Centre for Computational Finance and Economic Agents

WP062-12

Working Paper Series

Wing Lon Ng

Read for Greed - Read With Speed

Financial News Analytics - An Overview

2012

University of Essex

CCFEA

www.ccfea.net

Read for Greed - Read With Speed

Financial News Analytics - An Overview[1]

Dr Wing Lon Ng
Centre for Computational Finance and Economic Agents
School of Computer Science and Electronic Engineering
University of Essex, UK
wlng@essex.ac.uk

Abstract

It is well-known that news is essential in financial markets. The amount of news releases is continually growing due to the expansion of communication media in today's computerised world, and the ability to correctly and efficiently interpret financial information has become much more important in trading. Financial news analytics allow the trader to significantly reduce the latency time to react to all market or stock relevant events. The general idea is to automate human economic and financial reasoning when processing news streams at high speed, including collection, extraction, aggregation and categorization of data. Text analytical algorithms are then employed to transform the qualitative content of news feeds to quantitative investment relevant sentiment scores, allowing the trader to predict direction or volatility of price movements. News streams can also provide other financially relevant information to help investors to revise their strategy or rebalance their portfolio.

[1] Version: 23 July 2012

Introduction

In financial markets, trading is traditionally influenced by information, and, especially in computerised markets, trade execution is nowadays algorithmically triggered by certain signals or events. Market movements can be very sensitive to new information. Therefore, gathering and understanding news announcements of public, company-specific, macroeconomic or political sources is essential for successful investments in general, be it long term or short term. As the sources and volumes of financial news are growing every day, new technologies that automatically collect, extract and categorise news feeds become indispensable for the success of trading in electronic markets, particularly in the high-frequency domain, where orders are expected to be submitted and executed within seconds, or fractions thereof.

A human trader's capability to digest news "on the fly" is limited in speed and in quantity. However, with increasing computational capacities and faster data mining methods nowadays, new automated news reading and interpreting technologies are available to efficiently extract, aggregate and categorise large volumes of information streams in an instant. For instance, the company RavenPack states on its website[2] that their "… *system can process hundreds of news stories per second, delivered directly from news providers, storing it, analyzing it, and cross-referencing it, and deliver the results to* [their] *customers generally within 0.2 seconds of its arrival*".

The arrival of news influences the market's expectations of future price movements and has a big impact on investor sentiment and behaviour (see also Aldridge, 2010, Ch.12). Innovative messaging models that include textual news data are in demand. Most traders and market participants have access to some kind of news feeds. They have to assimilate news very fast, and, if necessary, revise their trading strategy immediately according to their updated expectations of future market movements. Investment banks and brokerage firms that do not have the means or the expertise to analyse news data can buy news streams and ready-made sentiment indices as provided by business intelligence firms such as Thompson Reuters. These news vendors apply text mining techniques to process financial information on a large scale.

Text mining refers to the application of artificial intelligence to automatically extract information from machine-readable sources (XML, html, JSON, csv, and most other electronic documents) by distinguishing and detecting linguistic patterns in written text. Using the accuracy and efficiency of a computer, text mining tools aim to mimic a human's ability to

[2] http://ravenpack.com/aboutus/technology.htm

comprehend contextual meaning by looking for repeated patterns, key terms, named entities, or similar subsets thereof in a vast collection of texts within seconds.

In recent years, more and more market participants consider the addition of financial news analytics into their algorithmic trading engine to better predict the direction or volatility of market movements before making an investment decision. Particularly in high-frequency trading, these decisions need to be made almost immediately to reduce latency. Financial news analytics, however, is not just limited to textual data mining exercises. It is a new interdisciplinary research area requiring knowledge and expertise from finance, economics, econometrics, statistical learning, artificial intelligence and natural language processing. The implementation of advanced machine-learning algorithms without the consideration of market microstructure effects and economics of financial markets will have only little value for traders and in real-world applications.

Financial news analytics combines methods from artificial intelligence, statistical learning, natural language processing, and financial econometrics to collect, categorise, interpret unstructured textual input data and convert this into metric output data, such as a financial sentiment score. In general, the implementation of news analytics in an investment strategy involves four major steps (see also Johnson, 2010, Ch. 14): news filtering (what is economically or financially relevant?), news association (what is interesting for which investor?), news interpretation (what does the news mean?), and strategy testing and implementation (how can the trader capitalise the information?).

News Filtering

Today's communication technology allows senders to disseminate their message almost instantly to a broad audience around the globe. Financial news agencies, as well as government institutions, stock exchanges and private corporations are supplying a plethora of financial information such as real-time stock prices, online news wires, accounting reports, or even online conferences through different networks, most prominently the internet (see also Ettredge et al., 2002).

Research in natural language processing has developed several concepts that allow computers to "understand" and analyse linguistic patterns in text. For example, information extraction techniques analyse unstructured text and associate repeating phrases and relationships in a text document via pattern matching. Sentence extraction, a common text summarisation method, isolates salient phrases by statistically weighting all the sentences in

the document, hence "summarising" the document (Fan et al., 2006). In general, a search engine (e.g. crawlers or scrapers) will be employed to scan different media sources and filter out financially or economically relevant information to narrow down the amount of information, which will then fed into new statistically learning dictionary (Das, 2011).

According to Mitra and Mitra (2011), financial news falls in one of the two categories with respect to their timing: (a) the arrival of announcements is either expected and thus termed "regular" or "synchronous", or (b) unexpected, termed "asynchronous". Both categories have in common that the analysed data is textual, non-numeric and qualitative in nature, which means that another sophisticated framework is needed to transform the data into a quantitative information before it can be considered in the financial decision making process. However, Groß-Klußmann and Hautsch (2011) stated that particularly non-scheduled news is traditionally considered "noisy", making it hard to differentiate between relevant information and irrelevant "noise".

As most news messages are released on a very precise schedule, many analysts only focus on news impacts based on pre-defined time intervals around the target event dates (see Figure 1), whereas other studies consider the complete trading process in the entire sample period. In the latter case, the algorithm must resist not to automatically classify non-scheduled news as relevant simply because of the "surprise factor". Failure to filter out ("false positive") noise will highly compromise the performance of the news engine and, hence, the profitability of the derived trading strategy.

In contrast to official news from established information providers, the quantity and quality of content from non-professional sources can vary significantly. While private blogs or microblogs from financial gurus are usually considered very insightful, posts on social media platforms are less reputable and often only provide inaccurate and biased information. Interestingly, however, recent studies found that although individual opinion on financially relevant topics may have no impact, the collective opinion emerging from those forums can be useful, as it has a better signal-to-noise ratio and can help assess market volatility (see also Wright, 2009; Leinweber, 2009, Ch.9). For example, the more people mention the word "euro crisis" in a certain period, the less erroneous the individual belief, and the more trustworthy the collective sentiment will be that financial products related to euro zone are or will become less attractive.

Although Twitter posts are less precise, they are available in real time and comparatively cheap to collect. Searching for particular words in the chats and monitoring their frequency in a certain time frame will give a first indication on a particular topic's relevance and novelty (see also Savage, 2011). In May 2011, the London-based investment firm Derwent Capital introduced

a hedge fund whose portfolio rebalance scheme is mainly based on sentiment in Twitter forums, reporting excess returns and an 87% accuracy of forecast.[3]

Today's Calendar » Tomorrow » This Week Filter

Date	4:13am	Currency	Impact		Detail	Actual	Forecast	Previous	Chart
Tue Jul 10	1:00am	JPY		Household Confidence		40.4	40.9	40.7	
	2:45am	EUR		French Industrial Production m/m		-1.9%	-0.9%	1.4%	
	4:00am	EUR		Italian Industrial Production m/m		0.8%	-0.3%	-2.0%	
	4:30am	GBP		Manufacturing Production m/m			0.1%	-0.7%	
	4:30am	GBP		Trade Balance			-9.0B	-10.1B	
	4:30am	GBP		Industrial Production m/m			-0.1%	0.0%	
	All Day	EUR		ECOFIN Meetings					
	8:15am	CAD		Housing Starts			203K	211K	
	10:00am	GBP		NIESR GDP Estimate				0.1%	
	10:00am	USD		IBD/TIPP Economic Optimism			46.9	46.7	
	7:00pm	AUD		RBA Deputy Gov Lowe Speaks					
	7:50pm	JPY		Tertiary Industry Activity m/m			0.2%	-0.3%	
	7:50pm	JPY		CGPI y/y			-0.9%	-0.5%	
	8:30pm	AUD		Westpac Consumer Sentiment				0.3%	
	9:30pm	AUD		Home Loans m/m			0.9%	0.2%	

Figure 1: *Sample of pre-scheduled news announcements and their estimated impact*[4]

News Association and Tagging

News stories usually have different impacts on different economic sectors and different classes of financial assets. It has become a challenging task for market participants to consider, choose and process the relevant news. Filtered information might be financially relevant, but not necessarily interesting for all investors.

Various concepts are applied to improve the streamlining of news categorisation, making sure that different kinds of information are forwarded to the corresponding group of interested recipients. Topic tracking systems that are fed with user-chosen keywords, preferences and usage profiles allow the practitioner to combine news related to specific companies and events to create his own set of trading signals. Special tagging approaches are used to compute correlations, identify exceptions and other trading signals that could potentially serve as useful triggers. Furthermore, clustering techniques, association methods and concept-linkage algorithms will be applied to detect possible secondary or tertiary subtopics with potential news announcement effects.

[3] See also: http://www.derwentcapitalmarkets.com/

4 Screenshot from www.forexfactory.com on 10th July 2012

Sometimes, even if several news messages related to the same company have similar conceptual meanings, they still might differ in their effect on particular financial variables depending on the timeliness of the announcement. For example, the asset's buy and sell prices and its liquidity is not always directly influenced by news but instead by co-movements of trading volumes, price volatility and related contagion effects (Groß-Klußmann and Hautsch, 2011). Clark (1973) first documented an increase of trading activity and price volatility (rather the price itself) after news announcements (this was at a time when financial markets were not yet computerised and far from today's standards in communication technology).

In general, for foreign exchange markets, the most influential type of information with far reaching impacts are macroeconomic news, which is regularly announced by official government bodies (particularly those from the major economies, e.g. US Federal Reserve), and then further disseminated by accredited news agencies. For trading in equity markets, corporate news is much more often and more intensively scrutinized. Events related to specific companies are often used as input variables to trigger implemented trading strategies.

News analytics information can be provided in any different machine-readable formats. Sophisticated tagging concepts allow news agencies to tailor their news feeds to customers' individual needs.

Sentiment Analysis and Interpretation

Having (a) filtered out the relevant information from the noise, and (b) sorted the condensed news feed into the corresponding category, the next step is to convert the textual information into a quantitative sentiment index. In computer science, sentiment analysis is the application of natural language processing and computational linguistics to analyse textual data with the aim to associate the (individual or collective) opinions of authors towards a certain topic. In its simplest form it focuses on sentences that indicate positive, negative, or neutral attitudes and then derives a numerical statistic from the pattern matching in order to "quantitatively summarise" the written content (for a wider discussion, see Karlgren et al., 2012).

Similarly, in financial news analytics, the researcher aims to identify and estimate the impact of news announcements obtained from different sources. The first step is again to determine the polarity (positive or negative) of a sentiment and its intensity and degree of objectivity (reliability). Computational text-analytical algorithms run linguistic analyses by scanning, extracting, and analysing keywords found in news releases and assign them numerical

values according to a predetermined thesaurus. For example, Tetlock (2007) and Tetlock et al. (2008) studied the textual content in archived daily Wall Street Journal articles and used the psychosocial dictionary Harvard-IV-4 to differentiate between positive and negative stories. In public media forums, the language used is sometimes not professional or tagged, and can be colloquial or even abbreviated, making the categorisation into the statistical learning dictionary difficult.

Furthermore, it is to be noted that different announcements can have contradicting implications (see sample announcements in Table 1 and the corresponding illustration in Figure 2). Again, news items can be weighted and aggregated to improve the signal-to-noise ratio. If a certain linguistic pattern is recognised more often, sentiment will be amplified and relevance strengthened. The news engine is expected to automate or semi-automate this approach. In particular, rich collection of different techniques and classifiers (latent semantic analysis, support vector machines, Bayes, word count, vector distance, discriminant-based, and adjective-adverb classifiers) for measuring sentiment can be considered as well as a set of metrics (sentiment error, aggregation level, phase-lag, etc.) to assess the performance on the fly (Das, 2011).

Time & Date [CET]	Currency	Expected Impact	Heading	Actual	Forecast	Source	Impli-cation
30/03/2012 08:00	EUR	Medium	German Retail Sales m/m	-1.20%	-1.10%	Destatis	↓
30/03/2012 08:45	EUR	Medium	French Consumer Spending m/m	-0.40%	3.00%	INSEE	↓
30/03/2012 11:00	EUR	Medium	CPI Flash Estimate y/y	2.70%	2.60%	Eurostat	↑
30/03/2012 11:00	EUR	Low	Italian Prelim CPI m/m	0.40%	0.50%	Istat	↓
30/03/2012 14:30	USD	Medium	Core PCE Price Index m/m	0.20%	0.10%	Bur. of Econ. Analysis	↑
30/03/2012 14:30	USD	Medium	Personal Spending m/m	0.40%	0.80%	Bur. of Econ. Analysis	↓
30/03/2012 14:30	USD	Low	Personal Income m/m	0.20%	<0.20%	Bur. of Econ. Analysis	↑
30/03/2012 15:55	USD	Low	Revised UoM Inflation Expect.	4.00%	3.90%	University of Michigan	↑

Table 1: *Sample of prescheduled news releases[5] directly related to the Euro or the US Dollar on 30th March 2012, indicated in Central European Time (see also Figure 2). In these selected announcements, scenarios where the actual value is higher than the corresponding forecast are usually considered "good" for the currency (i.e. the currency is likely to experience a subsequent appreciation).*

[5] Source: http://www.forexfactory.com

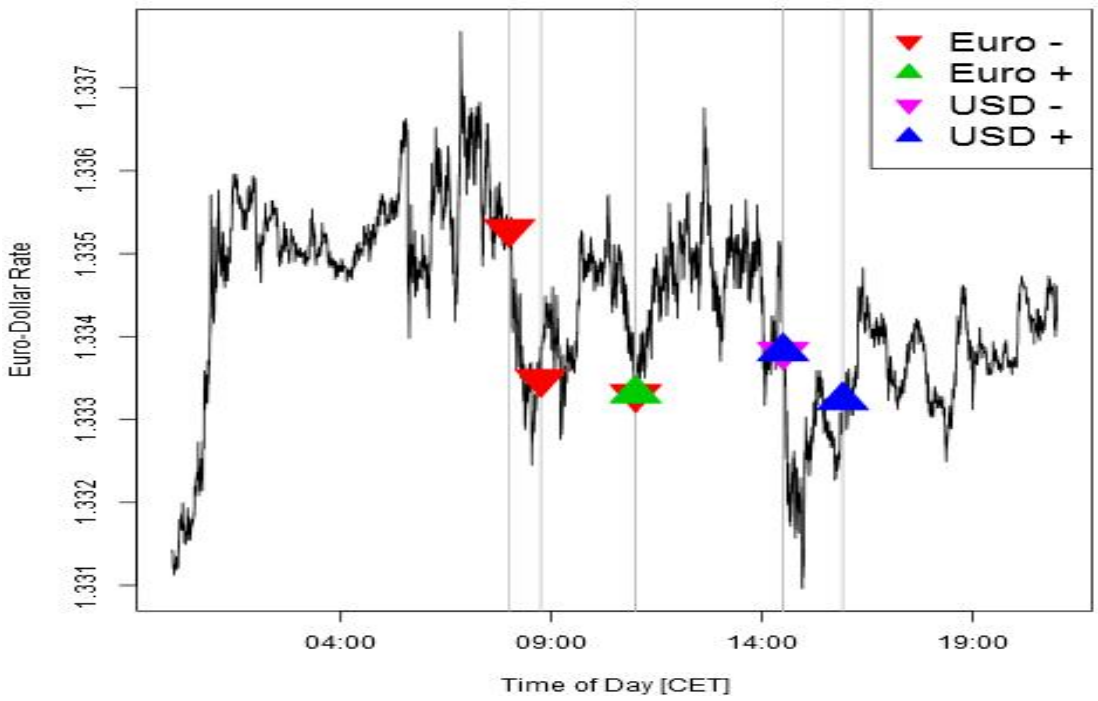

Figure 2: *Intraday High-frequency Euro/Dollar FX rates on 30th March 2012 (provided by OLSEN[6]) and the corresponding news releases (see Table 1) implying either an appreciation (+, triangle up) or a depreciation (-, triangle down) of the respective currencies.*

Testing and Implementation

Depending on the interested financial variable, the corresponding sentiment score will be fed into the relevant asset pricing or trading model for forecast purposes. If a profitable strategy can be detected, the subsequent order submission algorithm will be executed. The computerisation of this financial judgement process allows the human trader to consider a more diversified portfolio of assets, while traditional trading approaches only recommend a small number of stocks in high trading frequency (Barber and Odean, 2007).

The news-mining step should first analyse historical data from a news archive to find further or strengthen known relationships between the news and the market dynamics. More recent economic and financial news is then categorised and fed into the news archive to improve statistical learning patterns under different market environments.

Positive or negative news can help in predicting the short-term trends of the price movement (trend or contrarian), while other applications consider the prediction of intraday

[6] Source: http://www.olsen.ch/more/datasets/ . The author is grateful to Richard Olsen for providing the high-frequency FX data.

volatility, or assess which companies, industries, or entire sectors are affected. In the former case, it is assumed that news announcements can help the trader to track or to outperform a given benchmark (can be a single asset, a portfolio, or an index), using a relative return strategy. In the latter case, the trader first has to hedge his position(s) to have zero exposure; this can be achieved by buying an asset and a derivative, e.g. a put option. This allows him then to bet on temporary short-term price divergence of the associated financial products, which could be predicted by news feeds. The idea is similar to pairs trading, which is deemed market-neutral, as it does not depend on market trends but volatility, which in this kind of applications is often expressed in absolute returns.

Having correctly predicted the direction and magnitude of market movements does not yet guarantee trading success, especially if the strategy was executed poorly and revealed a certain repetitive behaviour to competitors. For example, a too aggressive selling mode will make buyers suspicious and provoke them to be more cautious. Decreased buying prices represent additional trading costs for sellers.

If the algorithm also accounts for market's over- and under- reaction with respect to the expected amount of news, be it company, sector or market specific, it can use this information to optimise order scheduling, submission and cancellation in order to reduce transaction costs. The simplest approach is to consider the number of reported relevant events, compare it with the expected number, and adjust the trading intensity depending on the difference.

Using news-based risk measures, other solutions look at the implementation of news feeds as potential circuit breakers in algorithms and procedures that influence the decision in the strategy. Triggered by a certain type, amount, or intensity of news, the circuit breaker would then change the set of rules for the strategy.

In finance, 'back-testing' refers to the *ex-post* assessment of how strategies would have performed in the past under different situations. It aims to help the machine to identify future market constellations faster and compare them with past successful or unsuccessful trading decisions under similar circumstances. For this purpose, foreign exchange and certain fixed income markets (some of which are the largest and most liquid markets) are most appropriate due to their long and well-documented history and allow for comprehensive back-testing. Also, the news database will not only help to improve the forecast accuracy in the computation of future sentiment scores, but also provide a realistic test environment for the development and assessment of new trading strategies.

Back-testing is not limited to assessing the accuracy of forecasts but can enhance post-trade analysis of execution costs. Being able to predict the intraday trading volume can help the

trader optimise his order splitting, i.e. match his own intended trading quantity against the market's dynamic supply and demand, avoiding excess exposure, reducing market impact and hence saving investment funds a lot of trading cost. Depending on the availability of additional indicators of news relevance and substantiveness, the program can also calibrate the intensity and the speed of the order submission.

Conclusion

Recent literature suggests that media influences investor sentiment, and hence market trends and movements. In today's financial sector's arms race, financial news analytics have gained a lot of attention from both buy-side and sell-side companies who intend to explore and exploit new sources of "alpha" (excess return). Particularly in high-frequency trading, news analytics are becoming more and more important, allowing traders to enhance their trading performance by utilising not only the quantitative financial market data but also the breaking news as textual data that would associate with market movements, be it price directions, volatility spikes or volume peaks.

Recent studies have demonstrated how information extraction from this kind of data can detect patterns of positive or negative stories and the corresponding reactions in the price movements for a specific asset, a market or an industry. Particularly in the case of unstructured non-pre-scheduled text, which arrives in irregular time intervals, one can take the relevance and intensity of news to predict fluctuation in volume (liquidity) and short-term volatility (potential risk).

References

Aldridge, I. (2010). High-frequency trading: a practical guide to algorithmic strategies and trading systems. Hoboken, N.J.: Wiley.

Barber, B. M., & Odean, T. (2007). All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. *Review of Financial Studies*, *21*(2), 785-818.

Clark, P. K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, *41*(1), 135–155.

Das, S. R. (2011). News Analytics : Framework , Techniques and Metrics. In G. Mitra & L. Mitra (Eds.), *Handbook of News Analytics in Finance* (pp. 43-71). Chicester: Wiley

Ettredge, M., Richardson, V. J., & Scholz, S. (2002). Timely financial reporting at corporate web sites? *Communications of the ACM*, *45*(6), 67-71.

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*, *49*(9), 76-82.

Groß-Klußmann, A., & Hautsch, N. (2011). When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. Journal of Empirical Finance, 18(2), 321-340.

Johnson, B. (2010). Algorithmic Trading and DMA: An introduction to direct access trading strategies. London: 4Myeloma Press.

Mitra, L., & Mitra, G. (2011). Applications of news analytics in finance: A review. In G. Mitra & L. Mitra (Eds.), *Handbook of News Analytics in Finance* (pp. 1-39). Chicester: Wiley.

Karlgren, J., Sahlgren, M., Olsson, F., Espinoza, F. & Hamfors, O. (2012). Usefulness of Sentiment Analysis. In R. Baeza-Yates et al. (Eds.), *Lecture Notes in Computer Science, 2012, Volume 7224, Advances in Information Retrieval* (pp. 426–435). Berlin Heidelberg: Springer.

Leinweber, D. J. (2009). Nerds on Wall Street: Math, Machines, and Wired Markets. Hoboken, N.J.: Wiley.

Savage, N. (2011). Twitter as medium and message. *Communications of the ACM*, *54*(3), 18-20.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Journal of Finance, 62(3), 1139-1168.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. Journal of Finance, 63(3), 1437-1467.

Wright, A. (2009). Our sentiments, exactly. *Communications of the ACM*, *52*(4), 14-15.