Predicting Financial Market Risk with Text Analytics: The Role of Intelligence and Readability

Tong Wu[1], Hao Liu[1], Liangbo Zhang[2], and Ge Zhan[3](✉)

[1] Department of Business and Management, Beijing Normal University-Hong Kong Baptist University (BNU - HKBU United International College), 2000 Jintong Road, Zhuhai, Guangdong, China

[2] School of Economics and Management, Harbin Institute of Technology (Shenzhen), Shenzhen, China

[3] AI Analytics Lab, Beijing Normal University-Hong Kong Baptist University (BNU - HKBU United International College), Zhuhai, China

garygezhan@uic.edu.cn

Abstract. Due to the impact of the COVID-19, the global financial market suffered serious losses since 2020. Digital transformation can not only help listed firms recover from the pandemic, but may also boost the total amount of consumption and sales both in online and offline settings. The purpose of this study is to develop a new model on the effect of digitalization (three components: intelligence, platform and information) on financial performance of listed firms. We seek to identify the drivers of firm market risk by drawing from annual-report text data. The text data together with financial data of tourism firms were analyzed with text mining and Python. This will help tourism companies to understand in a more intuitive way that the benefits of digital technology for the tourism industry and how to use digital technology protecting firms from financial market risks during challenging crisis.

Keywords: Text mining · digitalization · intelligence · readability · stock market

1 Introduction

The digitalization of tourism will help tourism companies analyze financial data through digital tools to identify weaknesses and strengths, which will help them improve the level of online and offline travel products and services, as well as significantly improve the level of personalization and diversification of enterprises. Dai Bin, president of the China Tourism Research Institute, believes that modern information technology represented by the internet has driven rounds of innovation in tourism services [8]. The accelerated application of big data, cloud computing, mobile communications, and smart terminals in the tourism industry has not only brought about changes in consumption patterns, but also changed the way in which tourism services are provided.

Z. Zeng et al. (Eds.): ECIT 2022, AHE 11, pp. 585–592, 2023.
https://doi.org/10.2991/978-94-6463-005-3_59

The purpose of this study is to investigate the influence of digitalization on financial performance of tourism enterprises. We seek to identify the drivers of firm performance by drawing from tourism text data. By searching literature, we found that few studies examined relationship between the performance of tourism companies and their text data. Most of works were conducted by collecting and analyzing consumer feedback.

2 Literature Review

Textual data were collected and content-analyzed by a text-mining program aided by human judgments [16]. One way of using text mining approach is to extract information from unstructured textual data and highlights the most frequently used keywords in a text paragraph, in order to discover new knowledge [5]. In the study of Amadio and Procaccino [2], since online review archives are too large to process manually, the value that online reviews deliver to managers is a function of the ability of text mining tools to support realistic analyses. By using the data analysis method of sentiment analysis, the understanding of a large data set of employee comments from an annual employee job satisfaction survey of a US hospitality organization was improved Young & Gavade [17]. Text mining techniques are also helpful in discovering interesting areas in the messages. Loh, Lorenzi and Saldana [12] presented a recommender system to support travel agents in discovering tourist options for customers. The software used text mining techniques to analyze textual messages exchanged between a travel agent and a customer. After that, the system searches a database and retrieves tourist options classified in these interesting areas. In the study of Guerreiro and Rita [5], positive and negative reviews or word of mouth gives a clear picture about the tourism destination to the consumers or tourists and provide a support to service providers to understand the tourists needs.

3 Methodology

Data mining approaches are well used to forecast demand in tourism industry. Traditionally, many empirical studies incorporate the ESI as a measure of economic sentiment in different econometric models. Altin and Uysal [1] used ARIMA and ARDL bound test approaches to cointegration for long- and short-run elasticities, by analyzing 15 tourist-generating countries in the EU in order to explain and predict changes in tourist demand. However, when forecasting tourists arrivals in Croatia, it was found that by applying a Machine Learning Method for Decision Support and Pattern Discovery such as ANN, represents an occasion to achieve a greater accuracy if compared to results usually obtained by other methods, such as Linear Regression, and the results showed that including also data from sentiment analysis, the neural network model to predict tourists arrivals outperforms the previous obtained results [3]. In real time forecasting of hotel arrivals, it was predicted using supply-side information from business surveys, and augmented models were found to be more accurate than baseline models. Thus, there is an opportunity to extend the current business confidence surveys carried out by provincial manufacturers' chambers to the tourism industry, thereby providing effective and timely management of local tourism markets where official information may be lacking or of poor quality [6].

```
for x in range(0,i):
    j=Idea['Ideas'][x]
    #print(j)
    FOG=textstat.gunning_fog(j)
    ARI=textstat.automated_readability_index(j)
    CLI=textstat.coleman_liau_index(j)
    mean=(FOG+ARI+CLI)/3
    content.append([FOG,ARI,CLI,mean])
    dd=pd.DataFrame(content,columns=['FOG','ARI','CLI','mean'])
```

Fig. 1. Sample Program Code.

Firstly, we will use Python to process annual report data of selected firms. Before doing the word processing, all the pictures and tables in annual reports will be deleted, leaving only text data. Secondly, the fast OCR function of Foxit Advanced PDF Editor will be used to edit the text of reports in PDF format. Foxit Advanced PDF Editor is an editing software specifically for PDF files, and it has the functions of text recognition and text editing. Thirdly, Pandas package and Natural Language Toolkit (NLTK 3.6.2 documentation) will be adopted to facilitate the text mining. The text data together with financial data of tourism firms will be analyzed with STATA.

In the big data environment, similarity analysis is an effective text data audit method at present [7, 9, 11, 18]. In the big data audit environment, a feasible method to analyze text data can be TF-IDF (term frequency - inverse document frequency) technology, which is a commonly used in natural language processing (NLP). The main idea of TF-IDF is to calculate the importance of a word in the whole text library according to the frequency of the word in the text and in the whole text library. If a word or phrase appears frequently in an article and rarely in other texts, it is considered that the word or phrase is well representative and suitable for classification. TF-IDF can be used to compare the similarity of two text files, text clustering, text classification and so on (see Fig. 1 for sample program code).

$$\text{CLI} = 0.0588\text{L} - 0.296\text{S} - 15.8$$

$$\text{FOG Level} = 0.4(\text{ASL} + \text{PHW})$$

4 Results

Regression analysis was used to test the relationship between financial performance and the three independent variables, i.e. intelligence, platform and information. In regression analysis, readability and internationalization stage were controlled (see Table 1 and 2 for measurement and descriptive statistics). Table 3 and table 4 show the results of regression analysis. It is found that the effect of internationalization stage on ROA_t3 and ROE_t3 was significant at levels of $p \leq 0.01$, whether for model 1 or model 2. For all results, the internationalization stage is negatively correlated with the firm's financial performance. For ROA_t3, we found that the effect of platform on ROA_t3 in both model 1 and model 2 was significant at the level of $p \leq 0.05$ and was negatively correlated. For ROE_t3, the influence of informationization on ROE_t3 of model 1 and model 2 was significant

Table 1. Operationalization of Key Variable.

Variables	Measure
Firm performance	Return on equity for t + 3 year, return on assets for t + 3 year
CLI	CLI readability index of each sentence in annual report
FOG	FOG readability index of each sentence in annual report
Intelligence	A sum of word frequency reflecting the level of intelligence in each sentence
Platform	A sum of word frequency reflecting the development of platform in each sentence
Information	A sum of word frequency reflecting the use of information in each sentence

Table 2. Descriptive Statistics of Variables.

Variable	Obs	Mean	Std. Dev.	Min	Max
Roa_t3	15393	0.011	0.091	−0.115	0.187
Roe_t3	15393	−0.045	0.115	−0.321	0.051
Intelligence	29761	0.082	0.356	0	11
Platform	29759	0.129	0.476	0	17
Information	29759	0.056	0.316	0	5
Internationalization stage	29759	1.725	35.083	0	6
Readability	29738	0.191	0.124	0	0.5

at the level of $p \leq 0.05$, and there was a positive correlation. For model 2, we found that the effect of intelligence * readability on ROA_t3 and ROE_t3 was significant at the $p \leq 0.05$ level. There is a positive correlation between Intelligence * readability and the company's financial performance, and when the intelligence-readability value higher, the firm's financial performance better.

Closer inspection of the Table 2 and Table 3 shows intelligence and information are positively correlated with DV, which means that ROA or ROE grows when words about intelligence and information appear more frequently in the company's annual report. Another interesting finding in the result is that platform is negatively correlated with DV, which means that ROA or ROE decreases when the word platforming appears frequently in the company's annual report.

Table 3. Regression Results (Dependent Variable: ROE t3).

	Model 1	Model 2
Constant	0.172 (0.002)	0.172 (0.002)
Intelligence	0.002 (0.000)	0.002 (0.000)
Platform	−0.000 (0.000)	−0.000 (0.000)
Information	0.001 (0.000)*	0.001 (0.000)*
Readability	−0.004 (0.001)	−0.005 (0.001)
Internationalization stage	−0.157 (0.002)**	−0.157 (0.002)**
Year fixed	Yes	Yes
Intelligence * readability		0.004 (0.000)*
N	15393	15393
R^2	0.907	0.907
Root MSE	0.035	0.035

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 4. Regression Results (Dependent Variable: ROA t3).

	Model 1	Model 2
Constant	0.251 (0.001)	0.251 (0.001)
Intelligence	0.002 (0.000)	0.002 (0.000)
Platform	−0.001 (0.000)*	−0.001 (0.000)*
Information	0.000 (0.000)	0.000 (0.000)
Readability	−0.007 (0.001)	−0.008 (0.001)
Internationalization stage	−0.062 (0.000)**	−0.062 (0.000)**
Year fixed	Yes	Yes
Intelligence* readability		0.007 (0.000)**
N	15393	15393
R^2	0.921	0.921
Root MSE	0.026	0.026

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

5 Discussion

With the development of information technology, digitalization plays an increasingly important role in the tourism industry. However, it remains unclear how digitalization of tourism companies will affect their financial performance. Most studies have shown that digital technology can help tourism companies improve their business capabilities and

expand their business. For example, in the hospitality industry, Singal [15] found that consumer sentiment can be predicted by analyzing comments on the Web, which in part predicts changes in the firm's share price. Others argue that investing in digitalization does not give businesses a bigger or lasting advantage than their competitors. So far, there has not been much literature on this issue. To solve this problem, we directly use text mining to extract keywords about digitalization from the annual reports of tourism companies for analysis. The basic idea is that when digital frequencies occur in a company's annual report, organizations invest more cost and effort in digitizing.

We found that intelligence and information are positively correlated with financial performance. This is consistent with our hypotheses, which suggests digitalization can have a positive impact on the company's financial performance. An interesting finding of this study is that platform is negatively correlated with the company's financial performance, which is contrary to common sense. One possible explanation is that the platform increases the administrative costs of travel companies. It is difficult for a company to develop a new way of providing services, often requiring multi-sectoral collaboration. The provision of travel-related services on digital platforms breaks the rules of the traditional tourism industry, which could lead to increased costs of service delivery or management confusion. Second, because some of the data were collected after 2004, digital platforms were not well established fifteen years ago, especially for tourism. Digital platforming of tourism was an emerging market in China in 2004, and different types of institutional voids are often characteristic of emerging markets, mostly manifested in legal and regulatory loopholes that often reflect unethical or illegal corporate behavior [14]. Institutional voids, market competition, market uncertainty and product specificity drive the risk of digital platforms. These risks of digital platform could lead to a negative correlation between platforming and the company's financial performance [4, 13].

The results of this study have some limitations. First, we focused on the digitalization of tourism, without considering about the digitalization in other industries which might be related with tourism, such as the development of smartphones and mobile communications. Second, our data include some special periods during which events can have a serious impact on the travel industry, such as the 2008 financial crisis, COVID-19. Finally, Liu et al. [9] indicator that the market features and preferences of Chinese tourists remain distinguished from international tourists. As our research background is in China, this study may not be suitable for other tourism markets.

Acknowledgement. This study is funded by Philosophy and Social Science of Guangdong Province (grant no. GD20XGL55), Guangdong College Enhancement and Innovation Program (UICR0400011-21), 2021 UIC DBM Student Research Grant, and the UIC Research Grant (grant no. R202027).

References

1. Altin M, Uysal M (2014) Economic sentiment indicator as a demand determinant. Tour Anal 19(5):581–597
2. Amadio WJ, Procaccino JD (2016) Competitive analysis of online reviews using exploratory text mining. Tour Hosp Manag 22(2):193–210

3. Folgieri R, Baldigara T, Mamula M (2018) Sentiment analysis and artificial neural networks-based econometric models for tourism demand forecasting. In Faculty of Tourism and Hospitality Management in Opatija. Biennial International Congress. Tourism & Hospitality Industry, pp 88–97
4. Guanglu X (2021) Research on financial risk early warning of listed companies in Guangzhou based on z-score model. In 2021 2nd international conference on e-commerce and internet technology (ECIT). IEEE
5. Guerreiro J, Rita P (2020) How to predict explicit recommendations in online reviews using text mining and sentiment analysis. J Hosp Tour Manag 43:269–272
6. Guizzardi A, Stacchini A (2015) Real-time forecasting regional tourism with business sentiment surveys. Tour Manag 47(Apr):213–223
7. He Y (2021) Enterprise financial risk intelligent control system based on artificial intelligence algorithm. In 2021 international conference of social computing and digital economy (ICSCDE). IEEE
8. Ke W, Zhihong T (2021) Digital technology has helped accelerate the recovery of the tourism industry. Internet Plus brings a new experience to tourism - Rolling News - Chinese government website. http://www.gov.cn/xinwen/2021-03/24/content_5595239.htm
9. Liu Q, Liu Z, Zhang H, Chen Y, Zhu J (2021) Mining cross features for financial credit risk assessment. In Proceedings of the 30th ACM international conference on information & knowledge management, pp 1069–1078
10. Liu Y, Huang K, Bao J, Chen K (2019) Listen to the voices from home: an analysis of Chinese tourists' sentiments regarding Australian destinations. Tour Manag 71:337–347
11. Li Z, Cai Y, Hu S (2021) Research on systemic financial risk measurement based on HMM and text mining: a case of China financial market. IEEE Access 9:22171–22185
12. Loh S, Lorenzi F, Salda R, Licthnow D (2003) A tourism recommender system based on collaboration and text analysis. Inf Technol Tour 6(3):157–165
13. Sawhney R, Mathur P, Mangal A, Khanna P, Shah RR, Zimmermann R (2020) Multimodal multi-task financial risk forecasting. In Proceedings of the 28th ACM international conference on multimedia, pp 456–465
14. Sheng S, Zhou KZ, Li JJ (2011) The effects of business and political ties on firm performance: evidence from china. J Mark 75(1):1–15
15. Singal M (2012) Effect of consumer sentiment on hospitality expenditures and stock returns. Int J Hosp Manag 31(2):511–521
16. Singh N, Hu C, Roehl WS (2007) Text mining a decade of progress in hospitality human resource management research: identifying emerging thematic development. Int J Hosp Manag 26(1):131–147
17. Young LM, Gavade SR (2018) Translating emotional insights from hospitality employees' comments: using sentiment analysis to understand job satisfaction. Int Hosp Rev 32(1):75–92
18. Zhang Y (2021) Application of data mining technology in financial risk management. In 2021 IEEE conference on telecommunications, optics and computer science (TOCS). IEEE