Decision Analytics Journal 10 (2024) 100422

Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

A financial anomaly prediction approach using semantic space of news flow on twitter

Amirhosein Bodaghi [a,b,*], Jonice Oliveira [b]

[a] Laboratory for AI-Powered Financial Technologies Limited, Hong Kong
[b] Department of Computing Science, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

ARTICLE INFO

Keywords:
Anomaly prediction
News flow
Twitter
Natural language processing
Complex networks
Semantic space

ABSTRACT

This study represents an initial endeavor to harness the potential of the semantic space within the Twitter news flow to forecast financial anomalies. In pursuit of this objective, approximately two million entities were extracted from the news text disseminated by the most widely followed news channels on Twitter. These entities were scrutinized over 12 years to explore potential correlations between their evolution and future stock market anomalies. The examination focused on the centrality measures of these entities within their daily semantic graphs, with particular emphasis on identifying the most correlated entities. Subsequently, these entities were employed to construct a logistic regression model capable of predicting the presence of future anomalies and their direction—whether indicative of an upward trajectory associated with a rise in stock prices or a downward trajectory associated with a decline in prices. The evaluation results demonstrate a remarkable level of accuracy for the prediction model, thereby holding promise for further advancements in this interdisciplinary research domain that encompasses natural language processing, complex networks, and artificial intelligence. Lastly, the findings are discussed in light of pertinent theories that furnish a robust foundation for future investigations.

1. Introduction

The advent of digital technologies has permeated numerous disciplines, ranging from smart cities [1] to healthcare [2,3], among others. Within this vast digital landscape, social media stands out as a prominent digital service, encompassing diverse applications such as self-presentation [4] and catalyzing significant social revolutions [5]. However, the widespread accessibility of social media platforms, coupled with their capacity to aggregate information from various sources, exposes users to a deluge of both genuine and false news on a daily basis [6,7]. Financial matters, in particular, captivate substantial attention among the diverse range of news topics [8], with the prediction of stock price anomalies emerging as a critical concern. Grounded in the efficient market theory, stock prices are believed to be influenced by all observable information and relevant news [9,10]. Consequently, this study endeavors to leverage news shared on Twitter as a basis for developing a novel model for predicting anomalies in stock prices. Within the literature, various definitions and terms have been proposed to describe anomalies, including notable outlier observations, surprises, aberrations, discordant observations, exceptions, peculiarities, or contaminants. Despite the diverse terminology employed, these concepts collectively refer to the identification of outliers within unlabeled time series data [11]. Fu et al. [12] defined anomalies as deviations from the efficient market hypothesis, a key area of investigation within the field of financial economics. In our study, an anomaly signifies an upward or downward deviation in a specific stock price that surpasses the established historical average.

1.1. Significance of the research

The significance of this research lies in its contribution to the field of finance and the prediction of stock price anomalies. Accurate prediction of anomalies in stock prices holds substantial importance for investors, asset managers, and stock exchange regulators. By leveraging news shared on Twitter, this study aims to develop a novel model for predicting anomalies in stock prices. Such a model would provide valuable insights into the behavior of stock prices, enabling investors to adjust their strategies and mitigate investment risks. Additionally, this research addresses a gap in the literature by focusing on social media as a unique and rich source of news, offering a broader time span and capturing investors' attitudes, which are crucial factors for predicting stock movements.

1.2. Research gap and novelty

The research gap addressed by this study is the lack of efficient approaches for utilizing news derived from social media, specifically

* Corresponding author at: Department of Computing Science, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.
E-mail addresses: bodaghi@ppgi.ufrj.br (A. Bodaghi), jonice@dcc.ufrj.br (J. Oliveira).

https://doi.org/10.1016/j.dajour.2024.100422
Received 23 June 2023; Received in revised form 24 January 2024; Accepted 9 February 2024
Available online 10 February 2024

Twitter, in predicting stock price anomalies. While existing literature has explored the use of specific financial media or general news sources for predicting financial markets, the potential of social media as a valuable source of news remains largely untapped. This study fills this research gap by harnessing the power of Twitter, a widely used social media platform with a high level of user engagement, as a unique and rich source of news. Furthermore, this research introduces several novel aspects that contribute to the advancement of knowledge in this field. Firstly, it adopts a distinct natural language processing methodology that focuses on extracting entities from news texts and capturing their semantic relationships with a higher level of granularity. This approach enables a more comprehensive analysis of the news data and its impact on stock price anomalies. Secondly, the study employs complex network analysis to model the daily semantic relationships among entities, providing a deeper understanding of the underlying dynamics. Lastly, the research contributes to reproducibility and knowledge sharing by introducing an open-source software package for the proposed anomaly prediction model based on news channels on Twitter. Overall, this study not only addresses a significant research gap but also offers novel insights and methodologies, enhancing the current understanding and prediction of stock price anomalies.

1.3. Research hypothesis

The behavior of stock prices adheres to a random walk pattern in terms of both magnitude and direction [13], rendering systematic forecasting inherently challenging [14]. The dynamics of stock prices are not solely influenced by market participants but are also subject to various external events [15], such as political factors [16–18], information security incidents [19], and specific news or announcements [20,21]. Furthermore, individuals are attuned to financial and economic news, as well as information pertaining to personal safety, terrorist activities, and other threats [8]. Consequently, it is plausible that decision-makers dynamically adjust their choices based on the latest news developments. The existing literature underscores the significance of the relationship between news and stock prices, thereby motivating the formulation of the following research hypothesis:

H: The integration of semantic relationship dynamics among entities extracted from news texts originating from the most widely followed news channels on Twitter enhances the predictive capability for anomalies within financial markets.

1.4. Objectives

The primary objective of this study is to identify news entities whose semantic relationships with other entities undergo an evolution that exhibits a high correlation with the future deviation rate of a specific stock. However, this objective raises the question of how the temporal evolution of a series of entities within their semantic graph relates to the likelihood of future anomaly occurrences in the price of the given stock. Answering this question proves challenging, as anomalous behavior in stock prices can be influenced by a wide array of factors, including social, economic, environmental, political, and global events. Nonetheless, by leveraging a comprehensive collection of news encompassing these events over an extended time period, it becomes feasible to extract a substantial number of entities along with their interconnectedness. From this pool of entities, it is possible to identify those whose temporal dynamics exhibit the highest correlation with the potential occurrence of anomalies in the future price of the targeted stock. This study seeks to take preliminary steps in this direction by introducing a methodology for detecting the most correlated entities for each specific stock. These identified entities are subsequently utilized to develop a predictive model capable of forecasting future anomalies for the given stock. The model will incorporate the entity features at the current time $En(t)$, and the anomaly records of the given stock in the past $An(t')$ to predict the presence of an anomaly for the given stock in the future $An(t'')$. Eq. (1) shows the relationship.

$$An\left(t''\right) = En\left(t\right) + An\left(t'\right) \quad where \quad t' < t < t'' \tag{1}$$

1.5. Theoretical background

Theories concerning the prediction of firm performance, such as information theory [22] and adaptive market theory [23], highlight the role of information in reducing uncertainty [24]. The adaptive market theory emphasizes the insufficiency of relying solely on summary records for predicting firm performance, necessitating the analysis of diverse information sources to mitigate uncertainty. Consequently, access to various information sources becomes crucial in financial predictions [25,26], with significant attention given to news stories in the literature [27]. However, there is a gap in efficient approaches for utilizing news derived from social media. Social media serves as a unique platform for disseminating news due to its high user engagement among news consumers. This characteristic makes social media a distinct source of news that can be leveraged for predicting financial market dynamics. Crowd psychology [28] offers a possible explanation, suggesting that as the size and diversity of the statistical population increase, the predictive accuracy of the model improves [29]. Moreover, behavioral finance and investment psychology theories posit that investor behavior is influenced by their prevailing sentiments of optimism or pessimism regarding future market conditions [30]. This study aims to leverage social media as a valuable source of news to develop improved prediction models for financial markets. Additionally, in line with the efficient market theory [10], three types of information are deemed relevant for financial prediction: historical information (consisting of stock price records), public information, and private information. While the weak form of the efficient market theory suggests that historical information alone does not provide predictive advantages for stock prices, it acknowledges the impact of public and private information. This highlights the importance of new information, especially for stocks characterized by short-term memory where technical analysis offers limited insights. Building upon this theoretical foundation, the present research focuses on a novel form of information that combines public and private perspectives. The public aspect arises from utilizing Twitter as a source of publicly available news, while the private aspect is derived from extracting latent patterns in the evolution of semantic relationships among news entities.

1.6. Methodological approach

The methodological approach employed in this study combines big data analysis, natural language processing, and complex network methods to develop a news-based anomaly prediction model. The data collection process involves gathering daily news posted by popular news channels on Twitter, spanning a 12-year period and encompassing over 713 million followers. This extensive corpus of news data ensures access to a substantial volume of published news over a long duration. However, existing literature primarily relies on specific financial media, general news sources for news data [31], or alternative sources like the Global Database of Events, Language, and Tone (GDELT) [32]. Nonetheless, GDELT covers news only from 2015 onwards, while social media offers a broader time span, providing valuable data on investors' attitudes, a crucial factor for predicting stock movements [33]. In terms of data processing, a distinct natural language processing methodology is adopted to extract entities from the news texts at a higher level of granularity and capture their semantic relationships. This results in the identification of approximately two million entities, whose daily semantic relationships are modeled using graphs and complex network analysis techniques. For anomaly prediction, while statistics-based methods are commonly used for time-based regression

prediction [34], this study tackles the prediction problem through classification, utilizing polynomial logistic regression with varying numbers of features. Unlike conventional text-to-numerical conversion methods, this approach retains the semantic relationships between entities and allows for a more nuanced analysis of the data. In terms of reproducibility, an open-source software package[1] is developed for the proposed anomaly prediction model, promoting transparency, accessibility, and collaboration within the research community. Especially, with the availability of numerous online open-source platforms [35], ensuring code reproducibility has become more feasible [36].

1.7. The structure

The paper is structured as follows. Firstly, a comprehensive literature review is presented, examining prior research on the topic. Subsequently, the section on data collection provides insights into the process, while the research methods employed are elaborated upon in detail. The findings of the study are then presented, accompanied by visual aids for better comprehension. Finally, the paper concludes with a thorough discussion of the results and their broader implications.

2. Literature review

This section presents a review of the literature pertaining to the key aspects closely related to the research conducted in this study. Initially, it explores the significance of news in the context of financial anomalies. Subsequently, existing studies focusing on anomaly prediction and detection techniques are delved into. Lastly, it investigates the role of social media and graph-based methods in this domain.

2.1. News and financial anomalies

News narrative studies and natural language processing have been developed with the advent of big data in recent years, particularly in the field of economic forecasts [37,38]. Boubaker et al. [39] proved the existence of a link between news diversity and financial market movements. They found that news diversity increases when the market progresses, but decreases when the market slips into a downward trend. Moreover, the complex relationships between news elements leave room for complex network methods to develop robust prediction models [32]. The domains of news narrative studies and natural language processing have witnessed notable advancements in recent years, largely attributed to the advent of big data, particularly in the realm of economic forecasts [37,38]. Remarkably, Boubaker et al. [39] have empirically demonstrated a correlation between news diversity and movements within financial markets. Their findings indicate that news diversity tends to increase during market upswings, whereas it decreases during periods of market downturn. Furthermore, the intricate interrelationships among various elements of news content present opportunities for the application of complex network methodologies in the development of robust prediction models [32].

Shaikh and Huynh [40] investigated the impact of the Coronavirus Disease of 2019 on the financial market, particularly focusing on investor fear during this period. Their research employed a time-series based regression model that incorporated dummy regressions to track the progression of the pandemic. The study revealed that news, regardless of its nature (positive, negative, or misleading), influenced market volatility. In another study, Engelberg et al. [41] examined a comprehensive set of 97 stock return anomalies. The researchers discovered that anomaly returns were 50% higher on days when corporate news was released and six times higher on days of earnings announcements. By comparing the predictability of returns between news days and non-news days, they found that anomaly returns were higher on news days. Importantly, these findings were robust across various types of anomalies and were not attributable to day-of-the-week effects or volatility. However, the authors acknowledged that the results primarily reflected firm-specific news, as anomaly returns did not exhibit the same increase on days featuring macroeconomic news announcements.

Tao et al. [42] sought to examine how news contributes to resolving information uncertainty in the stock market. Their findings provided empirical evidence supporting the pivotal role of news in reducing information uncertainty. Raman et al. [24] conducted sentiment analysis and quantitative assessments using Wall Street Journal articles for companies in the S&P 500 index. Their objective was to ascertain the potential effects of these articles on abnormal returns over time. Furthermore, Chen et al. [43] conducted a study in which they collected news content from selected official accounts on Sina Weibo. The extracted news content underwent sentiment feature extraction, and these features, along with technical indicators, were employed as inputs for a novel hybrid model aimed at forecasting stock volatility in the Chinese stock market.

2.2. Anomaly detection

Blázquez-García et al. [11] conducted a comprehensive review of outlier detection techniques in time series data, categorizing them based on input, outlier type, and method. A taxonomy based on this review is depicted in Fig. 1. The input data can either be a univariate time series, consisting of ordered real-valued observations recorded at specific times, or a multivariate time series, consisting of k-dimensional vectors with each dimension representing a real-valued observation recorded at a specific time. The outlier type refers to parts of a time series that significantly deviate from their expected values, which can be calculated using estimation or prediction models. The method employed can be univariate, considering only a single time-dependent variable, or multivariate, capable of simultaneously analyzing multiple time-dependent variables. In this research, a univariate approach is employed using input data based on estimation models to detect point outliers.

Qiu et al. [44] proposed a method based on Granger causality to detect anomalies in multivariate time series by identifying changes in dependency relationships. Aditya et al. [45] employed higher-order joint moment tensors to characterize anomalous events, utilizing the abnormal values observed in the distribution of certain features. They demonstrated the effectiveness of decomposing the cumulant fourth-moment tensor to detect outlier data. Zhang et al. [33] utilized the random forest algorithm for anomaly detection and introduced the concept of abnormal point scale, which measures the abnormality of a sample based on its similarity to other samples. ElBannan [46] investigated financial distress in Middle East and North African countries, revealing that mature, profitable, liquid, small firms with a high market-to-book ratio and low asset growth are less susceptible to financial distress.

2.3. Anomaly prediction

Yin et al. [47] classified abnormal predictions into three categories: (1) anomaly prediction as a classification problem, where a classifier is trained on abnormal and normal time series samples to detect and predict anomalies; (2) early detection of abnormal signs and continuous monitoring, which involves the continuous analysis of anomalous symptoms to enable prediction of anomalies in streaming data; and (3) prioritization and comprehensive analysis of multiple data streams for anomaly prediction, utilizing multivariate time series. Langone et al. [48] proposed a method for predicting the likelihood of future failures in an industrial system. They employed a three-stage approach using an imbalanced multivariate time-series dataset: (1) extraction of representative features from the raw time series, (2) selection of the most relevant predictors based on Kolmogorov–Smirnov distance

[1] https://codeocean.com/capsule/0910305/tree/v1.

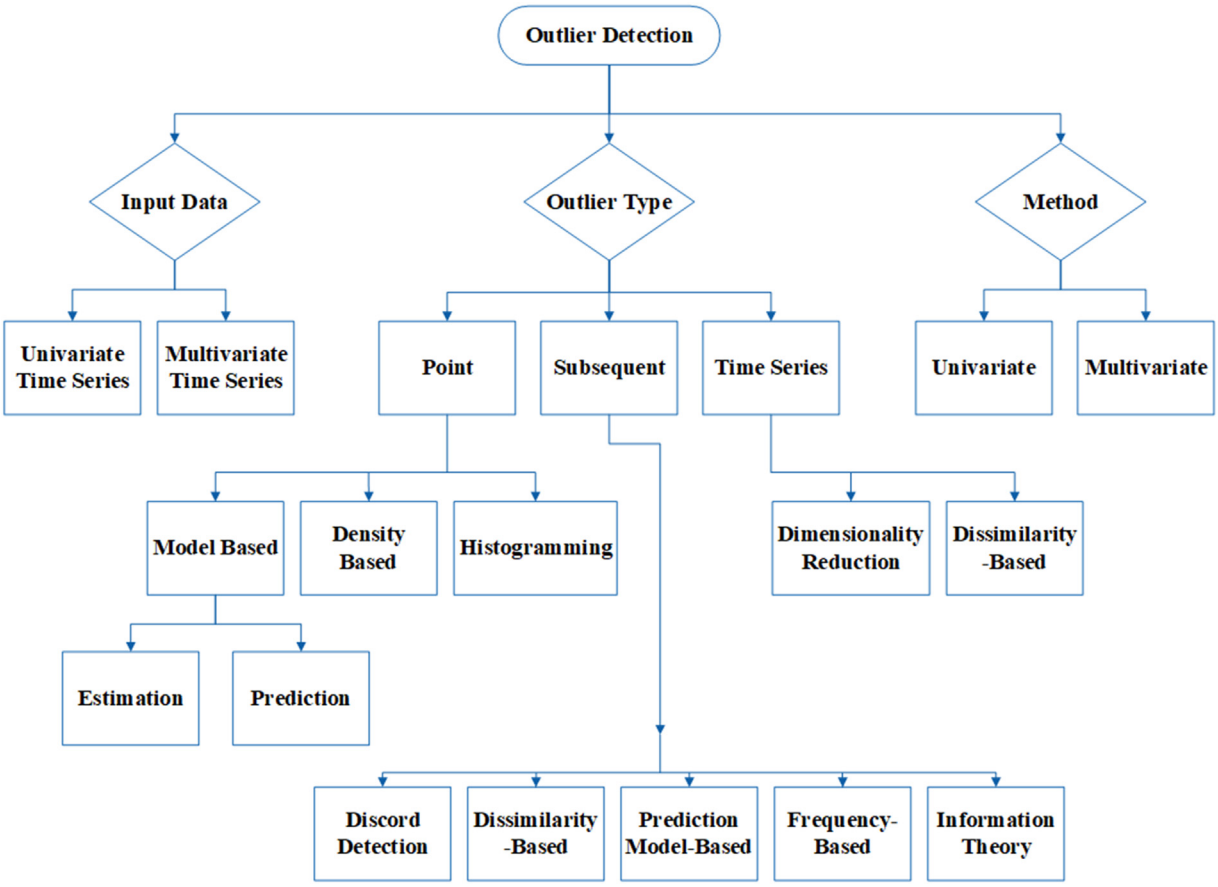

Fig. 1. Outlier detection techniques.

and elastic net regularization, and (3) training a regularized logistic regression model. This approach enables the prediction of failure occurrence up to one hour in advance. Wang et al. [49] introduced a novel machine learning-based framework for predicting the occurrence of system exceptions and failures in a financial information technology system. They combined time series data of key performance indicators with text data from system exception logs to build the anomaly prediction framework. Fu et al. [12] proposed an integrated stochastic optimization technique as a comprehensive method for market return prediction, utilizing a financial knowledge graph as part of the approach. The features used in their study are drawn from a financial knowledge graph and include financial ratios, trading volume, price data, industrial macro data, and public sentiments.

2.4. Social media in stock price prediction

In the domain of social media and stock market prediction, several studies have investigated the influence of social media influencers on market movements [50], as well as the discussions related to finance on social media and online platforms [51]. However, a significant portion of the literature has focused on analyzing social media sentiments toward specific companies or stocks to predict stock prices. Such sentiment metrics have been demonstrated to be valuable for stock market prediction [31]. Huang and Liu [51] developed a novel forecasting model by utilizing sentiment analysis of social media reviews and replies. Carosia et al. [52] conducted multiple machine learning experiments to analyze the effects of social media sentiments on the Brazilian stock market. Zhang et al. [53] presented a framework for forecasting based on extracted sentiments from stock-related posts on social media and web news from Chinese financial discussion boards. Li et al. [54] analyzed Twitter data to predict the stock price movements of 30 companies listed on NASDAQ or the New York Stock Exchange. Li et al. [55] developed a model combining four different sentiment dictionaries with technical indicators and textual news data from Hong Kong for prediction purposes.

Maqsooda et al. [56] examined sentiment analysis on Twitter data to investigate the impact of major events on stock markets during the period of 2012–2016. Their findings highlighted the significant influence of the US election on the stock markets of different countries. Moreover, Hu and Zhu [57] conducted a simulation study to explore the dynamic relationships among news media, social media, and public opinion. They found that interactions within social networks can reinforce the influence of media. Teti et al. [29] employed an ordinary least square model with tweet count and sentiment as variables, without considering autoregressions, to examine the predictive power of Twitter. Their results for the technology industry demonstrated a strong relationship between Twitter sentiment from the previous day and the stock price of the current day. Albarrak et al. [58] analyzed the impact of a company's Twitter activities in disseminating financial information to investors and found a significant reduction in the cost of equity. Chen et al. [59] focused on social media articles from the perspective of potential investors, inferred from commentaries written in response to these articles, and found that future stock returns can be predicted based on the expressed views in both articles and commentaries.

2.5. Graph-based prediction

Tilly and Livan [32] have posited that the conversion of news narratives into knowledge graphs can enhance traditional macroeconomic forecasting models. Sun et al. [60] constructed trading graphs for individual stocks based on investors' relationships and categorized nodes according to their connectivity features. They observed a robust

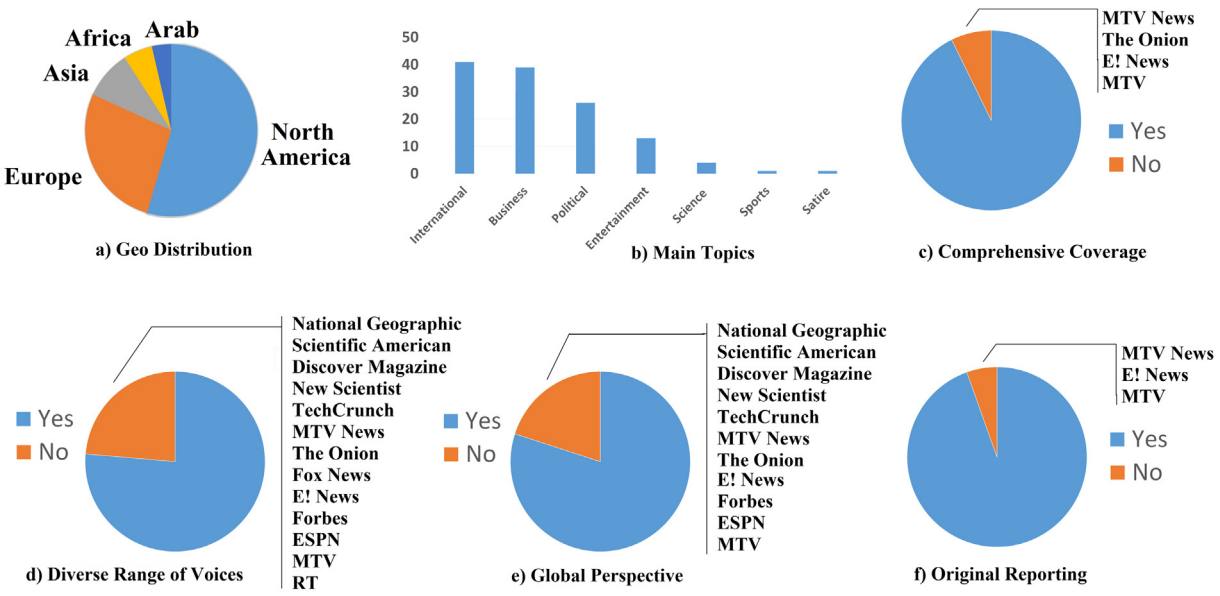

Fig. 2. Visualization of the news channels' characteristics.

Granger causality between these trading relationship indices and stock prices, which they utilized for stock price prediction. Li et al. [61] created a textual network based on the interplay of words and their co-occurrence patterns in documents to explore the relationships between the Shenwan index in China and analysts' research reports. Osipov et al. [8] developed an oriented graph using word-word pairs to forecast news feeds using text data. Adamic et al. [62] employed graph analysis tools to describe the dimensions of information and liquidity flows in the E-mini S&P stock index futures market. They demonstrated a strong correlation between financial volatilities and graph features. Additionally, other researchers [63] have indicated that graph features can provide deeper insights into entity characteristics. Tilly and Livan [32] introduced a graph-based approach to macroeconomic forecasting, leveraging various themes extracted from news narratives, which yielded improved predictions of industrial production (IP) in three major economies. Patil et al. [64] developed a graph-based method that incorporated spatial–temporal relationships between different stocks, coupled with historical stock price-based technical methods.

3. Data

The dataset used in this study consists of two primary components: news data and stock price data. The news data was acquired through web retrieving from the top 55 Twitter news channels, renowned for their substantial following. A comprehensive explanation of the news data will be provided in Sections 3.1 and 3.2. On the other hand, the stock price data was gathered from eight distinct companies and employed as the target variable for the analysis. A more detailed account of the stock price data will be presented in Section 3.3.

3.1. Selection of the news channels

After conducting an exploration of news channels on Twitter, a carefully chosen set of the most highly followed accounts has been selected and is presented in Table 1.

The most prominent news channels encompass a broad array of subjects, including international affairs, business updates, political developments, entertainment news, and scientific advancements. These channels have a global reach, covering all continents across the world. Apart from these fundamental characteristics, there are four key attributes associated with these news channels. Firstly, comprehensive coverage denotes the extensive range of topics addressed by a news channel. A news channel with comprehensive coverage offers news spanning a wide spectrum of subjects. Over 92 percent of these news channels exhibit this attribute. Secondly, original reporting refers to a news channel's capacity to independently gather news and information. Such channels have their own team of reporters and journalists who are actively involved in fieldwork to procure news stories. This distinguishes them from channels that merely repackage news from external sources. Approximately 94 percent of these news channels possess this quality. The third attribute is a global perspective, which relates to a news channel's ability to present news from various countries and cultures worldwide. Channels with a global perspective offer a broad international outlook, in contrast to those primarily focusing on news from a single country or region. More than 80 percent of these news channels embrace this global perspective. Lastly, the diverse range of voices pertains to a news channel's inclusion of multiple viewpoints. Channels with a diverse range of voices provide news from various sources, including politicians, experts, and ordinary individuals. This stands in contrast to channels that exclusively represent the views of a particular group or faction. More than 76 percent of the news channels exhibit this characteristic. The specifics of these news channels are visualized in Fig. 2.

In this research, it is recognized that financial markets are influenced not only by news explicitly related to finance but also by a wide array of factors that shape market sentiment and public discourse. By including non-financial news channels in the analysis, the aim is to capture the broader impact of news on financial markets. Investors and market participants rely on a diverse range of information sources to make informed decisions. These sources encompass not only dedicated financial news channels but also general news outlets, social media platforms, entertainment channels, and more. In today's interconnected world, information flows across various domains, and market participants often consider news and events from multiple perspectives when assessing investment opportunities. By incorporating non-financial news channels among the list of most followed news channels, such as ESPN, MTV, E! News, India Today, NTV Kenya, and others, the comprehensive range of information sources that investors access in the real world is acknowledged. These channels play a contributory role in shaping collective sentiment and discourse, thereby exerting significant influence on financial markets. This inclusion captures the multifaceted nature of information flow and its consequential impact on financial markets, facilitating a departure from rigid categorizations that strictly classify news channels as either financial or non-financial.

Table 1
Popular news channels on Twitter. The number of tweets encompasses all published tweets by the respective accounts from January 1, 2010, until January 1, 2022, excluding January 1, 2022. The data regarding the number of followers, followings, and tweets was collected on January 21, 2023.

No.	News agency	Twitter ID	#Fr	#Fw	#Tweets	Joined date	Host	The main news topics covered by the channel
1	CNN breaking news	cnnbrk	64M	122	78.9K	Jan 2007	US	International news, Business news, Entertainment news, Political news
2	CNN	cnn	61M	1093	395.1K	Feb 2007	US	International news, Business news, Entertainment news, Political news
3	The New York Times	nytimes	54M	863	496.8K	Mar 2007	US	International news, Business news, Political news
4	ESPN	espn	44M	436	135.5K	Mar 2007	US	Sports news, Entertainment news
5	BBC News (World)	bbcworld	39M	18	357.6K	Feb 2007	UK	International news, Business news
6	National Geographic	NatGeo	28M	195	68.8K	Nov 2008	US	Science news, Nature news
7	The Economist	theeconomist	27M	151	331.3K	May 2007	UK	Business news, International news
8	Reuters Top News	reuters	25M	1200	963.9K	Mar 2007	UK	International news, Business news
9	Fox News	FoxNews	23M	262	522.5K	Mar 2007	US	Political news, Business news, Entertainment news
10	The Wall Street Journal	wsj	20M	1085	395.6K	Apr 2007	US	Business news, International news
11	The Washington Post	washingtonpost	19M	1704	451.1K	Mar 2007	US	International news, Business news, Political news
12	Time	time	19M	530	407.3K	Apr 2008	US	International news, Business news, Entertainment news
13	Forbes	Forbes	18M	5038	449.4K	Nov 2009	US	Business news, International news
14	ABC News	abc	17M	447	405.5K	Apr 2009	US	International news, Business news, Political news
15	MTV	MTV	17M	29K	311.9K	Mar 2007	US	Entertainment news
16	NDTV	ndtv	17M	15	1000K	May 2009	India	International news, Business news, Political news, Entertainment news
17	The Associated Press	ap	16M	6697	341.4K	Jun 2009	US	International news
18	BBC News (UK)	BBCNews	14M	82	512K	Jan 2007	UK	International news, Business news
19	CGTN	CGTNOfficial	13M	75	251.5K	Jan 2013	China	International news, Business news, Political news
20	China Xinhua News	XHNews	12M	76	247.1K	Feb 2012	China	International news, Business news, Political news
21	E! News	enews	11M	108K	310K	Mar 2007	US	Entertainment news
22	The Onion	TheOnion	11M	6	95.3K	Mar 2008	US	Satire, Humor
23	The Huffington Post	huffpost	11M	5535	617K	Apr 2008	US	International news, Business news, Political news, Entertainment news
24	The Guardian	guardian	10M	1051	801.7K	Nov 2009	UK	International news, Business news, Political news
25	TechCrunch	TechCrunch	10M	426	248.9K	Mar 2007	US	Technology news
26	WIRED	WIRED	10M	457	157K	Mar 2007	US	Technology news
27	Breaking News	breakingnews	9M	537	107.2K	May 2007	US	International news
28	ABS-CBN News	ABSCBNNews	8M	1078	1000K	Aug 2008	Philippines	International news, Business news, Political news, Entertainment news
29	Sky News	skynews	8M	22	568.1K	Jul 2007	UK	International news, Business news
30	Al Jazeera News	AJEnglish	8M	242	323.8K	Apr 2007	Qatar	International news, Business news, Political news
31	Financial Times	financialtimes	7M	1043	326K	Apr 2007	UK	Business news, International news
32	IndiaToday	IndiaToday	6M	114	1100K	Feb 2009	India	International news, Business news, Political news, Entertainment news
33	Channels Television	channelstv	6M	189	173.5K	Mar 2010	Nigeria	International news, Business news, Political news, Entertainment news
34	CNBC	CNBC	5M	853	552.8K	Feb 2009	US	Business news, International news
35	MTV NEWS	MTVNEWS	5M	5849	134K	May 2009	US	Entertainment news
36	Sky News Breaking	SkyNewsBreak	4M	3	57.2K	Nov 2009	UK	International news
37	MSNBC	MSNBC	4M	786	292.8K	Mar 2007	US	Political news, Business news
38	POLITICO	politico	4M	1188	366.6K	Oct 2007	US	Political news
39	New Scientist	newscientist	4M	291	108.2K	Jan 2009	UK	Science news
40	Scientific American	sciam	4M	807	70.8K	May 2008	US	Science news
41	NTV Kenya	ntvkenya	4M	358	606.4K	Mar 2009	Kenya	International news, Business news, Political news
42	France 24	FRANCE24	4M	453	380.4K	Mar 2007	France	International news, Business news
43	Los Angeles Times	latimes	3M	6886	492.3K	Oct 2008	US	International news, Business news, Political news
44	eNCA	eNCA	3M	734	411.4K	May 2011	South Africa	International news, Business news, Political news
45	Guardian news	guardiannews	3M	1216	338.5K	Feb 2007	UK	International news, Business news, Political news
46	BBC News Africa	BBCAfrica	3M	1687	91.6K	Apr 2009	UK	International news, Business news
47	The Independent	Independent	3M	434	1300K	Oct 2008	UK	International news, Business news, Political news
48	Newsweek	Newsweek	3M	578	322.1K	Mar 2007	US	International news, Business news, Political news

(*continued on next page*)

Table 1 (*continued*).

49	CBC News	CBCNews	3M	1278	270.7K	May 2007	Canada	International news, Business news
50	CBC News	cbcnews	3M	1278	270.7K	May 2007	Canada	International news, Business news
51	The Telegraph	telegraph	3M	724	523.4K	Sep 2008	UK	International news, Business news
52	RT	rt_com	3M	686	434.4K	Aug 2009	Russia	International news, Political news
53	The Star	staronline	1M	251	516.6K	Mar 2009	South Africa	International news, Business news, Political news
54	Discover Magazine	DiscoverMag	1M	193	95.8K	Mar 2009	US	Science news
55	Al Arabiya English	AlArabiya_Eng	9M	46	334.6K	Feb 2009	Saudi Arabia	International news, Business news, Political news
SUM			760M	195K	22M			

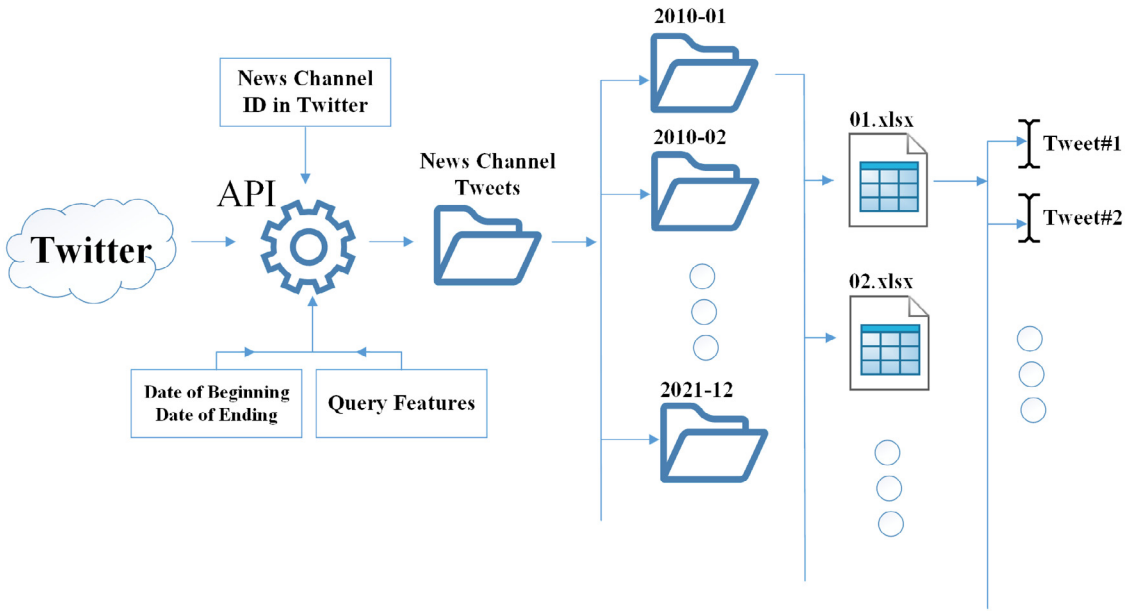

Fig. 3. Data Collection and Storage Workflow. The data collection and storage process involves specifying the start and end dates, as well as defining a set of desired query features, such as follower and following counts, tweet ID, and other relevant information. Through the retrieving process, all tweets published by the designated news channel within the specified time frame are retrieved and saved in individual Excel files categorized into date-labeled folders. Each Excel file contains separate rows presenting the information for each tweet, with each column representing the value of a specific query feature, such as tweet text, tweet ID, and more.

3.2. The news data collection

The retrieving procedure is executed by employing HTTP GET requests via the REST API version 2 approach. The implementation code is scripted in Python, leveraging the Twitter API version 2 by an academic account. The retrieving process requires input data consisting of several query parameters, including the start date (January 1, 2010) and end date (January 1, 2022). Subsequently, the program retrieves all the tweets published by the specified news channel and stores them in distinct Excel files. These files are appropriately labeled with the corresponding day number and meticulously organized within parent folders, which are labeled with the respective month and year numbers. The retrieving process and the structure of the dataset are visually illustrated in Fig. 3.

The implemented code generates a file for each day, even if the specified news channel did not publish any tweets during that particular day. For instance, within this research, it was observed that four out of the 55 news channels established their accounts after January 2010. Consequently, the code generated empty daily files for the time period preceding their membership commencement. During the 12-year study duration, each news channel only published a few hundred thousand tweets. This signifies that it is feasible to fetch a limited number of news channels within a month without surpassing the 10 million tweets per month restriction imposed by API v2 for academic accounts. However, due to Twitter's querying speed limitations, the process is slowed down, allowing only one news channel to be fetched per day. The complete retrieving process was executed on the backend of a Linux server.

3.3. The stock market data

Prior to delving into the process of selecting stock market data as the focal point of this phase, it is imperative to establish a clear definition of the term "anomaly".

3.3.1. Anomaly

In financial markets, selecting an appropriate time period for anomaly detection and prediction is crucial for practical decision-making. While shorter time intervals, such as daily predictions, may offer more frequent insights, they are susceptible to high levels of noise and volatility, making it challenging to distinguish meaningful anomalies from temporary fluctuations. Conversely, longer time intervals, like monthly or yearly predictions, may overlook important short-term dynamics, hindering timely decision-making. Recognizing this need for a suitable time frame, the research focuses on detecting and predicting anomalies on a weekly basis. The choice of a weekly time interval strikes a balance between capturing significant market developments and providing a practical timeframe for decision-making. A week is a widely accepted time period for evaluating financial data and making informed investment decisions, allowing for the aggregation of daily price movements and capturing meaningful trends and patterns influencing stock prices. Weekly predictions offer investors a more comprehensive view of market dynamics compared to daily predictions, facilitating timely reactions. Moreover, a weekly prediction model aligns well with the typical reporting and decision-making cycles in financial markets, meeting the practical needs of market participants. By focusing on weekly anomalies, the prediction model provides investors

Table 2
Selected companies as target data.

No.	Stocks	Company	Number of weeks active in the 12 years period of study	Number of weeks with anomalies in the 12 years period of study	Anomaly rate	Market cap	Category
1	AAPL	Apple	637	64	0.1	$2.461 T	Technology
2	AMZN	Amazon	637	81	0.13	$1.106 T	E_Commerce
3	TSLA	Tesla	611	237	0.38	$815.45 B	Industry
4	FB	Facebook	511	84	0.16	$560.96 B	Social Media
5	NVDA	NVIDIA	637	162	0.25	$422.39 B	Technology
6	TCEHY	Tencent	636	111	0.17	$411.74 B	Communication
7	600519.SS	Kweichow	637	70	0.11	$331.67 B	Beverages
8	BAC	Bank of America	637	97	0.15	$293.02 B	Banking

with a timely and practical tool for identifying potential irregularities or abnormal patterns in stock price behavior, impacting investment strategies, risk management, and overall portfolio performance. In line with the objective of predicting weekly anomalies, conducting anomaly detection on a weekly basis is imperative. The established methodology of calculating the moving average [65] is employed for detecting weekly anomalies. This involves determining the maximum deviation rate from the computed average and assessing whether it surpasses the predefined threshold. To compute weekly anomalies, the mean value of daily prices for a given week is calculated, and the maximum deviation from the mean is determined. This maximum deviation is divided by the mean to derive the deviation rate, which is then compared to a predefined threshold interval of [−Threshold, +Threshold] to ascertain the anomaly status. Weeks exceeding, falling below, or falling within the threshold interval are respectively assigned values of +1, −1, or 0, as illustrated in Eq. (2).

$$Anomaly = \begin{cases} +1: & Deviation_{Rate} > +Threshold \\ 0: & -Threshold < Deviation_{Rate} < +Threshold \\ -1: & Deviation_{Rate} < -Threshold \end{cases} \quad (2)$$

$where:$

$$Deviation_{Rate} = \frac{\max_{i=1}^{i=n} |V_i - M|}{M}, M = \frac{\sum_{i=1}^{i=n} V_i}{n}$$

$n = 7\,(due\ to\ the\ weekly\ analysis), V_i = the\ stock\ price\ in\ the\ day\ i,$

$$M = weekly\ average\ of\ the\ stock\ price$$

3.3.2. Election of the stocks

Upon conducting an analysis of the global financial markets with the highest capitalization, it was observed that 8 stock markets exhibited an anomaly rate surpassing 0.1. The details of these stocks are presented in Table 2.

The selection of these companies as the target data for the anomaly prediction model is justified by several factors. Firstly, they are among the highest capitalized entities in global financial markets, indicating their significant influence and market prominence. However, their selection is not solely based on market capitalization. These companies possess additional merits that make them compelling choices for the study. One notable aspect is the diverse range of industries they represent. Technology giants like Apple Inc., Amazon.com, Inc., and NVIDIA Corporation account for the dynamic and rapidly evolving nature of the technology sector. Furthermore, the presence of Tesla, Inc. highlights the growing importance of sustainable energy solutions and electric vehicles in the global market. Additionally, the selected companies have wide global coverage, transcending geographical boundaries. Companies like Facebook, Tencent Holdings Limited, and Amazon.com, Inc. have a significant international presence, serving millions of users worldwide. This global reach implies that their share prices in the stock market are subject to international news, economic developments, and geopolitical factors. Therefore, studying these companies provides valuable insights into the impact of international events and news on stock price movements. Moreover, the inclusion of companies from both the United States (e.g., Apple Inc., Amazon.com, Inc., Bank of America Corporation) and China (e.g., Tencent Holdings Limited, Kweichow Moutai Co., Ltd.) offers a balanced representation of two of the world's most powerful economies. This allows for a comparative analysis of how different market conditions, regulatory environments, and economic policies influence stock price anomalies.

4. Method

The methodology employed in this research starts by extracting entities from textual news data, followed by entity resolution to identify unique entities within the dataset. Subsequently, leveraging the semantic relationships among these entities, daily graphs are constructed, and various centrality measures are computed for each entity within these graphs. Finally, utilizing these centrality measures, the entities most correlated with the future deviation rate of each stock are identified and utilized to construct an anomaly prediction model specific to that particular stock. The subsequent subsections elaborate on each of these stages, providing a comprehensive explanation of their implementation.

4.1. Entity extraction

In preparation for entity extraction, a cleaning process was conducted to eliminate HTTP addresses, mentions, hashtags, and punctuation from the news texts acquired from the news channel tweets. Subsequently, the SpaCy open-source software library, renowned for its advanced natural language processing capabilities, was employed for entity extraction. Specifically, the en_core_web_lg pipeline provided by SpaCy, which has been extensively trained on a vast web-based dataset, was utilized for the precise labeling of entities. The precision, recall, and F-score values for this pipeline are reported to be 86%, 85%, and 86% respectively. However, in order to enhance accuracy, certain modifications were made to the pipeline. The focus was directed towards a restricted set of named entity labels, such as 'persons', and for frequently mentioned and widely recognized named entities, the labeling process was closely monitored by human referees.

4.2. Entity resolution

The entity resolution process aims to consolidate various variations of the same entity within the dataset. In this study, the method employed focuses on two indicators: the label assigned to entities and their position within the corresponding tweets. The entity resolution process consists of five distinct stages, as outlined below:

1. Resolution for entities labeled as "Person" within each tweet: The entity with fewer characters will be merged into the other entity unless the other entity ends with a suffix like Jr. or Sr.
2. Resolution for entities labeled as "non-Person" within each tweet: The entity with fewer characters will be merged into the other entity if the similarity score, based on Levenshtein distance, between the two entities exceeds the specified threshold.

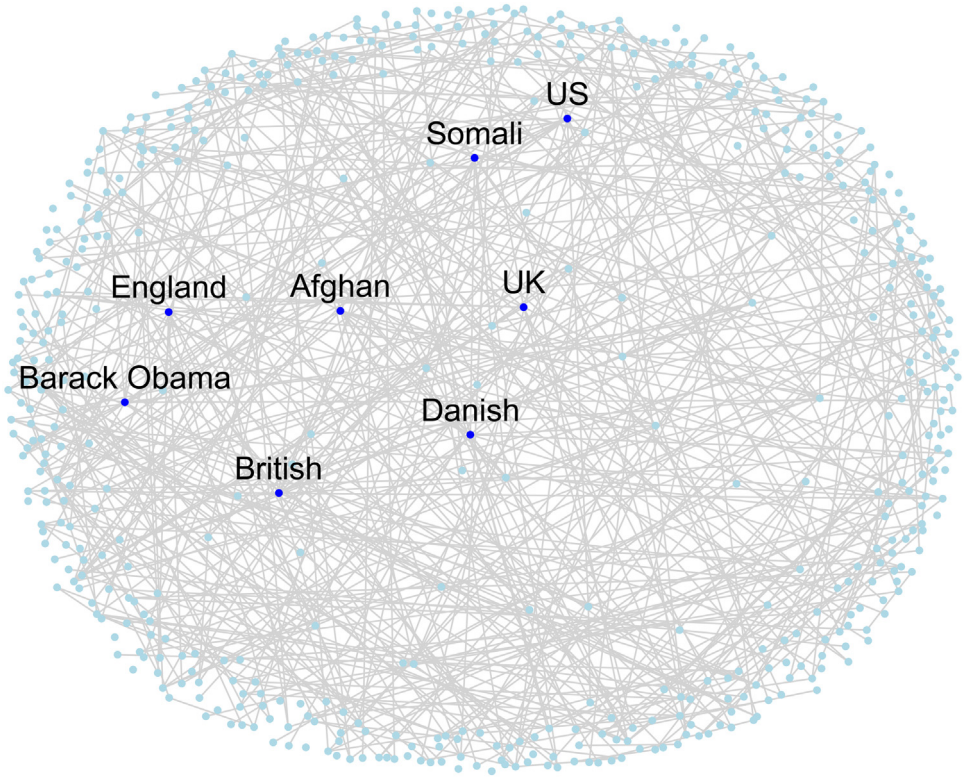

Fig. 4. A Merged Daily Graph: Aggregated Representation of Daily Graphs from All News Channels on January 2, 2010. Nodes in the graph represent entities, and an edge connecting two nodes signifies that these entities co-occurred in at least one tweet released by one of the 55 most widely followed news channels on that specific day. For clarity, only nodes with a high degree are labeled and highlighted in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Resolution for entities labeled as "Person" across different tweets: The entity into which other entities are merged is determined by its higher frequency within the dataset. In cases where two or more entities possess the same frequency, the entity with a greater number of characters will be selected.
4. Resolution for entities labeled as "non-Person" across different tweets: The "non-Person" entities with higher frequencies in the dataset serve as the anchors to which other "non-Person" entities with significant similarity and common words are merged.
5. Resolution for entities with the same characters but different labels: If there are entities with identical letters but differing labels, the "non-Person" entity will be merged into the entity labeled as "Person". However, if none of the entities bear the "Person" label, the entity with a lower frequency will be merged into the other entity.

Each of these stages contributes to the comprehensive entity resolution process, facilitating the unification of entities within the dataset.

4.3. Graph creation

Constructing daily graphs based on entity co-occurrence, as we have seen in Section 2.5, is a widely used technique in network analysis for exploring relationships among distinct entities. In this approach, graph nodes represent unique entities extracted from news channels, and links between nodes signify their co-occurrence within tweets. The decision to use co-occurrence in tweets for constructing daily graphs is motivated by its efficiency and adequacy in representing semantic information within the news ecosystem. Co-occurrence in tweets directly indicates the association between entities, capturing immediate semantic connections and reflecting information flow within the news ecosystem. Additionally, using co-occurrence for daily graph construction provides a scalable and computationally efficient approach, essential for processing large volumes of real-time tweets. This method aligns with the nature of news dissemination on platforms like Twitter, capturing relevant associations in a concise format. Generating a comprehensive daily graph involves merging individual graphs from various news sources, ensuring a holistic representation of entity interactions on a specific day. Over the 12-year study, 4338 daily graphs were generated by amalgamating data from different news channels, allowing for a thorough longitudinal analysis of entity relationships and semantic associations. This extensive dataset facilitates the exploration of dynamic evolution, detection of trends, and deeper insights into entity interactions across news channels over the study period. Fig. 4 shows an instance graph.

4.4. Extraction of centrality features

The dataset consisted of 1,930,053 unique entities, each evaluated for four centrality features: degree, betweenness, closeness, and clustering. These centrality measures were chosen for their ability to capture essential characteristics in the dynamic nature of news entities within the daily graph, offering insights into significance, influence, and positioning. Degree centrality quantifies an entity's connections, revealing highly connected entities that may wield greater influence in information propagation. Betweenness centrality identifies pivotal intermediaries between clusters, shedding light on entities crucial for information flow. Closeness centrality gauges an entity's proximity to others, indicating potential faster diffusion and greater visibility. Clustering centrality assesses an entity's tendency to form clusters, identifying those embedded within cohesive substructures. While additional centrality measures could enhance understanding, computational costs were considered. The selected measures strike a balance between capturing essential characteristics and maintaining manageable computational costs. Python libraries, including Vaex, were employed for efficient data manipulation. A multi-processing approach on a Linux

server with 22 cores and 1 Terabyte of memory facilitated timely computations. The analysis produced 4 time series, each containing 4338 data points, representing centrality features for approximately 2 million entities within the daily news entity graph. This yielded a total of around 8 million time series, capturing the evolution of daily news entity graphs. These time series form the foundation for subsequent computations and analyses, enabling detailed investigations into patterns, trends, and anomalies within the news ecosystem.

4.5. Finding the most correlated entities

Identifying the entities that exhibit the highest correlation with future deviation rates in stock prices involves a multi-step procedure. This process entails the following steps: (1) Conversion of the data on a weekly basis, (2) Establishment of criteria for selection, and (3) Calculation of correlations. By executing these sequential steps, the entities demonstrating the strongest correlation with future deviation rates can be identified.

4.5.1. Conversion into a weekly basis

Given that anomalies in stock prices are assessed on a weekly basis, the corresponding centrality values for each entity must also be converted to a weekly framework. Specifically, for every stock, the minimum, maximum, and mean closing prices are calculated on a weekly basis, commencing on Monday and concluding on Sunday. Subsequently, the deviation rate and anomaly status are determined utilizing Eq. (1), with the threshold value set to 0.03.

Additionally, to ensure consistency, the mean value of each entity's feature is computed on a weekly basis. To rectify the previous −1 bias assigned to the centrality values, which distinguished between absent and isolated entities within the daily graph, a value of +1 is added. Consequently, the previous daily values derived from the daily graphs are replaced with these updated weekly values.

4.5.2. Setting the correlation criteria

Before identifying entities most correlated with future deviation rates in stock prices, two key considerations must be addressed. Firstly, the prediction timeframe must be specified; in this study, predictions are made for the upcoming week. Entities with the highest correlation between their current week values in quadruple centrality features (betweenness, closeness, clustering, and degree) and the next week's deviation rate in stock prices are sought. Secondly, auto-correlation lags are considered, with four lags (+3, +2, +1, 0) representing deviation rate values from the previous weeks to the current week. For each centrality feature, correlation values are computed between the current week's 1,930,053 entities and the subsequent week's deviation rate in the stock price. Additionally, correlations between the next week's deviation rate and rates from the current week to the previous weeks are examined. Eq. (3) formalizes these criteria in mathematical notation, utilizing the EntityFeature and AutoCor components in the input data frame, along with the target data frame containing anomaly values for the subsequent week. The correlation process yields a data frame with the most highly correlated entity values for a given feature at week 't'. The number of top entities (e.g., 250, 500, 750, or 1000 entities) is predetermined, resulting in dataset sizes of 1K, 2K, 3K, and 4K, respectively, considering the four distinct features.

$$
\begin{gathered}
Correlation_c\left(D_{input}, D_{target}\right) => CorEntFea_c \\
where: \\
c \in \{degree, betwenness, closeness, clustering\} \\
D_{target} = \left[d_{w1}\right] \quad 5 \le w \le t, \quad t = total\ number\ of\ weeks\ in\ the\ study, \\
d_{w1} = deviation\ rate\ at\ week\ w, \quad D_{input} = [EntityFeature, AutoCor] \\
EntityFeature = \left[e_{km}\right] \quad 4 \le k \le t-1,\ 1 \le m \le f, \\
f = total\ number\ of\ entities, \\
e_{km} = the\ average\ value\ of\ feature\ c\ for\ entity\ number\ m\ at\ week\ k \\
AutoCor = [AutoCorLag_3, AutoCorLag_2, AutoCorLag_1, AutoCorLag_0] \\
AutoCorLag_i = \left[d_{j1}\right] \quad 4-i \le j \le t-i-1, \quad i = lag\ number \\
d_{j1} = deviation\ rate\ at\ week\ j, \quad CorEntFea_c = \left[e_{zh}\right] \quad 4 \le z \le t-1 \\
1 \le h \le n,\ n = the\ number\ of\ most\ correlated\ entity \\
feature\ to\ be\ found, \\
e_{zh} = the\ average\ value\ of\ feature\ c\ for\ the\ most\ correlated\ entity \\
number\ h\ at\ week\ z
\end{gathered}
\tag{3}
$$

4.5.3. Calculation of the correlations

After obtaining the most correlated entities in each feature, the top 250 correlated entities, which may include lagged deviation rates, are identified as those with the highest correlation values for their current week values (for entities) or previous week values (for lagged deviation rates) with respect to the deviation rate of the given stock in the next week. By repeating this process for all four centrality features, a set of 1000 entity centrality features is obtained, representing the entities whose values are most correlated with the deviation rate of a given stock in the next week. This process culminates in the generation of the ultimate data frame, as depicted in Eq. (4). This equation outlines the composition of the final data frame, which consists of two key components: 'CorEntFea', encompassing the feature values of the most highly correlated entities within each feature at week 't', as established in Eq. (3). The second component comprises 'Anomaly' and encompasses the classified anomaly values at week 't + 1'.

$$
\begin{gathered}
FinalDataframe = [CorEntFea, Anomaly] \\
where: \\
CorEntFea = \left[CorEntFea_{degree}, CorEntFea_{betweenness}, \right. \\
\left. CorEntFea_{closeness}, CorEntFea_{clustering}\right] \\
CorEntFea_* \ is\ defined\ in\ Eq.\ (3), \quad Anomaly = \left[a_{w1}\right], \\
5 < w < t, \\
t = total\ number\ of\ weeks\ in\ the\ study, \\
a_{w1} = anomaly\ at\ week w \quad \{+1, 0, -1\}
\end{gathered}
\tag{4}
$$

5. Results

As previously mentioned, the objective of this study is to predict the occurrence of anomalies in the next week. With the final data frame that has been created in the previous stage, this objective can be approached as a classification problem. The classification problem arises from the fact that the presence of an anomaly and its direction, whether it is an upward anomaly (indicating a jump in price) or a downward anomaly (indicating a plummet in price), can be categorized into non-continuous classes. In this study, the anomalies are classified into three classes: class 0 represents the absence of an anomaly in the next week, class +1 represents an upward anomaly in the next week (corresponding to a jump in the stock price), and class −1 represents a downward anomaly in the next week (indicating a plummet in the stock price).

5.1. Balancing the data

In anomaly detection or prediction tasks, imbalanced data is a common challenge, and this study is no exception. As shown in Table 2, anomalies occurred for each stock in approximately 0.1% of all weeks

during the 12-year study period. The class distribution becomes highly skewed when considering the breakdown of anomalies into upward and downward directions, with Class 0 (absence of anomalies) having a significantly larger number of samples compared to the other two classes. This severe class imbalance requires careful handling for reliable and accurate anomaly detection. To address this challenge, the Synthetic Minority Oversampling Technique (SMOTE), a well-established data augmentation method for minority classes, was employed. SMOTE generates synthetic samples specifically for the minority classes (+1 for upward anomalies and −1 for downward anomalies). This technique identifies the K nearest neighbors for each minority class sample and creates synthetic samples in close proximity to these instances. SMOTE excels in creating representative synthetic samples while preventing overfitting, a common issue with traditional methods like oversampling by duplication or undersampling. The following steps outline the SMOTE algorithm used to generate synthetic samples for addressing class imbalance:

1. Select k nearest neighbors from the minority class to A.
2. For each of the k nearest neighbors B:
    - Compute the difference between A and B for each feature, resulting in a feature vector difference, let us call it $diff(A, B)$.
3. Create synthetic samples by combining the feature vectors of A and its selected neighbors. For each synthetic sample:
    - Choose a random number r between 0 and 1.
    - Compute the synthetic sample's feature vector S as $S = A + r \times diff(A, B)$. This equation essentially performs a linear interpolation between A and B.
4. Repeat steps 2 and 3 for each of the k nearest neighbors.
5. The number of synthetic samples created increases the size of the minority class to match that of the majority class. Adjusting the size of the minority class to equate it with the majority class.

The application of SMOTE facilitates the rebalancing of the dataset by introducing synthetic samples, thereby mitigating the class imbalance issue. By generating additional instances for the minority classes, the dataset becomes more balanced, allowing the anomaly detection model to learn and generalize effectively across all classes. Notably, the synthetic samples produced by SMOTE closely resemble the characteristics and distribution of the real samples, ensuring that the augmented data reflects the underlying patterns and variations present in the original dataset. Through the utilization of SMOTE, the aim was to mitigate the adverse effects of imbalanced data on the performance of the anomaly detection model. This data augmentation technique plays a crucial role in creating a more representative and balanced dataset, enhancing the model's ability to accurately detect and classify anomalies in both upward and downward directions. By employing SMOTE as a practical solution to address the class imbalance challenge, a more comprehensive and reliable analysis of anomalies within the dataset is ensured.

5.2. Logistic regression model for polynomial classification

We used logistic regression for stock price classification due to its proven binary classification effectiveness, prioritizing interpretability and computational efficiency over more complex models like support vector machines and neural networks. This aligns with our goal of extracting insights from daily news and centrality measures to predict anomalies. To apply logistic regression for anomaly detection, especially with a limited number of classes, a systematic approach was followed. The dataset was initially divided into training and testing sets (0.8 and 0.2 ratios, respectively) for model training and independent evaluation. Standardization was applied exclusively to the training set, transforming features to exhibit zero mean and unit variance. This process mitigated the potential influence of variables with different scales, enhancing convergence during model fitting. The standardization transformation was then applied to the test set, ensuring consistency in feature scaling across datasets for fair evaluation. This approach simulates real-world scenarios where the model encounters previously unseen instances during deployment. The mathematical formula for multinomial logistic regression is adjusted accordingly. Given an input vector X and weight matrix W, where each column of W corresponds to a class, the multinomial logistic regression model calculates probabilities P for each class as shown in Eq. (5).

$$P_C = \frac{e^{(W_C . X)}}{\sum_{J=1}^{C} e^{(W_J . X)}} \tag{5}$$

where P_C is the probability that the samples belongs class C. C is the number of classes. W_C is the weight vector for class C, and . represents the doc product. In this case, we have three weight vectors W_C corresponding to each of the triple classes. To make a classification decision, we select the class with the highest probability. Indeed, If P_C is the highest for class C, then we classify the sample as class C. The model is trained by finding the optimal weight matrices W that maximize the likelihood of the observed data. The optimization (Eq. (6)) involves minimizing the multinomial logistic loss or cross-entropy loss, which is a generalization of the binary logistic loss to multi-class problems:

$$J(W) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} I(y^{(i)} = c) \log(P_C^{(i)}) \tag{6}$$

where N is the number of training samples, $y^{(i)}$ is the true class label of the ith sample, $P_C^{(i)}$ is the predicted probability that the ith sample belongs to class C, and finally, $I(y^{(i)} = c)$ is an indicator function that equals 1 if $y^{(i)} = c$ and 0 otherwise. The goal is to find the weight matrices W that minimize this cost function using optimization techniques.

5.3. Implementation and results

The logistic regression model was implemented using scikit-learn, a widely adopted and reliable Python machine learning library. Scikit-learn offers a comprehensive suite of tools for developing and evaluating machine learning models. Leveraging scikit-learn's logistic regression implementation, the model was trained on the designated training set, taking advantage of the library's robust optimization algorithms and efficient computation routines. Following training, the model's performance was rigorously evaluated on an independent test set. Evaluation metrics, including accuracy, precision, recall, and F1-score, were employed to assess the model's ability to accurately classify anomalies. These metrics offer valuable insights into the model's predictive capabilities and its capacity to identify anomalous instances. The combination of logistic regression and the well-established scikit-learn library ensured a robust and reliable approach to anomaly detection in the research. Adhering to established best practices and leveraging available tools aimed to provide a comprehensive and trustworthy analysis of the anomaly detection model's performance. Fig. 5 presents the evaluation results, including precision, recall, and F-score, for each stock.

The utilization of precision, recall, and F-score metrics offers a balanced evaluation of different aspects of the method's predictive capability. The relatively lower accuracy observed for TSLA, BAC, and FB compared to other companies could be attributed to several factors. Here are some potential explanations:

- **Internal Factors:** These companies may have significant internal factors that influence their stock price dynamics, which are not fully captured by the selected exogenous variables. For example,

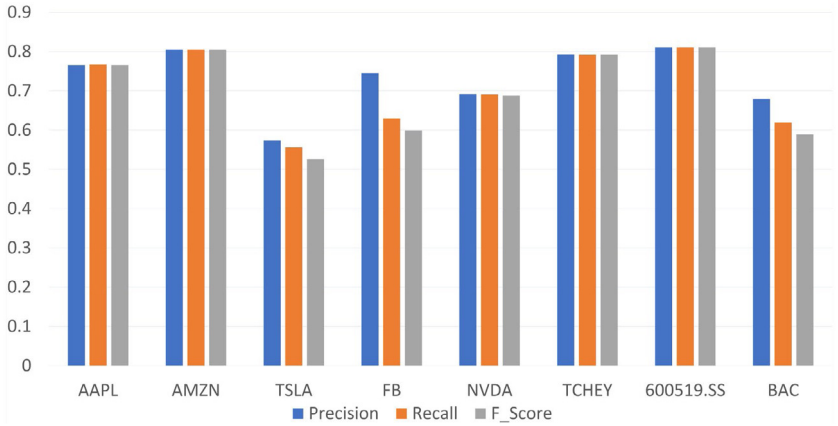

Fig. 5. Evaluation results for different stocks.

TSLA's stock price is highly influenced by factors such as production numbers, vehicle deliveries, and Elon Musk's statements, which may not be adequately reflected in the chosen news entity features. Similarly, BAC's stock price is influenced by factors specific to the financial industry, such as interest rates, economic indicators, and regulatory changes. FB's stock price, on the other hand, is impacted by user engagement, advertising revenues, and privacy concerns. Failure to incorporate these specific internal factors can result in lower accuracy in predicting anomalies.

- **Sector-specific Factors:** Different sectors have unique characteristics and drivers that can impact stock price behavior. TSLA operates in the highly competitive and rapidly evolving automotive industry, while BAC is in the finance sector and FB is in the social media sector. These sectors have their own distinct market dynamics, regulatory influences, and macroeconomic factors that can affect stock price movements. The selected exogenous variables may not adequately capture the sector-specific factors relevant to these companies, leading to lower prediction accuracy.
- **Noise and Market Volatility:** TSLA, BAC, and FB are all well-known, highly traded stocks, which can experience higher levels of noise and market volatility compared to other companies. The presence of noise in the stock price data can make it more challenging to accurately predict anomalies, especially when using exogenous variables. Market volatility, driven by factors such as news sentiment, market sentiment, and macroeconomic conditions, can introduce additional complexity and unpredictability, making it harder to achieve high accuracy in anomaly prediction.

It is important to note that these are potential explanations and further analysis would be necessary to precisely determine the reasons for the observed lower accuracy in these specific companies. Conducting more comprehensive studies, including exploring additional exogenous variables specific to these companies and considering a broader range of factors, may help improve the accuracy of anomaly prediction for TSLA, BAC, FB, and other similar companies.

5.4. The impact of features' number

In order to comprehensively evaluate the impact of feature selection on the accuracy of the model, a series of experiments were conducted. Each experiment employed a distinct set of features. These experiments were designed to provide an analysis of how varying the number of features influences the predictive performance of the model. The outcomes of these experiments are presented in Fig. 6. This figure serves as a visual representation of the relationship between the number of features and the corresponding accuracy achieved by the model. The presented findings reveal the trends and patterns observed in the model's accuracy as the number of features increases or decreases. Through this analysis, a better understanding of the optimal number of features that strike the right balance between accuracy and model complexity can be gained.

The observed behaviors in the accuracy of the predicted model with increasing the number of exogenous variables can be attributed to several factors, including the size and capitalization of the companies. It is important to note that individual company dynamics, market conditions, and specific characteristics can also influence behavior. However, based on the information provided, potential reasons for the similarities and differences in behavior among the companies can be explored.

- **NVDA, AMZN, and TECHY:** These companies display a similar behavior of increasing accuracy with the inclusion of more features. One common aspect among these companies is their large market capitalization and high levels of technological innovation. These companies operate in sectors that are influenced by various factors, such as technological advancements, market trends, and consumer behavior. Therefore, incorporating additional exogenous variables that capture these dynamic factors can contribute to improved accuracy in predicting anomalies.
- **TSLA, BAC, and FB:** These companies exhibit relatively indifferent behavior regarding the increase in the number of features. These companies belong to different sectors, namely the automotive industry, banking, and social media, respectively. The varying behavior might indicate that these companies' stock prices and anomalies are driven by factors that are not necessarily captured by the selected exogenous variables. It is possible that their stock price dynamics are influenced by internal company performance, industry-specific factors, or unique market dynamics, rather than the specific news entities considered as exogenous variables.
- **Apple and 600519.SS:** Apple and 600519.SS display distinct behaviors in terms of accuracy with increasing the number of features. For Apple, there is an initial decrease in accuracy followed by an increase as more features are added. This behavior suggests that the initially included features might have introduced noise or non-relevant information, impacting the accuracy. However, as more informative features are incorporated, the accuracy improves. On the other hand, 600519.SS exhibits an increase in accuracy initially, but with further inclusion of features, there is a subsequent decrease. This pattern may indicate a saturation point where additional features might introduce noise or complexity that adversely affects the model's performance.

Indeed, the observed behaviors in the accuracy of the anomaly prediction model can be attributed to various factors, including the size and capitalization of the companies, the specific sectors they operate in, and the relevance of the selected exogenous variables to their stock

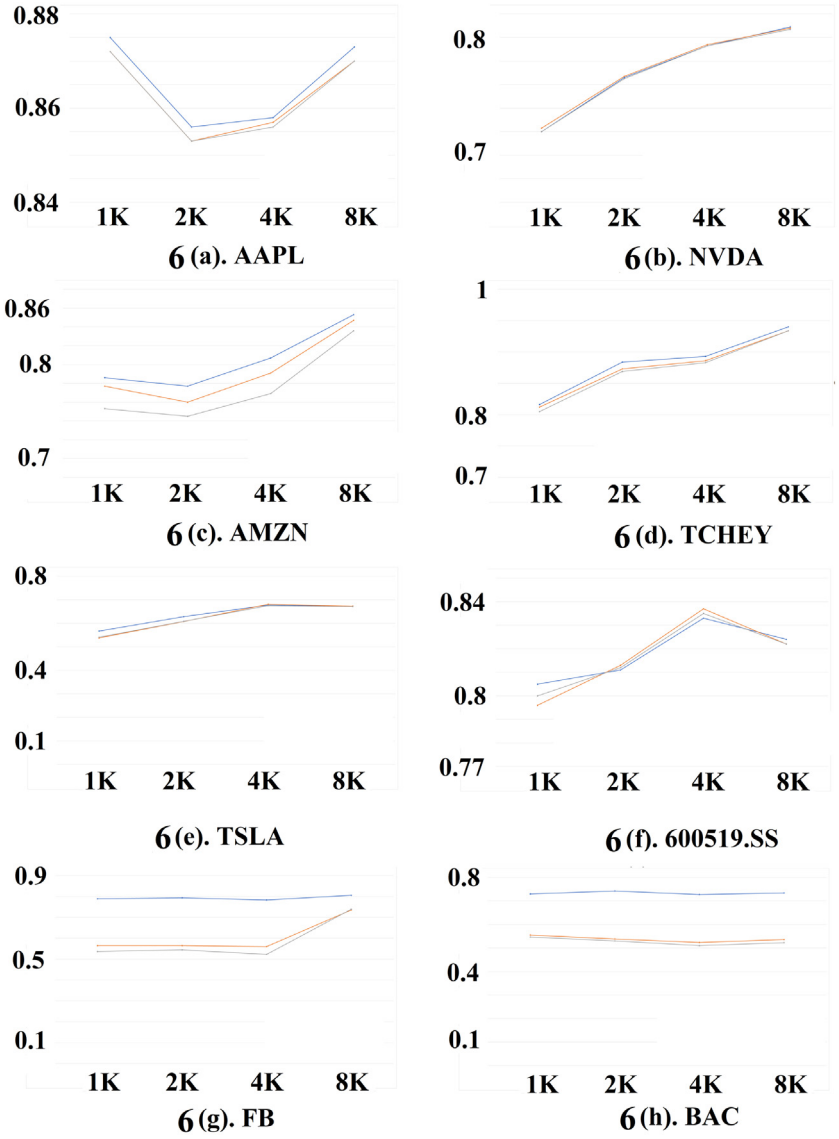

Fig. 6. Influence of the number of features on stock evaluation for different stocks. The graph illustrates the performance metrics, including precision, recall, and F-score, for four different values of feature numbers (1000, 2000, 3000, and 8000) across eight distinct stocks. The metrics are represented by the colors blue, orange, and gray, corresponding to precision, recall, and F-score, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

price dynamics. These factors contribute to the distinct behaviors seen among the different companies in terms of the impact of increasing the number of features on model accuracy.

5.5. Concerns about comparative analysis and overfitting

The primary objective of this study is to present a methodology for identifying highly correlated news entity features associated with future anomalies in stock prices, rather than developing the most optimized model specifically for anomaly prediction. The methodology's significance lies in its ability to extract relevant features, enhancing the accuracy of diverse prediction models when incorporated as exogenous variables. The study does not aim to comprehensively evaluate all possible advanced models; instead, it emphasizes showcasing the potential benefits of integrating identified exogenous features into various predictive frameworks. This approach ensures the methodology's practicality across a wide range of advanced models and contributes significantly to the field by underscoring the importance of feature selection and incorporation in augmenting prediction accuracy. Despite strides in enhancing model accuracy through incorporating highly correlated features, concerns about overfitting persist. To address this, broadening the dataset's scope by collecting data from additional weeks is essential. This approach captures a more diverse range of scenarios and patterns, mitigating the risk of overfitting. Leveraging time allows for continuous model training with new data, enabling adaptation and evolution over time. The iterative training process incrementally updates parameters and fine-tunes internal representations, safeguarding against overfitting and enabling adaptation to evolving trends. This dynamic framework fosters a model that progressively refines itself, ultimately becoming a reliable and accurate tool for prediction and analysis.

6. Discussion

The findings of this study highlight the significance of incorporating longitudinal semantic data from news flow for predicting anomalies in financial markets. In this section, the theoretical and practical implications of these findings will be delved into, elucidating their broader implications for the field.

6.1. Theoretical implications

The findings of this study contribute to existing theories by providing insights into the impact of news sourced from social media platforms. By highlighting the significance of incorporating semantic data from news flow, this research expands the understanding of the role of social media in influencing various aspects of society, including financial markets.

6.1.1. Behavioral finance and investment psychology theory

The impact of news flow on society has been a subject of academic inquiry for a considerable period of time. Recent studies have also examined the role of AI in journalism, highlighting its influence on news dissemination [66]. With the emergence of social media platforms, news flow has gained greater speed and reach, allowing news consumers to not only access news rapidly but also express their reactions to it. This interactive nature of social media can escalate attention and interest among individuals and investors alike. Additionally, cognitive dissonance resulting from the interplay between sentiment and culture has been identified as a key driver of anomalies [67].

Furthermore, factors beyond news flow can directly influence investor actions and subsequently impact anomaly patterns. For example, researchers have explored the relationship between air quality and financial market efficiency, finding that stock market anomalies are more pronounced following periods of severe pollution. However, in contrast, Andrikopoulos et al. [68] examined the impact of weather on stock and foreign exchange markets and found no significant effect on either market or investor mood for the stock exchange. The findings of this study contribute to understanding the impact of interconnected events on stock market anomalies. By shedding light on the hidden connections between different events and their influence on anomaly patterns, this research expands the knowledge in this area. It highlights the importance of considering social media and other external factors in the study of financial market anomalies.

6.1.2. Mispricing theory

The study conducted by Engelberg et al. [41] provides empirical evidence supporting the mispricing theory of anomaly returns. Their research suggests that anomaly returns are driven by biased expectations, which are subsequently corrected upon the release of news. This finding aligns with the notion that news has a greater impact on correcting mispricing in small stocks, given the higher levels of mispricing observed in this segment. The authors also found indications that investors exhibit overly optimistic expectations regarding the cash flows of certain firms while holding overly pessimistic expectations for others. Consequently, when new information becomes available, investors adjust their biased beliefs, leading to changes in prices and the observed predictability of returns.

Furthermore, Lazuarni and Asri [69] examined the influence of heuristic behavior on the emergence of fundamental and technical anomalies in the capital market. Their study revealed that many investors rely on heuristics and fail to consider existing anomalies in the market. These findings contribute to the mispricing theory by highlighting the impact of news on social media, as it can also shape investors' beliefs and impact market prices.

6.2. Practical implications

The study's practical implications encompass various aspects. Firstly, it underscores the emergence of news disseminated through popular social media channels as a novel source of information. The structured nature of text data and the extensive user base of platforms like Twitter render this data unique and valuable for analysis. Secondly, the study's approach, utilizing semantic-based graphs, holds broad applicability and can be adapted to track specific phenomena such as "bankruptcy", "inflation", or "economic recession". Incorporating new entity characteristics, as suggested by Bodaghi and Oliveira [70], has the potential to further enhance the outcomes of such analyses. Thirdly, the study's longitudinal approach, employing daily knowledge graphs based on semantic entities, offers a pathway for the development of evolutionary models for intelligence derived from Twitter data. The companies examined in this study are market giants, known for exhibiting lower immediate reactions to news and events compared to regular companies. Therefore, it is anticipated that the proposed model would yield even greater improvements for regular stocks, which are more susceptible to global signals such as political and economic events at national and international levels [71].

7. Conclusions and future research

The study introduces a novel approach to incorporate Twitter news in predicting stock market anomalies, leveraging natural language processing to extract entities from tweets. The method constructs daily knowledge graphs capturing semantic relationships among entities and employs complex network and time series analyses to identify correlated patterns between entities and stocks, enhancing prediction accuracy. Grounded in relevant theories, the findings have practical implications across disciplines. While the method poses computational challenges due to processing vast entities across daily knowledge graphs, the resource-intensive computations primarily concentrate on the preprocessing stage. Strategic management of these challenges has the potential to redefine real-time data analysis in financial markets, paving the way for transformative applications. The prospective utility extends to fintech, streamlining computational processes and enabling real-time application in streaming financial data. By mitigating the computational burden, our method facilitates a seamless transition into the fintech sector, offering opportunities for real-time anomaly prediction and enhancing decision-making in finance. Acknowledging potential limitations, the study focused on the 55 most-followed news channels on Twitter, suggesting the inclusion of national news channels for detecting impactful national events. Utilizing advanced versions of SpaCy for entity extraction, exploring additional graph features and centrality measures, and incorporating macroeconomic indicators could enrich the analysis. Additionally, optimized models considering influential events like elections or disasters may enhance prediction accuracy, warranting further research in these directions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to enhance the writing process. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgment

The work described in this paper was partially supported by the InnoHK initiative, The Government of the HKSAR, and Laboratory for AI-Powered Financial.

References

[1] N.F.M. Shari, A. Malip, State-of-the-art solutions of blockchain technology for data dissemination in smart cities: A comprehensive review, Comput. Commun. 189 (2022) 120–147.

[2] K. Krmpotic, J.R. Gallant, C. Zufelt, User-centred development of an mhealth app for youth with type 1 diabetes: the challenge of operationalizing desired features and feasibility of offering financial incentives, Health Technol. 12 (2022) 499–513.

[3] A. Bodaghi, A novel pervasive computing method to enhance efficiency of walking activity, Health Technol. 6 (2016) 269–276.

[4] A. Bodaghi, J. Oliveira, A longitudinal analysis on instagram characteristics of olympic champions, Soc. Netw. Anal. Min. 12 (1) (2022a).

[5] G. Wolfsfeld, E. Segev, T. Sheafer, Social media and the arab spring: Politics comes first, Int. J. Press/Politics 18 (2) (2013) 115–137.

[6] A. Bodaghi, S. Goliaei, A novel model for rumor spreading on social networks with considering the influence of dissenting opinions, Adv. Complex Syst. 21 (06n07) (2018) 185001.

[7] A. Bodaghi, S. Goliaei, M. Salehi, The number of followings as an influential factor in rumor spreading, Appl. Math. Comput. 357 (2019) 167–184.

[8] V. Osipov, S. Kuleshov, A. Zaytseva, D. Levonevskiy, D. Miloserdov, Neural network forecasting of news feeds, Expert Syst. Appl. 169 (2021) 114521.

[9] E.F. Fama, The behaviour of stock market prices, J. Bus. 38 (1) (1965) 34–105.

[10] E.F. Fama, L. Fisher, M.C. Jensen, R. Roll, The adjustment of stock prices to new information, Internat. Econom. Rev. 10 (1) (1969) 1–21.

[11] A. Blázquez-García, A. Conde, U. Mori, J.A. Lozano, A review on outlier/anomaly detection in time series data, ACM Comput. Surv. 54 (3) (2021) 1–33.

[12] X. Fu, X. Ren, O.J. Mengshoel, X. Wu, Stochastic optimization for market return prediction using financial knowledge graph, in: 2018 IEEE International Conference on Big Knowledge, 2018.

[13] E.F. Fama, Random walks in stock market prices, Financ. Anal. J. 51 (1) (1995) 75–80.

[14] S.A. Ross, Neoclassical Finance Neoclassical Finance, Princeton University Press, 2009.

[15] A. Thakkar, K. Chaudhari, A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions, Expert Syst. Appl. 177 (2021) 114800.

[16] S.W. Lee, H.Y. Kim, Stock market forecasting with super-high dimensional time-series data using convlstm, trend sampling, and specialized data augmentation, Expert Syst. Appl. 161 (2020) 113704.

[17] N. Maqbool, W. Hameed, M. Habib, Impact of political influences on stock returns, Int. J. Multidiscip. Sci. Publ. 1 (1) (2018) 1–6.

[18] A. Zussman, N. Zussman, Assassinations: Evaluating the effectiveness of an Israeli counterterrorism policy using stock market data, J. Econ. Perspect. 20 (2) (2006) 193–206.

[19] G. Spanos, L. Angelis, The impact of information security events to the stock market: A systematic literature review, Comput. Secur. 58 (2016) 216–229.

[20] R.P. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news: The AZFin text system, ACM Trans. Inf. Syst. 27 (2) (2009) 12.

[21] R.J. Shiller, Narrative economics, Amer. Econ. Rev. 107 (4) (2017) 967–1004.

[22] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.

[23] A.W. Lo, The adaptive markets hypothesis: Market efficiency from an evolutionary perspective, J. Portfolio Manag. 30 (5) (2004) 15–29.

[24] R. Raman, R. Aljafari, V. Venkatesh, V. Richardson, Mixed-methods research in the age of analytics, an exemplar leveraging sentiments from news articles to predict firm performance, Int. J. Inf. Manage. 64 (2022) 102451.

[25] O. Altinkilic, R.S. Hansen, On the information role of stock recommendation revisions, J. Account. Econ. 48 (1) (2009) 17–36.

[26] M. El-Haj, P. Rayson, M. Walker, S. Young, V. Simaki, In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse, J. Bus. Finance Account. 46 (3–4) (2019) 265–306.

[27] P.C. Tetlock, Giving content to investor sentiment: The role of media in the stock market, J. Finance 62 (3) (2007) 1139–1168.

[28] J. Surowiecki, The Wisdom of Crowds, Doubleday. Tammet, D., 2004, 2009.

[29] E. Teti, M. Dallocchio, A. Aniasi, The relationship between twitter and stock prices. Evidence from the US technology industry, Technol. Forecast. Soc. Change 149 (2019) 119747.

[30] J. Bollen, H. Mao, A. Pepe, Modeling public mood and emotion: twitter sentiment and socio-economic phenomena, in: ICWSM, Vol. 11, 2011, pp. 450–453.

[31] O. Bustos, A. Pomares-Quimbaya, Stock market movement forecast: A systematic review, Expert Syst. Appl. 156 (2020) 113464.

[32] S. Tilly, G. Livan, Macroeconomic forecasting with statistically validated knowledge graphs, Expert Syst. Appl. 177 (2021) 114800.

[33] Q. Zhang, C. Qin, Y. Zhang, F. Bao, C. Zhang, P. Liu, Transformer-based attention network for stock movement prediction, Expert Syst. Appl. 202 (2022) 117239.

[34] J. Durbin, Estimation of parameters in time-series regression models, Retriev. J. Royal Stat. Soc. Ser. B (Methodological) 22 (1) (1960) 139–153, http://www.jstor.org/stable/2983884.

[35] A. Bodaghi, J. Oliveira, J.J.H. Zhu, The rumor categorizer: An open-source software for analyzing rumor posts on Twitter, Softw. Impacts 12 (2022c) 100232.

[36] A. Bodaghi, J. Oliveira, J.J.H. Zhu, The fake news graph analyzer: An open-source software for characterizing spreaders in large diffusion graphs, Softw. Impacts 10 (2021) 100182.

[37] D. Buono, G. Kapetanios, M. Marcellino, G.L. Mazzi, F. Papailias, Evaluation of nowcasting/flash estimation based on a big set of indicators, 2018.

[38] M. Elshendy, A.F. Colladon, E. Battistoni, P.A. Gloor, Using four different online media sources to forecast the crude oil price, J. Inf. Sci. 44 (3) (2018) 408–421.

[39] S. Boubaker, Z. Liu, L. Zhai, Big data, news diversity and financial market crash, Technol. Forecast. Soc. Change 168 (2021) 120755.

[40] I. Shaikh, T.L.D. Huynh, Does disease outbreak news impact equity, commodity and foreign exchange market? Investors' fear of the pandemic COVID-19, J. Econ. Stud. (2021).

[41] J. Engelberg, R.D. McLean, J. Pontiff, Anomalies and news 2017, J. Finance (2017) Forthcoming, 6th Miami Behavioral Finance Conference.

[42] R. Tao, C. Brooks, A.R. Bell, When is a MAX not the MAX? How news resolves information uncertainty, J. Empir. Finance 57 (2020) 33–51.

[43] W. Chen, C.K. Yeo, C.T. Lau, B.S. Lee, Leveraging social media news to predict stock index movement using RNN-boost, Data Knowl. Eng. 118 (2018) 14–24, http://dx.doi.org/10.1016/j.datak.2018.08.003.

[44] H. Qiu, Y. Liu, N.A. Subrahmanya, et al., Granger causality for time-series anomaly detection, in: 2012 IEEE 12th International Conference on Data Mining (ICDM), IEEE, 2012, pp. 1074–1079.

[45] K. Aditya, H. Kolla, W.P. Kegelmeyer, T.M. Shead, J. Ling, W.L. Davis, Anomaly detection in scientific data using joint statistical moments, J. Comput. Phys. 387 (2019) 522–538, http://dx.doi.org/10.1016/j.jcp.2019.03.003.

[46] M.A. ElBannan, On the prediction of financial distress in emerging markets: What matters more? Empirical evidence from arab spring countries, Emerg. Mark. Rev 47 (2021) 100806.

[47] X.X. Yin, Y. Miao, Y. Zhang, Time series based data explorer and stream analysis for anomaly prediction, Wirel. Commun. Mobile Comput. (2022) 5885904, http://dx.doi.org/10.1155/2022/5885904.

[48] R. Langone, A. Cuzzocrea, N. Skantzos, Interpretable anomaly prediction: Predicting anomalous behavior in industry 4.0 settings via regularized logistic regression tools, Data Knowl. Eng. 130 (2020) 101850.

[49] J. Wang, J. Liu, J. Pu, Q. Yang, Z. Miao, J. Gao, Y. Song, An anomaly prediction framework for financial IT systems using hybrid machine learning methods, J. Ambient Intell. Humaniz. Comput. 14 (2019) http://dx.doi.org/10.1007/s12652-019-01645-z.

[50] K. Yuan, G. Liu, J. Wu, H. Xiong, Dancing with trump in the stock market, ACM Trans. Intell. Syst. Technol. 11 (2020) 1–22.

[51] J.Y. Huang, J.H. Liu, Using social media data mining technology to improve stock price forecast accuracy, J. Forecast. 39 (2020) 104–116.

[52] A.E.O. Carosia, G.P. Coelho, A.E.A. Silva, Analyzing the Brazilian financial market through portuguese sentiment analysis in social media, Appl. Artif. Intell. 34 (2019) 1–19.

[53] X. Zhang, Y. Zhang, S. Wang, Y. Yao, B. Fang, P.S. Yu, Improving stock market prediction via heterogeneous information fusion, Knowl.-Based Syst. 143 (2018) 236–247.

[54] B. Li, K. Chan, C.X. Ou, R. Sun, Discovering public sentiment in social media for predicting stock movement of publicly listed companies, Inf. Syst. 69 (2017) 81–92.

[55] X. Li, P. Wu, W. Wang, Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong, Inf. Process. Manage. 57 (5) (2020b) 102212.

[56] H. Maqsooda, I. Mehmood, M. Maqsood, M. Yasir, S. Afzal, F. Aadil, M.M. Selim, K. Muhammad, A local and global event sentiment based efficient stock exchange forecasting using deep learning, Int. J. Inf. Manage. 50 (2020) 432–451.

[57] H. Hu, J.J.H. Zhu, Social networks, mass media and public opinions, J. Econ. Interact. Coord. 12 (2) (2017) 393–411.

[58] M.S. Albarrak, M. Elnahass, S. Papagiannidis, A. Salama, The effect of twitter dissemination on cost of equity: A big data approach, Int. J. Inf. Manage. 50 (2020) 1–16.

[59] H. Chen, P. De, Y. Hu, B.H. Hwang, Wisdom of crowds: The value of stock opinions transmitted through social media, Rev. Financ. Stud. 27 (5) (2014) 1367–1403.

[60] X.Q. Sun, H.W. Shen, X.Q. Cheng, Trading network predicts stock price, Sci. Rep. 4 (2014) 3711.

[61] Y. Li, Z. Mi, W. Jing, Incorporating textual network improves Chinese stock market analysis, Sci. Rep. 10 (2020a) 20944.

[62] L. Adamic, C. Brunetti, J.H. Harris, A. Kirilenko, Trading networks, Econom. J. 20 (3) (2017) 126–149.

[63] A. Bodaghi, J. Oliveira, The theater of fake news spreading, who plays which role? A study on real graphs of spreading on Twitter, Expert Syst. Appl. 189 (2022b) 116110.

[64] P. Patil, C.S.M. Wu, K. Potika, M. Orang, Stock market prediction using ensemble of graph theory, machine learning and deep learning models, in: ICSIM '20: The 3rd International Conference on Software Engineering and Information Management, 2020.

[65] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Comput. Surv. 41 (3) (2009) 1–55, http://dx.doi.org/10.1145/1541880.1541882.

[66] A.L. Opdahl, B. Tessem, D.-T. Dang-Nguyen, E. Motta, V. Setty, E. Throndsen, A. Tverberg, C. Trattner, Trustworthy journalism through AI, Data Knowl. Eng. 146 (2023) 102182, http://dx.doi.org/10.1016/j.datak.2023.102182.

[67] A. Altanlar, J. Guo, P. Holmes, Do culture, sentiment, and cognitive dissonance explain the 'above suspicion' anomalies? Eur. Financial Manag. Eur. Financial Manag. Assoc. 25 (5) (2019) 1168–1195.

[68] A. Andrikopoulos, C. Wanga, M. Zheng, Is there still a weather anomaly? An investigation of stock and foreign exchange markets, Finance Res. Lett. 30 (2019) 51–59.

[69] S. Lazuarni, M. Asri, Does heuristic behavior leave anomalies in the capital market? J. Indonesian Econ. Bus. 34 (3) (2019) 217–228.

[70] A. Bodaghi, J. Oliveira, The characteristics of rumor spreaders on Twitter: A quantitative analysis on real data, Comput. Commun. 160 (2020) 674–687.

[71] R. Yang, L. Yu, Y. Zhao, H. Yu, G. Xu, Y. Wu, Z. Liu, Big data analytics for financial market volatility forecast based on support vector machine, Int. J. Inf. Manage. 50 (2020) 452–462.