amodeo et al. (Amodeo et al., 2011) proposed a hybrid model consisting of time-series analysis, to predict future events using the New York Times corpus. FBLG (Cheng et al., 2014) focused on discovering temporal dependency from time series data and applied it to a Twitter dataset mentioning the Haiti earthquake. Similar work by Luo et al. (Luo et al., 2014) showed correlations between real-world events and time-series data for incident diagnosis in online services. Other similar works like, Trend Analysis Model (TAM) (Kawamae, 2011) and Temporal-LDA (TM-LDA) (Wang et al., 2012) model the temporal aspect of topics in social media streams like Twitter. Structured data extraction from news have also been used for stock price prediction using techniques of information retrieval in (Ding et al., 2014; Xie et al., 2013; Ding et al., 2015; Chang et al., 2016; Ding et al., 2016). Vaca et al. (Vaca et al., 2014) used a collective matrix factorization method to track emerging, fading and evolving topics in news streams. WIN is inspired by such time series models and leverages the Granger causality detection framework for the trend prediction task.

3 Word Influence Network

Word Influence Network (WIN) addresses the discovery of *influence* between words that appear in news text. The identification of influence link between words is based on temporal co-variance, so that answers to questions of the form "Does the appearance of word x influence the appearance of word y after δ days?" can be addressed. The influence of one word on another is determined based on pairwise causal relationships and is computed using Granger causality test. Following the identification of Granger causal pairs of words, such pairs are combined together to form a network of words, where the directed edges depict potential influence between words. The network provides a more holistic view of the causal information flow by overcoming a common drawback of pair-wise Granger causality, when the true relationship involves three or more variables (Maziarz, 2015). In the final network an edge or a path between a word pair represents a flow of influence from the source word to the final word and this *influence* depicts an increase in the appearance of the final words when the source word was observed in news data.

The word influencer network can offer the following that can significantly increase the benefits of using news for analytics – (1) Detection of influence path, (2) Discovery of unknown facts, (3) Hypothesis testing and (4) Feature extraction for experiment design.

4 Methodology

Construction of WIN from the raw unstructured news data, finding pairwise causal links and eventually building the influence network involves numerous challenges. In the rest of the section we discuss the design methodologies used to overcome these challenges along with some properties of the network.

Selecting *Informative* Words: Only a small percentage of the words appearing in news can be used for meaningful information extraction and analysis. There are some words that are *too frequent* and some are *too rare* to establish any significant relationship(Manning et al., 1999; Hovold, 2005). Any word whose frequencies were in those range were removed. Specifically, we eliminated too frequent (at least once in more than 50% of the days) or too rare (appearing in less than 100 articles). These thresholds were determined empirically by looking at the temporal frequency distribution of the words. Many common English nouns, adjectives and verbs, whose contribution to semantics is minimal(Forman, 2003) carry very little significance were also removed from the vocabulary. However, named-entities were retained for their newsworthiness and a set of trigger words were retained that depicts events (e.g. flood, election) using an existing event trigger detection algorithm. The vocabulary set was enhanced by adding bigrams that are significantly collocated in the corpus, such as, 'fuel price' and 'prime minister' etc. after applying similar filtering methods as described for words.

Time-series Representation of News Data: Consider a corpus $\mathcal{D}$ of news articles indexed by time t, such that $\mathcal{D}_t$ is the collection of news articles published at time t. Each article $d \in D$ is a collection of words W_d, where i^{th} word $w_{d,i} \in W_d$ is drawn from a vocabulary V of size N. The set of articles published at time t can be expressed in terms of the words appearing in the articles as $\{\alpha_1^t, \alpha_2^t, ..., \alpha_N^t\}$, where α_i^t is the sum of frequency of the word $w_i \in V$ across all articles published at time t. α_i^t corresponding to $w_i \in V$ is defined as, $\alpha_i^t = \frac{\mu_i^t}{\sum_{t=1}^{T} \mu_i^t}$ where $\mu_i^t = \sum_{d=1}^{|D_t|} TF(w_{d,i})$. α_i^t is normalized by using the frequency distribution of w_i in the entire time period. $\mathcal{T}(w_i)$ represents the time series of the word w_i, where i varies from 1 to N, the vocabulary size.

4.1 Measuring Influence between Words

Given two time-series X and Y, the Granger causality test checks whether the X is more effective in predicting Y than using just Y and if this holds then the test concludes X "Granger-causes" Y (Granger et al., 2000). However, if both X and Y are driven by a common third process with different lags, one might still fail to reject the alternative hypothesis of Granger causality. Hence, in WIN, we explore the possibility of causal links between all word pairs and detect triangulated relations to eliminate the risk of ignoring confounding variables, otherwise not considered in Granger causality test.

Constructing WIN using an exhaustive set of word pairs can be computationally challenging and prohibitively expensive when the vocabulary size is fairly large. This is true in our case, where even after using a reduced set of words and including the collocated phrases, the vocabulary size is around $39,000$. One solution to this problem is considering the Lasso Granger method (Arnold et al., 2007) that applies regression to the neighborhood selection problem for any word, given the fact that the best regressor for that variable with the least squared error will have non-zero coefficients only for the lagged variables in the neighborhood. The Lasso algorithm for linear regression is an incremental algorithm that embodies a method of variable selection (Tibshirani, 1994). In our case, if we are determining the influence link between a word y to the rest, then,

$$\mathbf{w} = argmin \frac{1}{N} \Sigma_{(\mathbf{x},y) \in V} |\mathbf{w}.\mathbf{x} - y|^2 + \lambda ||\mathbf{w}|| \quad (1)$$

where V is the input vocabulary from the news dataset, N is the vocabulary size, x is the list of all lagged variables (maximum lag of 30 days per word) of the vocabulary and λ is a constant to be determined. To set λ, we use the method used in (Meinshausen and Bühlmann, 2006). We select the variables that have non-zero co-efficients and choose the best lag for a given variable based on the maximum absolute value of a word's coefficient. We then, draw an edge from all these words to the predicted word with the annotations of the optimal time lag (in days) and incrementally construct the graph as illustrated in Figure 3.

4.2 Topic Influence Compression

The number of nodes in this version of WIN corresponds to the vocabulary size and it can be hard to visualize the graph due to its size. To make information gathering from WIN easier, we make the graph coarser by clustering the nodes based on *topics*. Topics are learned from the original news corpus using Latent Dirichlet Allocation (LDA)(Blei et al., 2003). Influence is generalized to topic level by calculating the weight of inter-topic influence relationships as a total number of edges between vertices of two topics. The strength of this influence is defined as,

$$\Phi(\theta_u, \theta_v) = \frac{\text{\# Edges between u and v}}{(|\theta_u| \times |\theta_v|)} \tag{2}$$

where, θ_u and θ_v are two topics in our topic model and $|\theta_u|$ represents the size of topic θ_u, i.e. the number of words in the topic whose topic-probability is greater than 0.001. $\Phi(\theta_u, \theta_v)$ is termed as *strong* if its value is within the top 1% of Φ for all topics. Any edge in the original WIN is removed if there are no strong topic edges between the corresponding word nodes. This filtered topic graph has only edges between topics which have high influence strength.

5 Evaluation

5.1 Data

The news dataset[1] we used for stock price prediction contains news crawled from 2010 to 2013 using Google News APIs and New York Times data from 1989 to 2007. We construct WIN from the time series representation of its 12,804 unigrams and 25,909 bigrams, as well as the 10 stock prices[2] from 2013 we use for prediction. The prediction is done with varying step sizes (1,3,5), which indicates the time lag between the news data and the day of the predicted stock price in days. In order to qualitatively validate that latent inter-topic edges exist in the news stream, we also constructed WIN from the online archives of Times of India (TOI), the most circulated Indian English newspaper. This dataset contains all the articles published in their online edition between January 1, 2006 and December 31, 2015 containing 1,538,932 articles.

5.2 Inter-topic edges of WIN

The influence network we constructed from the TOI dataset has 18,541 edges and 7,190 unigrams and bi-gram vertices. We split the edges to inter-topic (9774) edges and intra-topic (8765) edges. We were interested in the inter-topic non-associative relationships that WIN is expected to capture. From Figure 1, we can see many topics (44) do not have inter-topic influence relationships, but a few topics (5) influence or are influenced by a large number of topics. Some of these highly influential topics are composed of words describing "Education", "Economics", "Politics", "Crime" and "Agriculture", and the maximum number of influencer relationships in WIN is from "Politics" to "Crime".

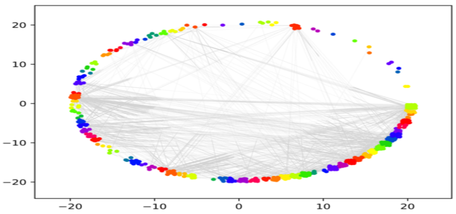

Figure 1: Inter-topic word relationships

5.2.1 Links of the network

Inspecting the links and paths of WIN gives us qualitative insights into the context in which the word-word relationships were established. Since WIN is also capable of representing other stock time series as potential influencers in the network, we can use this to model the propagation of shocks in the market as shown in Figure 2. WIN also highlights one of the limitations of granger causality by running on the entire vocabulary as shown in Figure 3, i.e if an underlying event (slum rehabilitation) causes two other events at different time lags (provided relief and coordinate committee), the link between the two lagged events can be pruned as it is dependent on the underlying cause.

5.3 Prediction using causal links

To evaluate the causal links generated by WIN, we use it to extract features for predicting stock prices using the exact data and prediction setting used in Kang et al. (2017). Note that the features and topics were not chosen in an unsupervised manner in Kang et al. (2017), but rather based on a semantic parser. Once the features are extracted from WIN, we use the past values of stock prices and time series corresponding to the incoming word edges of WIN to predict the future values

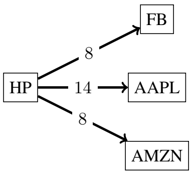

Figure 2: Inter-stock influencer links

[1] https://github.com/dykang/cgraph
[2] https://finance.yahoo.com

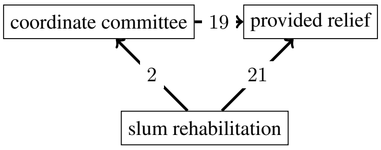

Figure 3: WIN highlighting the underlying cause

Table 1: Stock price prediction error using WIN

Step size	C_{best}	WIN_{uni}	WIN_{bi}	WIN_{both}
1	1.96	0.022	0.023	$\underline{0.020}$
3	3.78	0.022	0.023	$\underline{0.022}$
5	5.25	0.022	0.023	$\underline{0.021}$

of the stock prices using the multivariate regression equation used to determine Granger Causality. The results shown in Table 1 is the root mean squared error (RMSE) calculated on a 30 day window averaged by moving it by 10 days over the period and hence is directly comparable to (Kang et al., 2017)'s CGRAPH - C_{best}. The mean absolute error (MAE) for the same set of evaluations is within 0.003 of the RMSE, which indicates that the variance of the errors is also low. As compared to their best error, WIN from unigrams, bigrams or both obtain two orders lower error and significantly outperforms CGRAPH, which also includes features from topics and sentiments from tweets. We attribute this gain to the flexibility of WIN's Lasso Granger method to produce sparse graphs as compared to CGRAPH's Vector Auto Regressive model with exogenous variables which uses a fixed number (10) of incoming edges per node. This imposes an artificial bound on sparsity thereby losing valuable information. We overcome this in WIN using a suitable penalty term (λ) in the Lasso method.

The causal links in WIN are also more generic (Table 2) than the ones described in CGRAPH. The nodes of CGRAPH are tuples extracted from a semantic parser (SEMAFOR) based on evidence of causality in a sentence. WIN poses no such restriction and and derives topical (unfriended, FB) and inter-topical (healthcare, AMZN), sparse, latent and semantic relationships.

Table 2: Stock price predictive features from WIN

Stock symbol	Prediction indicators
AAPL	workplace, shutter, music
AMZN	healthcare, HBO, cloud
FB	unfriended, troll, politician
GOOG	advertisers, artificial intelligence, shake-up
HPQ	China, inventions, Pittsburg
IBM	64 GB, redesign, outage
MSFT	interactive, security, Broadcom
ORCL	corporate, investing, multimedia
TSLA	prices, testers, controversy
YHOO	entertainment, leadership, investment

5.4 Causal evidence in WIN

To validate the causal links in WIN, we extracted word pairs which depicted direct causal relationships in the news corpus. We narrowed down the search to words surrounding verbs which depict the notion of causality like “caused”, “effect” and manually verified that these word pairs were indeed causal. We then searched the shortest path in WIN between these word pairs. 67% of the word pairs which were manually identified to be causal in the news text through causal indicator words such as “caused”, were linked in WIN through direct edges, while the rest were linked through an intermediate relevant node. As seen in Table 3, the bigram involving the word in the path is relevant to the context in which the causality is established. The time lags in the path show that the influence between events are at different time lags. We also qualitatively verified that two unrelated words are either not connected or have a path length greater than 2, which makes the relationship weak.

Table 3: Comparison with manually identified influence from news articles

Word pairs	Words of the influence path
price, project	price-hike –(19)– power-project
land, budget	allot-land –(22)– railway-budget
price, land	price-hike –(12)– land
strike, law	terror-strike –(25)– law ministry
land, bill	land-reform –(25)– bill-pass

6 Conclusions

In this paper, we have presented WIN, a framework that learns latent word relationships from news streams in an unsupervised manner for stock price prediction. This prediction model considerably lowers the error as compared to a related causal graph method by capturing rich inter-topical features. In future work, we aim to extend the concept of *influencer network* for other types of text abstraction, like word embeddings and explore influencer network based econometric predictive models.

References

Giuseppe Amodeo, Roi Blanco, and Ulf Brefeld. 2011. Hybrid models for future event prediction. CIKM '11, pages 1981–1984.

Andrew Arnold, Yan Liu, and Naoki Abe. 2007. Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 66–75, New York, NY, USA. ACM.

Mehmet Balcilar, Rangan Gupta, and Clement Kyei. 2017. Predicting stock returns and volatility with investor sentiment indices: A reconsideration using a nonparametric causalityinquantiles test. *Bulletin of Economic Research*, 70(1):74–87.

Ivo Bernardo, Roberto Henriques, and Victor Lobo. 2018. Social market: Stock market and twitter correlation. In *Intelligent Decision Technologies 2017*, pages 341–356, Cham. Springer International Publishing.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Ching-Yun Chang, Yue Zhang, Zhiyang Teng, Zahn Bozanic, and Bin Ke. 2016. Measuring the information content of financial news. In *COLING*, pages 3216–3225. ACL.

Dehua Cheng, Mohammad Taha Bahadori, and Yan Liu. 2014. Fblg: A simple and effective approach for temporal dependence discovery from time series data. KDD '14, pages 382–391.

Ali F. Darrat, Maosen Zhong, and Louis T.W. Cheng. 2007. Intraday volume and volatility relations with and without public news. *Journal of Banking and Finance*, 31(9):2711 – 2729.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using structured events to predict stock price movement: An empirical investigation.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *IJCAI*, pages 2327–2333. AAAI Press.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2016. Knowledge-driven event embedding for stock prediction. In *COLING*, pages 2133–2142. ACL.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 294–303, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pegah Falinouss. 2007. Stock trend prediction using news events. *Masters thesis*.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar):1289–1305.

Clive WJ Granger, Bwo-Nung Huangb, and Chin-Wei Yang. 2000. A bivariate causality between stock prices and exchange rates: evidence from recent asianflu. *The Quarterly Review of Economics and Finance*, 40(3):337–354.

Michael Hagenau, Michael Liebmann, and Dirk Neumann. 2013. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, 55(3):685 – 697.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, and Jong-Hoon Oh. 2015. Generating event causality hypotheses through semantic relations. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2396–2403. AAAI Press.

Johan Hovold. 2005. Naive bayes spam filtering using word-position-based attributes. In *CEAS*, pages 41–48.

Joshi Kalyani, H. N. Bharathi, and Rao Jyothi. 2016. Stock trend prediction using news sentiment analysis. *CoRR*, abs/1607.01958.

Dongyeop Kang, Varun Gangal, Ang Lu, Zheng Chen, and Eduard Hovy. 2017. Detecting and explaining causes from text for a time series event. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2758–2767, Copenhagen, Denmark. Association for Computational Linguistics.

Noriaki Kawamae. 2011. Trend analysis model: trend consists of temporal words, topics, and timestamps. WSDM '11, pages 317–326.

Zornitsa Kozareva. 2012. Cause-effect relation learning. In *Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing*, TextGraphs-7 '12, pages 39–43, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chen Luo, Jian-Guang Lou, Qingwei Lin, Qiang Fu, Rui Ding, Dongmei Zhang, and Zhe Wang. 2014. Correlating events with time series for incident diagnosis. KDD '14, pages 1583–1592.

Christopher D Manning, Hinrich Schütze, et al. 1999. *Foundations of statistical natural language processing*, volume 999. MIT Press.

Huina Mao, Scott Counts, and Johan Bollen. 2011. Predicting financial markets: Comparing survey, news, twitter and search engine data. *Arxiv preprint*.

Mariusz Maziarz. 2015. A review of the granger-causality fallacy. *The Journal of Philosophical Economics*, 8(2):6.

Nicolai Meinshausen and Peter Bühlmann. 2006. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 64–75. The COLING 2016 Organizing Committee.

Kira Radinsky and Eric Horvitz. 2013. Mining the web to predict future events. WSDM '13, pages 255–264. ACM.

Tushar Rao and Saket Srivastava. 2012. Analyzing stock market movements using twitter sentiment analysis. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ASONAM '12, pages 119–123, Washington, DC, USA. IEEE Computer Society.

Y. Shynkevich, T. M. McGinnity, S. Coleman, and A. Belatreche. 2015. Stock price prediction based on stock-specific and sub-industry-specific news articles. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

X-Q Sun, Shen H-W, and Cheng X-Q. 2014. Trading network predicts stock price. *Scientific Reports. 2014;4:3711. doi:10.1038/srep03711.*

Robert Tibshirani. 1994. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Carmen K Vaca, Amin Mantrach, Alejandro Jaimes, and Marco Saerens. 2014. A time-based collective factorization for topic discovery and monitoring in news. WWW '14, pages 527–538.

Ishan Verma, Lipika Dey, and Hardik Meisheri. 2017. Detecting, quantifying and accessing impact of news events on indian stock indices. In *Proceedings of the International Conference on Web Intelligence*, WI '17, pages 550–557, New York, NY, USA. ACM.

Yu Wang, Eugene Agichtein, and Michele Benzi. 2012. Tm-lda: efficient online modeling of latent topic transitions in social media. KDD '12, pages 123–131. ACM.

Boyi Xie, Rebecca J. Passonneau, Leon Wu, and Germán Creamer. 2013. Semantic frames to predict stock price movement. In *ACL (1)*, pages 873–883. The Association for Computer Linguistics.