Financial Innovation

RESEARCH

Open Access

Predicting abnormal trading behavior from internet rumor propagation: a machine learning approach

Li-Chen Cheng[1], Wei-Ting Lu[1] and Benjamin Yeo[2*]

*Correspondence:
benjkyeo@gmail.com

[1] Department of Information and Finance Management, National Taipei University of Technology, Taipei 100, Taiwan
[2] Albers School of Business and Economics, Seattle University, 901 12th Ave, Seattle, WA 98122, USA

Abstract

In 2021, the abnormal short-term price fluctuations of GameStop, which were triggered by internet stock discussions, drew the attention of academics, financial analysts, and stock trading commissions alike, prompting calls to address such events and maintain market stability. However, the impact of stock discussions on volatile trading behavior has received comparatively less attention than traditional fundamentals. Furthermore, data mining methods are less often used to predict stock trading despite their higher accuracy. This study adopts an innovative approach using social media data to obtain stock rumors, and then trains three decision trees to demonstrate the impact of rumor propagation on stock trading behavior. Our findings show that rumor propagation outperforms traditional fundamentals in predicting abnormal trading behavior. The study serves as an impetus for further research using data mining as a method of inquiry.

Keywords: Fake news, Rumors, Data mining, Social media, Classification, Machine learning, GameStop, Reddit

Introduction

In January 2021, GameStop's stock price increased from $16 to $347 as a result of the trading behavior of many small investors rather than their underlying financials (Umar et al. 2021). The unusual, or abnormal, short-term stock price variations highlight the significant influence of the volume of Internet stock discussions on corresponding stock trading activities (Anderson et al. 2021; Lyócsa et al. 2021), prompting calls for the Securities and Exchange Commission to address the issue (Anderson et al. 2021). The efficient market hypothesis in traditional finance states that "unemotional" investors are constantly updating their beliefs about market directions as new information about national economies or firm fundamentals becomes available, causing stock prices to fluctuate around their intrinsic values (Wang et al. 2019). The health of stock markets is dependent on the accuracy, timeliness, and transparency of information. However, rumors spread through stock discussions interfere with investors' decision-making, influencing their sentiments (Wang et al. 2020) and, ultimately, their trading behavior. More importantly, as Internet technologies and electronic commerce tools advance,

rumors on digitized platforms make larger interferences with the stock market. Furthermore, rumors evolve, gain visibility, and become more influential as they spread through the media (Shin et al. 2018). These not only undermine information transparency, but also put the stock market at risk.

Today, investors obtain information about stocks online (Wen et al. 2019), and rumors are one of the most widely studied types of misinformation on the internet. They refer to information that has not been verified by official sources and may or may not be correct. Social media platforms are frequently used to quickly share information among users (Bondielli and Marcelloni 2019) and are a breeding ground for rumor propagation (Zubiaga et al. 2018; Wang et al. 2021). Social media platforms are also valuable sources of stock market information. Fake financial news causes temporary fluctuations in stock prices and an increase in trading volume, resulting in abnormal trading behavior (Kogan et al. 2019). As such, sentiments expressed in tweets have been used to forecast stock market reactions (Arif et al. 2016; Bustos and Pomares-Quimbaya 2020). Although they improve user information dissemination, they also have the potential to disrupt markets (Klein 2021), resulting in unexpected and unusually large trading volumes.

However, we identify two gaps in the literature. First, unlike traditional managerial and financial variables, the impact of Internet stock rumors and rumor propagation on trading behaviors has received little attention. There are no satisfactory explanations for how or why rumors influence stock prices (Fong 2021), despite the fact that their impact on stock markets (Prabhala and Bose 2019) is recognized (Majumdar and Bose 2018). Although several recent studies have used text mining for stock market prediction (cf. Nassirtoussi et al. 2015; Oliveira et al. 2017; Zhang et al. 2018; Vanstone et al. 2019; Bustos and Pomares-Quimbaya 2020; Gupta et al. 2020), studies that specifically focus on rumors and their propagation are lacking. Studies have demonstrated the impact of rumors on the GameStop squeeze (Anderson et al. 2021; Lyócsa et al. 2021), but Klein (2021) called for more research in this area that goes beyond this single incident. Second, machine learning algorithms, though increasingly used in stock market analysis today (Zhong and Enke 2019), are still comparably less used to predict trading behaviors despite their higher accuracy (Barboza et al. 2017). To the best of our knowledge, no study has used machine learning to predict abnormal stock trading behavior based on internet rumor propagation.

In our exploratory study, we use machine learning on social media and numerical data to address these two gaps in a novel way. Our goals are to compare the impact of internet rumor propagation and fundamental predictors, and to train a model that can predict abnormal stock trading behavior that is challenging and risky for firms (Anderson et al. 2021). The findings, specifically identifying rumor propagation variables that are missing in the literature, are useful to financial analysts and academics studying stock trading predictions, and they contribute to the growing body of knowledge on stock markets. Global stock market regulators can use a similar approach to forecast based on each country's definition of abnormal trading behavior, echoing a call for regulation to avoid market inefficiency (Anderson et al. 2021; Umar et al. 2021). Bankers and academics also advocate a need for financial regulation that can reduce systemic risks (Kou et al. 2019).

The remainder of the paper is structured as follows. The "Literature review" Section examines the literature on internet stock rumors and how they affect stock trading

behavior, as well as other traditional fundamentals. The "Method" Section discusses the data and the machine learning method. The "Results" Section summarizes the findings. Finally, the "Conclusionanddiscussion" Section concludes the paper.

Literature review

Rumor propagation

Information exchange is a critical factor in shaping public opinion and decisions. Access to electronic word of mouth has transformed the way Internet users communicate (Ma et al. 2019), allowing for closer social connections and interactions (Li et al. 2022). This exchange is propagated by new media sources, such as discussion boards and social media, among others (Wang et al. 2021), where peer opinions influence investors and their stock trading behavior (Li et al. 2018). Text mining methods on social media data have been used previously for stock market prediction because they can be used to analyze investor attitudes and opinions (Kou et al. 2019). Table 1 summarizes recent related articles that demonstrate the impact of web-based communication–news and social media–on stock markets. However, the majority of these are based in the United States, where abnormal stock trading behavior is not officially defined and is instead based on sample distributions (cf. Joseph et al. 2011). We focus on rumor propagation over time in the days preceding specific stock-day trades. This is not explicitly and thoroughly addressed in the literature. We quantify rumor propagation by investigating rumors in the 30-, 60-, and 90-day periods preceding each stock-day trade.

To define rumor propagation, we first examined how misinformation–false or unverified information–spread via the Internet, strongly influences stock prices (Fong 2021) and attracts significantly more attention to stock price reaction than legitimate information (Clarke et al. 2020). There is a scarcity of research on the propagation of misinformation (Jang et al. 2018) and how it influences human behavior (Bastick 2021), or in this case, stock trading behavior.

Uninformed traders who react to misinformation prefer to trade in equity markets rather than options markets (Brigida and Pratt 2017). Furthermore, heavy social media users gaining information from these platforms are four times more likely to blindly follow the actions of other traders. Misinformation has been found to increase abnormal stock trading activity and stock price volatility by more than 50% and 40%, respectively, in the United States (Kogan et al. 2019). These can also be used to promote specific stocks, resulting in stock trading speculative campaigns (Tardelli et al. 2020). Opinion leaders can influence the behavior of other users (Chen et al. 2021); in crowdsourced systems, they can have a greater influence in stock price manipulations through their posts (Wang et al. 2017), similar to how expert reviews via electronic word of mouth influence purchase decisions (Naujoks and Benkenstein 2020). Taken together, misinformation can lead to market instability and should be detected as soon as possible (Bondielli and Marcelloni 2019).

Official sources, opinion leaders, and crowdsourcing can all be used to validate information spread via the Internet (Chen et al. 2021). However, this takes time to happen. An efficient financial market should reflect accurate stock prices (Anderson et al. 2021), but misinformation comes at a high cost (Bondielli and Marcelloni 2019) that reduces market efficiency and stability because it undergoes more content modifications in the

Table 1 Recent articles using text mining for stock market prediction

Description	Data used	Country/ Region
Hájek (2018) used information—readability, sentiment categories, and bag-of-words (BoW)—from the annual reports of U.S. firms combined financial indicators to predict stock returns	Annual reports	U.S
Nassirtoussi et al. (2015) analyzed information from breaking financial news headlines to predict intraday directional-movements in the foreign exchange market	News	U.S
Feuerriegel and Prendinger (2016) analyzed news-based novel information entering a market to design trading strategies for automated decision making	News	U.S
Vanstone et al. (2019) used sentiment predictors in a neural network to show the number of news articles and twitter posts improve stock price predictions	Bloomberg	U.S. and Australia
Song et al. (2017) developed stock portfolio selection rules by applying learning-to-rank algorithms on news sentiments	Thomson Reuters News	U.S
Zhang et al. (2018) improved stock market prediction by using heterogeneous information fusion on correlated stocks	News and social media	U.S
Li et al. (2018) surveyed literature on web media and stock markets in finance, management information systems, and computer science to identify relationships and provide future research directions	News and social media	U.S
Wu et al. (2021) conducted a sentiment analysis of stock markets to predict intensities of stock dimensional valence–arousal	News and stock price	Taiwan
Jing et al. (2021) designed a hybrid model integrating deep learning with investor sentiment analysis for stock price prediction	Social media and stock price	China
Oliveira et al. (2016) created a stock market sentiment lexicon by using microblogging data and statistical measures	Twitter	U.S
Oliveira et al. (2017) studied the impact of microblogging on stock market predictions: returns, volatility, trading volume and survey sentiment indicators	Twitter	U.S
Kumar and Ravi (2016) conducted a survey of research papers from 2000 to 2016 on text mining applications in finance	News and social media	U.S
Bustos and Pomares-Quimbaya (2020) conducted a systematic review of forecasting techniques covered in studies on stock market movement prediction from 2014 to 2018	News and social media	U.S
Gupta et al. (2020) conducted a literature review on text mining applications in financial forecasting, banking, and corporate finance covered in recent studies	News and social media	U.S

propagation process than accurate information (Jang et al. 2018). Therefore, it is worthwhile to explore the impact of information propagation prior to verification. As shown in the recent GameStop phenomenon, the sheer volume of verified and unverified information can heavily influence stock trading behavior (Anderson et al. 2021; Lyócsa et al. 2021).

Thus, rather than examining the impact of misinformation, we examine rumor propagation. Rumors are a type of web-based communication that has the potential to influence stock markets. We define rumor propagation as the spread and growth of rumors about a particular topic over time. This goes beyond diffusion, which is defined by spreading. Stock market traders are more concerned with receiving information quickly than with verifying its accuracy (Oberlechner and Hocking 2004), emphasizing the importance of studying rumors. Rumors are circulating relevant information that has yet to be verified, which is consistent with longstanding literature (DiFonzo and Bordia

2007; Donovan 2007). Rumors become legitimate information or misinformation after being verified. The spread of rumors occurs in four stages: detection, tracking, stance, and veracity stages. The detection stage detects information and determines whether it is a rumor based on its veracity. In the tracking stage, the identified rumors are disseminated through various media platforms. Meanwhile, rumors are classified as supporting, denying, querying, or commenting on a specific topic in the stance stage. Finally, they are classified as true, false, or unverified at the veracity stage (Zubiaga et al. 2018). Our research focuses on rumors in the tracking stage because our goal is to develop an early-stage detection of abnormal trading behavior. At this point, rumor classifications like supporting or denying are irrelevant.

Rumors spread through strong ties among users in a social network (Zubiaga et al. 2018); specifically, social media networks that allow users, including retail and institutional investors, to share information (Umar et al. 2021). Stock rumors are the spread of information about stocks. Because of their influence on market movements, Majumdar and Bose (2018) recognized the importance of detecting financial rumors in stock market regulation (Prabhala and Bose 2019). We focus on three stock rumor propagation variables in this study: the number of "posts," "replies," and "likes," which amount to the extent to which stock rumors influence stock trading behavior (Anderson et al. 2021; Lyócsa et al. 2021). The number of posts represents the extent of stock-related discussions or rumors. This is similar to how Internet search volume reflects investor interest (Wen et al. 2019) define replies as direct responses to posts (O'Dea et al. 2018). The number of responses indicates the level of interest in the posts or rumors. "Likes" denote positive responses or appreciation (O'Dea et al. 2018), implying the sentiments of responses that have been shown to influence trading behavior (Sabherwal et al. 2011).

Management shocks

Unexpected news and events about a company have been said to influence stock market trading behavior (Sindhu et al. 2014; Prasad and Prabhu 2020). Some investors base their stock trading behaviors on stock rumors, whereas others pay attention to relevant company news and events to make informed trading decisions. Management shocks are among the unexpected circumstances that influence investor sentiments, which in turn, translate to stock trading decisions. We include management shocks for comparison to assess the impact of rumor propagation, which has received less attention in this regard.

We anticipate that stocks that have experienced management shocks, such as strikes and unexpected top management changes, will trade abnormally. Strikes have been found to negatively influence stock prices because they cause investors to lose confidence and perceive an increase in risk (Wisniewski et al. 2020). As a result, this loss of stability can trigger increases in trading behavior in the short term. Furthermore, senior executives have a significant impact on a company's strategic direction. In turn, their decisions affect how a company's stock trades in the market. Stock prices have been found to be significantly influenced by senior executive characteristics and changes in senior management in a company (Yilmaz and Mazzeo 2014). Thus, the unexpected deaths of these senior executives cause significant stock price reactions (Salas 2010). In the US, stock prices were found to decrease by an average of 0.85% following the death of a director (Nguyen and Nielsen 2010).

Other factors

Additionally, investment decisions may be influenced by industry wide events that trigger trading behavior across firms within the same industry. For example, bankruptcy protection of US airlines influences many of their stock prices (Gong 2007). It follows that the higher the shareholding, the larger the potential to affect stock prices via trading behavior. Thus, in this study, we include both industry type and shareholding variation, akin to the common inclusion of demographics in surveys.

Method

The main steps of our research methodology are depicted in Fig. 1. The following subsections provide explanations for these. In Stage 1, we identified 476 stocks listed on the Taipei Exchange (TPEx) and compiled stock-day records to determine whether any of them had been flagged for abnormal trading. Each stock was broken up into 365 stock-day records for September 2019–August 2020, with each record containing the result of whether it was flagged for abnormal trading. As abnormal trading flags apply to each trading day rather than specific stocks, we computed stock-day trades as records for the final data set in Stage 2. Then, using stock-trending keywords to identify posts related to trading, we compiled the corresponding stock discussions on these chosen stocks in Stage 3. This involves compiling the volume of posts, likes, and replies for 30, 60, and 90 days prior to each stock-day record. This enabled us to compare the extent of influence from rumors in the time period leading up to the normal/abnormal trading day for each stock. In Stage 4, we acquired corresponding management-related data on these same stocks. The result is a flat file with all the machine learning-related variables. In Stage 5, we use descriptive statistics to describe the data. Our initial model, which only includes management variables, is discussed in Stage 6. Stage 7 saw the addition of rumor propagation variables to demonstrate model performance gains. Finally, in Stage

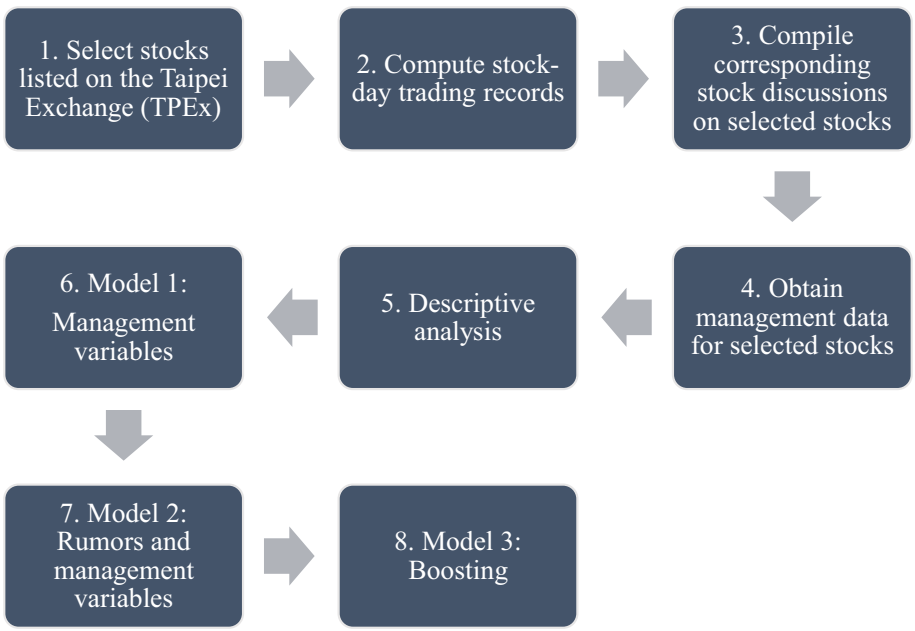

Fig. 1 Research method illustration

8, we built on our earlier findings by boosting the model with variables for management and rumor propagation to create a model that is useful for predicting abnormal trading behavior.

Stock trading in Taiwan

We chose stocks listed on the Taipei Exchange (TPEx) as the study's focus for two reasons. First, our study requires an established standard to define abnormal trading behavior, the target variable, in order to determine its predictors. Such standards can be defined by trade volume or price differentials relative to existing baselines, and they can vary across contexts. However, defining abnormal trading is difficult. Using arbitrary standards leads to poorly defined labels and, as a result, inaccurate classification models. In Taiwan, there is a legal system in place whereby stocks are flagged by the Taiwan Stock Exchange Corporation (TWSE) for abnormal trading by day as an advisory caution to investors (Law Source Retrieving System of Stock Exchange and Futures Trading 2021), because the same stock can be flagged multiple times in a given period of time. There are several conditions that legally define an abnormal stock-day trade under Taiwanese law, such as having a trading volume of at least 3,000 trading units, an intraday transaction price exceeding 9%, and exceeding the amplitude of change in the TWSE Capitalization-Weighted Stock Index by at least 5% (Law Source Retrieving System of Stock Exchange and Futures Trading 2021). Following that, stock days flagged for abnormal trading are kept in a public database that explains each legal violation for transparency's sake (Taiwan Stock Exchange Corporation 2022). For clarity, we used the availability of rigorously and legally defined labels of abnormal trading in our study. Second, rather than the offline grapevine, our study requires easy access to online stock trading discussions. Due to the widespread use of the Internet in Taiwan, it is common for investors to discuss and share relevant stock news and information online at CMoney, the go-to platform for these discussions. While Yahoo Finance, Reddit, Twitter, and Facebook are popular stock discussion platforms in the United States, CMoney is Taiwan's most popular social media platform for stock discussions. Using discussions from this primary source enriches the data, allowing us to capture the majority of stock discussions in Taiwan more comprehensively.

In Taiwan, stocks that are flagged for abnormal trading face severe penalties, including severely restricted trading. Furthermore, stock investors in Taiwan are not permitted to purchase them using margin accounts and must instead trade with existing cash balances. These discourage similar trading behaviors and imply the TWSE's commitment to healthy trading in a well-functioning market. This zealous policing also strengthens the definition of abnormal trading.

The results will help the TWSE predict abnormal trading activities before they occur, saving the authorities money in their efforts to curb such behavior. It is important to note, however, that these findings can also be applied to other countries.

Data

From September 2019 to August 2020, we used the 476 over-the-counter stocks listed on TPEx to compute the abnormal trading variable for each stock-day record to reflect whether its corresponding day trading activity was flagged for abnormal

trading (1 = "Yes," and 0 = "No"), as legally defined by TWSE (Law Source Retrieving System of Stock Exchange and Futures Trading 2021). For data on financial rumor propagation, we extracted 487,193 specific discussion posts from CMoney for the same period that were related to these 476 over-the-counter stocks. Following that, we calculated the corresponding variables for each stock-day based on discussion post attributes (volume, likes, and replies). In addition, for each stock-day record, we included the relevant financial and management predictors discussed in the literature review. These were obtained from the Taiwan Economic Journal (TEJ), a large database of financial data from the Greater China region. Table 2 summarizes the definitions of these variables. The computations of these variables are explained in the following subsections.

Stock discussion posts

Text mining can be used to analyze textual data from various media, such as social media, financial news, and reports (Kou et al. 2019). Social media is a breeding ground for rumors (Wang et al. 2021), and meaningful information has been extracted for financial analyses using text mining (Kumar and Ravi 2016; Gupta et al. 2020). We calculated the daily discussion volume of each TPEx-listed stock after performing word segmentation on CMoney discussion posts. The sums of daily posts, daily replies to posts, daily likes, and daily posts with stock-trending keywords comprise the daily discussion volume (Table 3). Because CMoney posts are written in Taiwanese Mandarin, some of the translated keywords are local colloquialisms that may not make sense when back-translated literally. We calculated the number of discussions for each stock-day record in the 30-, 60-, and 90-day periods preceding an abnormal trading day (Fig. 2). These times were chosen based on TWSE regulations for unusual stock-day trades (Law Source Retrieving System of Stock Exchange and Futures Trading 2021). Using official historical records, abnormal stock-day trades are flagged for legal violations (Law Source Retrieving System of Stock Exchange and Futures Trading 2021) within the preceding 30, 60, and 90 days of each stock-day trade. For example, Big Sun Shine Co. Ltd. (formerly known as Sumagh High Tech Corp.) was in violation on July 3, 2020, because its closing price had increased by 250.34% from its corresponding price 60 days earlier. Subsequently, on August 6, 2020, the increase was 192.60% over the previous 90-day price (Taiwan Stock Exchange Corporation 2022). To avoid endogeneity issues, we did not include discussion volume by day. Furthermore, we calculated variation as the difference in discussion volume between the two preceding time periods.

$V_{i,D}$ is the discussion volume for stock i within a day in CMoney on an abnormal trading day D. $V_{accumulation}$ is the summed daily discussion volume of a stock up to the preceding X days ($X = 30, 60, 90$) (Eq. 1):

Discussion volume accumulation computation

$$V_{accumulation} = \sum_{k=0}^{x} V_{i,D-k} \tag{1}$$

Table 2 Variable definitions

Attribute name		Notation	Description	Source/Calculation
Post				
Accumulation	In the preceding 90 days	PA_90	The summed daily post of a stock in CMoney up to the preceding X days. (X = 30, 60, 90)	CMoney
	In the preceding 60 days	PA_60		
	in the preceding 30 days	PA_30		
Variation	The preceding 60th and 90th day	PV_90	Between the preceding X_1th and X_2th day, the difference in the volume of posts within a day based on a stock in CMoney. ([X_1, X_2] = [0,30], [30,60], [60,90])	CMoney
	the preceding 30th and 60th day	PV_60		
	the preceding 30th day	PV_30		
Reply				
Accumulation	In the preceding 90 days	RA_90	The summed daily reply of all post for a stock in CMoney up to the preceding X days. (X = 30, 60, 90)	CMoney
	In the preceding 60 days	RA_60		
	In the preceding 30 days	RA_30		
Variation	The preceding 60th and 90th day	RV_90	Between the preceding X_1th and X_2th day, the difference in the volume of replies within a day based on all posts of a stock in CMoney. ([X_1, X_2] = [0,30], [30,60], [60,90])	CMoney
	The preceding 30th and 60th day	RV_60		
	The preceding 30th day	RV_30		
Like				
Accumulation	In the preceding 90 days	LA_90	The summed daily like of all post for a stock in CMoney up to the preceding X days. (X = 30, 60, 90)	CMoney
	In the preceding 60 days	LA_60		
	In the preceding 30 days	LA_30		
Variation	The preceding 60th and 90th day	LV_90	Between the preceding X_1th and X_2th day, the difference in the volume of likes within a day based on all posts of a stock in CMoney. ([X_1, X_2] = [0,30], [30,60], [60,90])	CMoney
	The preceding 30th and 60th day	LV_60		
	The preceding 30th day	LV_30		
Industry				
Biotechnology and medicine		BM	There are 6 industries including Biotechnology and Medicine, Cultural and Creative Industry, Electronics Manufacturing, Information and Communications Technology, Other Manufacturing and Other Service Industry	TEJ
Cultural and creative industry		CC		
Electronics manufacturing		EM		
Information and communications technology		IC		
Other manufacturing		OM		
Other service industry		OS		
Management shocks		FC	The variable was flagged ("1" = "Yes", and "0" = "No") whether the stocks have ever had management shocks such as strikes in the preceding year	

Table 2 (continued)

Attribute name	Notation	Description	Source/Calculation
Incorporation type	I_type	There are 2 types including General stock, and KY stock which is a stock registered abroad with initial public offering in Taiwan	
Shareholdings variation of major stockholders with above 600 lots			
The preceding 30th day	S600_30	The shareholdings of preceding Xth day for Major stockholders with above 600 lots. The major-stockholders definition is based on the multiple partitioning method. (X = 30, 60, 90)	
The preceding 60th day	S600_60		
The preceding 90th day	S600_90		
Shareholdings variation of individual stockholders with 20 lots and below			
The preceding 30th day	S20_30	The shareholdings of the preceding X^{th} day for individual stockholders with 20 lots and below. The individual-stockholders definition is based on the multiple partitioning method. (X = 30, 60, 90)	
The preceding 60th day	S20_60		
The preceding 90th day	S20_90		

Table 3 Stock-trending keywords on CMoney discussion posts

Trending	Keywords
Price correction	rebound (回**補**), settle (**結**算), absorb (吸籌碼), control (**控**盤)
Higher	Breakthrough (突破), spike (衝破), soar (升天), rocket (火箭), take off (**起**飛), limit up(漲停), skyrocket (噴射), stand on (站**上**), reaching with volume(帶**量上**攻), strong buying (**大**買), spurt (噴), get on (**上**車), surge (開飆), climbing (**上**山), keep holding (續抱), hang on tight (抱緊), overweight (**加**碼), confident (信**心**)
Lower	Plummet (跌破), big sell off (**大**賣), goodbye (說掰掰), run (快跑), run away (快逃), stuck (套牢), ready to break (準備破), plunge (跳**水**)

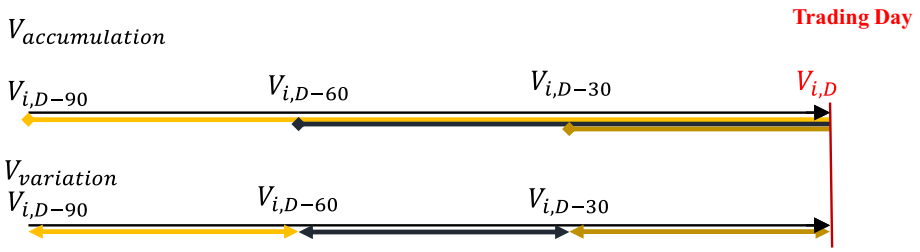

Fig. 2 Date selection of discussion volume in this dataset

$V_{variation}$ is the difference in the discussion volume of a stock between the preceding X_1th and X_2th day ($X_1, X_2 = (0, 30), (30, 60), (60, 90)$ (Eq. 2):

Discussion volume variation computation

$$V_{variation} = \left| V_{i,D-x_1} - V_{i,D-x_2} \right| \tag{2}$$

Management shocks

We include three types of management data in addition to rumor propagation (Table 1). Managerial shocks assess whether the company had a strike or a change in CEO in the previous year. TWSE defines these shocks in Taiwan, and corresponding data for each company is publicly available. We used a binary variable to indicate whether a company had a managerial shock the previous year.

Other data

In terms of shareholding data, we used the Taiwan Depository and Clearing Corporation's Distribution of Shareholdings Proportion to distinguish between individual and institutional ownership of shares. Due to the continuous variables in the distribution of shareholdings proportion, we divided it into five intervals using two data partitioning methods, including the first multiple partitioning method. Individual investors are those who own up to 20 lots, while major investors own more than 600 lots (each lot is equivalent to 1000 shares).

The Percentage of Centrally Deposited Securities represents the share capital of a company as a percentage of a specific population. If major shareholders continue to hold large amounts of company stock, it indicates that they are bullish on the company's future. We computed $S_{i,D}$ as the percentage of centrally deposited securities for stock i in the week of an abnormal trading day D. $S_{variation}$ denotes the shareholding of the week for the Xth day ($X = 30, 60, or 90$) (Eq. 3):

Shareholding variation computation

$$S_{variation} = S_{i,D} - S_{i,D-x} \tag{3}$$

Finally, stock trading activity trends may occur across the industry (Chan et al. 2013). For example, negative news about the retail industry can have a negative simultaneous effect on the stock prices of several retail companies over the same time period. Taiwan has 25 TEJ industry classifications. These are categorized into six industry groups in this study: biotechnology and medicine, cultural and creative, electronics manufacturing, information and communications technology, other manufacturing, and other service industries. In addition, pertinent to the Taiwanese context, stocks are classified according to their incorporation type: General or KY. The former includes stocks from Taiwan-based companies such as Hi-Lai Foods and Mornsun, and the latter includes companies incorporated outside of Taiwan that had their initial public offering in Taiwan.

Decision tree induction

Among data-driven approaches, machine learning algorithms are increasingly used to make accurate predictions on financial risks (Kou et al. 2019), and they can be used to investigate the impact of online misinformation (Choudrie et al. 2021), and are becoming increasingly popular in stock market analysis (Zhong and Enke 2019). In financial analyses, they have been shown to be more accurate than traditional

statistical models (Barboza et al. 2017). Furthermore, due to the non-linearity of the data, traditional linear or continuous statistical models may not be appropriate (Antunes 2021).

We trained a decision tree, a supervised learning technique, to predict abnormal trading activities among Taiwanese stock-day trades in this study. Decision trees are a popular supervised machine learning technique (Brieman et al. 1984) and are recommended for analyzing website comments (Tan 2015). They classify outcomes using if–then rules (Osei-Bryson 2004) and can handle non-linear relationships without making assumptions about the data distribution (Pal and Mather 2003). Decision trees have a tree-like structure (cf. Lin and Fan 2019), where the if–then rules partition data repeatedly until the subsets are as homogenous as possible with respect to the outcome variable (Ture et al. 2009; Osei-Bryson and Ngwenyama 2011). They are simple to interpret and utilize (Murphy and Comiskey 2013), without sacrificing performance (Du et al. 2020), giving them a competitive advantage over other machine learning algorithms that lack transparency and are difficult to understand or apply (Du et al. 2020). Tree-based models, by the way, have been found to be more stable than multilayer artificial neural networks (Addo et al. 2018). Decision trees have been used in medicine (Ture et al. 2009; Murphy and Comiskey 2013; Kobayashi et al. 2013) and marketing (Kim et al. 2011; Amir et al. 2015; Legohérel et al. 2015; Díaz-Pérez and Bethencourt-Cejas 2016). Although used less often in finance, Bondielli and Marcelloni (2019) observed a clear trend in supervised classification approaches to fake news and rumor detection for a variety of projects. Although deep learning models outperform tree-based models in terms of predictive accuracy, they are black boxes (Skrede et al. 2020) and do not explicitly identify the relevance of each predictor (Guo et al. 2016). In contrast, rule-based models are more interpretable (Kou et al. 2021), meeting the need for interpretability when analyzing complex financial data (Li et al. 2021). The goal of this study is to identify specific predictors of abnormal trading behavior in order to compare the impact of rumor propagation versus traditional financial and management predictors. This is analogous to the difficulty in determining priorities in complex issues (Kou et al. 2022). Hence, tree-based models are preferable in this study. Following the identification of these predictors, we discuss the use of deep learning as a future research direction.

The final dataset included 241,332 stock-day records, with only 6110 classified as abnormal trading. Given that data imbalance can have an impact on the performance of classifying algorithms (Budhi et al. 2021), we used an oversampling algorithm to balance the data set, resulting in a 50–50 split of records with normal and abnormal trading. We used these data to perform an 80%–20% training-validation split for supervised learning.

To avoid bias caused by relying on specific training and validation sets, we used ten-fold cross validation in our decision trees. Parsimony is essential in machine learning because highly complex models may be ineffective (Shmueli 2016). By limiting the size of decision trees, researchers can simplify them (Esposito et al. 1997). Hence, when two classification trees have similar accuracy rates, the simpler one with fewer leaves and a smaller size is preferred. We limited the decision tree growth for this study by setting the maximum depth to 8 and the minimum terminal node size to 300. We trained two classification trees to identify abnormal versus normal stock trading behavior in stock-day records. The first tree contains only the financial and managerial variables examined

Table 4 Descriptive statistics

Variable	Mean	Median	Std. Dev	Skewness	Kurtosis	AD
PA_90	439.40	37.00	1624.28	10.72	150.30	1393.20*
PA_60	352.80	27.00	1402.82	11.18	158.83	1454.94*
PA_30	220.14	16.50	939.10	12.04	185.72	1519.90*
PV_90	86.60	5.00	283.91	9.11	127.76	1363.38*
PV_60	132.67	7.00	522.89	12.98	254.49	1444.16*
PV_30	132.52	7.00	638.60	15.23	300.93	1592.26*
RA_90	3548.32	106.00	18,073.39	11.54	164.12	1709.99*
RA_60	3048.56	80.00	16,185.16	11.47	160.97	1756.40*
RA_30	2008.42	47.00	10,884.81	11.53	165.74	1792.91*
RV_90	499.76	9.00	2473.34	13.99	268.77	1692.36*
RV_60	1040.13	16.00	6089.55	14.23	269.71	1792.09*
RV_30	1179.16	15.00	7032.63	13.7	241.72	1842.37*
LA_90	7,235,764.00	269,581.00	31,439,316.00	11.01	161.28	1660.47*
LA_60	6,115,036.00	189,712.50	27,998,175.00	11.54	174.58	1684.70*
LA_30	4,033,419.00	92,133.50	19,796,518.00	12.59	207.17	1758.51*
LV_90	1,120,728.00	7755.50	5,008,407.00	12.44	254.21	1582.99*
LV_60	2,081,616.00	14,585.50	10,092,019.00	13.55	285.50	1622.77*
LV_30	2,346,409.00	17,171.00	13,394,405.00	15.76	319.85	1671.99*
S600_30	−0.26	0.00	2.37	−0.04	23.04	406.13*
S600_60	−0.27	0.00	3.46	0.36	18.83	336.44*
S600_90	−0.34	0.00	4.10	0.30	13.95	358.16*
S20_30	0.25	−0.01	2.22	0.66	11.54	407.94*
S20_60	0.35	−0.02	3.07	0.69	9.12	402.70*
S20_90	0.43	−0.02	3.77	0.97	9.10	354.13*

*$p < 0.001$

in this study. The second adds variables for rumor propagation. Our use of these two classification trees highlights the differences in impact between rumor propagation and company financial and managerial characteristics.

Finally, we boosted the combined classification tree using a random forest (cf. Brieman et al. 1984), which has been used in finance for things like predicting bank lending (Ozgur et al. 2021). A random forest is made up of several random trees that have been trained using bootstrapped observations from the training data (Shmueli 2016). In general, such ensemble models outperform individual classifiers (Kou et al. 2021). In this study, where the relationships between the rumor propagation variables and abnormal trading behavior are presumably complex, random forests can help learn the decision boundaries for better performance (Bacham and Zhao 2017).

Results

Descriptive analysis

Table 4 summarizes the numerical data in a descriptive manner. The medians of the stock discussion variables differ significantly from their respective means, indicating highly skewed and non-normal distributions. Further evidence is provided by the corresponding skewness and kurtosis statistics. The shareholding variation variables are less skewed, with closer medians and means, but they continue to be non-normally

Table 5 Performance of decision trees

Tree	Accuracy	Sensitivity	Specificity	Precision	AUC ROC
1	0.61	0.61	0.67	0.61	0.63
2	0.70	0.70	0.70	0.70	0.74
3	0.84	0.84	0.85	0.84	0.91

distributed. We performed an Anderson–Darlin (AD) test on all variables, and the AD statistics show that the distributions of all variables are significantly non-normal.

Pertaining to post accumulations (PA_30, PA_60, and PA_90), the average number of posts 30, 60, and 90 days after the stock behavioral outcome date increased. In terms of post variations (PV_30, PV_60, and PV_90), the averages increased to roughly comparable levels between the 10th and 30th days, and between the 30th and 60th days from the stock behavioral outcome date. Similar trends can be found in replies (RA_30, RA_60, and RA_90), reply volume variations (RV_30, RV_60, and RV_90), likes (LA_30, LA_60, and LA_90), and like volume variations (LV_30, LV_60, and LV_90). There were more replies and less variation, as well as more likes and less "like" volume variation, further away from the stock behavioral outcome date. These outcomes are expected because there are more accumulated posts and rumor propagation about a stock over time. The large standard deviations and skewness indicate that these rumor variables vary significantly across stock-day trades.

Major and individual stockholders' shareholdings varied less than their rumor counterparts. Corresponding computations for the 30th, 60th, and 90th days preceding a stock-day trading behavioral outcome show significantly lower standard deviations and skewness. This means that there were no significant changes in stock trading among major or individual stockholders during this time period.

A total of 200 records in the data set that are classified as General Stock. The data set contained 1164 records in biotechnology and medicine, 220 records in the cultural and creative industry, 2198 records in electronics manufacturing, 806 records in information and communications technology, 1178 records in other manufacturing, and 544 records in other manufacturing. In the previous year, there were 1664 records that experienced management shocks.

Predictive analysis

Table 5 displays the outcomes of three decision trees based on five evaluation metrics. Accuracy is frequently used as an aggregate measure to demonstrate how well a classification model correctly categorizes positive and negative outcomes. A model may outperform another in predicting one outcome over another. Sensitivity and specificity indicate how well the model predicts positive (abnormal trading) and negative (normal trading) outcomes. Precision measures the accuracy with which a model predicts a positive outcome. The Receiver Operator Characteristics (ROC) curve plots all possible combinations of the True Positive Rate versus the False Positive Rate for various classification cutoffs. The area under this curve (AUC) represents how well the classification

Table 6 Performance metrics definitions

Metric	Definition and Computation
Accuracy	The proportion of correctly classified records $\frac{Numberofcorrectlyclassifiedrecords}{TotalNumberofrecords}$
Sensitivity	The True Positive Rate (TPR); the proportion of correctly classified positive records (i.e., abnormal trading) among all positive records $\frac{Numberofcorrectlyclassifiedpositiverecords}{TotalNumberofpositiverecords}$
Specificity	The True Negative Rate (TNR); the proportion of correctly classified negative records (i.e., normal trading) among all negative records $\frac{Numberofcorrectlyclassifiednegativerecords}{TotalNumberofrecordsasnegative}$
Precision	The positive predictive value; the proportion of records correctly classified as positive (i.e., abnormal trading) among all records that were classified as positive $\frac{Numberofcorrectlyclassifiedpositiverecords}{TotalNumberofrecordsclassifiedaspositive}$
AUC ROC	Area under the Receiver Operator Characteristics (ROC) curve

model distinguishes between the two classes (i.e., abnormal vs. normal trading) and is one of the most widely used methods for comparing classification models (Kou et al. 2021). Table 6 shows the definitions and corresponding computations (Novaković et al. 2017). Overall, they predict abnormal trading fairly accurately, with accuracies ranging from 61.40 to 84.40%. Notably, performance improved across the board from Tree 1 (which only included shock and shareholding variables) to Tree 2 (which also included rumor propagation variables) to the boosted random forest with all variables.

Financial characteristics tree (Tree 1)

Figure 3 depicts the decision tree for financial characteristics that predict abnormal trading behavior. The names of the predictor variables correspond to the notations in Table 1. The terminal nodes display the number of records in each outcome variable class. The tree predicts abnormal trading behavior reasonably well (accuracy = 0.61). It does a good job of distinguishing between the two classes (AUC = 0.63), but it does a better job of predicting normal trading behavior (sensitivity = 0.61, specificity = 0.67). In addition, its precision is only moderate (0.61), implying that using this tree to detect abnormal trading behavior is not ideal.

Table 7 summarizes the decision tree rules. Records that meet the rules specified in the Description column are likely to have the corresponding Outcome column result. The variation in stockholding of individual stockholders with less than or equal to 20 lots in the previous 30 days is the best predictor. Those with a value greater than 2.29 and less than or equal to − 1.03 are flagged for abnormal trading behavior (Rules #1 and #2), while those with a value greater than − 1.03 are subject to other predictors, including other shareholding variation predictors. It is worth noting that the electronics manufacturing industry contributes to the classification of abnormal trading behavior (Rules #6, #7, and #8). For example, in Rule #8, records in the electronics manufacturing industry are not flagged for abnormal trading behavior despite meeting the conditions related to shareholding variations.

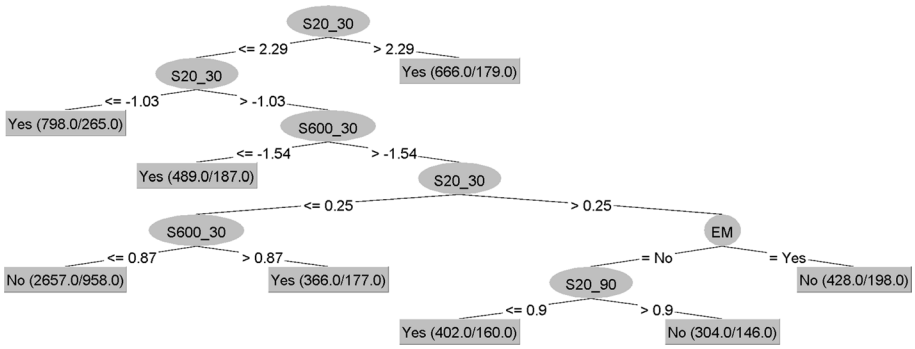

Fig. 3 Financial characteristics tree (Tree 1)

Table 7 Decision rules for financial characteristics tree (Tree 1)

Rule	Description	Outcome
1	S20_30 < = 1.03	Yes
2	S20_30 > 2.29	Yes
3	S20_30 < = 2.29, S20_30 > − 1.03, S600_30 < = 1.54	Yes
4	S20_30 < = 2.29, S20_30 > − 1.03, S600_30 > 1.54, S20_30 < = 0.25, S600_30 < = 0.87	No
5	S20_30 < = 2.29, S20_30 > − 1.03, S600_30 > 1.54, S20_30 < = 0.25, S600_30 > 0.87	Yes
6	S20_30 < = 2.29, S20_30 > − 1.03, S600_30 > 1.54, S20_30 > 0.25, Industry_Electronics = "No", S20_90 < = 0.9	Yes
7	S20_30 < = 2.29, S20_30 > − 1.03, S600_30 > 1.54, S20_30 > 0.25, Industry_Electronics = "No", S20_90 > 0.9	No
8	S20_30 < = 2.29, S20_30 > − 1.03, S600_30 > 1.54, S20_30 > 0.25, Industry_Electronics = "Yes"	No

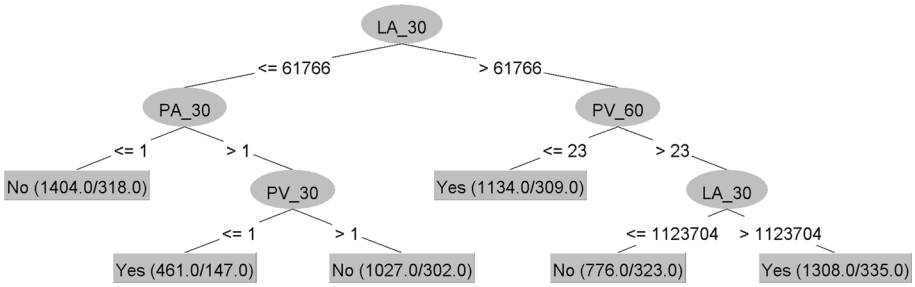

Fig. 4 Combined tree (Tree 2)

Combined tree (Tree 2)

Figure 4 depicts the decision tree with rumor propagation variables as predictors as well as all variables from Tree 1. Similarly, the names of the predictor variables correspond to the notations in Table 1. The terminal nodes display the number of records in each outcome variable class.

Overall, Tree 2 outperforms Tree 1 (accuracy = 0.70), representing an 8.79% improvement over the preceding tree. Its sensitivity, specificity, and precision have all improved, and it can predict both classes more accurately than Tree 1 (sensitivity = 0.70 and

Table 8 Decision rules for combined tree (Tree 2)

Rule	Description	Outcome
1	LA_30 < = 61,766, PA_30 < = 1	No
2	LA_30 < = 61,766, PA_30 > 1, PV_30 < = 1	Yes
3	LA_30 < = 61,766, PA_30 > 1, PV_30 > 1	No
4	LA_30 > 61,766, PV_60 < = 23	Yes
5	LA_30 > 61,766, PV_60 > 23, LA_30 < = 112,374	No
6	LA_30 > 61,766, PV_60 > 23, LA_30 > 112,374	Yes

specificity = 0.70). Interestingly, despite being better positioned to distinguish the classes (AUC = 0.74), Tree 2 lacks management shock and shareholding variation variables. In classifying abnormal trading behavior, only accumulated "likes," post accumulation, and post variations are relevant (Table 7).

Table 8 summarizes the Tree 2 decision tree rules. The number of "likes" in the previous 30 days is the best predictor. Following that, the splits are determined by the accumulation of posts and volume variations in posts over the previous 30 and 60 days, respectively. For example, among stock-day records with more than 61,766 "likes" in the previous 30 days and post variations of more than 23, those with less than 112,374 "likes" are not flagged for abnormal trading behavior (Rule #5), while those with more than 112,374 are flagged for abnormal trading behavior (Rule #6). According to these rules, the favorability, volume, and variations of CMoney posts, which appear in the classification rules, primarily in the preceding 2 months are sufficient to predict whether a stock will exhibit abnormal trading behavior. The rules also show that the relationships between these predictors and abnormal trading behavior are not linear, bolstering our decision tree's use.

Random forest (Tree 3)

The addition of rumor propagation variables improved the tree's predictive ability. To improve prediction, we trained a random forest using the same variables as Tree 2. It shows a significant improvement in all aspects of prediction, as expected (Table 2). At 0.84 and 0.84, respectively, the accuracy and precision are higher than the preceding trees. Furthermore, the random forest can almost equally well classify both classes—abnormal and normal trading (sensitivity = 0.84 and specificity = 0.85). With an AUC ROC of 0.91, the random forest clearly distinguishes between the two classes. These findings imply that the random forest can be used effectively to distinguish abnormal stock trading behavior from financial rumors, and that ensemble models outperform single models (Kou et al. 2021).

Robustness checks

We performed robustness checks on our results using other machine learning techniques (cf. Kou et al. 2021), including k-Nearest Neighbors (kNN), logistic regression (Logistic) and support vector machines (SVM). Their performance metrics are

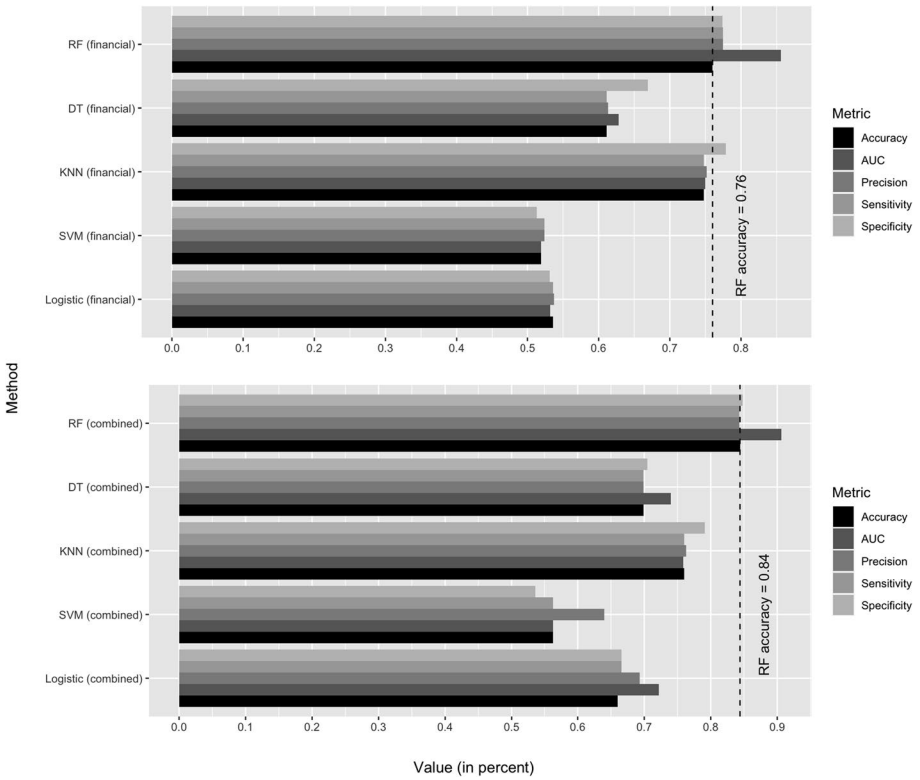

Fig. 5 Robustness checks

compared to those of decision trees (Tree 1, financial characteristics variables, and Tree 2, combined) and random forest (RF, financial and RF combined). A summary of the comparison is shown in Fig. 5: The top chart compares all models that include only the financial and managerial characteristics variables (similar to Tree 1), while the bottom chart compares all models that include all variables (comparable to Trees 2 and 3). The comparisons are based on the models' five performance metrics: accuracy, sensitivity, specificity, precision, and AUC ROC. Two findings emerged from the robustness tests. First, in all five models, the combined versions outperform their corresponding financial and managerial characteristics variables-only counterparts in terms of performance metrics. For example, the RF has the highest accuracy among the combined models, at 0.84, which is 11.05% higher than its financial characteristics-only counterpart, which has an accuracy of 0.76. The AUC ROC of the logistic regression comprising all variables (0.72) is 35.85% higher than its financial and managerial characteristics-only counterpart (0.53). Thus, financial rumors evidently have a greater influence on abnormal trading behavior than the financial and managerial characteristics variables included. Second, in both sets of comparisons, RFs outperform the other models. These are the strengths of using decision trees in financial predictions (cf. Barboza et al. 2017; Addo et al. 2018), which support our methodological rationale.

Conclusion and discussion

Summary

This study adopted an innovative approach using social media data (posts, likes, and replies) and decision tree induction to predict abnormal stock trading behavior in the context of trending data-driven approaches to studying financial contexts (Kou et al. 2019). In addition, we used the TWSE's official legal definition of abnormal trading. We addressed two gaps in the literature: the lack of studies on the influence of rumors despite their relevance (Arif et al. 2016) and the lack of analyses in this area using machine learning despite its higher accuracy (Zhong and Enke 2019). We included variables related to rumor propagation, managerial shocks, and shareholding characteristics using data from September 2019 to August 2020 in our decision trees. Our results show that rumor propagation outperforms management shocks and other variables in predicting abnormal trading behavior. In particular, the extents of "likes" and post volume are critical. Furthermore, the random forest improves prediction in all aspects and can be readily used by academics, financial analysts, and governing bodies seeking to achieve and maintain market efficiency and stability. We performed robustness checks using logistic regression, k-Nearest Neighbors, and SVM; all of these methods produced similar results, but our decision trees outperformed them.

Rationality and the predictors of abnormal trading activities

New media content influences stock trading behavior (Li et al. 2018). False or unverified information can easily spread through these new media channels (Bondielli & Marcelloni 2019), strongly influencing investor behavior and thus stock prices (Clarke et al. 2020; Fong 2021). These may take the form of rumors (DiFonzo and Bordia 2007; Donovan 2007). Our findings support previous research that stock rumor propagation influences stock trading behavior (Anderson et al. 2021; Lyócsa et al. 2021). Furthermore, we demonstrate that rumor propagation can be used to predict whether a stock will exhibit abnormal trading behavior in the subsequent 2 months.

A multitude of variables have been shown to influence stock trading behavior, including firm strikes (Wisniewski et al. 2020) and changes in senior management (Yilmaz and Mazzeo 2014). Rational decisions based on these shocks reflect investor confidence losses. In contrast to the literature, our findings from management and other fundamental variables (Tree 1) indicate that the extent of shareholding variation is more important in predicting abnormal trading behavior.

Meanwhile, rumors are unverified (DiFonzo and Bordia 2007; Donovan 2007), and our results show that rumor propagation is far more influential in prediction. This demonstrates the irrationality of stock trading behavior, in which investors' irrational emotions and errors in judgment can cause market volatility (Verma and Verma 2007). Our findings provide empirical explanations for GameStop's recent short-term volatility, correlating with the heavy influence of stock rumors (Anderson et al. 2021; Lyócsa et al. 2021) on large volumes of irrational trading behavior.

Contributions and applications

Our study makes three contributions to the literature on abnormal trading behavior. First, we present empirical evidence that rumor propagation is far more important

in predicting abnormal trading behavior than fundamental management shocks and traditional financial variables. This lends support to the notion of investor irrationality in stock trading and steers attention toward the overlooked rumor propagation in the literature on stock trading prediction. Second, we show that machine learning can be used to study financial outcomes, such as abnormal trading behavior. We chose a machine learning approach because it is more accurate than traditional statistical approaches (Barboza et al. 2017). Furthermore, it advances our understanding of how humans react to rumors via new technologies such as social media (Choudrie et al. 2021). The non-linearity of the resulting relationships offers additional methodological justifications (Antunes 2021). Third, our model is a direct response to calls for the Stock Exchange Commission to address stock market volatility (Anderson et al. 2021). This applies to other stock trading regulatory bodies.

Accurate, timely, and open information is essential for healthy stock market trading. With the advent of social media, stock discussions and rumors proliferate, interfering with investors' decision-making and sentiments (Wang et al. 2020) and thus triggering irrational and abnormal behavior in volatile markets. Our results show that within 2 months, rumor propagation can be used to predict abnormal trading behavior. Because rumors can be monitored and measured, stock trading regulatory bodies can use our model to flag stocks for abnormal trading in advance, preserving market efficiency and controlling potential volatility. Financial analysts can also use our model to investigate stock discussions and trading behavior.

Future research directions

Although our study has many applications, three limitations should be addressed in future research. First, our data were restricted to the period between September 2019 and August 2020. The global pandemic may have influenced stock trading decisions during this time period. Future research can thus use the same approach to investigate rumor propagation and trading behavior in a different time frame. Second, we conducted the investigation in Taiwan, primarily because TWSE flags abnormal trading behavior and provides an authoritative means to define trading behavior. Future research could investigate stock trading in other countries and compare the impact of rumor propagation. Third, although decision trees are more stable than artificial neural networks (Addo et al. 2018), future research can explore into the differences between the two methods to determine the extent to which either one is superior. After identifying the predictors of abnormal trading behavior, future studies can then better design and train deep learning models based on these predictors for improved prediction performance. Overall, with additional research, the novel application of machine learning to analyze stock rumors and trading behavior has the potential to become a more established research program.

Abbreviations

TPEx	Taipei exchange
TWSE	Taiwan Stock Exchange Corporation
TEJ	Taiwan Economic Journal
AD	Anderson–Darlin
ROC	Receiver operator characteristics
TPR	True positive rate
FPR	False positive rate

AUC	Under this curve
kNN	k-nearest neighbors
Logistic	Logistic regression
SVM	Support vector machines
RF	Random forest

Acknowledgements
The authors are very grateful to the anonymous referees and Editor for their helpful comments and valuable suggestions in improving the earlier versions of the paper. This study was supported in part by the National Science and Technology Council, Taiwan, under Grants MOST 108-2410-H-027-020, MOST 109-2410-H-027-009-MY2 and MOST 111-2410-H-027-011-MY3.

Author contributions
LC and WL collected and analyzed the data. BY reviewed the literature, described the data, and discussed the results and implications. LC and BY contributed substantially and collaborated to frame the study, read and approve the final manuscript.

Funding
This study was supported by the National Science and Technology Council, Taiwan, under grants MOST 108-2410-H-027-020, MOST 109-2410-H-027 -009 -MY2 and MOST 111-2410-H-027 -011 -MY3.

Availability of data and materials
Data were collected by the authors per the description in the paper and will not be available to public.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 26 January 2022 Accepted: 16 November 2022
Published online: 03 January 2023

References

Addo PM, Guegan D, Hassani B (2018) Credit risk analysis using machine and deep learning models. Risks 6(2):38
Amir S, Osman MM, Bachok S, Ibrahim M (2015) Understanding domestic and international tourists' expenditure pattern in Melaka, Malaysia: result of CHAID analysis. Contemp Issues Manag Soc Sci Res 172:390–397
Wang J, Xie Z, Li Q, Tan J, Xing R, Chen Y, Wu F (2019) Effect of digitalized rumor clarification on stock markets. Emerg Mark Financ Trade 55(2):450–474
Arif A, Shanahan K, Chou F-J, Dosouto Y, Starbird K, Spiro ES (2016) How information snowballs: exploring the role of exposure in online rumor propagation. In: Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing. pp 466–477
Anderson JP, Kidd J, Mocsary GA (2021) Social media, securities markets, and the phenomenon of expressive trading Secur Mark Phenom expressive trading. Lewis Clark L Rev 25:1223
Antunes JAP (2021) To supervise or to self-supervise: a machine learning based comparison on credit supervision. Financ Innov 7(1):1–21
Bacham D, Zhao J (2017) Machine learning: challenges, lessons, and opportunities in credit risk modeling. Moody's Anal Risk Perspect 9:30–35
Barboza F, Kimura H, Altman E (2017) Machine learning models and bankruptcy prediction. Expert Syst Appl 83:405–417
Bastick Z (2021) Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. Comput Hum Behav 116:106633
Bondielli A, Marcelloni F (2019) A survey on fake news and rumour detection techniques. Inf Sci 497:38–55
Brieman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth Inc 37(15):237–251
Brigida M, Pratt WR (2017) Fake news. North Am J Econ Financ 42:564–573
Budhi GS, Chiong R, Wang Z, Dhakal S (2021) Using a hybrid content-based and behaviour-based featuring approach in a parallel environment to detect fake reviews. Electron Commer Res Appl 47:101048
Bustos O, Pomares-Quimbaya A (2020) Stock market movement forecast: a systematic review. Expert Syst Appl 156:113464
Chan K, Hameed A, Kang W (2013) Stock price synchronicity and liquidity. J Financ Mark 16(3):414–438
Chen J, Kou G, Wang H, Zhao Y (2021) Influence identification of opinion leaders in social networks: an agent-based simulation on competing advertisements. Inf Fusion 76:227–242
Choudrie J, Banerjee S, Kotecha K, Walambe R, Karende H, Ameta J (2021) Machine learning techniques and older adults processing of online information and misinformation: a covid 19 study. Comput Hum Behav 119:106716
Clarke J, Chen H, Du D, Hu YJ (2020) Fake news, investor attention, and market reaction. Inf Syst Res 32(1):35–52
Díaz-Pérez FM, Bethencourt-Cejas M (2016) CHAID algorithm as an appropriate analytical method for tourism market segmentation. J Destin Mark Manag 5(3):275–282
DiFonzo N, Bordia P (2007) Rumor, gossip and urban legends. Diogenes 54(1):19–35
Donovan P (2007) How idle is idle talk? One hundred years of rumor research. Diogenes 54(1):59–82
Du M, Liu N, Hu X (2020) Techniques for interpretable machine learning. Commun ACM 63(1):68–77

Esposito F, Malerba D, Semeraro G, Kay J (1997) A comparative analysis of methods for pruning decision trees. IEEE Trans Pattern Anal Mach Intell 19(5):476–491
Feuerriegel S, Prendinger H (2016) News-based trading strategies. Decis Support Syst 90:65–74
Fong B (2021) Analysing the behavioural finance impact of 'fake news' phenomena on financial markets: a representative agent model and empirical validation. Financ Innov 7(1):1–30
Gong SXH (2007) Bankruptcy protection and stock market behavior in the US airline industry. J Air Transp Manag 13(4):213–220. https://doi.org/10.1016/j.jairtraman.2007.03.003
Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS (2016) Deep learning for visual understanding: a review. Recent Dev Deep Big vis 187:27–48. https://doi.org/10.1016/j.neucom.2015.09.116
Gupta A, Dengre V, Kheruwala HA, Shah M (2020) Comprehensive review of text-mining applications in finance. Financ Innov 6(1):1–25
Hájek P (2018) Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns. Neural Comput Appl 29(7):343–358
Jang SM, Geng T, Li J-YQ, Xia R, Huang C-T, Kim H, Tang J (2018) A computational approach for examining the roots and spreading patterns of fake news: evolution tree analysis. Comput Hum Behav 84:103–113
Jing N, Wu Z, Wang H (2021) A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. Expert Syst Appl 178:115019
Joseph K, Wintoki MB, Zhang Z (2011) Forecasting abnormal stock returns and trading volume using investor sentiment: evidence from online search. Int J Forecast 27(4):1116–1127
Kim SS, Timothy DJ, Hwang J (2011) Understanding Japanese tourists' shopping preferences using the decision tree analysis method. Tour Manag 32(3):544–554. https://doi.org/10.1016/j.tourman.2010.04.008
Klein T (2021) A note on GameStop, short squeezes, and autodidactic herding: an evolution in financial literacy? Financ Res Lett 46:102229
Wang D, Zhou Y, Qian Y, Liu Y (2021) The echo chamber effect of rumor rebuttal behavior of users in the early stage of COVID-19 epidemic in China. Comput Hum Behav 128:107088
Kobayashi D, Takahashi O, Arioka H, Koga S, Fukui T (2013) A prediction rule for the development of delirium among patients in medical wards: chi-square automatic interaction detector (CHAID) decision tree analysis model. Am J Geriatr Psychiatry 21(10):957–962. https://doi.org/10.1016/j.jagp.2012.08.009
Kogan S, Moskowitz TJ, Niessner M (2019) Fake news: evidence from financial markets. Available SSRN 3237763
Kou G, Chao X, Peng Y, Alsaadi FE, Herrera-Viedma E (2019) Machine learning methods for systemic risk analysis in financial sectors. Technol Econ Dev Econ 25(5):716–742
Kou G, Xu Y, Peng Y, Shen F, Chen Y, Chang K, Kou S (2021) Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. Decis Support Syst 140:113429. https://doi.org/10.1016/j.dss.2020.113429
Yilmaz N, Mazzeo MA (2014) The effect of CEO overconfidence on turnover abnormal returns. J Behav Exp Financ 3:11–21. https://doi.org/10.1016/j.jbef.2014.07.001
Kou G, Yüksel S, Dinçer H (2022) Inventive problem-solving map of innovative carbon emission strategies for solar energy-based transportation investment projects. Appl Energy 311:118680. https://doi.org/10.1016/j.apenergy.2022.118680
Kumar BS, Ravi V (2016) A survey of the applications of text mining in financial domain. Knowl-Based Syst 114:128–147
Law Source Retrieving System of Stock Exchange and Futures Trading (2021) Taiwan stock exchange corporation directions for announcement or notice of attention to trading information and dispositions. In: Law source retrieving Syst. Stock Exch. Futur. Trading. http://www.selaw.com.tw/LawArticle.aspx?LawID=G0100247. Accessed 14 Jul 2022
Legohérel P, Hsu CHC, Daucé B (2015) Variety-seeking: Using the CHAID segmentation approach in analyzing the international traveler market. Tour Manag 46:359–366
Li Q, Chen Y, Wang J, Chen Y, Chen H (2018) Web media and stock markets: a survey and future directions from a big data perspective. IEEE Trans Knowl Data Eng 30(2):381–399. https://doi.org/10.1109/TKDE.2017.2763144
Li T, Kou G, Peng Y, Yu PS (2021) An integrated cluster detection, optimization, and interpretation approach for financial data. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2021.3109066
Li Y, Kou G, Li G, Peng Y (2022) Consensus reaching process in large-scale group decision making based on bounded confidence and social network. Eur J Oper Res 303(2):790–802. https://doi.org/10.1016/j.ejor.2022.03.040
Lin C-L, Fan C-L (2019) Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan. J Asian Archit Build Eng 18(6):539–553. https://doi.org/10.1080/13467581.2019.1696203
Lyócsa Š, Baumöhl E, Výrost T (2021) YOLO trading: riding with the herd during the GameStop episode. Financ Res Lett 46:102359
Ma H, Kim JM, Lee E (2019) Analyzing dynamic review manipulation and its impact on movie box office revenue. Electron Commer Res Appl 35:100840
Majumdar A, Bose I (2018) Detection of financial rumors using big data analytics: the case of the Bombay stock exchange. J Organ Comput Electron Commer 28(2):79–97
Murphy EL, Comiskey CM (2013) Using chi-squared automatic interaction detection (CHAID) modelling to identify groups of methadone treatment clients experiencing significantly poorer treatment outcomes. J Subst Abuse Treat 45(4):343–349
Wen F, Xu L, Ouyang G, Kou G (2019) Retail investor attention and stock price crash risk: evidence from China. Int Rev Financ Anal 65:101376. https://doi.org/10.1016/j.irfa.2019.101376
Nassirtoussi AK, Aghabozorgi S, Wah TY, Ngo DCL (2015) Text mining of news-headlines for FOREX market prediction: a multi-layer dimension reduction algorithm with semantics and sentiment. Expert Syst Appl 42(1):306–324
Naujoks A, Benkenstein M (2020) Who is behind the message? The power of expert reviews on eWOM platforms. Electron Commer Res Appl 44:101015
Nguyen BD, Nielsen KM (2010) The value of independent directors: evidence from sudden deaths. J Financ Econ 98(3):550–567

Novaković JDj, Veljović A, Ilić SS, Papić Ž, Milica T (2017) Evaluation of classification models in machine learning. Theory Appl Math Comput Sci 7(1)
Oberlechner T, Hocking S (2004) Information sources, news, and rumors in financial markets: insights into the foreign exchange market. J Econ Psychol 25(3):407–424. https://doi.org/10.1016/S0167-4870(02)00189-7
O'Dea B, Achilles MR, Larsen ME, Batterham PJ, Calear AL, Christensen H (2018) The rate of reply and nature of responses to suicide-related posts on Twitter. Internet Interv 13:105–107. https://doi.org/10.1016/j.invent.2018.07.004
Oliveira N, Cortez P, Areal N (2016) Stock market sentiment lexicon acquisition using microblogging data and statistical measures. Decis Support Syst 85:62–73
Oliveira N, Cortez P, Areal N (2017) The impact of microblogging data for stock market prediction: using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Syst Appl 73:125–144
Osei-Bryson K-M (2004) Evaluation of decision trees: a multi-criteria approach. Comput Oper Res 31(11):1933–1945
Wisniewski TP, Lambe BJ, Dias A (2020) The influence of general strikes against government on stock market behavior. Scott J Polit Econ 67(1):72–99
Osei-Bryson K, Ngwenyama O (2011) Using decision tree modelling to support Peircian abduction in IS research: a systematic approach for generating and evaluating hypotheses for systematic theory development. Inf Syst J 21(5):407–440
Ozgur O, Karagol ET, Ozbugday FC (2021) Machine learning approach to drivers of bank lending: evidence from an emerging economy. Financ Innov 7(1):1–29
Pal M, Mather PM (2003) An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sens Environ 86(4):554–565
Wu J-L, Huang M-T, Yang C-S, Liu K-H (2021) Sentiment analysis of stock markets using a novel dimensional valence–arousal approach. Soft Comput 25(6):4433–4450
Prasad K, Prabhu N (2020) Does earnings surprise determine the timing of the earnings announcement? Evidence from earnings announcements of Indian companies. Asian J Acc Res 5(1):119–134
Prabhala M, Bose I (2019) Do emotions determine rumors and impact the financial market? The case of demonetization in India. In: 2019 IEEE international conference on industrial engineering and engineering management (IEEM), pp 219–223
Sabherwal S, Sarkar SK, Zhang Y (2011) Do internet stock message boards influence trading? Evidence from heavily discussed stocks with no fundamental news. J Bus Financ Acc 38(9–10):1209–1237
Salas JM (2010) Entrenchment, governance, and the stock price reaction to sudden executive deaths. J Bank Financ 34(3):656–666
Zhang X, Zhang Y, Wang S, Yao Y, Fang B, Philip SY (2018) Improving stock market prediction via heterogeneous information fusion. Knowl-Based Syst 143:236–247
Zhong X, Enke D (2019) Predicting the daily return direction of the stock market using hybrid machine learning algorithms. Financ Innov 5(1):1–20
Shin J, Jian L, Driscoll K, Bar F (2018) The diffusion of misinformation on social media: temporal pattern, message, and source. Comput Hum Behav 83:278–287
Shmueli G (2016) Business analytics, statistics, teaching. http://www.bzst.com/. Accessed 1 May 2016
Sindhu MI, Bukhari SMH, Sub-Campus BB, Hussain A (2014) Macroeconomic factors do influencing stock price: a case study on Karachi stock exchange. J Econ Sustain Dev 5:114–124
Skrede O-J, De Raedt S, Kleppe A, Hveem TS, Liestøl K, Maddison J, Askautrud HA, Pradhan M, Nesheim JA, Albregtsen F, Farstad IN, Domingo E, Church DN, Nesbakken A, Shepherd NA, Tomlinson I, Kerr R, Novelli M, Kerr DJ, Danielsen HE (2020) Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet 395(10221):350–360. https://doi.org/10.1016/S0140-6736(19)32998-8
Zubiaga A, Aker A, Bontcheva K, Liakata M, Procter R (2018) Detection and resolution of rumours in social media: a survey. ACM Comput Surv CSUR 51(2):1–36
Song Q, Liu A, Yang SY (2017) Stock portfolio selection using learning-to-rank algorithms with news sentiment. Neurocomputing 264:20–28
Tan L (2015) Chapter 17: code comment analysis for improving software quality. In: Bird C, Menzies T, Zimmermann T (eds) The art and science of analyzing software data. Morgan Kaufmann, Boston, pp 493–517
Taiwan Stock Exchange Corporation (2022) Announcement of attention securities. In: Taiwan stock exch. Corp. https://www.twse.com.tw/zh/page/announcement/notice.html. Accessed 17 Jul 2022
Tardelli S, Avvenuti M, Tesconi M, Cresci S (2020) Characterizing social bots spreading financial disinformation. In: International conference on human-computer interaction. Springer, pp 376–392
Ture M, Tokatli F, Kurt I (2009) Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients. Expert Syst Appl 36(2, Part 1):2017–2026
Umar Z, Gubareva M, Yousaf I, Ali S (2021) A tale of company fundamentals vs sentiment driven pricing: the case of GameStop. J Behav Exp Financ 30:100501
Vanstone BJ, Gepp A, Harris G (2019) Do news and sentiment play a role in stock price prediction? Appl Intell 49(11):3815–3820
Verma R, Verma P (2007) Noise trading and stock market volatility. J Multinatl Financ Manag 17(3):231–243
Wang J, Alfosool AM, Su J, Fu X, Tan J (2020) An intelligent system for rumor recognition and rumor sentiment judgment. In: 2020 International conference on computing, networking and communications (ICNC). IEEE, pp 309–313
Wang T, Wang G, Wang B, Sambasivan D, Zhang Z, Li X, Zheng H, Zhao BY (2017) Value and misinformation in collaborative investing platforms. ACM Trans Web TWEB 11(2):1–32

Publisher's Note