Automating Financial Surveillance

Maria Milosavljevic[1], Jean-Yves Delort[1,2], Ben Hachey[1,2], Bavani Arunasalam[1], Will Radford[1,3], and James R. Curran[1,3]

[1] Capital Markets CRC Limited, 55 Harrington Street, Sydney, NSW 2000, Australia
[2] Centre for Language Technology, Macquarie University, NSW 2109, Australia
[3] School of Information Technologies, University of Sydney, NSW 2006, Australia
{maria,jydelort,bhachey,bavani,wradford,james}@cmcrc.com

Abstract. Financial surveillance technology alerts analysts to suspicious trading events. Our aim is to identify explainable false positives (e.g., caused by price-sensitive information in company news) and explainable true positives (e.g., caused by ramping in forums) by aligning these alerts with publicly available information. Our system aligns 99% of alerts, which will speed the analysts' task by helping them to eliminate false positives and gather evidence for true positives more rapidly.

Key words: Financial Surveillance, Document Categorisation, Machine Learning, Sentiment Analysis

1 Introduction

Systems for detecting trading fraud are currently used by exchanges to help manage market integrity and by trading houses to help manage compliance. These systems raise alerts based on trading history and heuristic patterns. For example, a rapid change in price with respect to historical trends might indicate market manipulation (e.g., ramping through forum posts or spam emails encouraging trades). On the other hand, these unexpected changes may be caused by legitimate price-sensitive information (e.g., earnings announcements to the exchange, macro-economic news). Exchanges and trading houses incur substantial expense employing analysts to determine whether alerts indicate unsanctioned trading that should be flagged for investigation or prosecution.

The vast majority of surveillance alerts are "explainable" via publicly-available information such as a company announcement, a news article or a post on a forum. That is, there is a high likelihood that particular information is responsible for causing the price change which led to the alert. For alerts that are not explainable, an analyst must decide whether the matter requires further investigation and is cause for prosecution.

We explore the extent to which information in the marketplace can be used to explain behavior which is identified by current alerting software. We find that approximately 29% of short-term price alerts are potentially explained by company announcements. A further 13% of alerts are aligned with company-specific news or forum postings. By analysing the relationships between companies, both in terms of sector influences and other forms of relationships, we can successfully align information to alerts in 99% of cases.

2 Background and Motivation

In an efficient market, informed investors must act on information quickly to be rewarded for their attentiveness. It has long been established that information drives investment decisions [6, 8] and that informed individuals are compensated [7]. There is a recent growth of interest in measuring the impact of information on the financial markets, both retrospectively [2, 13] and for prediction [9, 14]. Language technologies such as sentiment detection ([4, 3]) have become a popular area of research in this domain. In such cases, time is critical because stock prices effectively convey information from informed investors to the uninformed, that is, when informed investors observe information which they believe will drive the price up, they bid its price up [7]. Uninformed investors may observe this price change and act accordingly or may completely miss the opportunity to trade.

Surveillance analysts attempt to identify people behaving inappropriately with information in the marketplace. On the one hand, insiders trade on information which is not yet public which in turn affects the stock price prior to the public announcement [1]. On the other hand, investors manipulate the market by circulating unfounded information such as rumors [10]. Forums are a common venue for publishing inappropriate content and [5] has demonstrated the impact of such content on the market. Surveillance software (such as SMARTS[1]) identifies suspicious patterns in trading data and reports alerts to analysts.

We aim to automate some of the tasks which a surveillance analyst performs. A successful solution to this problem would involve supporting the analyst by:

- explaining false positive alerts, e.g. movement due to company announcements or macro-economic news, to eliminate the time spent by analysts on these;
- explaining true positive alerts, e.g. ramping in forums or spam emails, to expedite the collection of relevant information for further investigation;
- identifying market manipulation in text that cannot be detected from anomalous trading behaviour, e.g. unsuccessful or subtle ramping in forums.

We focus here on addressing the first two problems in the Australian market.

3 Data

A substantial component of our activities has been federating and processing the many sources of trade and text data, and meta-data, available in the finance domain into an experimental framework. This turned out to be surprisingly difficult because of the need to combine text and trading data at fine granularity and over such large scales. The remainder of this section describes the main data sources used in the experiments reported in this paper.

Alerts Alerts represent unusual trading activity for a given financial instrument compared to an historical benchmark. We use Australian Securities Exchange (ASX) trading data from SIRCA's Taqtic service[2], which includes aggregated price

[1] http://www.smartsgroup.com/
[2] http://www.sirca.org.au/

and volume information for best bids (the price a purchaser is willing to pay) and best asks (the price a seller is willing to accept) at any given time. The alerts are generated using the SMARTS tool suite. In particular, short term price movements are generated if a price change over 15 minutes exceeds certain thresholds. This price change value is compared to 1) a minimum threshold (3%), 2) a scaled standard deviation threshold (4σ) based on historical data from the preceding 30 calendar days, and 3) a reissue threshold that governs when an alert is re-shown to the analysts. If these thresholds are exceeded, then an alert is generated indicating unusual price movement. The issue time associated with an alert is the same as the trade that triggered the alert.

Company Announcements The first source of textual information we use is ASX company announcements. As a condition of listing on the ASX, companies are required to comply with various listing rules aimed at protecting market integrity. Among these is the continuous disclosure rule,[3] which states: "Once a company is or becomes aware of any information concerning it that a reasonable person would expect to have a material effect on the price or value of the company's securities, the entity must immediately tell ASX that information." Therefore, any unusual price-movement based on information from within a company should be preceded by an announcement. ASX announcements are obtained through SIRCA and have meta-data including broadcast time, associated ticker(s), and the announcement category, e.g. a change in directors notice. The ASX also labels announcements as price sensitive. However, we believe this labelling is oriented towards high recall because the ASX would not want to mark an announcement incorrectly as not being price sensitive. In Section 5, we report results on reproducing this labelling.

Reuters Newswire The second source of textual information we use is news from the Reuters NewsScope Archive (RNA),[4] also obtained through SIRCA. Each RNA story is coded with extensive meta-data [12] including Reuters instrument codes (RICs), which are used to identify stocks, indices and tradeable instruments mentioned in a document. For instruments traded on the ASX, RICs are created by adding ".AX" to the end of the ASX ticker code (e.g., BHP.AX for BHP Billiton traded on the ASX). Each RNA story also has meta-data that indicates its relevance to Reuters topics (e.g., interest rates, corporate results), products (e.g., commodities) and entities (e.g., US equities diary). RNA stories comprise multiple broadcast events. For a typical story, this may consist of a news alert containing a concise statement of the key information followed by a story headline and body text, followed by further updates as the story unfolds [11].

Hot Copper Forum The third source of information we use is content from Hot Copper, a discussion forum for the Australian stock market that currently

[3] http://www.asx.com.au/ListingRules/chapters/Chapter03.pdf

[4] http://thomsonreuters.com/products_services/financial/financial_products/event_driven_trading/newsscope_archive

Year	Alerts	Announcements	RNA Events	Forum Posts
2003	5 365	65 233	92 419	
2004	7 043	80 570	86 955	246 338
2005	8 773	90 484	84 537	
2006	12 110	102 235		
2007		117 469		

Table 1. Size of the alert, announcement, news and forum datasets by year.

has over 80,000 active members and more than 4,000 posts per day. We scraped the Hot Copper web site to obtain meta-data for each post including the time it was submitted, the ticker it is about, the poster and the thread it belongs to.

Table 1 shows the document counts for each type of data. The growth in market activity is evident from the substantial increases in alerts and official announcements between 2003 and 2007. This growth will need to be matched by greater resourcing of surveillance operations or smarter technology.

4 Aligning Information to Alerts

A document is aligned to an alert if there is a possibility that it is responsible for causing the price change which led to the alert. In other words, alignment characterises a potential causality relationship between a document and an alert. If causality is established then the document is said to contain "market sensitive" information or to be "price sensitive" for the ticker associated with the alert.

As noted previously, an efficient market adjusts to new information quickly, meaning that the price of a stock changes rapidly. The time period between the information being released and a resulting price movement is termed the document's "decay period". We have calculated that a one-hour decay period covers the behaviour of most stocks, so we use this as a cutoff for aligning documents to alerts. Our alignment strategies include the following:

Ticker alignment A document and an alert are aligned if the document meta-data contains the alert ticker. This is the baseline alignment method.

Sector alignment Many of the documents in our corpora do not have specific tickers listed in their meta-data. For example, 59% of RNA events which include the topic "Australia" are not associated with any ticker. We use statistical analysis to identify significant pairwise χ^2 correlations ($p < 0.01$) between RNA topic codes and sectors. Then, RNA documents which do not have tickers are labelled with multiple sectors according to the resulting rules. 5-fold cross-validation on the 2004 RNA data showed this technique achieves 90% precision, 94% recall, and 91% F-score. This resulted in 198 rules. The top four are:

Telecommunications Services $\mapsto$ Telecommunication Services
Pharmaceuticals, Health, Personal Care $\mapsto$ Health Care
Non-Ferrous Metals $\mapsto$ Materials
Banking $\mapsto$ Financials

A document and an alert are aligned if the document sector matches the sector of the alert ticker.

Alignment scheme	Document type	Coverage (%)	Cumulative coverage (%)
Ticker	A	29	29
Ticker	R	2	29
Ticker	F	28	42
Sector	R	77	85
Firm	A+R+F	96	99

Table 2. The coverage indicates the number of alerts that are aligned to at least one document of the given type following the given alignment strategy. The document types are company announcements (A), Reuters news (R) and Hotcopper forum posts (F).

Firm Relationships aligning alerts to documents which refer to related firms. Two firms can be related in many ways (e.g., partners, competitors, producer/consumer, having board members in common). Consequently, a price sensitive announcement for a firm may impact the price of related firms. To date we have focused on identifying common sector (industry group) membership. A document and an alert are aligned if the document ticker and the alert ticker have the same sector.

The results for our alignment strategies are shown in Table 4. We can link 42% of alerts to ticker-specific documents. Adding in sector influences results in alignment of 85% of alerts to documents. Finally, by combining all three approaches, we can identify at least one document in the preceding hour which may be responsible for causing the market changes which led to an alert in 99% of cases. It is also worth mentioning that while Reuters news stories with tickers cannot be aligned to many alerts, Reuters news without tickers can be aligned to 50% of alerts using the sector-level information scheme.[5]

5 Price sensitivity

We conducted an experiment on reproducing the price sensitivity labels for ASX announcements issued in 2004. We used Weka's Naïve Bayes classifier with unigram and bigrams from the title and body of the announcements as features. Infogain was used to select the top 2000 features that best discriminate between the price sensitive and non price sensitive announcements. A separate classifier was trained and tested for each of the ASX ANNOUNCEMENT TYPES. Results from 5-fold cross-validation are shown in Table 3. The overall F-measure achieved was 0.901, with good recall on the minority YES class.

6 Conclusion

This paper has presented some preliminary results towards our goal of automated financial surveillance. Our analysis demonstrates that automation will be critical

[5] Evaluation of alignment accuracy depends on annotation of true positive and false positive alerts as well as annotation of alert-document alignments, which is a matter for future work.

Sensitive	Precision	Recall	F-Measure
YES	0.787	0.868	0.826
NO	0.950	0.914	0.931

Table 3. Results for price sensitivity classification on 2004 ASX announcements

for timely investigation as information sources and trade volumes in capital markets continue to grow rapidly.

We have identified the primary sources of textual information that can potentially explain, with up to 99% coverage, the alerts presented to ASX surveillance analysts. We have also shown that price sensitivity labels on ASX announcements can be reliably reproduced automatically. These are key stages in demonstrating that (semi-)automated financial surveillance is accurate and efficient.

References

1. Aitken, M., Czernkowski, R.: Information Leakage Prior to Takeover Announcements: The Effect of Media Reports. Accounting and Business Research, 23(89) 3–20 (1992)
2. Antweiler, W., Frank, M.Z.: Is all that talk just noise? The information content of internet stock message boards. Journal of Finance, 59(3) 1259–1294 (2004)
3. Chua, C.C., Milosavljevic, M., Curran, J.R.: A Sentiment Detection Engine for Internet Stock Message Boards. In Proceedings of the Australasian Language Technology Workshop (ALTW) (2009)
4. Das, S.R., Chen, M.Y.: Yahoo! for Amazon: Sentiment extraction from small talk on the web. Management Science, 53(9) 1375–1388 (2007)
5. Delort, J-Y., Arunasalam, B., Milosavljevic, M., Leung, H.: The Impact of Manipulation in Internet Stock Message Boards. Submitted (2009)
6. Fama, E.: Efficient Capital Markets: A Review of Theory and Empirical Work. Journal of Finance, 25(2) 383–417 (1970)
7. Grossman, S.J., Stiglitz, J.E.: On the Impossibility of Informationally Efficient Markets. The American Economic Review, 70(3) 393–408 (1980)
8. Mitchell, M.L., Mulherin, J.H.: The Impact of Public Information on the Stock Market. Journal of Finance, 49(3) 923–950 (1994)
9. Mittermayer, M.: Forecasting Intraday Stock Price Trends with Text Mining Techniques. In Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04), pp. 30064.2 (2004)
10. Pound, J., Zeckhauser, R.: Clearly Heard on the Street: The Effect of Takeover Rumors on Stock Prices. Journal of Business, 63(3) 291–308 (1990)
11. Radford, W., Hachey, B., Curran, J.R., Milosavljevic, M.: Tracking Information Flow in Financial Text. In Proceedings of the Australasian Language Technology Workshop (ALTW) (2009)
12. Reuters NewsScope Archive v2.0: User Guide (2008)
13. Robertson, C., Geva, S., Wolff, R.: What types of events provide the strongest evidence that the stock market is affected by company specific news? In Proceedings of the fifth Australasian conference on Data mining and Analytics, 145–153 (2006)
14. Schumaker, R.P., Chen, H.: Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems (TOIS), 27(2) 1–19 (2009)