Tracking interactions across business news, social media, and stock fluctuations

Double Blind and Peer Review

No Institute Given

Abstract. In this paper we study the interactions between how companies are mentioned in news, their presence on social media, and daily fluctuation in their stock prices. Our experiments demonstrate that for some entities these time series can be correlated in interesting ways, though for others the correspondences are more opaque. In this study, social media presence is measured by counting Wikipedia page hits. This work is done in a context of building a system for aggregating and analyzing news text, which aims to help the user track business trends; we show results obtainable by the system.

1 Introduction

The nature of the complex relationships among traditional news, social media, and stock price fluctuations is the subject of active research. Recent studies in the area demonstrate that it is possible to find some correlation between stock prices and news, when the news are properly classified [9, 1]. A comprehensive overview of market data prediction from text can be found in [7]. Joint analysis of news and social media has been previously studied, *inter alia*, by [3, 8, 4, 2], the approach followed in these papers, as well as our approach, has two interrelated goals: to find information complementary to what is found in the news, and to control the amount of data that needs to be downloaded from social media. In particular, [6, 5] reported an increase in Wikipedia views for company pages and financial topics before stock market falls.

We study the interplay among business news, social media, and stock prices. We believe that the combined analysis of information derived from news, social media and financial data can be of particular interest for specialists in various areas: business analysts, Web scientists, data journalists, etc. We collect on-line news articles from multiple sources and use our text analysis system[1] to identify the business entities mentioned in the news texts, e.g., companies and products, and the associated event types such as "product launch," "recall," "investment." Using these entities we then construct queries to get the corresponding social media content and its metadata, such as, Twitter posts, YouTube videos, or Wikipedia pages. We focus on analyzing the activity of users of social media in numerical terms, rather than on analyzing the content, polarity, sentiment, etc.

The main contributions of this paper: we combine NLP with social media analysis, and discover interesting correlations between news and social media.

[1] References to our natural language processing (NLP) system withheld for anonymity.

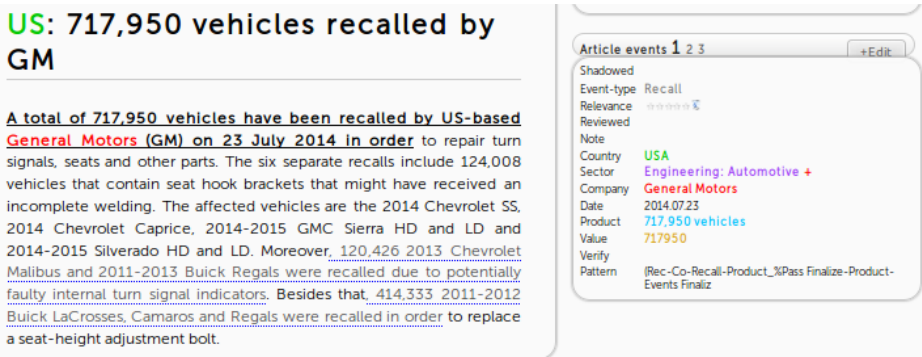


Fig. 1: A news text and a product recall event produced by our IE engine.

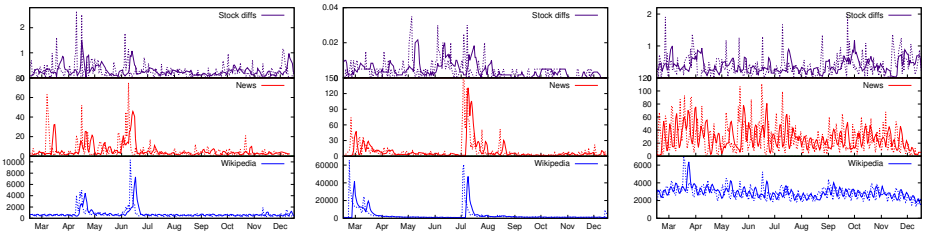
2 Process Overview

We now present the processing steps. First, the system collects unstructured text from multiple news sources on the Web. The system uses over a thousand websites which provide news feeds related to business (Reuters Business News, New York Times Business Day, etc.). Next, the NLP engine is used to discover, aggregate, and verify information obtained from the Web. The engine performs Information Extraction (IE), which is a key component of the platform that transforms facts found in plain text into a structured form.

An example event is shown in Figure 1. The text mentions a product recall event involving General Motors, in July 2014. For each event, the system extracts a set of *entities*: companies, industry sectors, products, location, date, and other attributes of the event. This structured information is stored in the database, for querying and broader analysis. The engine performs deeper semantic analysis and uses machine learning to infer some of the attributes of the events, providing richer information than general-purpose search engines.

Next, using the entities aggregated from the texts, the system builds queries for the social media sources. The role of the social media component is to enable investigation of how companies and products mentioned in the news are portrayed on social media. Our system supports content analysis from different social media services. In this paper, we focus on numerical measurement and analysis of the content. We count the number of Wikipedia views of the company and the number of its mentions in the news and then use time series correlation to demonstrate the correspondence between news and Wikipedia news. We also then correlate these with upward vs. downward stock fluctuations.

Our system has complete Wikipedia page request history for all editions, starting from early 2008, updated daily. We can instantaneously access the daily hit-count history for any Wikipedia article. Mapping a name of an entity to a Wikipedia article is not always trivial to do automatically, but the mapping appears to be easy in the vast majority of cases. Thus, we have used the Wikipedia



(a) Alstom (b) Malaysia Airlines (c) General Motors

Fig. 2: Daily differences in stock prices, number of mentions in news and number of Wikipedia hits in 2014 for three companies.

data to explore and demonstrate visibility in social media in the results presented in the following section.

3 Results

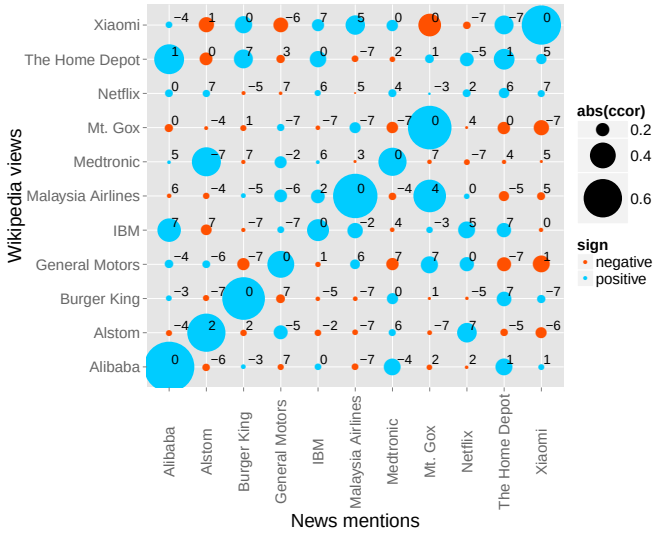
In this section we demonstrate results that can be obtained using this kind of processing. We present two types of results: A. visual analysis of correspondence between Wikipedia views, news hits and stock prices, and B. time-series correlations between news hits and Wikipedia views.

In the first experiment we chose three companies—Alstom, Malaysia Airlines, and General Motors. We present the number of mentions in the news, the number of views of the company's English-language Wikipedia page, and stock data, using data from March to December 2014.

In each figure, the top plot shows the daily *difference* in stock price—the absolute value of the opening price on a given day minus price on the previous day, obtained from Yahoo! Finance. The middle plot shows the number of mentions of the company in news. The bottom plot shows the number of hits on the company's Wikipedia page. In each plot, the dashed line represents the daily values and the bold line is the value smoothed over three days.

Figure 2a plots the data for the French multinational Alstom. The company is primarily know for its train-, power-, and energy-related products and services. In the plot we can see a pattern where the stock price and news mentions seem to correlate rather closely. Wikipedia page hits show some correlation with the other plots. The news plot shows three major spikes, with two spikes in Wikipedia hits. The March peak corresponds to news about business events (investments), whereas the other peaks had a political aspect, which could trigger activity in social media; e.g., in June, the French government bought 20% of Alstom shares, which caused an active public discussion.

Malaysia Airlines suffered two severe incidents in 2014. On March 8, they lost one aircraft over the Indian Ocean, and on July 17 another was shot down in Eastern Ukraine. Strong correlation in the patterns between news mentions



Circle width represents strength of correlation; color represents sign of correlation: blue is positive, red is negative; the numbers indicate the time lag (in days) at which the correlation with the greatest magnitude is obtained for the given pair: positive lag means that Wikipedia views follow news mentions.

Fig. 3: Cross-correlation between Wikipedia views and news mentions for 11 companies.

and Wikipedia hits is clearly visible in Figure 2b. The correlation with the stock price is less clear.

Figure 2c plots the data for General Motors, which was affected by numerous product recalls throughout the year. The company has been mentioned in the news and has been looked up on Wikipedia throughout the covered period. The stock price also oscillates over the entire year.

Although most of the local oscillations are due to normal fluctuations in the weekly flow of data on the Internet (with regular dips corresponding to the weekends), some broader-range correspondence is also discernible from the plots. Note, that the IE system automatically assigns sentiment polarity to the news, classifying events as "positive" (e.g., investments, contracts, acquisitions) or "negative" (e.g., bankruptcies, layoffs, product recalls). This will form the basis for more detailed analysis of correlations with stock fluctuations in the future.

In the second experiment, we choose eleven big companies from different industry sectors, namely Alibaba, Alstom, Burger King, General Motors, IBM, Malaysia Airlines, Medtronic, Mt. Gox, Netflix, The Home Depot, and Xiaomi. For each of these companies we collect two time series: daily news mentions and Wikipedia views during time period from March to December 2014. Then we calculate the cross-correlation between all possible pairs in these dataset, for a total of 121 cross-correlations[2]. We limit the lag between time series by seven

[2] We use standard R ccp function to calculate cross-correlation.

days, based on the assumption that if there exists a connection between news and Wikipedia views it should be visible within a week.

The results of this experiment are presented in Figure 3, where the circle size represents correlation strength, the colors represents correlation size: blue means positive correlation, red negative; the numbers mean the time lag at which the highest correlation for a given company pair was obtained: positive lag means that Wikipedia views followed news mentions, negative lag means that news followed Wikipedia views.

It can be seen from the figure that the largest correlations and the lowest lags can be found on the diagonal, i.e., between news mention for a company and the number of views of the company Wikipedia page. Among the 11 companies there are two exceptions: The Home Depot and Netflix. For Netflix, news mentions and Wikipedia views do not seem to be strongly correlated with any time series. News about Alibaba show a surprising correlation with Wikipedia hits on Home Depot on the following day. At present we do not see a clear explanation for these phenomena; these can be accidental, or may indicate some hidden connections (they are both major on-line retailers).

The lag on the diagonal equals to zero in most cases, which means that in those cases the peaks occur on the same days. At a later time, we can investigate finer intervals (less than one day). We believe it would be interesting if a larger study confirmed that we can observe regular patterns in the correlations and the lags are stable—e.g., if a spike in the news regularly precedes a spike the Wikipedia views—since that would confirm that these models can have predictive power.

4 Conclusion and Future Work

We have presented a study of the interplay between company news, social media visibility, and stock prices. Information extracted from on-line news by means the of deep linguistic analysis is used to construct queries to various social media platforms. We expect that the presented framework would be useful for business professionals, Web scientists, and researchers from other fields.

The results presented in Section 3 demonstrate the utility of collecting and comparing data from a variety of sources. We were able to discover interesting correlations between the mentions of a company in the news and the views of its page in Wikipedia. The correspondence with stock prices was less obvious. We continue work on refining the forms of data presentation. For example, we have found that plotting (absolute) differences in stock prices may in some cases provide better insights than using raw stock prices.

In future work, we plan to cover a wider range of data sources and social platforms, general-purpose (e.g., YouTube or Twitter) and business-specific ones (e.g., StockTwits). We plan to analyze the social media content as well, e.g., to determine the sentiment of the tweets that mention some particular company. Covering multiple sources is important due to the different nature of the social media. Tweets are short Twitter posts, where usually a user shares her/his

impression about an entity (company or product), or posts a related link. Wikipedia, on the other hand, is used for obtaining more in-depth information about an entity. YouTube, in turn, is for both the consumption and creation of reviews, reports, and endorsements.

This phase faces some technical limitations. For example, while Twitter data can be collected through the Twitter API in near-real time, the API returns posts only from recent history (7-10 days). This means that keyword extraction and data collections should be done relatively soon after the company or product appears in the news; combined with Twitter API request limits, this poses challenges to having a comprehensive catalogue of the posts.

Our research plans include building accurate statistical models on top of the collected data, to explore the correlations, possible cause-effect relations, etc. We aim to find the particular event types (lay-offs, new products, lawsuits) that cause reaction on social media and/or in stock prices. We also aim to find predictive patterns of visibility on social media for companies and products, based on history or on typical behaviour for a given industry sector.

References

1. Boudoukh, J., Feldman, R., Kogan, S., Richardson, M.: Which news moves stock prices? A textual analysis. Tech. rep., National Bureau of Economic Research (2013)
2. Du, M., Kangasharju, J., Karkulahti, O., Pivovarova, L., Yangarber, R.: Combined analysis of news and Twitter messages. In: Joint Workshop on NLP&LOD and SWAIE: Semantic Web, Linked Open Data and Information Extraction (2013)
3. Guo, W., Li, H., Ji, H., Diab, M.T.: Linking tweets to news: A framework to enrich short text data in social media. In: Proceedings of ACL-2013 (2013)
4. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on World Wide Web. ACM (2010)
5. Moat, H.S., Curme, C., Avakian, A., Kenett, D.Y., Stanley, H.E., Preis, T.: Quantifying wikipedia usage patterns before stock market moves. Scientific reports 3 (2013)
6. Moat, H.S., Curme, C., Stanley, H., Preis, T.: Anticipating Stock Market Movements with Google and Wikipedia. In: Matrasulov, D., Stanley, H.E. (eds.) Nonlinear Phenomena in Complex Systems: From Nano to Macro Scale, pp. 47–59 (2014)
7. Nassirtoussi, A.K., Aghabozorgi, S., Wah, T.Y., Ngo, D.C.L.: Text mining for market prediction: A systematic review. Expert Systems with Applications 41(16), 7653–7670 (2014)
8. Tanev, H., Ehrmann, M., Piskorski, J., Zavarella, V.: Enhancing event descriptions through Twitter mining. In: Sixth International AAAI Conference on Weblogs and Social Media (2012)
9. Tetlock, P.C.: Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance 62(3) (2007)