Domain Motivation and Datasets for Text Analytics in Financial Domain – Literature Review

Siang Lee Khaw
P-COM0023/19
CDS522 Text And Speech Analytics
School of Computer Sciences
University Science Malaysia
Penang, Malaysia
sianglee.khaw@student.usm.my

***Abstract*— In recent years, Text Analytics has had a strong impact on different industries. With the support of Text Analytics, structure and unstructured data could be harvested to provide more insides for the financial domain. Therefore, reviewing the recent literature on text analytics in finance can be useful for further research. This study aims to answer the research question of the domain motivation and datasets used for text analytics.**

Keywords— text analytics, financial domain, domain motivation, datasets.

I. INTRODUCTION

The financial sector includes banking, credit, financial information, investing, due diligence and trading markets produce and consume very large volumes of structured transactional data, "semi-structured" forms and filings, and "unstructured" news and communications, together with external data, like social media data and data from websites that can use text analytics to solve problems and support decision making in the financial domain.

Text analytics play a significant role in the financial sector, especially for stock predictions, customer relationship management, and cybersecurity issues, among others. Text mining extracts relevant words (N-grams) and relationships between them to categorize them and make conclusions relevant to a business problem or scientific inquiry. In other words, the goal of text mining is the extraction of knowledge and patterns from various text documents [1]. To highlight the various aspects of the use of textual mining in banking and finance, this study aims to answer the below research questions: Which datasets are mostly used for text analytics in the financial domain and for which motivation or purposes?

The paper is structured as follows. The introductory part examines the impact of text analytics in the financial domain. The second chapter will present a summary of findings taking into consideration three major areas of the sector: financial forecasting, banking, and corporate finance. The third chapter will be discussed upon the finding and conclusion in the fourth chapter.

II. LITERARY REVIEW

As mentioned in earlier sections, this paper focuses on text analytics in three sectors of the finance domain, namely financial predictions, banking, and corporate finance. The various study will be reviewed in each sector and summaries in the subsection below.

A. Financial Prediction

Utilizing textual data to improve modeling of the financial market dynamics has long been the tradition of trading practice. The growing volume of financial reports, press releases, and news articles also galvanize the wish to run this analysis automatically to keep a competitive business advantage. [2]

Wu et al. [3] mention that many researchers use dictionaries by expert definition only in text analytics for stock price prediction. So, the researcher proposed a stock price prediction model which is a combinational feature from technical analysis and sentiment analysis (SA). They focus on sentiment analysis on Stock news articles as stock prices also depend on the decisions of investors who read stock news articles. The results show that the use of sentiment analysis and technical analysis achieves higher performance than that without sentiment analysis in predicting stock price.

Khadjeh Nassirtoussi et al. [4] proposed to predict intraday directional movements of a currency pair in the foreign exchange market based on the text of financial news headlines. Their research is more emphasis on the text-mining methods and tackles some specific aspects thereof that are weak in previous works performed by other approaches. The second part of the motivation is to research the foreign exchange market, which seems not to have been researched in the previous works based on predictive text-mining. The results of their work successfully demonstrate a predictive relationship between specific market-type and the textual data of news.

Al-rubaiee et al. [5] analyzed the relationship between Saudi Twitter posts and Saudi Stock Market using machine learning methods. According to him, this is the first study performed on Saudi tweets and the Saudi stock market. The analysis shows promising results in SVM and KNN algorithms. The researchers mention that they intend to add the Saudi stock market closing values and the sentiment feature on tweets to look for patterns between the Saudi stock index and public opinion on Twitter.

Li et al. [6] acknowledged that financial volatility implies financial risk. Therefore, accurate prediction of financial volatility has been one of the unsolved issues. The study focuses on both the information volume and the information sentiment. They investigate the correlations between financial volatility, that is, asset price volatility and trading volume volatility, and the financial information. Sentiment analysis is employed to probe into online financial information to determine correlations between information sentiment and asset price volatility.

B. Banking Application

Banking is one of the fastest-growing industries in this era of globalization. Due to competitive reasons, the industries led to the rising importance of implementation of various techniques for risk controls and smooth flow of transactions over electronic mediums. [7]

Gao & Ye [8] propose a framework for data mining based on anti-money laundering, using transaction histories of customers which recognize the rare transactional pattern and by identifying suspicious data from various textual reports from law enforcement agencies. The paper also proposed some improvements to the current methodologies.

Fritz & Töws [9] used automated text mining measures to assess the quality of the annual risk report of banks in terms of customer fulfillment of regulatory requirements and identify its main drivers in a panel regression. A close reading of risk reports is expensive in terms of time and personnel and is also prone to error and subjectivity. The author finds that automated textual analysis performs quickly with the introduction of a model to consistently assess the degree to which a report fulfills the high-quality requirement, there has been no automated quality evaluation of risk reports to date.

Cook & Herron [10] quote that most financial institutions have yet to apply text analytics to suspicious activity reports, money laundering cases from the Internal Revenue Service, wire data, Transaction Review Memos, Negative News, Trade Documents, Email/Phone/Chat and Law Enforcement Requests to uncover new patterns and trends that might not surface in traditional structured data. The paper explores the potential use cases for text analytics in anti-money laundering and provides examples of entity and fact extraction and document categorization of unstructured data.

Chaturvedi & Chopra [11] quote that it is a difficult task to find and monitor opinion sites on the Web and the average human reader will have difficulty identifying relevant sites and extracting and summarizing the opinions in them. Thus, automated sentiment analysis systems are needed. The author focuses on extracting the reviewer's opinions toward banks from mouthshut.com and myBankTracker.com.

C. Corporate Finance

Corporate finance is an important aspect of the financial domain. Annual reports, corporate disclosures of a company have a lot of hidden financial context. Text-mining techniques can be employed to extract valuable information and also to predict the company's future financial sustainability. [7]

Chan & Franklin [12] Most of the existing research of text mining in financial reports or knowledge discovery from the text (KDT) relies on the identification of a predefined set of keywords. It is laborious activity and difficult to construct the event extraction templates which can anticipate all of the possible combinations of events. The author proposed a new decision-support system to predict the occurrence of an event by analyzing patterns and extracting sequences from financial reports.

Holton [13] consider disgruntled employees or dissatisfied employee are a major contributor to corporate financial fraud. Furthermore, there is less research in detecting motivational factors like employee disgruntlement. A highly accurate naïve Bayes model was proposed to predict whether email messages sent and received by the corporate employee may contain disgruntled communications which help to detect and deter fraud by an employee.

Lee et al. [14] analyze the annual reports of US-listed companies 10-K format, intending to help Korean SMEs absorb foreign knowledge and utilize both internal knowledge and external knowledge to strengthen SME competitiveness. In the paper, various text or data-mining techniques were introduced to extract the risk factors and sales information in the annual reports and examine the correlation between these text patterns and sales.

Humpherys et al. [15] implemented various text mining on the Management's Discussion and Analysis section of corporate annual financial reports. The research achieves the highest accuracy for classifying 10-K reports into fraudulent and non-fraudulent. The author suggested the potential use of linguistic analyses by auditors to flag questionable financial disclosures and to assess fraud risk under Statement on Auditing Standards No. 99.

III. SUMMARY OF FINDING

Below are summaries of the finding in table format.

Sector	Motivation	Dataset
Financial prediction	Use of text analytics to automate the process to keep the competitive business advantage. Prediction of financial volatility has been one of the unsolved issues. The combination of sentiment analysis and technical analysis achieved higher performance in predicting stock market price.	Stock news articles, financial news headlines, social media information, Twitter posts, online financial information, online newspaper articles
Banking	Improve anti-money laundering detection which recognizes the rare transactional pattern by identifying suspicious data from various textual reports. It is time-consuming and difficult to manually monitor web site or reading of risk reports and it is also prone to error and subjectivity performed by a human.	Annual risk report of banks, suspicious activity reports, cases report from IRS, transaction review memos, customer opinions on web site, social media review and Twitter review.
Corporate finance	There is less research in detecting motivational factors like employee disgruntlement. Most research on text mining in financial reports used a predefined set of keywords. Text mining also helps to strengthen local SME competitiveness from the gain of external knowledge.	Financial reports, email messages, company annual reports, Management's Discussion and Analysis section of corporate annual financial reports, social media data.

IV. Conclusion

This paper aims to conduct a literature review about domain motivation and datasets for text analytics in the financial domain. It is further separated into three specific sectors of finance. First, the research paper shows the growing importance of financial prediction, and text mining had helped improve the performance of unpredictable financial trends. Secondly, text analytics played a key role in the constant growth of banking digitization. Thirdly, the importance of text mining on financial reports and statements for corporate finance that support corporate sustainability goals.

Researchers are showing more interest in text analytics in the various domain which motivate them to constantly seek methods to improve current studies. There are still many unexplored possibilities in the financial domain, and the related research will give more opportunities for accurate predictive and more robust analytic systems. [7]

V. References

[1] Arner, D.W.; Barberis, J.; Buckley, R.P. The evolution of Fintech: A new post-crisis paradigm. Georget. J. Int. Law. 2015, 47, 1271.

[2] F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: a survey," Artificial Intelligence Review, vol. 50, no. 1, pp. 49–73, 2018.

[3] J.-L. Wu, C.-C. Su, L.-C. Yu, and P.-C. Chang, "Stock price predication using combinational features from sentimental analysis of stock news and technical analysis of trading information," International Proceedings of Economics Development and Research, 2012.

[4] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment," Expert Systems with Applications, vol. 42, no. 1, pp. 306–324, Jan. 2015, doi: 10.1016/j.eswa.2014.08.004.

[5] H. al-rubaiee, R. Qiu, and D. Li, "Analysis of the relationship between Saudi twitter posts and the Saudi stock market," Dec. 2015, pp. 660–665, doi: 10.1109/IntelCIS.2015.7397193.

[6] N. Li, X. Liang, X. Li, C. Wang, and D. Wu, "Network Environment and Financial Risk Using Machine Learning and Sentiment Analysis," Human and Ecological Risk Assessment, vol. 15, pp. 227–252, Mar. 2009, doi: 10.1080/10807030902761056.

[7] A. Gupta, V. Dengre, H. A. Kheruwala, and M. Shah, "Comprehensive review of text-mining applications in finance," Financ Innov, vol. 6, no. 1, p. 39, Nov. 2020, doi: 10.1186/s40854-020-00205-1.

[8] Z. Gao and M. Ye, "A framework for data mining-based anti-money laundering research," Journal of Money Laundering Control, 2007.

[9] D. Fritz and E. Töws, "Text Mining and Reporting Quality in German Banks - A Cooccurrence and Sentiment Analysis," Universal Journal of Accounting and Finance, vol. 6, pp. 54–81, May 2018, doi: 10.13189/ujaf.2018.060204.

[10] A. Cook and B. Herron, "Harvesting Unstructured Data to Reduce Anti-Money Laundering (AML) Compliance Risk," p. 10.

[11] D. Chaturvedi and S. Chopra, "Customers Sentiment on Banks," *International Journal of Computer Applications*, vol. 98, pp. 8–13, Jul. 2014, doi: 10.5120/17242-7578.

[12] S. W. K. Chan and J. Franklin, "A text-based decision support system for financial sequence prediction," Decision Support Systems, vol. 52, no. 1, pp. 189–198, Dec. 2011, doi: 10.1016/j.dss.2011.07.003.

[13] C. Holton, "Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem," Decision Support Systems, p. 12, 2009.

[14] B. Lee et al., "About relationship between business text patterns and financial performance in corporate data," Journal of Open Innovation: Technology, Market, and Complexity, vol. 4, Dec. 2018, doi: 10.1186/s40852-018-0080-9.

[15] S. Humpherys, K. Moffitt, M. Burns, J. Burgoon, and W. Felix, "Identification of fraudulent financial statements using linguistic credibility analysis," Decision Support Systems, vol. 50, pp. 585–594, Feb. 2011, doi: 10.1016/j.dss.2010.08.009.