Forecasting out-of-the-ordinary financial events[1]

Marco Brambilla[2] and **Davide Greco**[3] and **Sara Marchesini**[2] and **Luca Marconi**[2] and **Mirjana Mazuran**[2] and **Martina Morlacchi Bonfanti**[2] and **Alessandro Negrini**[2] and **Letizia Tanca**[2]

Abstract. Being able to understand the financial market is very important for investors and, given the width and complexity of the topic, tools to support investor decisions are badly needed. In this paper we present Mercurio, a system that supports the decision-making process of financial investors through the automatic extraction and analysis of financial data coming from the Web. Mercurio formalizes the knowledge and reasoning of an expert in financial journalism and uses it to identify relevant events within financial newspapers. Moreover, it performs automatic analysis of financial indexes to identify relevant events related to the stock market. Then, sequential pattern mining is used to predict exceptional events on the basis of the knowledge of their past occurrences and relationships with other events, in order to to warn investors about them.

1 Introduction

Financial data are daily produced and made available on the Web, therefore the possibility to process them allows us to model and study a world that is inherently complex due to the rules governing the financial market and to the internal and external factors influencing it. Investors constantly read financial news and analyze financial indexes, using their knowledge and experience to predict market events and make profitable investments. Our research aims at developing Mercurio, a decision support system to help investors during these activities.

Mercurio identifies relevant financial events, understands how they are related to each other and exploits this knowledge to predict future happenings. It uses: (i) the knowledge of an expert in financial journalism, whose deep understanding of the news does not consist of sole natural language processing and (ii) financial indicators that provide an objective overview of the stock and, more in general, of the companies' performances. On one hand, a domain expert knows "how to" read an article and understand its meaning, especially since its literal inspection might not coincide with the real meaning of what has happened. On the other hand, financial indicators provide an impartial overview of the past and current financial situation of companies. Financial happenings are all about signals and indications that companies leave behind along their life, and that the system must capture and interpret. Investment decisions are still made by human investors, and Mercurio provides them with more knowledge, possibly hidden to human observers, to improve their decision-making process.

Among the many financial data available on the web, Mercurio looks for those that convey "important" happenings, i.e., happenings that influence and possibly shake the market: we call them *events*. Some of them are more relevant because they represent considerable changes of the financial market: we call them *catastrophes*, and they coincide with extraordinary financial moves (not necessarily negative, though), e.g. merger and acquisition, or other significant moves of the company management, or stockprice variations. The occurrence of a catastrophe is usually anticipated by "symptoms" that we call *signals*. For example, an investor might observe that often, before a crash, a company gives an interview stating that profits are increasing; from now on, whenever such an interview is published the expert will expect the related stock to fall in the stock market. Thus, an article containing an interview about increasing profit is a signal, while a stock crash is a catastrophe.

The paper is organized as follows: Section 2 briefly describes some proposals with aims similar to ours, Section 3 gives the details of the Mercurio system, Section 4 provides the current implementation state and, finally, Section 5 draws the conclusions we have currently reached and future research directions.

2 Related work

Market prediction always receives high interest in the financial literature: mostly, only numerical data are used, but some approaches exploit also textual information to increase the quality of input data and improve predictions.

Works in [3, 4, 5, 6, 7] use Automated Text Categorization techniques to predict short-term market reactions to news. Articles are categorized depending on the influence their publication has on financial indexes, and then correlated with financial trends and different approaches use different types of classifiers. Our approach differs from these as we use expert knowledge to determine the relevance of articles. Among the examined works, [8] has a similar goal as Mercurio, to find sequences of articles that anticipate a changing trend. Once again the focus is on numerical data, while we are interested in predicting strategically extraordinary financial moves.

Existing works are primarily data driven, however some proposals use a-priori knowledge about the application domain. Works in [9, 10] analyze financial articles and create a handcrafted thesaurus containing words that drive the stock prices and that are later used to predict stock prices. Similarly, [11] uses a-priori domain knowledge to predict interest rates: a cognitive map represents cause-effect relationships among the events in the domain and is used as the basis to retrieve the relevant news; these are then classified as either positive or negative according to the way they influence the rates. A work similar to ours is [12], where the objective is to predict the Tokyo stock exchange price using a-priori knowledge in the form of rules. Domain rules are defined eliciting non-numerical factors that influence the stock price, however these rules differ from ours as they

[1] This research is partially supported by the IBM Faculty Award "SOFIA: Semi-autOmatic Financial Information Analytics"
[2] Politecnico di Milano
[3] Accento

convey general knowledge about political and international events. On the contrary, we focus on financial and economic events typical of a company's life. The latter approaches differ from ours either in the way knowledge is represented or in the kind of knowledge adopted as background; we are currently trying to find a basis for an effective comparison, since the systems are not available and thus an experimental comparison on the same corpus is for the moment impossible.

To the best of our knowledge, a comprehensive system that makes use of both textual and numerical information to predict strategically extraordinary financial moves is still missing.

3 The Mercurio system

We envision an integrated and modular system that draws information from various sources and uses them appropriately with the final aim of predicting the happening of extraordinary financial events, that is, catastrophes. Finance is a kind of domain in which the key to successful data analysis is the integrated analysis of heterogeneous data, where time-dependent and highly frequent numerical data (e.g., price and volume) and textual data (e.g., news articles) should be considered jointly [13]. Both categories might encompass various data sources that can be easily added to the system (as shown in Figure 1). Each of the textual data sources is managed by an Event Recognizer that is able to extract events from the data and feed them into Mercurio. Events can be *catastrophes* (i.e. they convey considerable changes of the financial market) or *signals* (i.e. symptoms anticipating a catastrophe). Event recognition strategies vary depending on the type and nature of the managed data, for instance, each financial market (Italian, British, etc.) has its own language and dynamics, and there are differences also among financial newspapers of the same country.

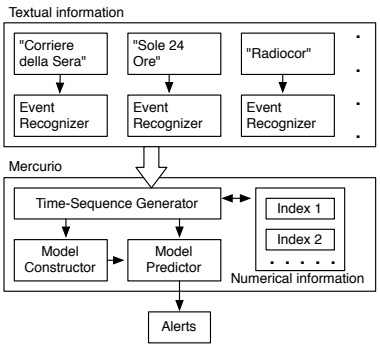

Figure 1. Mercurio architecture

In Mercurio, the events extracted from the financial news are received by the Time-Sequence Generator that arranges them on one or more timelines depending on the use the system has to make of them. If the aim is to construct a model from them all, then the Time-Sequence Generator creates a single timeline where all the received events are placed and provides this timeline as input to the Model Constructor. On the other hand, if the aim is to predict the future happenings related to specific companies, each created timeline contains only events related to a specific company, and inputs these data to the Model Predictor.

The Model Constructor module takes a sequence of events and uses Sequential Pattern Mining techniques to find frequent subsequences of events and thus creates a model of the data represented in terms of a set of sequential patterns. These patterns, together with the timeline of a company are taken as input by the Model Predictor module that uses them to forecast the happening of a certain catastrophe with respect to a certain company. The output provided by the Model Predictor is composed by a set of alerts such as "there is a P% probability that company A will encounter catastrophe C within X timeslots".

The most challenging and crucial aspect of the project is thus the process of event recognition and sequencing; however, as a side analysis, the time series generated by the Time-Sequence Generator can be compared with numerical data (indexes), arranged on their own timeline, in order to understand correlations between them.

3.1 Textual information

Events can be recognized inside textual information through text analysis; in Mercurio we propose the use of three different approaches:

- *Semantic approach*: events are recognized by means of semantic rules that formalize the knowledge and experience of our domain expert.
- *Automatic approach*: events are identified by applying clustering algorithms to financial news.
- *Hybrid approach*: a combination of the previous approaches where catastrophes are recognized with semantic rules and signals by means of clustering.

In the semantic approach, in particular, rules define a relationship between sentence structures and corresponding events. This is one of the innovative features of Mercurio and can be further improved by introducing different formalization strategies.

Some rules are independent of each other in the sense that they represent events that do not interact in any way. Other rules instead might represent events that are somehow related, e.g., one event might be a composition of two different events. Moreover, some rules are related to events that involve only one company while others might represent an interaction among different financial players. These considerations generate a rule categorization that also introduces the need for rule ordering. Such ordering is needed during the phase when rules are applied to the financial news in order to ensure the correct event recognition.

An interesting idea is to organize and formalize the semantic rules into an ontology. The concepts in the ontology would represent events, and relationships among concepts would describe how events are related to each other and how they interact and depend on each other. Each concept should be related to a set of words (or sentence structures): those that express the corresponding rule. These words could be defined ad-hoc according to the semantic rules in Mercurio, but can also originate from external ontologies describing the financial scenario or others. This addition helps to enrich the semantic formalization by taking into account both synonyms and new terms.

The use of an ontology would also allow us, through the use of inference, to discover novel information about the formalized data, possibly stimulating the discovery of new events.

3.2 Numerical information

Time-dependent series such as financial indexes are represented as values on a timeline. Each timeslot (e.g. hour or day) is associated – according to the index – with a value, e.g. an opening value, price, closing value, average and so on. The timeline containing these values can be used, in addition to the timeline containing events coming

from textual data, to enrich our data representation for the user. This is possible not only by taking into consideration single values but also by looking at some patterns inside the index.

A first technique is based on Bollinger Bands [4] that, given a numerical series, provide an upper and lower band such that the observed values usually oscillate within them. Whenever a value goes beyond these bands, it means that an unusual oscillation is happening. Thus, a trend that goes below the lower band is an unexpected price fall while a trend that goes above the upper band is an unexpected price rise.

A second technique that has been applied in the financial context is the detection of specific patterns, in terms of curve shape, inside financial time series (rather than single interesting points). The financial domain comprises some well known and meaningful trend patterns [14] such as "double top", "spike bottom", "wedge" and so on.

Another interesting approach is to approximate financial time series through the use of segments, for example by using piecewise segmentation [8]. In such way each segment represents a trend in the series, thus, we might have segments representing increasing, stable or decreasing volumes or prices.

Yet another segmentation technique specifically adopted in the financial scenario is based on Turning Points (TP) [15]. TPs are local minimum and maximum points from the historical data and are widely used in technical analysis for predicting the movement of a stock. In fact, they represent the trend of the stock change and can be used to identify the beginning or end of a transaction period.

4 Current implementation

Currently. our system predicts catastrophes by taking into consideration the information coming from financial news, while the part allowing the comparison with financial indexes is not implemented yet. The system comprises three main phases:

1. *Data acquisition and management*: financial news are extracted from web sources, structured and stored into a relational database; their contents are then cleaned and pre-processed;
2. *Event recognition*: articles are analyzed to identify both catastrophes and signals. Mercurio adopts the three different approaches introduced in Section 3: (i) semantic approach, (ii) automatic approach and (iii) hybrid approach.
3. *Model construction*: the events found in the previous step are used in combination with sequential pattern mining to learn a model, represented by means of temporal patterns, to predict the arrival of catastrophes.

4.1 Data acquisition and management

Mercurio currently monitors 250 Italian mid-cap companies and the information about them is gathered from important Italian financial and economic web sources such as "Il Sole 24 Ore", "Radiocor", "La Repubblica" and "Il Corriere della Sera". Articles about companies are extracted directly from the newspaper websites and stored into a MySQL database (our initial data contains about 14,000 articles, from year 2010 to 2015) keeping only those that: (i) are part of financial and economic sections and (ii) refer to one of the chosen companies. After this phase the article texts are cleaned by tokenization, stopword elimination and word stemming.

[4] http://www.investopedia.com/terms/b/bollingerbands.asp

Two different text pre-processing strategies are adopted, one used during the semantic event recognition and the other for the automatic event recognition. In the first strategy we kept all special characters, symbols, punctuation marks, numbers, words, company names and persons details because they are needed by the expert's rules. In the second strategy these data are not significant, sometimes even misleading when applying clustering algorithms, thus they are eliminated from the texts.

4.2 Event recognition

Events are detected through text analysis of the financial news. Mercurio implements three event recognition approaches; all of them output a temporal sequence containing the recognized events.

4.2.1 Semantic event recognition.

Mercurio uses a set of rules that formalize the recognition of relevant events inside financial news. Rules define a relationship between some keywords, regular expressions (in general, sentence structures), and corresponding events (e.g. "take" is a keyword related to an acquisition event). An article that contains the expressions defined in a rule is assigned a label corresponding to the event formalized by the rule. Each article is assigned zero, one or more labels depending on the rules it triggers.

Rules capture meanings that go beyond the sole natural language processing. For example, financial newspapers, usually, publish interviews when requested by a company. The question is: why would a company want to be interviewed? When this breaks a trend of non-communication it must be a signal. Also, an article that mentions the gross profit of a company is not a good sign because this indicator does not provide the amount of real revenue of the company, thus it could hide a negative trend of the company, whereas the net profit is not ambiguous, so this is a positive financial communication.

Currently, Mercurio encompasses 30 semantic rules, 7 of which identify catastrophes while the rest formalize signals.

4.2.2 Automatic event recognition.

This approach does not use any a-priori knowledge but relies on the detection of events by only applying clustering algorithms to the pre-processed financial news. Articles are represented in the Vector Space Model [1] where the weight of each term is the TF-IDF frequency of its occurrences in the article. Then, articles are clustered using the K-means algorithm and each article is assigned one label, corresponding to the cluster it belongs to.

The process of article clustering has proven to be quite challenging because at the end of the clustering phase we tried to interpret the results and found it impossible to distinguish between clusters representing signals and those representing catastrophes. This was a big drawback from our point of view since we were not able to understand how to predict catastrophes.

4.2.3 Hybrid event recognition.

To overcome the problem exposed above, we "added some semantics" to the automatic approach, obtaining what we called the hybrid one. In this approach, catastrophes are found by using the semantic rules that formalize catastrophic events, while the other signals are obtained by clustering all those articles that were not isolated by the rules defining catastrophic events.

4.3 Model construction

The output of the event recognition phase is a sequence of events, each associated with a timestamp that corresponds to the date and time of publishing of the article in which the event was found. Based on this sequence, Mercurio uses Sequential Pattern Mining to find "recurring" temporal patterns in the input data which are then used to predict future catastrophes.

This step is performed by using AIDA [2], a tool that encompasses both the model creation and prediction features. The tool is applied in two phases: (i) given as input a temporal sequence of events, a specific event e from the sequence and a minimum support threshold, it finds all temporal patterns that end with e and whose support is above the threshold; (ii) given the found model and a real-time flow of previously unseen articles, it predicts the happening of the learned events within a certain time span. In particular, during the prediction phase, each incoming new article is processed and labeled according to the events it triggers. Then, the system tries to match each event to the ones in the patterns of the model. If this happens, it waits for another event that would match the next event in the pattern. This process is repeated until a pattern expires because of time constraints or its last but one event is reached. When this happens, we can predict the happening of the next event, which is the one corresponding to the last node of the pattern, which, by construction, is always a catastrophe.

4.4 Experiments

Let us briefly discuss on the performance of our prototype and compare the semantic approach (SA) and hybrid approach (HA). First of all, let us recall the differences between the two approaches, in terms of article-event relationships: (i) in SA an article might contain both catastrophes and signals, while in HA this is not possible because clustering is computed only on those articles that do not trigger any catastrophe; (ii) in SA an article might not trigger any rules thus not generate any event; in HA all the articles are associated with exactly one event, either a catastrophe or a cluster label; (iii) in SA an article might trigger more than one signal, while in HA each article belongs to only one cluster, thus, it is related to only one signal. These differences make it difficult to qualitatively compare the results of the two approaches, articles that trigger the same events in SA often belong to different clusters in HA.

In the semantic approach we considered 2549 instances of events (556 of catastrophes, 1993 of signals) and, for each catastrophe, built a model to predict it. The constructed models contain an average of 9 patterns whose lengths vary between 2 and 7. In the hybrid approach we consider 3283 articles (438 catastrophes, 2845 are clustered). The constructed models contain an average of 13 patterns whose lengths vary between 2 and 6. The hybrid approach allows us to obtain a greater number of patterns w.r.t. the semantic approach and results in an increase of the average number of patterns for each catastrophe. All the constructed models were tested on previously unseen data to determine the precision and recall of the predictions. We recall that low precision means that there are many wrong predictions, i.e. many times the system predicts a catastrophe which does not actually happen, and a low recall means that there are many missed predictions, i.e. many times the system does not predict a catastrophe and the catastrophe actually happens.

The results obtained by applying the two methods vary depending on the catastrophe: (i) some catastrophes cannot be predicted because their model has only one pattern which does not appear in the testing set; (ii) some catastrophes have maximum precision and maximum recall thus they are perfectly predicted, i.e., there are only right predictions and not wrong or missed ones; (iii) other catastrophes have always maximum precision because the system makes only right predictions about them, however (iv) some have low recall which means that many times the catastrophe happens and the system was not able to predict it.

These results strongly depend on the minimum support threshold: the higher the support threshold, the higher the precision and the lower the recall; conversely, the lower the support threshold, the lower the precision and the higher the recall. In general, we noticed that both approaches offer satisfactory performances, however we are working at making the models more accurate, so that the final prototype will be based on more training data and on an integration of the two techniques.

5 Conclusion

In this paper we discussed Mercurio, a system that supports the decision-making process of investors through the automatic extraction and analysis of financial data, with the aim of predicting extraordinary financial moves. Current results are encouraging but leave space for many improvements, especially related to enrichments of the current model, such as introducing weights and polarity to each event and the use of statistical information about the whole financial market, its different sectors and each monitored company.

REFERENCES

[1] G. Salton, A. Wong, C. S. Yang. A vector space model for automatic indexing. Commun. ACM 18, 11 (November 1975), 613-620.

[2] M. Mazuran, M. Simoni, L. Tanca. AIDA: Automatic Indexing based on DAta mining. SEBD 2015. pp.176-183.

[3] S. Bacher. Mining Unstructured Financial News to Forecast Intraday Stock Price Movements. PhD Thesis. University Mannheim. 2012.

[4] G.P.C. Fung, J. Xu Yu, W. Lam. News Sensitive Stock Trend Prediction. PAKDD 2002 pp.481-493.

[5] G. Gidofalvi. Using News Articles to Predict Stock Price Movements. Department of Computer Science and Engineering, University of California, San Diego. 2001.

[6] M.A. Mittermayer. Forecasting Intraday Stock Price Trends with Text Mining Techniques. HICSS 2004.

[7] D. Peramunetilleke, R.K. Wong. Currency exchange rate forecasting from news headlines. ADC 2002. pp.131-139.

[8] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, J. Allan. Language models for financial news recommendation. CIKM 2000. pp.389-396.

[9] M.A. Mittermayer, G. F. Knolmayer. NewsCATS: A News Categorization and Trading System. ICDM 2006. pp.1002-1007.

[10] B. Wuthrich, V. Cho, S. W. Leung, D. Permunetilleke, K. Sankaran, J. Zhang. Daily stock market forecast from textual web data. ICSMC 1998. pp.2720-2725.

[11] T. Hong, I. Han, Knowledge-based data mining of news information on the Internet using cognitive maps and neural networks. Expert Systems with Applications 2002, 23(1):1-8.

[12] K. Kohara, T. Ishikawa, Y. Fukuhara, Y. Nakamura. Stock Price Prediction Using Prior Knowledge and Neural Networks. Intelligent Systems in Accounting, Finance and Management 1997, 6(1):11-22.

[13] F. Wanner, T. Shreck, W. Jentner, L. Sharalieva, D. A. Keim, Relating Interesting Quantitative Time Series Pattern with Text Events and Text Features. SPIE 2013.

[14] T. Fu, F. Chung, V. Ng, R. Luk. Evolutionary Segmentation of finalcial time series into subsequences. Evolutionary Computation 2001.

[15] J. Yin, Y. Si, Z. Gong. Financial Time Series Segmentation Based On Turning Points. ICSSE 2011.