

Enabling Sophisticated Financial Text Mining

Calum Robertson¹

¹ Sirca, Sydney, Australia, calum.robertson@sirca.org.au

INTRODUCTION

A popular theory is that financial markets are efficient: all available information is factored into the price of an asset [1]. Academics, commercial researchers, and investors devote considerable time analysing markets with the hope of improving efficiency and/or exploiting inefficiencies for profit. The sheer volume of real time information makes it difficult for an individual to keep abreast of all information related to an asset, and impossible to keep abreast of all information related to all available assets [2].

Information commonly used to analyse market efficiency can be broadly categorised as numerical or textual. Trading data (e.g., the time series of a share price), is numerical information that is commonly used to analyse market efficiency, and consumed by algorithmic trading models so a computer can automatically trade when certain conditions are met. News, such as macroeconomic and company announcements, are sources of textual information that can have significant impact on the price of an asset [3].

Recent years has seen increasing interest in the role of news in financial markets [4-7]. The increase in the availability of significant datasets of news, and advances in text mining technologies have helped drive this research. The future is bright for financial text mining, though there are several obstacles which must be overcome to promote research in this field. In this paper we describe common data sources, research strategies in this domain, and address some of the ways we are overcoming obstacles in this field.

DATA SOURCES

A common source of news has traditionally been newspapers, and whilst early research relied on this type of news [8], recent years have seen a shift to electronic forms of news. This includes macroeconomic news [3], company announcements [4,5], online newspaper columns [6] and news distributed by financial data providers [7]. Macroeconomic news and company announcements are considered to have more impact than other types of news, such as some newspapers. Financial data providers, such as Bloomberg and Thomson Reuters provide macroeconomic news, company announcements, in addition to other sources, and are therefore preferable sources for this type of research.

Currently Sirca handles roughly 50 GB of data transmitted by Thomson Reuters a day. Virtually all of this is trading data, though there are hundreds of thousands of news messages a day. News is transmitted as a series of small messages which have to be assembled in the correct order to produce the story, and subsequent messages can alter the story if a journalist requires corrections to be made. Furthermore, redundancy is built into the system so clients can assemble stories correctly even when they have experienced communication failure.

RESEARCH STRATEGIES

Researchers of financial news have varying objectives requiring differing strategies. In this paper we will cover the three most common strategies and the requirements for each approach.

The classic use of news is in **event studies**, where the researcher seeks to find a correlation between news and abnormal trading behaviour after a certain type of news became available to the market [3,4,8]. This requires large datasets of aggregated trading data, and news. Typically, researchers are only concerned with the number of news articles of a certain type which occurred during a given period, so the content is disregarded. However, researchers have found that third party tagging provided with the news can be utilised to further restrict the type of news under investigation and subsequently improve the correlation with abnormal trading behaviour [4].

More recently researchers have become interested in **text classification**, whereby abnormal trading behaviour is predicted utilising the content of news [5, 7]. The same techniques used for event studies are used to categorise news before machine learning algorithms can be trained and tested. Typically, the content of news is aggregated by counting the number of times each term appears in a document. Dimensionality is further reduced by stemming terms, such that “finance”, “finances”, “financed” and “financing” are represented by the same stem. The choice of how terms/stems are utilised, and how machine learning is performed varies but this type of research requires both aggregated trading data and news content. Therefore these researchers require a superset of the data used for event studies, as the content of each news article is important.

Another popular strategy is to find positive words/phrases (e.g., “rise”, “better than expected”), and negative words/phrases (e.g., “fall”, “lower than expected”), in the news to infer the author **sentiment** [6]. The theory is that when investors are subjected to overwhelming volumes of news with negative sentiment they take a negative view of the asset, and conversely if there is excessive positive news then investors may be more optimistic about the prospects of the asset. Some researchers are only concerned with the linguistics of the document and therefore do not require aggregated trading data. However, those who are concerned with trading behaviour tend to be interested in the cumulative sentiment from many news articles, as opposed to the text classification approach where researchers are concerned with the individual article.

OVERCOMING OBSTACLES

There are several obstacles to promoting financial text mining research, though in this section we will briefly address how we have overcome these and how we intend to proceed in the future.

The major obstacle is the sheer **volume of data** which is transmitted by financial data providers. Sirca has a proud history of delivering both aggregated and raw historical trading data directly to academic clients, and corporate clients via Thomson Reuters. In recent years we have begun providing historical news messages, which has provided further challenges as we have hundreds of millions of message parts which combine to form tens of millions of unique stories.

Providing the functionality to perform an organised **text search** is complicated by the fact that news is distributed in 19 languages, and contains numerous tags to help users find specific types of news. To handle the multi-lingual problem we have embraced the International Components of Unicode [9] tools provided by IBM. To handle the volume of data and the numerous possible searchable criteria, we developed our own querying language and developed efficient indexing tools.

The existing search capability is sufficient to support researchers who wish to perform event studies, and for researchers who want the raw data to perform text classification and/or sentiment analysis. However, many of our clients are from finance backgrounds and have few if any tools for performing text mining. Therefore we build and maintain lists of term counts for all news stories, and are developing tools to allow researchers to acquire the necessary data for performing text classification and/or sentiment analysis. This includes the capability to obtain aggregated term/stem counts for the given search criteria.

We are developing a series of graphical tools which will be made available through web services to allow our clients to perform a series of complex, though routine, operations over both the trading and news data that meets their requirements. Our intention is to provide support for clients to interface with existing services such as OpenCalais [10] for building additional tags for the news, and Harvard Inquirer to perform sentiment analysis [11].

CONCLUSIONS

In this paper we have discussed the issues affecting large scale financial text mining, and detailed the steps we have taken to promote research in this field. We hope that feedback from our clients will enable us to develop even more sophisticated tools which will further promote financial text mining.

REFERENCES

1. Shiller, R J (2003) From Efficient Markets ... to Behavioral Finance. *Journal of Economic Perspectives* **17**, 83-104.
2. Oberlechner, T and S Hocking (2004) Information Sources, ... Market. *Journal of Economic Psychology* **25**, 407-24.
3. Ederington, L and J Lee (1995) The Short-Run *Journal of Financial & Quantitative Analysis* **30**, 117-34.
4. Kalev, P S, *et al.* (2004) Public Information Arrival and *Journal of Banking and Finance* **28**, 1441-67.
5. Mittermayer, M-A (2004) Forecasting Intraday Stock Price Trends with Text In *Proceedings of HICSS'04*.
6. Tetlock, P C (2007) Giving Content to Investor Sentiment: The Role of ... *Journal of Finance* **62**, 1139-68.
7. Robertson, C S, *et al.* (2007) News Aware Volatility Forecasting: Is the In *Proceedings of AusDM*, 157-66.
8. Cutler, P C *et al.* (1989) What Moves Stock Prices? *Journal of Portfolio Management* **15(3)**, 4-12.
9. *International Components for Unicode*. Available from: <http://site.icu-project.org/>, accessed 25 Jun 2009.
10. *OpenCalais*. Available from: <http://www.opencalais.com/>, accessed 25 Jun 2009.
11. *Harvard Inquirer*. Available from: <http://www.wjh.harvard.edu/~inquirer/>, accessed 25 Jun 2009.