Analyzing Stock Market Fraud Cases Using a Linguistics-Based Text Mining Approach

Mohamed Zaki[1,*] and Babis Theodoulidis[2]

[1] Cambridge Service Alliance, Department of Engineering, University of Cambridge, Cambridge, UK
mehyz2@cam.ac.uk
[2] Manchester Business School, University of Manchester, Manchester, UK
b.theodoulidis@manchester.ac.uk

ABSTRACT. The paper proposes a linguistics-based text mining approach to demonstrate the process of extracting financial concepts from the Security Exchange Commission (SEC) litigation releases (LR). The proposed approach presents the extracted information as a knowledge base to be used in market monitoring surveillance systems. Also, it facilitates users' acquisition, maintenance and access to financial fraud knowledge and improves search results in the SEC enforcement portal. Answering questions such as: who are the agents involved in the manipulation? Which patterns are associated with this manipulation? When was this manipulative action performed? This paper used the financial ontology for fraud purposes introduced by [19] to provide underlying framework for the extraction process and capture financial fraud concepts from the SEC-LR. In particular, text mining analysers have been developed to extract metadata concepts (e.g. 'LR No.', 'dates') and stock market fraud concepts (e.g. agents and manipulation types) from the actual SEC fraud case.

Keywords

Text mining, stock market fraud, stock market fraud ontology, stock market monitoring and surveillance system.

1 Introduction

The finance domain has suffered from a lack of efficiency in managing vast amounts of financial data, a lack of communication and knowledge sharing between analysts. Particularly, with the growth of fraud in financial market, cases are challenging, complex, and involve huge information that needs to be analyzed similarly to other legal cases. Gathering facts and evidence from the information available is often a very complex process. The impetus to effectively and systematically address stock financial market efficiency, including factors such as stock price manipulation, has long presented a very dynamic challenge to academia, the industry and relevant authorities. Interestingly, since 1960 the Security Exchange Commission (SEC) has prosecuted more than 24,525 fraud cases.

Many authors [1,10,12,15] have tested the impact of non-competitive behaviour in the stock market, and verified the possibilities of market manipulation. [1] presented a theoretical framework for profitable market manipulations, and provided empirical evidence using a comprehensive dataset of manipulation cases which occurred in the US stock markets and were published in SEC litigation releases from 1990 to 2001. Manipulators can artificially increase securities prices and make profits using various strategies, from classic manipulative trading practices that influence prices, to the sophistication of spam and scam manipulation using various internet channels [1]. Since the passing of the Securities Act in 1930, there is evidence that market manipulation has a significant impact on the efficiency of the securities market [11].

Despite the existence of many authoritarian regulations such as the Securities Exchange Act of 1934 and European Market Abuse Directive 200, prosecutors find it challenging to prepare a case with appropriate evidence, and the gain for an exchange of a successful prosecution would be small in comparison to the efforts and resources necessary to bring a criminal case. Fraud cases can be extremely complex and difficult to demonstrate to juries. That is why regulators only select certain cases for prosecution and prioritize instances of organized manipulation. Furthermore, few courts have experience in trying securities fraud cases.

There is an urgent need to develop a novel approach that could help regulators and relevant authorities in managing vast quantities of financial data, providing better communication and knowledge sharing among analysts, providing a mechanism to demonstrate knowledge of the processes of financial fraud, understanding and sharing financial fraud logic operations, managing relevant facts gathered for case investigations, and allowing reuse of these knowledge resources in different financial contexts [8].

Therefore, the continuous improvement and development of financial market monitoring and surveillance systems with high analytical capabilities to capture the fraud is essential to guarantee and preserve an efficient market [14]. Currently, these systems are used in limited fashion and act as a reporting archive for multiple functions within the exchange. Thus, this paper aims to provide significant cross-fertilization between financial research studies and information technology as it attempts to incorporate text mining techniques for the analysis as one of the most appropriate technological area, allowing analysis of stock market fraud documents through the development of linguistic and non-linguistic patterns. In this context, text mining is used to extract financial concepts from the SEC litigation releases and thus, provide an appropriate knowledge base about financial market manipulations. The paper provides empirical evidence of how text mining could help the market monitoring surveillance systems to explore the potential efficiency and effectiveness benefits for analysing litigation releases.

2 Previous Work

This section provides a review on existing market monitoring surveillance systems and fraud detection studies that used data mining and text mining to investigate and

detect fraudulent behaviors and ascertaining evidence of potential cases of fraud within different financial markets. In fact, these systems could support financial organizations to proactively detect transactions where market abuse is suspected.

Focusing specifically on the market manipulations domain, [9] described how the National Association of Securities Dealers (NASD) Inc. used a fraud detection system [called Advanced Detection System (ADS)] to monitor trades, to detect and identify any suspicious trading behavior for further investigation in the NASDAQ stock market. [6] work describe the Securities Observation, News, Analysis and Regulation system (SONAR), also developed by NASD. The system's main purpose is monitoring NASDAQ transactions in the stock market to detect and identify any potential insider trading and any falsification of news stories for the purposes of fraud. The work of [13] introduced a monitoring system used on the Thai Bond Market, which was commissioned by the Thai Bond Market Association (ThaiBMA). The market uses a real time approach to monitor transactions, investigate any unusual ones, and notify regulators where enforcement action is required. [4] proposed a market monitoring framework, comprising of the analysis components, tasks and flows of information of a complete financial market monitoring system. The framework is designed to have a past time or reactive monitoring engine, which is fed with either structured or unstructured data sources.

Many studies show how data mining and text mining techniques can be used in such domain. For example, [5] utilized data mining techniques (C4.5, decision tree, neural network, K-mean clustering, and logistic regressions) for the early detection of insider trading manipulation schemes before the news broke within the option market. [16] generated a conceptual framework to identify the individuals (and their communities) involved in trade-based manipulation using data mining such as Euclidian Distance (ED), Shared Nearest Neighbour (SNN), density-based algorithm (DBSCAN) and graph-partitioning algorithm (METIS). [2] employed a data mining approach to analyze two cases of manipulation in the New York Stock Exchange. The researcher used the decision trees technique to distinguish between manipulations and normal trading and to improve organizational fraud detection systems. [18] described a case study on fraud detection using data mining techniques that help analysts to identify possible instances of touting based on spam emails in the Pink Sheets market. Various data mining techniques such as decision trees, neural networks and linear regression are utilized in this emerging domain. [17] presents an exemplar case study of text mining and data mining to analyze the impact of 'stock-touting' spam e-mails and misleading press releases on trading data a real case from the over-the-counter (OTC) market, and which was prosecuted by the SEC. [3] presents a high frequency trading analysis of a particular trading scenario and discusses how quote stuffing can affect the function of trading systems.

This research contributes to the development of a comprehensive domain ontology for stock markets. Currently, the existing market monitoring systems lack a comprehensive financial knowledge base [19]. This research contributes to the development by providing an additional context for the evaluation of the domain ontology and furthermore, demonstrates how data sources such as the SEC litigation releases could be analysed using a text mining approach and could support fraud analysts in the in-

vestigation process. Finally, the paper uses SEC cases as the data source that has not been addressed in previous work.

3 Methodology

This paper used the financial ontology for fraud purposes introduced by [19] to provide underlying framework for the extraction process and capture financial fraud concepts from the SEC litigation releases. The ontology has a comprehensive financial concept system for fraud purposes, which utilized to semantically rich the knowledge base of market monitoring surveillance systems to potentially help fraud analysts to understand different manipulation patterns from prosecuted cases. In this context, this paper evaluates this ontology through a specific text mining instantiation that demonstrates the published prosecuted case in an appropriate fraud knowledge base. The role of the analysers is to identify the knowledge that lies in the prosecuted cases to be able to answer questions similar to those asked by users and analysts reading the cases themselves. Some key questions that the text mining analysers should answer include: Who is (are) the agent(s) involved in the manipulation? Which asset is being targeted? In which venue is the manipulation taking place? Which action has been performed or is planned? Which patterns are associated with this manipulation? When was this manipulative action performed? Where is the manipulator getting his profit?

3.1 Data Source

This paper used the SEC published litigation releases which are prosecuted fraud cases for the US stock markets as a main data source of fraud knowledge to be analyzed. The litigation releases are concerning civil lawsuits brought by the Commission in federal court. Each litigation release has a release number, release publication date, and action that include the defendants' names; and most of the releases have an external link to SEC complaint.

The SEC complaint documents are the reports produced by the US district court that describe violations cases of securities laws. In these documents the court describes in detail all the evidence and facts that make its decisions to prohibit the acts or practices that were the results of violation of the law or commission rules. The SEC complaint document structure consists of five main sections: 'Civil case action no' which is the file number the court allocates to the document; 'district court name', 'court clerk's office stamp' which includes his name, signature, title and filing date, 'title' including the names of the defendant and the plaintiff, and the main 'document sections' which generally include a summary of the complaint, list of defendants and relevant person entities involved in the violation, jurisdiction and the venue of the court, facts that describe the manipulation scheme and all the evidence that has been collected by the Commission to prosecute the defendants, 'fraud for reliefs and violation' includes all the acts and laws that defendants violated in such case.

Different textual sources were collected from the SEC website, such as RSS format, HTML and PDF files. The RSS format is used to download the recent litigation releases published in the SEC website. Most of the litigation releases have an HTML link that contains a short description as a summary of the case followed by a detailed description of the cases (SEC Complaints). The SEC complaints are PDF documents produced by the district courts to provide a full description of the prosecuted cases. Based on the case, the average size of these documents could vary from 10 to 60 pages. In particular, the text-mining application analyzed the third quarter of 2012 that contains 62 litigation releases with total size of the corpus is 185.1 MB.

3.2 Text Mining Application Design

This section demonstrates the design of text mining, which is constructed for the SEC fraud cases. As shown in figure 1, this study adapted the information extraction application layer introduced by the stock market fraud ontology [19]. The application layer can process all document types such as text, audio and video. Thus the TextualFormat class will have concepts related to documents used in the application, such as the SEC case study, the SEC litigation release and RSS feed. In addition, each document instance could be related to one or more annotations used to develop the linguistic patterns. Two main types of resource are used in the text-mining components, namely language resources and processing resources. Language resources contain resources such as a thesaurus, list of terms, concepts, synonyms, and types (semantic groupings of concepts). Processing resources incorporate analyzers, generators, recognizers (e.g. speech transcribers, handwriting recognizers), and retrievers (e.g. search engines).

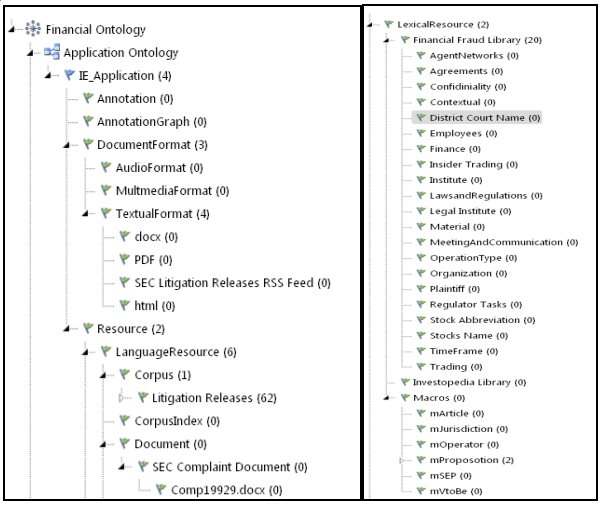

Fig 1 Text Mining Application Layer [19]

The 'corpus' concept has the actual litigation releases of the third quarter of 2012 that have been analyzed in the application, such as 'LR-22420' and 'LR-22421'. The

'document' class contains the 'insider trading' case study of Bio-Medicus, Inc.[1], which was originally a PDF file but has been converted to the 'docx' extension. This case will be used to demonstrate how text mining could automate the process of classifying financial concepts in the proposed classes in the financial fraud ontology.

The developed text-mining application uses different components to develop the advanced linguistic patterns, which could contain the following components: sub-classes from the developed library, synonyms, macros, and word gaps. 'FinancialFraud Library' is another sub-class added to the 'lexicalResource' concept which has 20 classes.This library was developed to help the text-mining application to extract concepts from the litigation releases, especially the fraud-related concepts. Based on the [19], the library has over 223 concepts that are classified and mapped to classes.

This library is used to develop and construct the advanced linguistic patterns. For example, "Confidentiality" includes a list of terms related to confidential information such as "confidential information, confidentiality, confidential advice, confidentiality agreement, confidentiality policy, code of ethical conduct, confidential, etc". "OperationType" contains terms related to trading operation such as "acquired, sold, obtained recommended, sell, buy, purchase, etc". "AgentNetwork" includes all terms representing the type of relatives who help manipulators to execute the fraud, such as "son, cousin, friend's wife, friend, relative, etc". "Employees" has all terms related to employees such as "employee, manager, chairman, board, office manager". "District Court Name" holds a list of different state courts such as "federal district court, eastern district of New York, middle district of Florida, district of New Jersey".

Furthermore, the library includes synonyms, which associate two or more concepts that have the same meaning. In particular, synonyms have been used to resolve the issue of misspelled concepts, and concepts having the same meaning, e.g. "Securities and Exchange Commission" and "Commission".

The macros is another class added to the lexicalResource concept, which represents reusable patterns; it is used to simplify the appearance of literals and word strings needing to be extracted, e.g. prepositions, articles, and verbs. Overall, the application includes six macros that support the extraction process and pattern development. For example, "mPreposition" includes a list of tokens related to prepositions such as "to, from, for, of, on, at, with, about, into, etc". "mArticles" includes a list of tokens related to English language articles such as "a, the, an, etc". "mVtobe" is another macro that holds tokens related to concepts concerning the verb 'to be, e.g. "is, are, was, were", , etc.

Process resource concept demonstrates the text-mining process and the advanced linguistic pattern approach employed on the basis of natural language processing (NLP) in order to linguistically analyze the litigation releases. 60 advanced linguistic patterns were developed to analyze the litigation releases sentence-by-sentence and to apply focus group participants' recommendations. This section demonstrates the 'Insider Trade' analyzers developed to automate the process of extracting information in the context recommended by the ontology to explain the fraud cases

[1] Case web link at http://www.sec.gov/litigation/complaints/2006/comp19929.pdf

4 Text Mining Analysis

This paper used the IBM-SPSS Modeler14 data and text mining workbench to develop the text mining analyzers [7]. The text mining application contains two components metadata analysers and SEC Complain Document Analysers.

4.1 Metadata Analysers

The text-mining metadata analysers aim to extract key metadata information from the data source. The target concepts are 'litigation release number', 'release publication dates', 'agents' which are the defendants' (individual or organization) names, 'document format type', the 'document link', 'civil case no.' which is the allocation number issued by the court, 'district court no.' including state courts or federal courts, and the 'plaintiff'. The text-mining application used some of the predefined libraries incorporated in the IBM PASW 14 software, such as date, time, person, organization. However, these patterns did not capture all related concepts within the document. Thus, extra patterns have been developed to cover the gap and to increase the accuracy of extraction. Furthermore, other analyzers were developed from scratch to match the specific structure of linguistic patterns.

Table 1 explains the patterns developed by the 'Civil Case No' analyzer. Each court has a different pattern and in order to capture most of them, 18 linguistic patterns using regular expressions were developed. These patterns have been numbered sequentially from 1–18 to avoid any break in numbering which might cause suspension or conflict when processing the document. For example, in the first pattern in Table 1, "6-12-CV-00932-JA-GJK", the regular expression was developed as [0-9]{1, 2} to match a digit repeated exactly one or two times [e.g. 6], followed by specific character "–", similarly, s[0-9]{1, 2} to match the two digits, followed by "-", followed by two character [case sensitivity is considered] [a-zA-Z] {1,2}, followed by five numbers [0-9]{3, 5}, followed by 2 characters [a-zA-Z]{1,2}, followed by three character [a-zA-Z]{1,3}. The "Civil Case No" analyzer successfully captured 100% from the 62 litigation releases used in the application.

From the 62 litigation releases, the analyzers' 'litigation release number', 'release publication dates', 'actions', 'document format type', 'document link', 'short description', 'detailed description', and 'plaintiff' have a high accuracy level of almost 98 % precision. Regarding 'Civil case no.' the analyzer achieved 98.93% precision, and for 'District court Name' 93.55% precision. In five releases the court names were not included due to the pending status of the allegation or administrative proceeding status, or were missed.

Table 1. 'Civil Case No' Analyzer

Item	Regular Expression developed Patterns	Examples of 'Civil case no.' Patterns
1	regexp1=[0-9]{1,2}-[0-9]{1,2}-[a-zA-Z]{1,2}-[0-9]{3,5}-[a-zA-Z]{1,2}-[a-zA-Z]{1,3}	Civil Action No. 6-12-CV-00932-JA-GJK
2	regexp2=[0-9]{1,2}:[0-9]{1,2}-[a-zA-Z]{1,2}-[0-9]{3,5}-[a-zA-Z]{1,3}-[a-zA-Z]{1,3}	Case No. 2:12-cv-03794-JLL-MAH
3	regexp3=[0-9]{1,2}-[a-zA-Z]{1,2}-[0-9]{3,5}-[a-zA-Z]{1,3}-[a-zA-Z]{1,3}	Case No. 09-cv-7594-KBF-THK
4	regexp4=[0-9]{1,2}-[a-zA-Z]{1,2}-[0-9]{4,5}-[a-zA-Z]{1,3}	Civil Action No.07-CV-1643-D
5	regexp5=[0-9]{1,2}-[a-zA-Z]{1,2}-[0-9]{4,5}	Case No.12-CV-6421
6	regexp6=[0-9]{1,2}:[0-9]{1,2}-[a-zA-Z]{1,2}-[0-9]{3,5}-[a-zA-Z]{1,3}	Civil Action No. 1:12-CV-02831-ODE
7	regexp7=[0-9]{1,2}:[0-9]{1,2}-[a-zA-Z]{1,2}-[0-9]{3,5}	Civil Action No. 3:12-CV-519
8	regexp8=[0-9]{1,2} [a-zA-Z]{1,3}. [0-9]{4,5}	Civil Action No. 12 Civ. 5751
9	regexp9=[a-zA-Z]{1,2}-[0-9]{1,2}-[0-9]{1,4} [a-zA-Z]{1,3}	Civil Action No. CV-11-0137 WHA
10	regexp10=[a-zA-Z]{1,2}[0-9]{1,2}-[0-9]{1,4} [a-zA-Z]{1,3}	Civil Action No. CV12-1179 Jst
11	regexp11=[a-zA-Z]{1,2} [0-9]{1,2}-[0-9]{1,4} [a-zA-Z]{1,3}	Civil Action No. CV 10-8383 DSF
12	regexp12=[a-zA-Z]{1,4}-[0-9]{1,6}[a-zA-Z]{1,3}-\([a-zA-Z]{1,4}\)	Civil Action No.SACV-1213271JST-(JPRX)
13	regexp13=[a-zA-Z]{1,2} [0-9]{1,2}-[0-9]{1,5}	Civil Action No. CV 12-3200/cv 11-09859
14	regexp14=[0-9]{1,2}-[a-zA-Z]{1,3}.?-[0-9]{3,4}	Civil Action No.12-CIV-5100/12-CIV.-5550
15	regexp15=[0-9]{1,2} [a-zA-Z]{1,2} [0-9]{3,4}	Civil Action No.07 CV 3444
16	regexp16=[0-9]{1,2}:[0-9]{1,2}-[0-9]{1,3}	Civil Action No. 1:09-361
17	regexp17=[0-9]{1,2} [0-9]{1,4}	Civil Action No. 08 2457
18	regexp18=[0-9]{1,2}-[0-9]{1,3}	Civil Action No. 12-134

Despite the size of the documents, the analyzers are good enough to demonstrate how text mining could be used for automating the analysis of SEC litigation releases. Table 2 shows the final output of the target concepts captured by the data source analyzers. For example, in litigation release number 'LR-22396' published on 20th June 2012, the defendants are 'Gary J.Mortal', 'Martel Financial Group', and 'MFG Funding'. The user can find short or detailed descriptions of the release via the link http://www.sec.gov/litigation/litreleases/2012/lr22396.htm. Therefore, the only information provided by the SEC is through the link as the release does not yet have a complain file (See also has null value) issued by the respective court. The Security and Exchange Commission, the plaintiff in this release, sent the case to the federal district court. The civil case number of the release is '12-cv-11095'.

Table 2 Metadata Data Source Ontology Text Mining Analyzers' Output

	Release No.	Date	Action (Defendants)	Short Description	Detailed Description
1	LR-22396	Wed, 20 Jun 2012 16:03:29 EDT	Gary J. Martel, d/b/a Martel Finan...	SEC CHARGES MASSACHUSETT...	Gary J. Martel, d/b/a Martel Fi...
2	LR-22398	Mon, 25 Jun 2012 13:59:04 EDT	Ralph R. Cioffi and Matthew M. T...	Court Approves SEC Settlements ...	Ralph R. Cioffi and Matthew ...
3	LR-22399	Mon, 25 Jun 2012 16:26:34 EDT	Gurudeo Persaud	The Securities and Exchange Co...	Gurudeo Persaud: Lit. Rel. N...
4	LR-22400	Mon, 25 Jun 2012 16:26:34 EDT	Manuel M. Bello, Ayuda Equity F...	The Securities and Exchange Co...	Manuel M. Bello, Ayuda Equity...
5	LR-22401	Wed, 27 Jun 2012 10:43:55 EDT	Tai Nguyen	SEC Charges Founder of Equity R...	Tai Nguyen | U.S. Securities ...
6	LR-22402	Wed, 27 Jun 2012 10:43:55 EDT	AMMB Consultant Sendirian Ber...	SEC SUES FUND ADVISER FOR ...	AMMB Consultant Sendirian ...
7	LR-22403	Thu, 28 Jun 2012 12:41:52 EDT	Harbinger Capital Partners LLC ...	SEC Charges Philip A. Falcone a...	Harbinger Capital Partners L...
8	LR-22404	Thu, 28 Jun 2012 13:27:09 EDT	H. Clayton Peterson et al.	SEC Obtains Final Judgments On...	H. Clayton Peterson et al. | ...
9	LR-22405	Thu, 28 Jun 2012 13:27:09 EDT	FalconStor Software, Inc.	FalconStor Software, Inc.	FalconStor Software, Inc. | U...

	Release No.	Document Link	See Also	Document Type	Civil Case No.	DistrictCourtName	Plaintiff
1	LR-22396	http://www.sec.gov/litigation/litreleases/2012/lr22396.htm	$null$	htm	12-cv-11095	federal district court	securities and exchange commission
2	LR-22398	http://www.sec.gov/litigation/litreleases/2012/lr22398.htm	$null$	htm	08 2457	eastern district of new york	securities and exchange commission
3	LR-22399	http://www.sec.gov/litigation/litreleases/2012/lr22399.htm	$null$	htm	6-12-cv-00932-ja-gjk	middle district of florida	securities and exchange commission
4	LR-22400	http://www.sec.gov/litigation/litreleases/2012/lr22400.htm	$null$	htm	2:12-cv-03794-jll-mah	district of new jersey	securities and exchange commission
5	LR-22401	http://www.sec.gov/litigation/litreleases/2012/lr22401.htm	$null$	htm	12-cv-5009	southern district of new york	securities and exchange commission
6	LR-22402	http://www.sec.gov/litigation/litreleases/2012/lr22402.htm	$null$	htm	1:12-cv-01052	district of columbia	securities and exchange commission
7	LR-22403	http://www.sec.gov/litigation/litreleases/2012/lr22403.htm	$null$	htm	12-cv-5029	southern district of new york	securities and exchange commission
8	LR-22404	http://www.sec.gov/litigation/litreleases/2012/lr22404.htm	$null$	htm	11 civ. 5448	southern district of new york	securities and exchange commission
9	LR-22405	http://www.sec.gov/litigation/litreleases/2012/lr22405.htm	$null$	htm	cv 12-3200	eastern district of new york	securities and exchange commission

4.2 SEC Complain Document Analysers

The text-mining application analyzed the SEC complaint document produced by the US district courts. 11 analyzers have been developed to extract the annotated financial concepts. The developed analyzers analyzed the document sentence-by-sentence. In total, 60 advanced linguistic patterns were developed to extract infor-

mation related to financial fraud and classify this information in the appropriate ontology classes, as guided by the financial ontology [19].

In particular, manipulation participants' analyzer represents the manipulator who performs the manipulation, and whether the manipulator acts by him or has networks of other agents who helped him to execute the manipulation. Furthermore, the analyzer extracts the information describing benefits behind such manipulation, whether they accrue to the manipulator or to others. Finally, it checks whether the manipulator has any previous records or history of manipulations or violations. In total, the analyzer has 9 linguistic patterns to answer these questions and describe the manipulator and his social network profile.

The first three patterns show the agent who performed the violation and the manipulation activity type, as shown in figure 2. The patterns automatically analyze the sentence, extract the concept 'Robert J. Gallivan' and classify it under the <Person> sub-category. The concept 'defendant' is classified under the <LegalTitle> sub-category, the concepts 'breached a duty of trust and confidence' and 'insider trading activity' under the <Insider Trading> sub-category, and the concept 'the C&B consulting Firm' under 'organization'. Using the regular expression the analyzer retrieves the dates corresponding to the manipulation activity. The patterns automatically classify these sentences as the manipulator who performed the manipulation and map it to the 'ManipulationParticipants\Agent\AgentCharacteristics\Individual'. The line width and node sizes in a concept graph represent the global frequency counts of the extracted concepts from the document. For example, apparently the concepts 'Robert J. Gallivan' and 'breached a duty of trust and confidence' were mentioned in the document several times, represented by the thickness (Global count 5) of the line as shown in figure 2. In order to check whether the manipulator has a previous violation record, three patterns are developed to automatically analyze the complaint document and extracts the concepts that describe the manipulation history of the manipulator. The Commission found that Gallivan, who was affiliated with a broker-dealer at the time of the scheme, wilfully violated Section 17(a) of the Securities Act of 1933, in 1975, without admitting or denying the Commission's findings, Gallivan consented to the entry of a Commission order against him in the Proceeding File No. 3-4425.

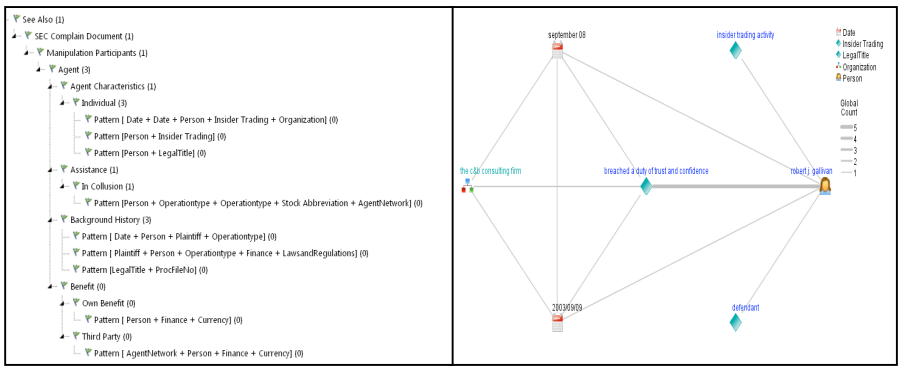

Fig 2 Manipulation Participant Analyzer

The last three patterns in the manipulation participants' analyzer are developed to extract the information that addresses the manipulator's social network, which helped him to violate the securities Harbour, Mid valley and Valencia securities (Target Assets). In this case Gallivan recommended the purchase of different stocks to his relatives, friend, and cousin to gain unlawful and combined profits reaching $58,453. The patterns automatically classified these sentences to the ontology class as the agent was in collusion with other agent networks (Assistance\In collusion). Furthermore, both the manipulator and his networks received benefits from the manipulation (Benefit\Own Benefit and Benefit\Third Party), as shown in figure 3.

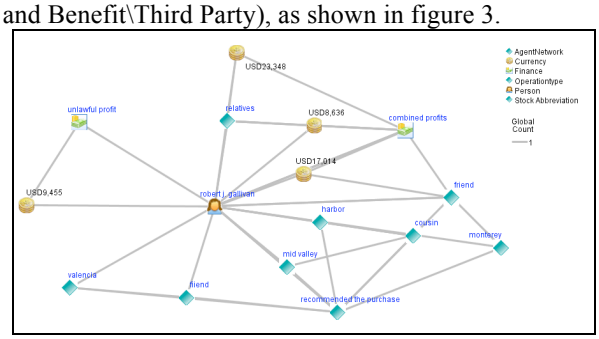

Fig 3 Agents' Social Network with Benefits

Timeline Manipulation Events and Actions Analyzer combine three analyzers 'Actions', 'Effects', and 'Time'. This analyser demonstrates the patterns developed for the three analyzers. The analysis indicates a strong relationship between the three analyzers because they describe the facts and nature of the manipulation activity executed by fraudsters. These actions could be related information-based activities such as obtaining non-public information and breach of confidence or trust, or could be trade-based activities such as buying and selling stocks to stimulate the market and violate prices. In this case, each action is associated with the temporal dimension and explains the period in which manipulator performed these manipulative activities.

These actions have an effect on the manipulated assets represented in direct or indirect benefits and unlawful profits. In this case, patterns and evidence such as legal, financial and economic unstructured information are used to trace the behavior of manipulators and show the consequences of their behavior on the market.

This analyzer contains 50 advanced linguistic patterns to extract information related to facts and the actions executed by the manipulator associated with the timeline. nine patterns are classified under 'Actions' classes , 20 patterns extract concepts related to the 'Effects' of manipulation , and 21 patterns are classified under the 'Time' class which describe the events before, during and after the fraud.

Figure 4 demonstrates the output of the 50 patterns developed for this analyzer applied to the 'insider trading' case study. The output shows the complexity of the manipulated activities executed by the agent. Most of the events are interconnected and interrelated, such as date, agreements and confidentiality, stocks prices, amount of purchases, manipulated assets, agent and his networks, amount of combined profits collected by the agent, the way of communication and meeting to obtain the non-

public material to take advantage, other evidence and patterns either legal or related to economic structure or trading used by the agent to violate the market.

In this case, the manipulator violated the prices of four stocks: Valencia Stock, Sun Country Stock, Mid valley Stock, and Harbor Stock. The manipulation actions between the four stocks are similar, as the manipulator 'Robert J. Gallivan' has breached a duty of trust and confidence of his company 'C&B Consulting Firm'. Based on his role in the company, the manipulator attended meetings and set up calls with companies and investors which gave him an opportunity to obtain non-public material. 'Robert J. Gallivan' used this information to buy these stocks and recommended them to his social network to combine profits from these trading transactions. Indeed, the agent agreed and signed that he would keep the matter confidential, but this was not the case and he breached this trust and violated the securities based on the insider information he acquired.

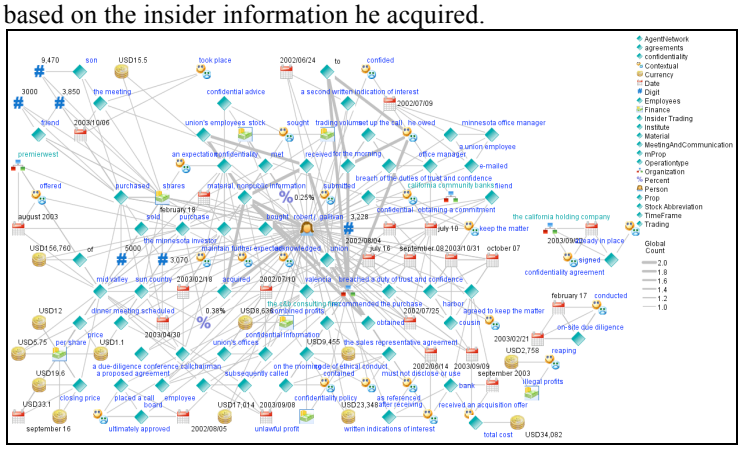

Fig 4 Timeline Manipulation Event and action Analyser Output

5 Conclusions

This paper contributes to market monitoring surveillance systems work. A linguistic based text mining approach is demonstrated for different market manipulation types based on the SEC litigation releases. The approach provides empirical evidence of how text mining could be integrated with the financial fraud ontology to improve the efficiency and effectiveness of extracting financial concepts. However, focusing on only a few case studies can potentially skew the outcome. This paper has limitations regarding the coverage of the cases and datasets.

In terms of future work, it is still possible to enhance and expand the cases to evaluate the text mining model by including new manipulation schemes and corresponding concepts on the basis of the SEC and other possible sources. For example, high frequency trading and stuff-quoting examples of possible stock-market manipulation cases should be included. Future work will pursue the full deployment of the text mining solution for the SEC litigation use as fraud knowledge management portal.

References

1. Aggarwal, R.K. and Guojun, W.U. Stock Market Manipulation -- Theory and Evidence. *Working Papers (Faculty) -- University of Michigan Business School*, 2003,
2. Diaz, D., Theodoulidis, B., and Sampaio, P. Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices. *Expert Systems with Applications 38*, 10 (2011), 12757–12771.
3. Diaz, D. and Theodoulidis, B. Financial Markets Monitoring and Surveillance: A Quote Stuffing Case Study. *Available at SSRN 2193636*, (2012).
4. Diaz, D., Zaki, M., Theodoulidis, B., and Sampaio, P. A Systematic Framework for the Analysis and Development of Financial Market Monitoring Systems. *2011 Annual SRII Global Conference*, Ieee (2011), 145–153.
5. Donoho, S. Early detection of insider trading in option markets. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.
6. Goldberg, H., Kirkland, J., Lee, D., Shyr, P., and Thakker, D. The NASD Securities Observation, News Analysis & Regulation System (SONAR). *Proceedings of the Fifteenth Conference on Innovative Applications of Artificial Intelligence*, (2003).
7. IBM. SPSS Modeler. 2013. http://www.ibm.com/software/analytics/spss/products/modeler/.
8. Kingston, J., Schafer, B., and Vandenberghe, W. Towards a financial fraud ontology: A legal modelling approach. *Artificial Intelligence and Law 12*, 4 (2004), 419–446.
9. Kirkland, J.D., Senator, T.E., and Hayden, J.J. Advanced-Detection System (ADS). *20*, 1 (1999), 55–68.
10. Laroque, R.B.; G. Using Previleged information to manipulate markets.pdf. *The quarterly Journal of Economics*, (1992).
11. Martin A. Rogoff. Legal Regulation of OTC market manipulation criticq proposal. *Finance 1*, (2012).
12. Mirchandani, V.K. and R. Increasing the ROI of Social Media Marketing REPRINT NUMBER. *MIT Slogan Management Review 54*, 1 (2012).
13. Mongkolnavin, J. and Tirapat, S. Marking the Close analysis in Thai Bond Market Surveillance using association rules. *Expert Systems with Applications 36*, 4 (2009), 8523–8527.
14. Polansky, S., Kulczak, M., and Fitzpatrick, L. NASD Market Surveillance Assessment and Recommendations. Final Report. *Achievement of Market Friendly Initiatives and Results Program (AMIR 2.0 Program)*, 2004. http://pdf.usaid.gov/pdf_docs/PNADB391.pdf.
15. Vila, J. Simple games of market manipulation. *Economics Letters 29*, (1989), 21–26.
16. Xia. Applying data mining in market abuse detection. 2007.
17. Zaki, M., Diaz, D., and Theodoulidis, B. Financial Market Service Architectures: A "Pump and Dump" Case Study. *2012 Annual SRII Global Conference*, (2012), 554–563.
18. Zaki, M., Theodoulidis, B., and Solís, D.D. "Stock-touting" through spam e-mails: a data mining case study. *Journal of Manufacturing Technology Management 22*, 6 (2011), 770–787.
19. Zaki, Mohamed and Theodoulidis, Babis. An Ontology for Monitoring and Surveillance in Financial Markets. *SSRN Electronic Journal*, (2013)