information

Article

From Text Representation to Financial Market Prediction: A Literature Review

Saeede Anbaee Farimani [1], Majid Vafaei Jahan [1,*] and Amin Milani Fard [2]

[1] Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad 91871-47578, Iran
[2] Department of Computer Science, New York Institute of Technology, Vancouver, BC V5M 4X3, Canada
* Correspondence: vafaeijahan@mshdiau.ac.ir

Abstract: News dissemination in social media causes fluctuations in financial markets. (Scope) Recent advanced methods in deep learning-based natural language processing have shown promising results in financial market analysis. However, understanding how to leverage large amounts of textual data alongside financial market information is important for the investors' behavior analysis. In this study, we review over 150 publications in the field of behavioral finance that jointly investigated natural language processing (NLP) approaches and a market data analysis for financial decision support. This work differs from other reviews by focusing on applied publications in computer science and artificial intelligence that contributed to a heterogeneous information fusion for the investors' behavior analysis. (Goal) We study various text representation methods, sentiment analysis, and information retrieval methods from heterogeneous data sources. (Findings) We present current and future research directions in text mining and deep learning for correlation analysis, forecasting, and recommendation systems in financial markets, such as stocks, cryptocurrencies, and Forex (Foreign Exchange Market).

Keywords: text mining; sentiment analysis; financial market prediction; big data analytics; news; social media

Citation: Anbaee, S.A.; Jahan, M.V.; Milani Fard, A. From Text Representation to Financial Market Prediction: A Literature Review. *Information* **2022**, *13*, 466. https://doi.org/10.3390/info13100466

Academic Editor: Diego Reforgiato Recupero

Received: 14 August 2022
Accepted: 20 September 2022
Published: 29 September 2022

1. Introduction

The efficient market hypothesis [1], introduced by Fama in 1965, showed that financial time series are influenced by the information available to investors. This motivated behavioral economics [1,2], which studies the psychological behavior of investors and the role of social, cultural, and emotional factors in their decision making to justify market anomalies [3]. Most investors are influenced by news of political, economic, social, or emotional events that are posted on social media. Thus, information on web newsgroups, social networks, and stock chat boards are examples of important sources of information that can be used toward better business decision making. In the past decades, the field of behavioral economics was motivated by technical analysis, and a great deal of research was conducted recently through the fundamental analysis of unstructured textual data using embedding techniques [4–6].

In a technical analysis, indices such as World Bank reports on the GDP, analyzed by human experts, and also the technical analysis of market indicators are used [7–9], while in a fundamental analysis, text mining methods are applied to identify important events that influence investors and cause market fluctuations. To understand fundamental subjects that affect the market, one needs to represent text according to the contextual information in a document as well as the proximity of the information in news streams that report various aspects of events; however, most works that were published before 2006 only analyzed the market response to simple parameters, such as news counts. In this survey, we study methods for identifying the contextual information published in social media related to financial

markets. Text mining techniques, such as a sentiment analysis [10–13], part of speech tagging (POS) [14,15], text representation, such as transformer-based word embedding [16–22], and machine learning techniques [23–31], have been used in this area after 2006. Recently, researchers have focused on using deep learning-based natural language processing (NLP), such as Bidirectional Encoder Representations from Transformer (BERT) [18,21,32–34] or seq2seq architecture with an attention mechanism [20,35–38], to structure textual web data. BERT-contextualized word embedding, announced by Google in 2018, is used as a word sense disambiguation technique for summarizing and selecting important news for investors' behavior analysis [21,25]. Recent works have investigated the use of an attention mechanism in a deep encoder–decoder network that applies transfer learning from pre-trained models or calculating the impact of each news feed as well as assigning heavy weights for important time intervals in a time series [19,39,40]. Other recent research is on the implementation of graph embedding and knowledge graph mining with deep learning-based predictive models [20,24,41–43].

We study the literature that was published in the intersection of behavioral economics and the NLP field and analyzed various data, such as market indicators and textual sources. We review the ways of structuring text and the retrieval of attractive patterns that are extracted from unstructured text or market time-series data using deep feature fusion strategies and machine learning approaches. Figure 1 depicts our survey organization. We review the usage of the text-structuring approach from Bag of Words (BoW) as a traditional text representation method to contextualize transformer-based text representation in a deep neural network. We then study the sentiment analysis, graph mining, and deep feature fusion methods for information retrieval. Finally, we explore studies that present new ways in one of the three areas of forecasting models, investor behavior analysis techniques, or trading strategy recommendation systems that jointly use financial data and media information.

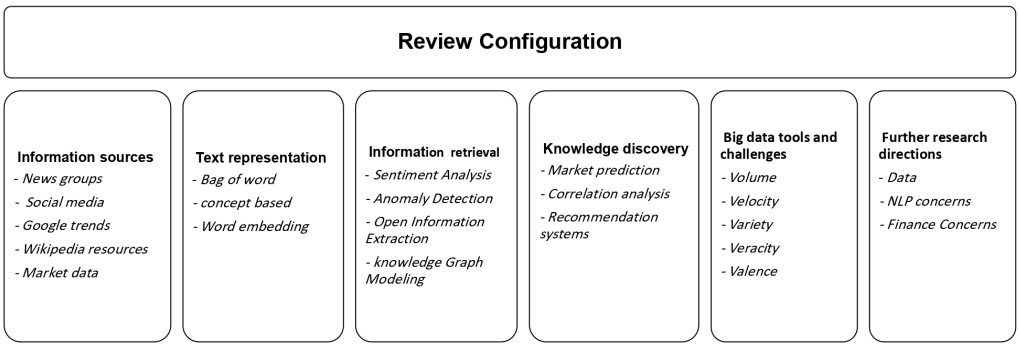

Figure 1. Organization of our review process.

Most of the reviewed works survey fundamental or technical analysis methods [44–48]. We review the literature that processed various heterogeneous data sources for financial decision support by exploring prestigious journals and top-rank conferences to find representative works during the past 13 years that cover the intersection of fundamental and technical analysis approaches. A distinctive aspect of our review is covering the literature from three mainstream viewpoints, including predictive models, correlation analysis methods, and recommendations for trading strategies. Another aspect of our work is studying big data challenges, such as the veracity of textual information and the impact of fake news on the stock market or the valence of groups of investor's behavior, and introducing tools for the huge amount of real-time data storage and processing.

Previous review papers, such as [44,47], present some hypotheses and earlier efforts in the domain. Compared to Li et al. [45], published in 2017, we include other aspects, such as covering the recent deep learning-based NLP approaches. Unlike [46,49], we not only cover a sentiment analysis but also other information retrieval approaches from the text. Compared to [48,50–52], we only focus on interactions between web media and the stock market. The authors in [53,54] review the literature from an information science perspective, while we review the technical characteristics of predictive models or trading strategy recommendations methods.

Our main contributions in this survey are as follows:

- We provide state-of-the-art use cases of fundamental analysis studies in financial market prediction, trading strategy recommendation, and correlation analysis, which distinguishes this study from the existing surveys.
- We systematically review the structuring heterogeneous data sources, such as knowledge graph mining or tensor decomposition techniques, and discuss the lag analysis or significance indication methods, which are often neglected in other surveys.
- We organize our review into four main categories of heterogeneous data sources, text structuring, analysis of information, and knowledge discovery methods.
- We present the future research directions in all of these categories.
- We discuss various big data aspects, such as variety, veracity, volume, valence, and velocity, as well as the tools and challenges in this field.

Outline: In Section 2, we discuss the background concepts and overview of the research methodology. Section 3 describes the review organization. Section 4 presents the big data, tools, and challenges. Section 5 presents the future directions of the research, and the conclusion is presented in Section 6.

2. Background and Methodology

In this section, we first explain our review methodology for selecting the studied resources and then present the motivation from the three viewpoints of information sources, textual representation, and financial predictive models.

2.1. Methodology

Figure 2 depicts the distribution of published papers in terms of years and text representation approaches that they investigated. The growing number of papers in recent years shows the importance of applied studies in the intersection of behavioral finance and artificial intelligence. This figure clearly shows the investigation of text mining methods after 2006 in the reviewed literature and the increasing number of embedding techniques used in the literature over the past two years. We explore papers published in prestigious journals and conferences by searching for keywords such as "financial market prediction", "financial sentiment analysis", "news text mining". We select papers that contribute to the intersection of behavioral finance and financial newsgroups and social media analysis. Table 1 depicts the most cited related works published in journals and conferences. The citation statistics were extracted from Google Scholar as of 10 September 2022. Most journals do not have commercial or advertiser relationships, while some conferences are sponsored by financial companies and stakeholders. The *Expert Systems and Applications*, with 14 publications, has the greatest number of related publications among other sources, and the *Association for Computational Linguistics (ACL)* conference with 4 publications has the most papers among other conferences. Bollen et al. [55], who proposed stock market mood estimation, from Twitter, have the most cited work since 2010.

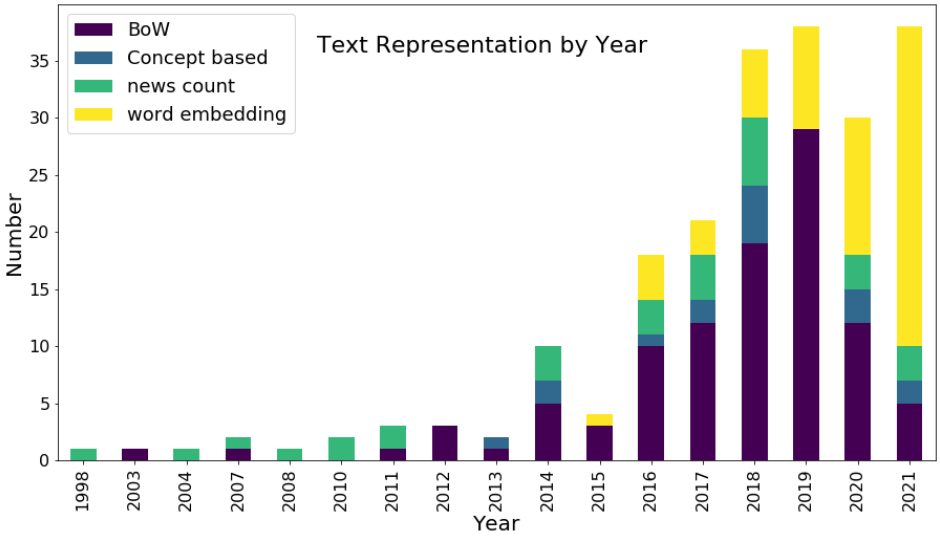

Figure 2. Distribution of reviewed works based on the text representation methods.

Table 1. Information sources in terms of citations.

Source	Papers	No. of Papers	Citation
Journal of Finance	[56–58]	3	6137
Journal of Computational Science	[55]	1	2801
Expert Systems with Applications	[9,22,47,48,52,59–66]	14	1229
Decision Support Systems	[24,67–74]	9	785
Information Processing & Management	[28,75–78]	5	406
Knowledge-Based Systems	[12,23,79–82]	7	162
Neurocomputing	[17,83–85]	4	158
The Journal of Finance and Data Science	[86,87]	2	111
International Review of Financial Analysis	[3,88]	2	100
International Journal of Data Science and Analytic	[50,51,89,90]	4	28
Neural Computing and Applications	[13,91,92]	3	99
Applied Soft Computing	[20,93,94]	2	77
Physica A: Statistical Mechanics and its Applications	[95,96]	3	50
IEEE Transactions	[45,97,98]	3	45
AAAI Conference on Artificial Intelligence	[99,100]	2	1080
Empirical Methods in Natural Language Processing (EMNLP)	[101,102]	2	320
Association for Computational Linguistics (ACL)	[21,25,103,104]	4	232

2.2. Information Sources

Applied studies that we reviewed in this work mainly used various information sources, including market- and media-based sources. Figure 3 depicts the distribution of market based on corresponding countries. Refs. [10,88,91,105] examined the predictability of US stock market, including Dow Jones, NYSE, and S&P 500 stocks. Refs. [15,21,60,95,96,106] have studied the predictability of the Foreign Stock Exchange (Forex) and cryptocurrencies market, where a currency is traded based on the ratio of two currency pairs, such as the EUR/USD, UDS/JPY, or BTC/USD. News [13,86,89,107]; social network data, such as Twitter [55,108,109], Stocktweet [88,108,110], or Sina Weibo [111]; as well as trends in search

engines [95,112]; referring to Wikipedia page statistics [59]; and online reviews of customers about firm's products [113] are the types of media-based sources available to investors.

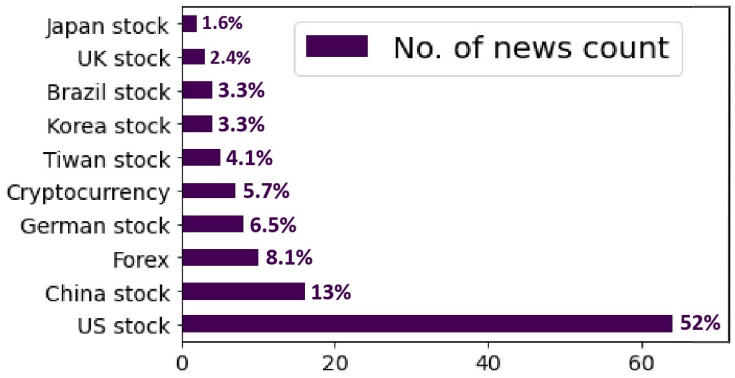

Figure 3. Distribution of reviewed works with respect to the analyzed market. Stock markets are mentioned with country names.

Thomson Reuters and Bloomberg newsgroup are the mainstream news resources in this category; however, it is a basic challenge to utilize relevant texts to the target market because of redundancy and noise in the associated texts. Most of the review papers have manually selected the target market-related news based on manually predefined rules [15,60,114]; however, there are some researchers that proposed a method for selecting the stock-related news, including summarizing the news, measuring the news information gain analysis, and knowledge graph techniques [21,39,67,115]. Farimani et al. [12,16] explored 12 specialized newsgroups that specifically publish for Forex, cryptocurrencies, and commodities and select relevant news to the target market based on judgments of specialized authors. They made public RESTFul APIs (https://robonews.robofa.cscloud.ir/Robonews/v1/, accessed on 2 December 2021) and *MarketNews* (https://figshare.com/articles/dataset/MarketData_for_MarketPredict_RESTFul_API_including_News_and_Market_Data/14754966, accessed on 21 September 2018) datasets so that the research community can access these news sources. For example, Figure 4 presents the number of news indicated for Bitcoin via 25 specialized authors among 592 authors in Cryptocurrency category of *MarketNews* dataset.

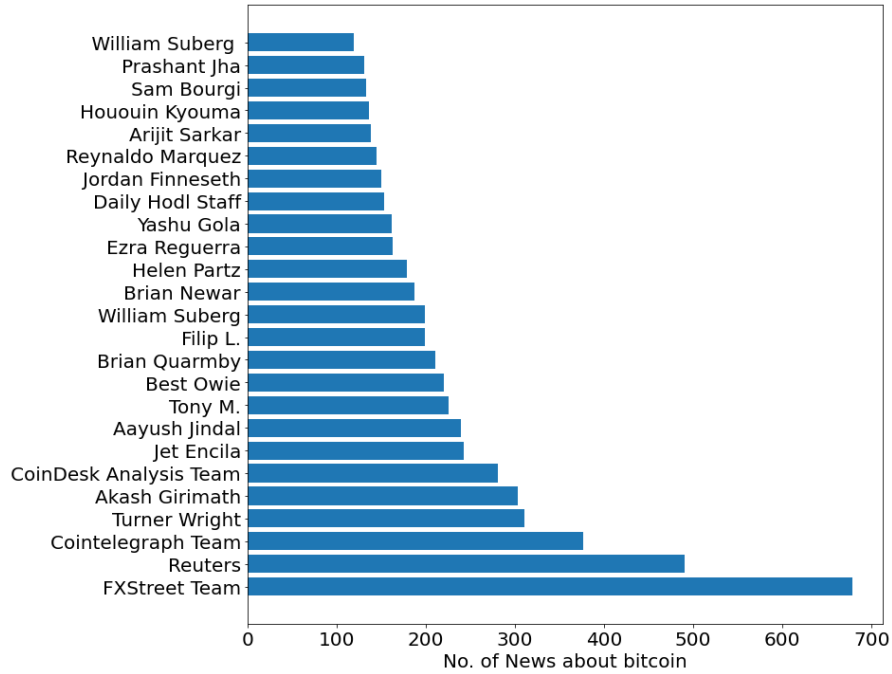

Figure 4. Number of news related to Bitcoin in *MarketNews* dataset in terms of authors.

2.3. Text Representation

The distributional semantic hypothesis [116] states that words that occur in similar contexts tend to have similar meanings, and words with similar distributional properties often have similar meanings. Salton presented a vector space model (VSM) for representing text documents in vector space called Bag of Words based on word occurrences in a text corpus [117]. This method suffers from the curse of dimensionality and sparse representation of TFIDF matrix with the growth of the vocabulary size. Latent Semantic Analysis (LSA) [118] decomposes the singular values of the TFIDF matrix for constructing some latent concepts and reducing the dimension size of word vector representation in a linear space. Feuerriegel and Gordon used LSA for stochastic concept modeling in the process of forecasting macroeconomic indices [119]. Another dimension reduction approach is Latent Dirichlet Allocation (LDA) which factorizes the TFIDF matrix with 3 levels Bayesian approximation [120,121]. Atkin et al. used LDA to extract some topics and each news document is vectorized through these topics for sentiment analysis in financial news [86]. However, with the growth of vocabulary size, it is relatively computationally time expensive.

Recently, neural network-based word embedding techniques have emerged. The word2vec method proposes a word vector representation that reflects both syntactic and semantics roles of words [5]. The main idea of word2vec is to predict a word vector based on its surrounding words instead of directly capturing co-occurrence counts [107]. Predictive methods, such as [18,36,122], consider the semantic relationships between words and use the word embedding technique [5] with long short-term memory predictive model to forecast S&P500 stock market. Lutz et al. [123] used Doc2vec embedding technique [4] for predicting the sentence-level sentiment for developing their financial decision support system. Devlin and his colleagues at Google fine-tuned Bidirectional Encoder Representations from Transformers (BERT) [32] with bidirectional pre-training model for language representations. Ref. [21] proposed a news summarization method for Forex market prediction based on the BERT [32]. They summarized news with SOTA extractive summarization method proposed in [115] and selected news based on the attention score of news headlines to some embedded financial indices. By using BERT, their method captures the deep semantic information effectively. Authors in [124] used BERT for calculating sentiment index of the market using news published in Chinese social media.

2.4. Predictive Models

A broad variety of predictive models have analyzed diverse portfolios [125–128] with an average of 5 years duration. Figure 5 depicts the duration of analysis in reviewed articles in the year. For example, bar '0–3' shows the number of methods with less than three years of duration of the analysis stacked based on the trading timeframe that the scholars focused on. There are two mainstream predictive methods. One predicts the real value of price based on regression models and the other uses a classifier for predicting the trend direction. Researchers have studied the change in the close price and have examined the different duration of time for analyzing the impact of news on the market for short term (minutes to a day) or long term (weekly to monthly). The latter are those [18,36,79,122] that focused on trend (up/down) prediction and formulated the problem as the binary or multiclass classification that was determined based on the change in close price due to two consequent time frames. In terms of news labeling, a few methods labeled news by considering asymmetric time intervals before and after news publishing. For example, Chen et al. [21] analyzed the influence of input time and prediction time on Forex trading in an asymmetric manner. They labeled news based on the change in close price within t1 minutes before news input timestamp and t2 minutes after news announcement.

Different methods have adopted different data types. Some works on predicting stock markets only use text information from daily news [19,39,61,99,101], while others [18,25,34,124] consider a combination of textual and market-based features. Traditional models split data into train and test in chronological order for model training, while in the recursive model, training in rolling window strategy is used. In the rolling window

training strategy, dynamic updating of the model parameters is investigated with respect to structural change in market, whereas in traditional train and test split, the training process is static. In some cases [23,28,61,91,129], cross-validation is used for evaluating the accuracy of prediction. Figure 6 shows the distribution of predictive models in the reviewed methods by publication year categorized into 4 groups of deep learning, machine learning, price regression, and statistical analysis of inter-dependencies between media and price fluctuations.

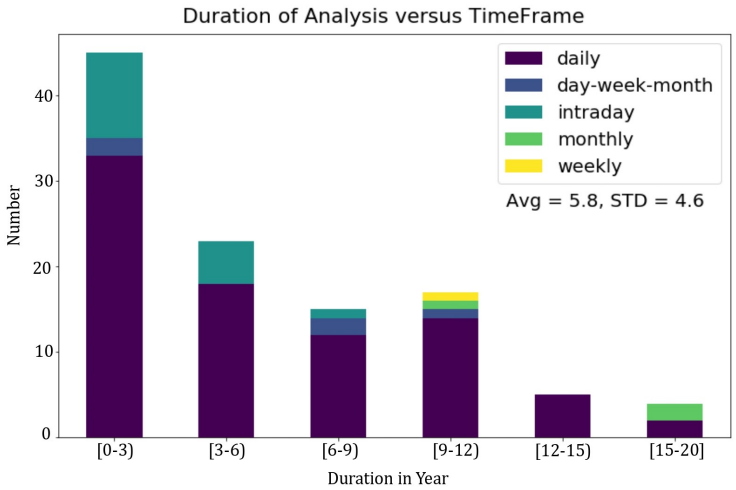

Figure 5. Histogram plot of duration of analysis in the surveyed papers versus time frames.

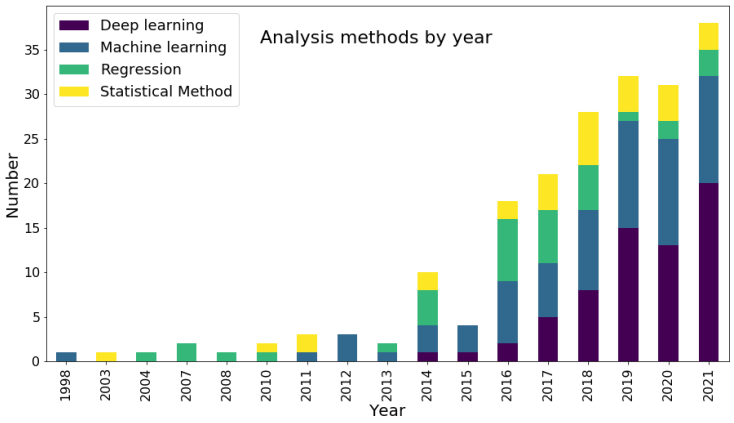

Figure 6. Distribution of predictive models used in the reviewed methods by publication year.

3. Review Configuration

We explore the prominent methods that evaluate some hypotheses related to behavioral economies via a sentiment analysis or text-based features. In Section 3.1, we study text representation methods, including Bag of Words, concepts-based approaches, and word embedding-based techniques. Then, in Section 3.2, we discuss in detail the sentiment analysis, the popular lexicons in the finance domain, the application of the sentiment analysis in the reviewed methods, and finally, focus on OpenIE and knowledge graph-based methods for real-world event detection. In Section 3.3, we address the popular taxonomy of the traits of market analysis models and some graceful issues for evaluating the hypothesis, predictive models, and trading strategies proposed in the literature.

3.1. Textual Representation

In this section, we investigate the main methods in the literature that contributed to the text representation step of the financial analysis, from earlier efforts by Wuthrich et al. [9] in *1998* to the GPT3 [130] presentation in *2020* as a powerful text-structuring method.

3.1.1. Bag of Words

The first leading studies, such as [56,57,131,132], provide regression models based on parameters, such as the amount of news, to examine the rate of the stock return. With the development of text mining techniques and the increasing volume of textual data on the Internet, various methods have been employed to organize documents based on term frequencies by the Bag-of-Words (BoW) matrix. The BoW approach suffers from the existence of noisy and rare terms as issues of scalability, especially in the case of growing words in a dictionary, and challenges such as high dimensionality or the sparse matrix problem that reduce the accuracy of the sentiment analysis. The scholars in [60,68–70] used feature selection methods or proposed term weighting approaches to increase the importance of the word which is unevenly distributed between the up and down classes. In all of the methods mentioned, the feature selection process will be effective as long as we do not face the big data challenge. However, as the use of the Internet grows, investors are facing a huge amount of data.

3.1.2. Concepts-Based Approaches

The Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) are two matrix factorization techniques that can also be used for dimensionality reduction. These methods produce some interpretable latent variables in low dimensions. The main purpose of these methods is to reduce the dimensions of the problem features to overcome the sparse matrix problem in the vocabulary integration model. The proposed methods in [86,91,105,125] extracted some topics from textual news with LDA to validate the hypothesis that news is a powerful proxy for market prediction. Ref. [119] proposed a latent variable extraction method, which develops a semantic path model, together with an estimation technique based on regularization in order to yield the full interpretability of the forecasts. Although the concepts in an LSA generally capture a better representation compared to the BoW, they fail to incorporate the nonlinear semantics between the words due to linear space matrix decomposition [83]. Moreover, the LDA distribution estimation has a limitation of time complexity.

3.1.3. Word Embedding

The word embedding technique [5] used a neural network to produce a dense vector for each word based on its occurrences within a contextual window of words. Recently, scholars [18,22,79,122,129,133] have used the word2vec [5] or Doc2vec [4] representation of news headlines and content in their financial decision support systems; however, transformer-based language models [6,130] have shown better results in reflecting deep semantic and syntactic news information compared to traditional word embedding in the financial domain. Liu, in [35], used long-term and short-term event embedding methods which contain the stack ELMO embedding of the t-days set of news headlines for the prediction of the S&P500 index. The authors of [21,25,35] used BERT-contextualized word embedding representation for market prediction. Farimani et al. [16] proposed context-aware conceptual document representation to model the relevance between the news based on all the information in financial news titles and bodies via the clustering of contextualized BERT word embedding. The BERT-BoEC method in [16] leveraged the conceptual relationship in the news that outperforms other concept-based approaches as well as other embedding techniques, such as BERT-based [cls] embedding.

3.1.4. Discussion

Introducing word embedding in 2013 by Mikolov [5] has led to a significant progress in NLP, especially for generating new language models, such as ELMO, BERT, GPT2, and GPT3. The main challenging task for representing a word in a vector space is the ability to reflect the proximity information of the semantic and syntactic roles, in which the transformer-based models can effectively capture the deep semantic and syntactic roles of words by presenting the contextualized word embedding technique. Applying

transfer learning from pre-train models, such as ELMO or BERT, in the financial domain is still a challenging problem. The need for semantic similarity modeling between the news documents that report various aspects of an important event remains debatable in the financial domain. Table 2 shows a comparison of methods based on text representation techniques and describes the data and analysis modes.

Table 2. Comparison of methods based on text representation manner.

		Data Attributes				Analysis		
	Literature	**Data**	**Market**	**Media Source**	**Duration**	**Feature Selection**	**Time Frame**	**Machine Learning**
BOW	[69]	news	S&P500	Yahoo Finance	October 2005–November 2005	POS	20 min	SVR
	[70]	news	DGAP	DGAP and EuroAdhoc	1997–2011	two-word combination	Intraday	SVM
	[60]	news	Forex	MarketWatch.com	2008–2011	Ontology	2 h	SVM
	[15]	news	Forex	MarketWatch.com	2008–2011	POS	Intraday	SVM
	[91]	news	NSYE and NASDAQ	MarketWatch.com	2013	BOW	Daily	NN
Concept Based	[125]	social media	2008 America bank crisis	Yahoo! Finance	January 2008–December 2008	LDA	20 min	Regression
	[134]	News	CSI100	Hexun	January 2015–December 2015	LDA	Daily	Naive Bayes
	[86]	News	NASDAQ	Reuters	September 2011–September 2012	LDA	Minutes	Naive Bayes
	[105]	News	Forex	Reuters	2012–2016	LDA	Daily	MLP
	[119]	News	CDAX	Website of the EQS Group	July 1996–April 2006	Latent Variable	monthly	Lasso Regression
Word Embedding	[122]	News	S&P500	Thomson Reuters, Bloomberg	2006–2014	word2vec	Day–week	LSTM
	[18]	News	S&P500	Reuter	2006–2013	word2vec	Daily	HCAN
	[79]	news, social media	HK and CSI100	Xueqiu and Guba and Sina and Hexun	January 2015–December 2015	word2vec	Daily	MFC
	[133]	News	TWSE	TWSE Official Website	2007–2017	word2vec	Daily	CNN-LSTM
	[22]	News	CDAX	DGAP	January 2001–September 2017	Doc2vec	Daily	LSTM
	[129]	News	Indian stock market	Indian news wires	January 2013–December 2016	pharagraph 2vec	Daily	LSTM
	[35]	News	S&P 500	Reuters and Bloomberg	October 2006–October 2013	Word embedding	Daily	LSTM
	[21]	News	Forex	Reuters	2013–2017	BERT Word embedding	Daily	LSTM
	[25]	News	S&P500	Google Trends	January 2004–December 2015	BERT Word embedding	Weekly	NN

3.2. Information Retrieval

We study the representative methods in the two categories of sentiment analysis and open information extraction for organizing information. For the sentiment analysis, we focus on the adopted methods for in-domain or cross-domain sentiment analysis processes for the finance-specific domain, and for the open information extraction, we investigate the challenge of finding structural dependencies between textual resources that report different aspects of real-world events with knowledge graph mining.

Sentiment Analysis. There have been two lines of work predicting stock markets using a sentiment analysis. In the first one, researchers extract the sentiment (positive, negative, or neutral) from documents as features [23,79], and in the other line of work, the public mood time series is calculated by aggregating a sentiment score for each time interval [25,71,108]. From the market perspective, there are two categories, coarse grained and fine grained, in the domain-specific usage of the financial sentiment analysis (FSA). The fine-grain sentiment analysis embraces tasks on a sub-sentence level related to a specific firm, while the coarse-grained SA often analyzes the whole document or a sentence-level

analysis. From an NLP viewpoint, the types of approaches applied for the sentiment analysis can be divided into two categories, lexicon-based and machine learning-based approaches. The lexicon-based methods analyze the text sentiment based on the high quality of the emotion dictionary and word polarity, while machine learning approaches rely on supervised learning and word features. In lexicon-based approaches, a domain-specific or domain adaptation lexicon is constructed based on pre-defined manual rules. Lexicons such as Sentiwordnet [135] and Loughran–McDonald [136] are common domain adaptation and finance-specific lexicons in the financial sentiment analysis, respectively. However, studies such as [11,72] presented a method for constructing a financial domain-specific lexicon with machine learning-based techniques. In machine learning approaches, the supervised methods are investigated for a sentiment analysis, and then, based on the document-, sentence-, or aspect-level sentiment, the in-domain classifier is trained to be used in the sentiment analysis phase. However, with some pre-trained models, more embedding methods, and significant improvements in attention mechanisms, deep learning methods are growing in the FSA [75,80–82,137,138].

Lexicon-based methods. Most of the lexicon-based methods in the FSA [28,60,69,139,140] focus on statistically evaluating the correlation between different moods extracted from lexicon-based SA and the fluctuations in target market indices. Most domain-specific lexicons, such as Loughran–McDonald [136] and RavenPack [140], are usually provided by human expert intervention, while the process of manually labeling words' sense is a time-consuming process. Challenges, such as defining the context-aware polarity of a certain word, especially in the financial domain, still exist for an accurate lexicon-based sentiment analysis. Issues such as existing Out-of-Vocabulary Words (OOV), especially in social media texts, cause more frequent updates in lexicons and a poor performance in financial prediction.

Machine Learning-based Methods. In this line of work, a classifier, such as naive Bayes, SVM, or Random Forest, are used to determine the sentiment of features extracted from a textual document [141,142]. There are two strategies for training-set labeling. The first one is training classifiers based on a seed of manually labeled features and the other that usually labels based on the alignment of the lag correlation between headlines' release time and the rate of the stock return. For example, if the stock returns at time t are lower than the time $t - \Delta$, then the news released at time t will be labeled negative. These methods [60,90] often suffer from the problem of market endogeneity [143]. In [11,26], a piece of news is considered good/bad if the market volatility before and after the news publishing is positive/negative, while [128] shows investors overreact to bad news in a short time even in a bullish market and under-react to good news in bad times; thus, the classification is endogenous to the market volatility.

Deep learning-based methods. The main difference between deep learning-based methods and the traditional machine learning models is the better performance of such methods using more data. General domain transformer models, such as BERT, have been used for a sentiment analysis in the literature [16,34,35,90]. In [124], the authors fine-tuned an uncased version of the BERT model for an FSA based on manually labeled firm-specific data from Weibo and then measured the time series of emotions from BERT. This work tried to use a cross-domain sentiment analysis based on a pre-trained BERT model and applied transfer learning for the sentiment analysis through financial news documents. The result indicated that the BERT-based SA outperforms other embedding techniques, such as FastText, Bidirectional Long Short-Term Memory, and Multichannel CNN. The cross-domain sentiment analysis refers to investigating a general domain pre-trained model for finance sentiment analysis problems. Araci et al. fine-tuned BERT for an FSA based on the human expert-labeled TRC2 dataset [144]. The scholars in [12,145] showed the superiority of the FinBERT model [144] as a fine-tuned version of BERT for an FSA.

3.2.1. Discussion about Sentiment Analysis

The empirical findings in the literature show that news and social media messages exhibit a persistent predictive power on financial market movement [18,55,85]. Twitter posts are very short messages less than 140 characters, averaging 11 words per message, and most of the tweets have simple meaning and complicated sentiment structure due to one or more aspects, while news stories contain both a news headline and article body which are longer than tweets. In addition to the linguistic differences between them, there are also differences in the timing behavior of their impact. Tweets are faster than news in exhibiting new market information, whereas news is broadly considered a more trustworthy source of information than tweets. The rumor-essence word-of-mouth statements that investors post on social media may cause earlier fluctuations in the market, while the news has a persistent effect. The ability of a transformer-based sentiment analysis, especially in the recently proposed finer-grained sentiment analysis approaches [63,76,78,146,147], can help scholars to analyze the different aspects of news and social media posts on investors' trading behaviors. Table 3 demonstrates a comparison between news and social media sentiment analysis methods.

Table 3. Comparison of news and social media sentiment analysis methods.

Category			Data Attributes				Analysis		
		Literature	Media Source	Market	Duration	Feature Selection	Time Frame	Machine Learning
News	Lexicon Based	[62]	Belgian financial newspaper *De Tijd*	Dutch Company	May 2012	BoW	Daily	SVM
		[72]	Twitter		December 2012–October 2015	LSA	Daily	Regression
		[58]	*Wall Street Journal*	DJIA	1935–1961	News count	Daily	Regression
	Machine Learning	[23]	Sina Weibo	Shanghai Stock SSEC	December 2014–April 2016	BoW	Intra- day	Logistic regression
		[148]	Reuters and Moneycontrol	NIFTY	Five month	POS	Daily	SVM
		[14]	LexisNexis	NIFTY		BoW	Monthly	ARM
	Deep Learning	[142]	Yahoo Finance	NASDAQ-100	2008–2018	BoW	Daily	MLP
		[22]	Deutsche Gesellschaft (DGAP)	CDAX	January 2001–September 2017	Doc2vec	Daily	LSTM
		[12]	Fxstreet, NewsBTC, Cointelegraph	Forex, Cryptocurrency	October 2018–July 2021	FinBERT	Hourly	LSTM
Social Media	Lexicon Based	[149]	Indian Financial	NIFTY	January 2009–December 2009	BoW	hourly	SVM
		[90]	Yahoo Finance, Reuters	Tokyo Stock Exchange	September 2015 and January 2007–December 2016	Word embedding	Daily	MLP
		[79]	Sina, Hexuan websites	HK, CSI100	January 2015–December 2015	Word embedding	Daily	Tensor decomposition
		[139]	twitter	NYSE	October 2011–March 2012	BoW	Daily	ARM
	Machine Learning	[55]	Twitter	Dow Jones	February 2008–December 2008	News count	Daily	Fuzzy neural network
		[110]	Reuters and Forbes	Amazon, Google	2017	BoW	Daily	Lasso regression, SVR
		[141]	Baidu news	Shanghai 50ETF	2008–2015	BoW	Daily	Naive Bayes, LSTM
	Deep Learning	[124]	Tencent, Ping An, CCB, Weibo	Hong Kong Market	January 2016–December2018	Word embedding	Daily	LSTM, VAR

3.2.2. Event Detection

The mainstream works [10,150] identify important events based on sudden changes in factors, such as the volume of the tweet, the sentiment score, word frequency, or distinct pattern in the occurrence of the words being reported, using the likelihood and point-wise mutual information. The other category of works apply knowledge graph mining approaches [24,39,42,61] and open information extraction (OpenIE) [19,40,99].

Knowledge Graph Modeling. Various knowledge graph-based techniques have been used in the literature to determine the structural relations between news items for a fundamental market analysis. Ref. [102] proposed a Semantic Stock Network (SSN) based on the stock-related cash-tag on Twitter and use it to predict stock based on the summarization of the latent semantics of stocks from social discussions. In their semantic stock graph, stocks are nodes and the frequency of multiple occurrences of two stocks in tweets is the weight of edges. Ref. [61] proposed a graph construction method to leverage both the bi-gram features from news content words and the information structures hiding in the relations between them for predicting the market direction. However, this method leverages a similar subjective relation between news items, and it suffers from the sparsity problem with respect to extracted bi-gram features. The GRU-based attention mechanism in [39] modeled a weighted knowledge graph based on the attention score between similar word embedding representations of news disclosures and corresponding movements in the target market. Ren et al. [24] presented a news recommendation system based on a target company and interested users using a knowledge graph that includes users, news, companies, concepts, and industry categories. The node2vec representation of this knowledge graph is used for predicting the relations between interested users, companies, and related news. Ref. [42] proposed a multimodality graph neural network (MAGNN) to learn from these multimodal inputs for financial time-series prediction, which facilitate the way of handling heterogeneity.

Open Information Extraction. OpenIE aims at extracting structural triples in the form of (subject, relation, object) from textual resources. In [101], entity-relation extraction in the form (Actor, Action, Object) was applied over web-scale data for an event based S&P 500 prediction. Ding [99] employed the event embedding as input features of a deep learning network. They proposed tensor embedding techniques and then applied a deep convolution neural network that processes an event embedding sequence in chronological order for combining the long-term event and short-term event on stock price movements. Following [99], Ding, in [19], used the YAGO knowledge base for incorporating a knowledge graph in the learning process of event embedding. The result indicated the incorporation of knowledge graph-based information, and OpenIE-based structured events to encode background knowledge help with a better prediction of stock movement. Deng [40], after the extraction of events in the form of tuples, constructed a knowledge graph based on Freebase and Wikipedia, and then used graph embedding for the event representation and estimation of the power of an event to move the market. In this article, the author proposed an equation for measuring the impact of events in their predictive models.

Discussion. Table 4 depicts the comparison of the above-mentioned methods. While there has been substantial progress in OpenIE for NLP tasks [151,152] and deep learning-based knowledge graph modeling [100], event detection from financial text remains challenging. Important events with long-term effects and minor effects with short-term effects can influence the market differently. The growth of relation extraction methods based on transformer models and knowledge graph embedding [97,103] can help to facilitate the way for extracting important information from heterogeneous sources in the finance domain.

3.3. Knowledge Extraction

In this section, we investigate a text-based market analysis and discuss issues from three viewpoints: predictive models, the statistical analysis of stakeholders' behavior, and recommendations for portfolio selection. Figure 7 depicts the taxonomy of traits researchers considered while designing their market analysis models. Some aspects in Figure 7 were explained in the background section and here we discuss the other characteristics.

Table 4. Comparison of event detection methods.

			Data Attributes			Analysis		
		Literature	Media	Market	Duration	Feature Selection	Time Frame	Machine Learning
Anomalous changes		[10]	Twitter	Dow Jones, S&P500	September 2013–September 2015	Hashtag	Daily	SVR
		[150]	Twitter	Tesco, Booker	January 2017	Hashtag	Minutes	Naive bayes
		[153]	*Irish Farmers Journal*	Irish Beef Market	2005–2015	BOW	Day–week–month	SVM, log-likelihood
OpenIE		[101]	Reuters, Bloomberg	S&P 500	October 2006–November 2013	OpenIE-based tuples	Day–week–month	MLP
		[99]	Reuters, Bloomberg	S&P500	October 2006–November 2013	OpenIE-based tuples	Day–week–month	MLP
		[19]	Reuters, Bloomberg	S&P500	October 2006–November 2013	Event embedding	Daily	SVM
Knowledge Graph		[24]	GF Securities	Airbnb Market	May 2017–May 2018	word2vec	News release time	Bi-LSTM
		[39]	Reuters, Bloomberg	S&P 500	October 2006–November 2013	Word embedding	Daily	Bi-LSTM
		[61]	ifeng.com (Financial China)	SZ002424	September 2012–March 2017	2-gram	Daily	Kernel SVM
		[40]	Reddit WorldNews Channel	DJIA	July 2008–January 2016	Price vector and event embedding	Daily	Temporal CNN
		[102]	Twitter	S&P500	January 2013	BoW	Hourly	Vector auto-regression model

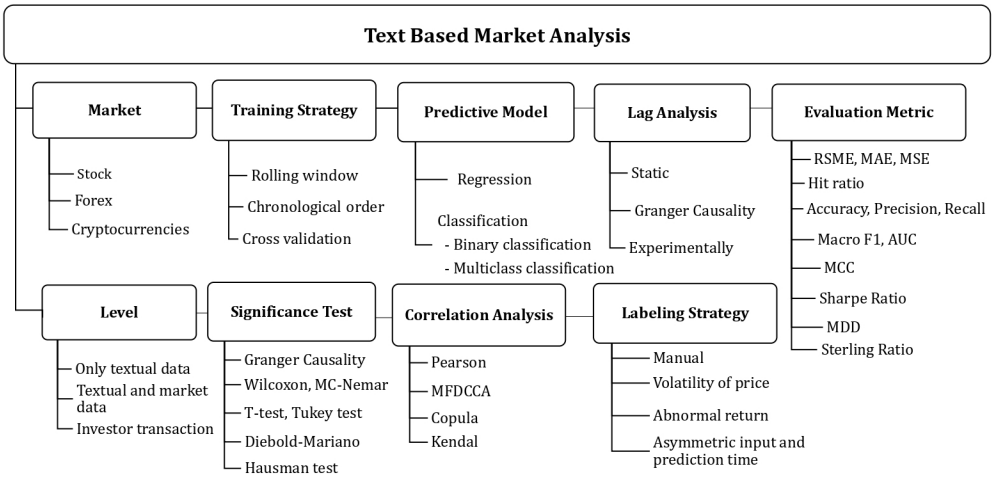

Figure 7. Taxonomy of trait for text-based market analysis models.

Lag analysis. One of the essential characteristics in a time-series analysis is the lag determination. Lag refers to a duration of time between the announcement of new information and the time the stock market adapted itself to a new equilibrium. Some research methods [15,23,92,154] have used a fixed lag size based on related works, while others [21,61,85,104,119] experimentally determine the lag size or use a Granger analysis test [129,142,155]. The Granger causality test [156] is a statistical hypothesis test for determining whether a time series X is useful in forecasting another time series Y. Time series X is considered as impacting Y, if it can be shown that X values provide significant information about future values of Y.

Evaluation Metrics. Accuracy, precision, and recall are three common metrics for evaluating the classifier for *trend prediction* (up/down classes). In the case of an unbalanced dataset due to an up or down trend in the market volatility regime, the weighted Macro F1, Area Under Curve (AUC), Receiver Operating Characteristic (ROC), and Matthews Correlation Coefficient (MCC) are more reliable metrics. The RMSE, MSE, MAE, and hit ratio are traditional metrics for the residual evaluation of *regression-based* predictive models. Some predictive models proposed a trading strategy and evaluated the effectiveness of such models based on profitability and investment metrics, such as the cumulative profit, Sharpe ratio, and Maximum Drawdown (MDD). In Section 3.3.3, we discuss these metrics and methods. Figure 8 depicts the distribution of the evaluation metrics used in publications, which shows about 60 papers focused on the trend prediction evaluation with an accuracy metric. Moreover, most predictive methods consider the trend (up/down) prediction and are less focused on price regression.

Significance test. Various statistical tests have been used in the literature. Most of the research indicated the effect of news information based on the regression-based Granger causality test and used the unit root test or Dickey–Fuller test for satisfying the stationary behavior of the market during the analysis period [55,129,141,157].

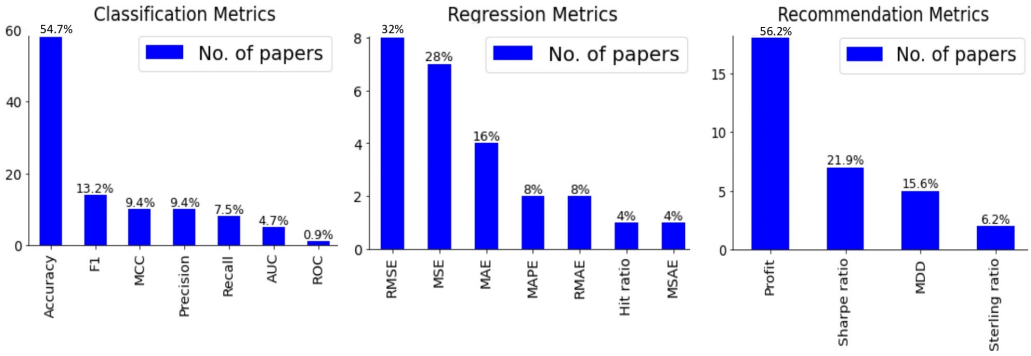

Figure 8. Distribution of evaluation metrics.

3.3.1. Predictive Models Based on Heterogeneous Data Fusion

In this section, we focus on studies about the insider influence of heterogeneous information on investors. Tensor decomposition has been used for handling the heterogeneity of information sources [23,59,79,154]. Tensor-based approaches have shown good results in handling heterogeneous data sources, while the iterative estimation of parameters for each mode of data is a time-consuming process and makes it difficult for models to learn the lead–lag effect of the news on market movements. The concatenation of extracted features from a deep neural network was also used to handle various information sources in market decision support [12,16,21]. Ref. [104] presented a novel deep generative model that jointly exploits text and price signals for assessing continuous latent variables for the superior treatment of stochasticity. Their deep model includes three components, the Market Information Encoder, Variational Movement Decoder, and Attentive Temporal Auxiliary, for the prediction of 80 stocks listed in S&P500. Ref. [17] proposed a hierarchical attention network for handling information fusion from news titles, bodies, and sentiment as well as market based information for stock prediction. Table 5 shows the trait of these reviewed predictive models.

Table 5. Detailed trait of predictive models based on chaotic data.

Literature	Media	Data Attributes Market	Duration	Analysis Feature Selection	Analysis Time Frame	Machine Learning
[92]	Caihua	HIS	2001	BoW	Daily	MLP
[154]	Sina, eastmoney	CSI100	2011	BoW	Minute	High-order tensor regression
[23]	Sina Weibo	Shanghai Stock SSEC	September 2014–April 2016	Tensor decomposition	Intraday	Logistic regression
[21]	Thomson Reuters	Forex	2013–2017	BERT word embedding, MLP-based feature extraction	Intraday	LSTM
[104]	Twitter	NASDAQ	2014–2016	Word embedding, latent variable extraction	Daily	Bi-GRU
[20]	CITIC Securities, GF Securities, China Pingan	Chinese stock	March 2012–July 2018	Graph embedding	Daily	Bi-LSTM

3.3.2. Statistical Analysis of Investors' Behavior

Various cross-correlation analyses have been used for revealing the correlation between investor sentiment and market returns on the non-stationary behavior of the market. Traditional correlation analysis methods, such as Pearson, cannot capture nonlinear relations. Ref. [158] adopted the Kendall correlation for the analysis of the news effect on FBOVESPA, because unlike the Pearson, instead of measuring linear relationships, it measures the monotonic relationship between two variables. Ref. [159] presented a nonlinear dynamic model to study the fractal influence of good and bad news and the actions of various market participants (fundamentalists, chartists, and insiders) on the change in market prices. Nonlinear long-term correlation analysis methods, such as the Multifractal Detrended Cross-Correlation Analysis (MF-DCCA) [160], in [95,161], were used in a rolling window-based strategy and revealed that there is a decreasing trend for the cross-correlations between the change in Google Trends and Bitcoin returns over time. The conditional Copula test was used in [128] to evaluate the nonlinear and asymmetric market effects of economic news that may vary with news intensity. Adopting the Granger causality test for the analysis of a linear causal relationship between the sentiment index and the market is a common task, while [67] used transfer entropy for the analysis of the nonlinear causal relationship between the news and the Korean stock market. Moreover, researchers, such as [162–164], have studied the simultaneous co-movements of the financial markets due to the disclosure of relevant news, macroeconomic announcements, and trending market-related words in Google searches for the short-term analysis. Table 6 depicts the finding that indicates the statistical analysis of the research hypothesis.

Table 6. Findings in behavior analysis methods.

Literature	Sentiment Analysis	Market	Analysis Method	Time Frame	Parameter	Finding
[96]	Yes	Forex	Mathem- atical model	5 min	volatility	Their findings show that the model with multiplicative noise can reproduce the dynamics observed in the real financial market affected by the arrival of high-impact news.
[158]	No	FBOVE- SPA	Kendall cross correlation	15 min	average price, trading volume	Besides demonstrating that sector is indeed influenced contrastingly by news, they also show that the machine learning models utilized performed better than random and other less complex baselines for all sectors in FBOVESPA, in this manner giving proof that news information conveys in reality a significant signal for comprehension of BM and FBOVESPA dynamics.

Table 6. *Cont.*

Literature	Sentiment Analysis	Market	Analysis Method	Time Frame	Parameter	Finding
[126]	Yes	NYSE	Regression	Daily	trading volume, volatility	Their results suggest that significant price discovery related to news stories occurs through institutional trading before the news announcement date.
[128]	No	US	Copulas Statics	Daily	equity returns	The finding shows the market reacts strongly and negatively to the most unfavorable macroeconomic news but appears to largely discount the good news.
[165]	Yes	S&P 500	Correlation	Week	volatility	Their findings show a statistically and economically significant relationship between stale news stories on unemployment and next week's S&P 500 returns. This effect is then completely reversed during the following week.
[95]	No	Bitcoin market	MF-DCCA	Daily	trading volume	By employing the Multifractal Detrended Cross-Correlation Analysis method, they find that the change in Google Trends (CGT) and the Bitcoin market, i.e., returns and changes of volume, is an overall higher degree of multifractal in the long term and weak multifractal in the short term.
[166]	Yes	S&P 500	DCCA-MIDAS	Daily	return	The results show that the composite index of investor sentiment has a significantly positive influence on the long-term stock–bond correlation, and the shock of crises significantly decreases the average correlation, but the effect of sentiment does not change significantly.
[87]	Yes	FTSE 100	Correlation	Daily	close price	Their findings show there is evidence of causation between public sentiment and the stock market movements, in terms of the relationship between MOOD and the daily closing price, and the time-lag findings of MOOD and PRICE.
[163]	No	DAX	Impulse response analysis	5 min	volatility	Show that 50 percent of the total accumulated impact of US macroeconomic news on the DAX 30 and CAC 40 volatilities is attained after 90 min.
[164]	No	Forex	Impulse response analysis	Intra-day	volatility	News surprise moderates extreme pure news effects that have nearly all positive (negative) coefficients in both regimes for volatility and depth (spread). In addition, volatility and depth respond positively to good and bad unscheduled news in both states (with more intensity during the expansion), while spread decreases in both states.

3.3.3. Recommendation Trading Strategies

Automated portfolio selection and designing trading strategies are challenging tasks due to differences between short-term and long-term investment risks. Short-term investments carry a heavy risk. The Sharpe ratio, Maximum Drawdown (MDD), and Sterling ratio are three measures for evaluating the utility of trading strategies based on risk adjustment. The authors of [85] proposed an optimal trading strategy using investors' sentiment feedback strength through news and tweets with the objective to maximize a risk-adjusted return measured by the Sterling ratio. Ref. [84] designed a business strategy recommendation system based on the public mood index and trend indicators and then according to rankings produced by a neural network-based learning-to-rank algorithm. Ref. [133] exploited the word embedding representation of news items and the local feature extraction capacity of CNN, and the temporal and long conditions of prices and news are taken care of by long short-term memory (LSTM) networks. Scholars have used a deep Q-network for training trading agents, in [27,64–66,94,167], based on financial market data. It seems the application of leveraging news information in the deep Q-learning approaches is not well explored. Table 7 depicts the detailed characteristics of the reviewed methods in strategy recommendation.

Table 7. Comparison of trading strategies models.

Literature	Media	Data Attributes Market	Duration	Feature Selection	Analysis Machine Learning	Evaluation Metrics
[84]	Thomson Reuters	S&P500	2006–2014	BoW	ListNet, RankNet	Sharpe ratio, MDD
[133]	TWSE Official Website	TWSE	2007–2017	word2vec	CNN, LSTM	RSME, profit
[85]	Twitter	NASDAQ, S&P 500	1 August 2012–30 January 2015	BoW	Genetic programming	Sterling ratio, Sharpe ratio, profit
[73]	Reuters	S&P500	September 2006–August 2007	BOW	NN, DT, SLR	Profit
[74]	DGAP	CDAX	2004–2011	BOW	Reinforcement learning	Profit

4. Big Data, Tools, and Challenges

In this section, we review some big data issues in the literature on financial market prediction based on textual data.

Volume: To meet the challenge of the huge volume of data, organizations need to look at distributed platforms, such as distributed Hadoop Eco-Systems, as big data management distributed systems and also use the No-SQL MongoDB database. For example, in SMeDA-SA [139], a Hadoop 1.1.2 cluster was utilized that comprised four nodes distributed through a WAN for each group in order to create separate processes for the classification tasks. Farimani et al. [16] proposed a restful microservice architecture for market prediction based on news and market data, where news data came from MongoDB-Flask-based APIs. MongoDB was also used in [168] to store around 31 million collected messages from Twitter, holding hashtags of all the traded stocks.

Velocity: The authors in [110] leveraged Apache Spark to deal with user-generated big data and to have a scalable system for stream processing. Most deep learning-based methods are implemented with TensorFlow [12,18,122,129,133] or PyTorch [21].

Variety: In terms of a market analysis, the methods are in three levels of data variety. The first category only uses text-based features for the stock prediction problem, in which the market information is often explicitly used for labeling the training set [19,39,61,99,101]. The other group presents mood criteria based on the firm, category, or market-related information via a sentiment analysis of text-based resources, especially stock board discussions [12,18,55,143]. Finally, the third group is methods that extract features based on the combination of market data, mood analysis, and feature selection from unstructured news documents [16,154]. Our analysis shows more than 30% of the works explored in this work use Python programming languages and Tensorflow APIs for implementing their decision support systems.

Veracity:The first step in big data analytics is to assure the originality of the data and production source [169]. One of the important areas of behavioral economics research is considering the impact of fake news and rumors on the market and investors. Kiymaz [170] examined the impact of rumors published in Turkish newspapers on the Istanbul Stock Exchange. Rumor identification in this research has been carried out by a human expert, while in recent years, research has been conducted on identifying rumors and fake news on social networks [77], and of course, the automatic identification of fake rumors and news in the behavioral economics domain is an open field of research.

Valence: Yang [85] used the big data valence concern among users with similar interests in the financial domain, extracting from the community of Twitter users and the messages diffused between members. Moreover, the authors in [171] proposed community detection for extracting relevant tweets for the target firms.

Value: The consideration of the different levels of these six V features depends on the type of business that faces the big data, and the organization determines the level of variety according to its cost and needs.

5. Open Research and Future Directions

Based on our quantitative analysis, some possible research questions for the future are:

- What is the effect of famous authors and influential persons working for financial newsgroups on market fluctuations?
- Do investors respond to the news, social media posts, or other information resources in the same way?
- Does the representation of contextual information in news documents and the proximity between a sequence of news help with improving financial decision supports?
- How do rumor and fake news diffusion patterns in financial social media correlate with market fluctuation?

We present possible research directions from three perspectives: data, NLP applications, and finance-related concerns.

5.1. Data

The methods presented over the past two years have been focusing on the use of multiple sources of information, as discussed in Section 3.3.1, and hence the integration of such data from multiple sources as well as data variety need to be investigated more. While our study (Figure 3) shows that the US and China stock markets have been well studied, the Forex market issues, such as pair selection based on news influence, are not well explored to the best of our knowledge. Cryptocurrency fluctuations under the news effect are also disregarded, which might be due to a lack of relevant news datasets. Recently, Farimani et al. gathered the Forex and Cryptocurrency news datasets in [12] which can facilitate the research in this domain. The early detection of rumors and fake news is of a great importance in behavioral economics. Very little research is currently conducted on identifying rumors and fake news in the behavioral economics domain which can be due to the complexity of this issue and the lack of appropriate datasets in this domain.

5.2. NLP Applications

Our survey shows that the use of word embedding methods for text structuring in behavioral economics has been growing over recent years. Future research could be incorporating other recent language models, such as RoBERTA, GPT3, and PEGASUS, presented in [172,173] and [174], respectively, for applying pre-training on a large corpus of text followed by the fine-tuning of a financial domain or for summarizing news. An accurate financial sentiment analysis can lead to better market mood identification, even in making soft real-time trading decisions. Usually, financial news is reported with ambivalence terms, hence focusing on a finer-grained sentiment analysis, such as recent deep learning-based methods [63,76,78,146,147], can be investigated. In terms of OpenIE, higher-level event detection methods, such as multiple event extraction [37], can extract more accurate features for market prediction. Deep feature extraction from various data sources, such as candle chart images, news text, news authors' roles, the trading transaction of investors, and social media comments, is recommended for research in financial market prediction.

5.3. Finance Applications

Deep learning approaches have been proposed in stock data and news releases over the past few years. Another interesting domain of study is investigating the impact of the information stream from news and social networks on investment transfers from one financial market to another. Investors move their investments from one market to another, influenced by news of some events for greater security. Modeling the impact of news events on markets that have co-movement is therefore a possible research direction. For instance, techniques based on information theory can be used to estimate the impact of news on financial markets. Portfolio selection based on textual information also requires more attention from the research community.

6. Conclusions

In this study, we reviewed the research on information retrieval from data sources, such as news and social networks, and the impact of such information on financial markets from the computer science perspective. During the review process, qualitative and quantitative analyses of behavioral economics knowledge extraction were performed in the four domains of heterogeneous data, contextual text representation and sentiment analysis, predictive models, and the correlation analysis of media behavior and investors' trading behavior. The purpose of this review is to cover different aspects of text mining interaction and behavioral economics and to identify research pathways in this domain. We studied the latest text mining techniques used in predicting financial markets, analyzing investor behavior, and reporting the trading strategy recommendation systems. This review can serve as a source of study for computer science researchers from the applied perspective in adopting long-term strategies and utilizing contextual text mining and behavior analysis. It is also desirable for market investors to reduce investment risk by leveraging decision support systems. From the financial perspective, the application of financial news recommendations and automatic trading agents leveraging media information in real-time decision support systems is another line of future work.

Author Contributions: S.A.F.: conception and design of study, acquisition of data, software, interpretation of data, and writing—original draft. M.V.J.: conception and design of study, supervision, review, and editing. A.M.F.: conception and design of study, supervision, and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: We declare that the authors have no competing interests as defined by MDPI or other interests that might be perceived to influence the results and/or discussion reported in this paper.

References

1. Fama, E.F. The behavior of stock-market prices. *J. Bus.* **1965**, *38*, 34–105. [CrossRef]
2. Shiller, R.J. From efficient markets theory to behavioral finance. *J. Econ. Perspect.* **2003**, *17*, 83–104. [CrossRef]
3. Ramiah, V.; Xu, X.; Moosa, I.A. Neoclassical finance, behavioral finance and noise traders: A review and assessment of the literature. *Int. Rev. Financ. Anal.* **2015**, *41*, 89–100. [CrossRef]
4. Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1188–1196.
5. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
6. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Minneapolis, Minnesota, 2019; pp. 4171–4186. [CrossRef]
7. Cutler, D.M.; Poterba, J.M.; Summers, L.H. What moves stock prices? Technical report; National Bureau of Economic Research: Cambridge, MA, USA, 1988.
8. Barber, B.M.; Loeffler, D. The "Dartboard" Column: Second-Hand Information and Price Pressure. *J. Financ. Quant. Anal.* **1993**, *28*, 273–284. [CrossRef]
9. Wuthrich, B.; Cho, V.; Leung, S.; Permunetilleke, D.; Sankaran, K.; Zhang, J. Daily stock market forecast from textual web data. In *SMC'98 Conference Proceedings, Proceedings of the 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218), San Diego, CA, USA, 14 October 1998;* IEEE: Piscataway, NJ, USA, 1998; Volume 3, pp. 2720–2725. [CrossRef]
10. Daniel, M.; Neves, R.F.; Horta, N. Company event popularity for financial markets using Twitter and sentiment analysis. *Expert Syst. Appl.* **2017**, *71*, 111–124. [CrossRef]
11. Sun, Y.; Fang, M.; Wang, X. A novel stock recommendation system using Guba sentiment analysis. *Pers. Ubiquitous Comput.* **2018**, *22*, 575–587. [CrossRef]
12. Anbaee Farimani, S.; Vafaei Jahan, M.; Milani Fard, A.; Tabbakh, S.R.K. Investigating the informativeness of technical indicators and news sentiment in financial market price prediction. *Knowl.-Based Syst.* **2022**, *247*, 108742. [CrossRef]
13. Passalis, N.; Avramelou, L.; Seficha, S.; Tsantekidis, A.; Doropoulos, S.; Makris, G.; Tefas, A. Multisource financial sentiment analysis for detecting Bitcoin price change indications using deep learning. *Neural Comput. Appl.* **2022**, 1–12. [CrossRef]

14. Krishnamoorthy, S. Sentiment analysis of financial news articles using performance indicators. *Knowl. Inf. Syst.* **2018**, *56*, 373–394. [CrossRef]
15. Seifollahi, S.; Shajari, M. Word sense disambiguation application in sentiment analysis of news headlines: An applied approach to FOREX market prediction. *J. Intell. Inf. Syst.* **2019**, *52*, 57–83. [CrossRef]
16. Anbaee Farimani, S.; Vafaei Jahan, M.; Milani Fard, A.; Haffari, G. Leveraging Latent Economic Concepts and Sentiments in the News for Market Prediction. In Proceedings of the 8th IEEE International Conference on Data Science and Advanced Analytics (DSAA), Porto, Portugal, 6–9 October 2021.
17. Chen, X.; Ma, X.; Wang, H.; Li, X.; Zhang, C. A hierarchical attention network for stock prediction based on attentive multi-view news learning. *Neurocomputing* **2022**, *504*, 1–15. [CrossRef]
18. Vargas, M.R.; dos Anjos, C.E.; Bichara, G.L.; Evsukoff, A.G. Deep leaming for stock market prediction using technical indicators and financial news articles. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; IEEE:Piscataway, NJ, USA, 2018; pp. 1–8. [CrossRef]
19. Ding, X.; Zhang, Y.; Liu, T.; Duan, J. Knowledge-Driven Event Embedding for Stock Prediction. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics, Osaka, Japan, 11–16 December 2016; Technical Papers; The COLING 2016 Organizing Committee: Osaka, Japan, 2016; pp. 2133–2142.
20. Long, J.; Chen, Z.; He, W.; Wu, T.; Ren, J. An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in Chinese stock exchange market. *Appl. Soft Comput.* **2020**, 91, 106205. [CrossRef]
21. Chen, D.; Ma, S.; Harimoto, K.; Bao, R.; Su, Q.; Sun, X. Group, Extract and Aggregate: Summarizing a Large Amount of Finance News for Forex Movement Prediction. In *Proceedings of the Second Workshop on Economics and Natural Language Processing*; Association for Computational Linguistics: Hong Kong, 2019; pp. 41–50. [CrossRef]
22. Lutz, B.; Pröllochs, N.; Neumann, D. Predicting sentence-level polarity labels of financial news using abnormal stock returns. *Expert Syst. Appl.* **2020**, *148*, 113223. [CrossRef]
23. Wang, H.; Lu, S.; Zhao, J. Aggregating multiple types of complex data in stock market prediction: A model-independent framework. *Knowl.-Based Syst.* **2019**, *164*, 193–204. [CrossRef]
24. Ren, J.; Long, J.; Xu, Z. Financial news recommendation based on graph embeddings. *Decis. Support Syst.* **2019**, *125*, 113115. [CrossRef]
25. Yang, L.; Xu, Y.; Ng, J.; Dong, R. Leveraging BERT to improve the FEARS index for stock forecasting. In Proceedings of the First Workshop on Financial Technology and Natural Language Processing, Macao, China, 12 August 2019; ACL: Stroudsburg, PA, USA, 2019.
26. Zhang, Z.; Zohren, S.; Roberts, S. Deep reinforcement learning for trading. *J. Financ. Data Sci.* **2020**, *2*, 25–40. [CrossRef]
27. Gao, Z.; Gao, Y.; Hu, Y.; Jiang, Z.; Su, J. Application of Deep Q-Network in Portfolio Management. In Proceedings of the 2020 5th IEEE International Conference on Big Data Analytics (ICBDA), Xiamen, China, 8–11 May 2020; pp. 268–275. [CrossRef]
28. Li, X.; Wu, P.; Wang, W. Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. *Inf. Process. Manag.* **2020**, *57*, 102212. [CrossRef]
29. Moews, B.; Ibikunle, G. Predictive intraday correlations in stable and volatile market environments: Evidence from deep learning. *Phys. A Stat. Mech. Appl.* **2020**, 124392. [CrossRef]
30. Teng, X.; Wang, T.; Zhang, X.; Lan, L.; Luo, Z. Enhancing Stock Price Trend Prediction via a Time-Sensitive Data Augmentation Method. *Complexity* **2020**, *2020, 6737951.* [CrossRef]
31. Wong, S.Y.K.; Chan, J.S.K.; Azizi, L.; Xu, R.Y.D. Time-varying Neural Network for Stock Return Prediction. *Int. J. Intell. Syst. Account. Financ. Manag.* **2022**, *29*, 3–18. [CrossRef]
32. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805
33. Lu, Z.; Du, P.; Nie, J.Y. VGCN-BERT: Augmenting BERT with Graph Embedding for Text Classification. In Proceedings of the European Conference on Information Retrieval, Lisbon, Portugal, 14–17 April 2020; Springer: Berlin, Germany, 2020; pp. 369–382. [CrossRef]
34. Wang, T.; Yuan, C.; Wang, C. Does Applying Deep Learning in Financial Sentiment Analysis Lead to Better Classification Performance? *Econ. Bull.* **2020**, *40*, 1091–1105.
35. Liu, X.; Huang, H.; Zhang, Y.; Yuan, C. News-Driven Stock Prediction With Attention-Based Noisy Recurrent State Transition. *arXiv* **2020**, arXiv:2004.01878.
36. Liu, Q.; Cheng, X.; Su, S.; Zhu, S. Hierarchical complementary attention network for predicting stock price movements with news. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Torino, Italy, 22–26 October 2018; pp. 1603–1606. [CrossRef]
37. Liu, X.; Luo, Z.; Huang, H. Jointly multiple events extraction via attention-based graph information aggregation. *arXiv* **2018**, arXiv:1809.09078.
38. Zhao, R.; Deng, Y.; Dredze, M.; Verma, A.; Rosenberg, D.; Stent, A. Visual attention model for cross-sectional stock return prediction and end-to-end multimodal market representation learning. *arXiv* **2018**, arXiv:1809.03684.
39. Yang, L.; Zhang, Z.; Xiong, S.; Wei, L.; Ng, J.; Xu, L.; Dong, R. Explainable text-driven neural network for stock prediction. In Proceedings of the 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), Nanjing, China, 23–25 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 441–445. [CrossRef]

40. Deng, S.; Zhang, N.; Zhang, W.; Chen, J.; Pan, J.Z.; Chen, H. Knowledge-Driven Stock Trend Prediction and Explanation via Temporal Convolutional Network. In Proceedings of the Companion Proceedings of The 2019 World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; Association for Computing Machinery: New York, NY, USA, 2019; WWW '19 ; pp. 678–685. [CrossRef]
41. Prokhorov, V.; Pilehvar, M.T.; Collier, N. Generating Knowledge Graph Paths from Textual Definitions using Sequence-to-Sequence Models. *arXiv* **2019**, arXiv:1904.02996.
42. Cheng, D.; Yang, F.; Xiang, S.; Liu, J. Financial time series forecasting with multi-modality graph neural network. *Pattern Recognit.* **2022**, *121*, 108218. [CrossRef]
43. Kim, R.; So, C.H.; Jeong, M.; Lee, S.; Kim, J.; Kang, J. HATS: A Hierarchical Graph Attention Network for Stock Movement Prediction. *arXiv* **2019**, arXiv:1908.07999.
44. Kumar, B.S.; Ravi, V. A survey of the applications of text mining in financial domain. *Knowl.-Based Syst.* **2016**, *114*, 128–147. [CrossRef]
45. Li, Q.; Chen, Y.; Wang, J.; Chen, Y.; Chen, H. Web media and stock markets: A survey and future directions from a big data perspective. *IEEE Trans. Knowl. Data Eng.* **2017**, *30*, 381–399. [CrossRef]
46. Man, X.; Luo, T.; Lin, J. Financial sentiment analysis (fsa): A survey. In Proceedings of the 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), Taipei, 6–9 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 617–622. [CrossRef]
47. Nassirtoussi, A.K.; Aghabozorgi, S.; Wah, T.Y.; Ngo, D.C.L. Text mining for market prediction: A systematic review. *Expert Syst. Appl.* **2014**, *41*, 7653–7670. [CrossRef]
48. Jiang, W. Applications of deep learning in stock market prediction: Recent progress. *Expert Syst. Appl.* **2021**, *184*, 115537. [CrossRef]
49. Rajendiran, P.; Priyadarsini, P. Survival study on stock market prediction techniques using sentimental analysis. *Mater. Today Proc.* **2021.** [CrossRef]
50. Saha, S.; Gao, J.; Gerlach, R. A survey of the application of graph-based approaches in stock market analysis and prediction. *Int. J. Data Sci. Anal.* **2022**, 14, 1–15. [CrossRef]
51. Cao, L.; Yang, Q.; Yu, P.S. Data science and AI in FinTech: An overview. *Int. J. Data Sci. Anal.* **2021**, *12*, 81–99. [CrossRef]
52. Kumbure, M.M.; Lohrmann, C.; Luukka, P.; Porras, J. Machine learning techniques and data for stock market forecasting: A literature review. *Expert Syst. Appl.* **2022**, *197*, 116659. [CrossRef]
53. Agarwal, S.; Kumar, S.; Goel, U. Stock market response to information diffusion through internet sources: A literature review. *Int. J. Inf. Manag.* **2019**, *45*, 118–131. [CrossRef]
54. Xing, F.Z.; Cambria, E.; Welsch, R.E. Natural language based financial forecasting: A survey. *Artif. Intell. Rev.* **2018**, *50*, 49–73. [CrossRef]
55. Bollen, J.; Mao, H.; Zeng, X. Twitter mood predicts the stock market. *J. Comput. Sci.* **2011**, *2*, 1–8. [CrossRef]
56. Tetlock, P.C. Giving content to investor sentiment: The role of media in the stock market. *J. Financ.* **2007**, *62*, 1139–1168. [CrossRef]
57. Tetlock, P.C.; Saar-Tsechansky, M.; Macskassy, S. More than words: Quantifying language to measure firms' fundamentals. *J. Financ.* **2008**, *63*, 1437–1467. [CrossRef]
58. Baker, M.; Wurgler, J. Investor sentiment and the cross-section of stock returns. *J. Financ.* **2006**, *61*, 1645–1680. [CrossRef]
59. Weng, B.; Ahmed, M.A.; Megahed, F.M. Stock market one-day ahead movement prediction using disparate data sources. *Expert Syst. Appl.* **2017**, *79*, 153–163. [CrossRef]
60. Nassirtoussi, A.K.; Aghabozorgi, S.; Wah, T.Y.; Ngo, D.C.L. Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Syst. Appl.* **2015**, *42*, 306–324. [CrossRef]
61. Long, W.; Song, L.; Tian, Y. A new graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity. *Expert Syst. Appl.* **2019**, *118*, 411–424. [CrossRef]
62. Van de Kauter, M.; Breesch, D.; Hoste, V. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Syst. Appl.* **2015**, *42*, 4999–5010. [CrossRef]
63. Do, H.H.; Prasad, P.; Maag, A.; Alsadoon, A. Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review. *Expert Syst. Appl.* **2019**, *118*, 272–299. [CrossRef]
64. Shavandi, A.; Khedmati, M. A multi-agent deep reinforcement learning framework for algorithmic trading in financial markets. *Expert Syst. Appl.* **2022**, *208*, 118124. [CrossRef]
65. Chen, J.; Luo, C.; Pan, L.; Jia, Y. Trading strategy of structured mutual fund based on deep learning network. *Expert Syst. Appl.* **2021**, *183*, 115390. [CrossRef]
66. Carta, S.; Ferreira, A.; Podda, A.S.; Reforgiato Recupero, D.; Sanna, A. Multi-DQN: An ensemble of Deep Q-learning agents for stock market forecasting. *Expert Syst. Appl.* **2021**, *164*, 113820. [CrossRef]
67. Nam, K.; Seong, N. Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market. *Decis. Support Syst.* **2019**, *117*, 100–112. [CrossRef]
68. Shynkevich, Y.; McGinnity, T.M.; Coleman, S.A.; Belatreche, A. Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. *Decis. Support Syst.* **2016**, *85*, 74–83. [CrossRef]
69. Schumaker, R.P.; Zhang, Y.; Huang, C.N.; Chen, H. Evaluating sentiment in financial news articles. *Decis. Support Syst.* **2012**, *53*, 458–464. [CrossRef]

70. Hagenau, M.; Liebmann, M.; Neumann, D. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decis. Support Syst.* **2013**, *55*, 685–697. [CrossRef]
71. Ho, C.S.; Damien, P.; Gu, B.; Konana, P. The time-varying nature of social media sentiments in modeling stock returns. *Decis. Support Syst.* **2017**, *101*, 69–81. [CrossRef]
72. Oliveira, N.; Cortez, P.; Areal, N. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decis. Support Syst.* **2016**, *85*, 62–73. [CrossRef]
73. Geva, T.; Zahavi, J. Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. *Decis. Support Syst.* **2014**, *57*, 212–223. [CrossRef]
74. Feuerriegel, S.; Prendinger, H. News-based trading strategies. *Decis. Support Syst.* **2016**, *90*, 65–74. [CrossRef]
75. Xiang, C.; Zhang, J.; Li, F.; Fei, H.; Ji, D. A semantic and syntactic enhanced neural model for financial sentiment analysis. *Inf. Process. Manag.* **2022**, *59*, 102943. [CrossRef]
76. Yang, C.; Zhang, H.; Jiang, B.; Li, K. Aspect-based sentiment analysis with alternating coattention networks. *Inf. Process. Manag.* **2019**, *56*, 463–478. [CrossRef]
77. Zhang, X.; Ghorbani, A.A. An overview of online fake news: Characterization, detection, and discussion. *Inf. Process. Manag.* **2020**, *57*, 102025. [CrossRef]
78. Ghorbanali, A.; Sohrabi, M.K.; Yaghmaee, F. Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks. *Inf. Process. Manag.* **2022**, *59*, 102929. [CrossRef]
79. Zhang, X.; Zhang, Y.; Wang, S.; Yao, Y.; Fang, B.; Philip, S.Y. Improving stock market prediction via heterogeneous information fusion. *Knowl.-Based Syst.* **2018**, *143*, 236–247. [CrossRef]
80. Liang, B.; Su, H.; Gui, L.; Cambria, E.; Xu, R. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowl.-Based Syst.* **2022**, *235*, 107643. [CrossRef]
81. Consoli, S.; Barbaglia, L.; Manzan, S. Fine-grained, aspect-based sentiment analysis on economic and financial lexicon. *Knowl.-Based Syst.* **2022**, *247*, 108781. [CrossRef]
82. Wu, H.; Zhang, Z.; Shi, S.; Wu, Q.; Song, H. Phrase dependency relational graph attention network for Aspect-based Sentiment Analysis. *Knowl.-Based Syst.* **2022**, *236*, 107736. [CrossRef]
83. Kim, H.K.; Kim, H.; Cho, S. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing* **2017**, *266*, 336–352. [CrossRef]
84. Song, Q.; Liu, A.; Yang, S.Y. Stock portfolio selection using learning-to-rank algorithms with news sentiment. *Neurocomputing* **2017**, *264*, 20–28. [CrossRef]
85. Yang, S.Y.; Mo, S.Y.K.; Liu, A.; Kirilenko, A.A. Genetic programming optimization for a sentiment feedback strength based trading strategy. *Neurocomputing* **2017**, *264*, 29–41. [CrossRef]
86. Atkins, A.; Niranjan, M.; Gerding, E. Financial news predicts stock market volatility better than close price. *J. Financ. Data Sci.* **2018**, *4*, 120–137. [CrossRef]
87. Nisar, T.M.; Yeung, M. Twitter as a tool for forecasting stock market movements: A short-window event study. *J. Financ. Data Sci.* **2018**, *4*, 101–119. [CrossRef]
88. Sun, A.; Lachanski, M.; Fabozzi, F.J. Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *Int. Rev. Financ. Anal.* **2016**, *48*, 272–281. [CrossRef]
89. Li, Y.; Pan, Y. A novel ensemble deep learning model for stock prediction based on stock prices and news. *Int. J. Data Sci. Anal.* **2022**, *13*, 139–149. [CrossRef] [PubMed]
90. Ito, T.; Sakaji, H.; Izumi, K.; Tsubouchi, K.; Yamashita, T. Ginn: Gradient interpretable neural networks for visualizing financial texts. *Int. J. Data Sci. Anal.* **2020**, *9*, 431–445. [CrossRef]
91. Hájek, P. Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns. *Neural Comput. Appl.* **2018**, *29*, 343–358. [CrossRef]
92. Li, X.; Xie, H.; Wang, R.; Cai, Y.; Cao, J.; Wang, F.; Min, H.; Deng, X. Empirical analysis: Stock market prediction via extreme learning machine. *Neural Comput. Appl.* **2016**, *27*, 67–78. [CrossRef]
93. Jahan, M.V.; Akbarzadeh-T, M.R. Extremal optimization vs. learning automata: Strategies for spin selection in portfolio selection problems. *Appl. Soft Comput.* **2012**, *12*, 3276–3284. [CrossRef]
94. Shi, Y.; Li, W.; Zhu, L.; Guo, K.; Cambria, E. Stock trading rule discovery with double deep Q-network. *Appl. Soft Comput.* **2021**, *107*, 107320. [CrossRef]
95. Zhang, W.; Wang, P.; Li, X.; Shen, D. Quantifying the cross-correlations between online searches and Bitcoin market. *Phys. A Stat. Mech. Appl.* **2018**, *509*, 657–672. [CrossRef]
96. Ochiai, T.; Nacher, J. A model for the dynamic behavior of financial assets affected by news: The case of Tohoku–Kanto earthquake. *Phys. Lett. A* **2011**, *375*, 3552–3556. [CrossRef]
97. Rodrigues, F.B.; Giozza, W.F.; de Oliveira Albuquerque, R.; García Villalba, L.J. Natural Language Processing Applied to Forensics Information Extraction With Transformers and Graph Visualization. *IEEE Trans. Comput. Soc. Syst.* **2022**, 1–17. [CrossRef]
98. Jahan, M.V.; Akbarzadeh-T, M.R. From local search to global conclusions: Migrating spin glass-based distributed portfolio selection. *IEEE Trans. Evol. Comput.* **2010**, *14*, 591–601. [CrossRef]
99. Ding, X.; Zhang, Y.; Liu, T.; Duan, J. Deep learning for event-driven stock prediction. In Proceedings of the Twenty-fourth international joint conference on artificial intelligence, Buenos Aires, Argentina, 25–31 July 2015.

100. Liu, W.; Zhou, P.; Zhao, Z.; Wang, Z.; Ju, Q.; Deng, H.; Wang, P. K-BERT: Enabling Language Representation with Knowledge Graph. In Proceedings of the AAAI, New York, NY, USA, 7–12 February 2020; pp. 2901–2908. [CrossRef]
101. Ding, X.; Zhang, Y.; Liu, T.; Duan, J. Using Structured Events to Predict Stock Price Movement: An Empirical Investigation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1415–1425. [CrossRef]
102. Si, J.; Mukherjee, A.; Liu, B.; Pan, S.J.; Li, Q.; Li, H. Exploiting social relations and sentiment for stock prediction. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1139–1145.
103. Papaluca, A.; Krefl, D.; Suominen, H.; Lenskiy, A. Pretrained Knowledge Base Embeddings for improved Sentential Relation Extraction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Dublin, Ireland, 22–27 May 2022; Association for Computational Linguistics: Dublin, Ireland, 2022; pp. 373–382. [CrossRef]
104. Xu, Y.; Cohen, S.B. Stock movement prediction from tweets and historical prices. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 1970–1979. [CrossRef]
105. Hajek, P.; Barushka, A. Integrating Sentiment Analysis and Topic Detection in Financial News for Stock Movement Prediction. In Proceedings of the 2nd International Conference on Business and Information Management, ICBIM '18, Barcelona, Spain, 20–22 September 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 158–162. [CrossRef]
106. Jin, F.; Self, N.; Saraf, P.; Butler, P.; Wang, W.; Ramakrishnan, N. Forex-foreteller: Currency trend modeling using news articles. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013; pp. 1470–1473. [CrossRef]
107. Sontag, D.; Roy, D. Complexity of inference in latent dirichlet allocation. In Proceedings of the Advances in neural information processing systems, Granada, Spain, 12–15 December 2011; pp. 1008–1016.
108. Zhang, X.; Fuehres, H.; Gloor, P.A. Predicting stock market indicators through twitter I hope it is not as bad as I fear. *Procedia-Soc. Behav. Sci.* **2011**, *26*, 55–62. [CrossRef]
109. Wang, Q. Cryptocurrencies asset pricing via machine learning. *Int. J. Data Sci. Anal.* **2021**, *12*, 175–183. [CrossRef]
110. Atzeni, M.; Dridi, A.; Recupero, D.R. Using frame-based resources for sentiment analysis within the financial domain. *Prog. Artif. Intell.* **2018**, *7*, 273–294. [CrossRef]
111. Zhang, K.; Zi, J.; Wu, L.G. New event detection based on indexing-tree and named entity. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands, 23–27 July 2007; pp. 215–222. [CrossRef]
112. Da, Z.; Engelberg, J.; Gao, P. The sum of all FEARS investor sentiment and asset prices. *Rev. Financ. Stud.* **2015**, *28*, 1–32. [CrossRef]
113. Wu, J.; Xu, K.; Zhao, J. Online reviews can predict long-term returns of individual stocks. *arXiv* **2019**, arXiv:1905.03189.
114. Wu, G.G.R.; Hou, T.C.T.; Lin, J.L. Can economic news predict Taiwan stock market returns? *Asia Pac. Manag. Rev.* **2019**, *24*, 54–59. [CrossRef]
115. Liu, Y. Fine-tune BERT for extractive summarization. *arXiv* **2019**, arXiv:1903.10318.
116. Harris, Z.S. Distributional Structure. *WORD* **1954**, *10*, 146–162. [CrossRef]
117. Salton, G. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*; Addison-Wesley: Boston, MA, USA, 1989; Volume 169.
118. Dumais, S.T. Latent semantic analysis. *Annu. Rev. Inf. Sci. Technol.* **2004**, *38*, 188–230. [CrossRef]
119. Feuerriegel, S.; Gordon, J. News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions. *Eur. J. Oper. Res.* **2019**, *272*, 162–175. [CrossRef]
120. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
121. Anbaee Farimani, S.; Tabatabaee, H.; Kaffashan, M. An Investigation into the Process of Organizing and Retrieving Web Texts Based on the Integration of Semantic Concepts In order to organize knowledge. *Iran. J. Inf. Process. Manag.* **2019**, *34*, 1879–1904.
122. Huynh, H.D.; Dang, L.M.; Duong, D. A new model for stock price movements prediction using deep neural network. In Proceedings of the Eighth International Symposium on Information and Communication Technology, Nha Trang City, Viet Nam, 7–8 December 2017; pp. 57–62. [CrossRef]
123. Lutz, B.; Pröllochs, N.; Neumann, D. Sentence-Level Sentiment Analysis of Financial News Using Distributed Text Representations and Multi-Instance Learning. *arXiv* **2018**, arXiv:1901.00400.
124. Hiew, J.Z.G.; Huang, X.; Mou, H.; Li, D.; Wu, Q.; Xu, Y. BERT-based Financial Sentiment Index and LSTM-based Stock Return Predictability. *arXiv* **2019**, arXiv:1906.09024.
125. Jiang, C.; Liang, K.; Chen, H.; Ding, Y. Analyzing market performance via social media: A case study of a banking industry crisis. *Sci. China Inf. Sci.* **2014**, *57*, 1–18. [CrossRef]
126. Hendershott, T.; Livdan, D.; Schürhoff, N. Are institutions informed about news? *J. Financ. Econ.* **2015**, *117*, 249–287. [CrossRef]
127. Gupta, K.; Banerjee, R. Does OPEC news sentiment influence stock returns of energy firms in the United States? *Energy Econ.* **2019**, *77*, 34–45. [CrossRef]

128. Medovikov, I. When does the stock market listen to economic news? New evidence from copulas and news wires. *J. Bank. Financ.* **2016**, *65*, 27–40. [CrossRef]
129. Verma, I.; Dey, L.; Meisheri, H. Detecting, quantifying and accessing impact of news events on Indian stock indices. In Proceedings of the International Conference on Web Intelligence, Sogndal, Norway, 25–27 May 2011; pp. 550–557. [CrossRef]
130. Dale, R. GPT-3: What's it good for? *Nat. Lang. Eng.* **2021**, *27*, 113–118. [CrossRef]
131. Tetlock, P.C. Does public financial news resolve asymmetric information? *Rev. Financ. Stud.* **2010**, *23*, 3520–3557. [CrossRef]
132. Tausch, F.; Zumbuehl, M. Stability of risk attitudes and media coverage of economic news. *J. Econ. Behav. Organ.* **2018**, *150*, 295–310. [CrossRef]
133. Lee, C.Y.; Soo, V.W. Predict Stock Price with Financial News Based on Recurrent Convolutional Neural Networks. In Proceedings of the 2017 Conference on Technologies and Applications of Artificial Intelligence (TAAI), Taipei, 1–3 December 2017; pp. 160–165. [CrossRef]
134. Rao, Y.; Zhong, X.; Lu, S. Research on News Topic-Driven Market Flucatuation and Predication. In Proceedings of the 2016 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI), Beijing, China, 20–21 October 2016; pp. 559–562. [CrossRef]
135. Baccianella, S.; Esuli, A.; Sebastiani, F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Paris, France, 17–23 May 2010; European Language Resources Association (ELRA): Valletta, Malta, 2010.
136. LOUGHRAN, T.; MCDONALD, B. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *J. Financ.* **2011**, *66*, 35–65. [CrossRef]
137. Zhang, H.; Li, Z.; Xie, H.; Lau, R.Y.; Cheng, G.; Li, Q.; Zhang, D. Leveraging statistical information in fine-grained financial sentiment analysis. *World Wide Web* **2022**, *25*, 513–531. [CrossRef]
138. Zhao, M.; Yang, J.; Zhang, J.; Wang, S. Aggregated graph convolutional networks for aspect-based sentiment classification. *Inf. Sci.* **2022**, *600*, 73–93. [CrossRef]
139. Li, B.; Chan, K.C.; Ou, C.; Ruifeng, S. Discovering public sentiment in social media for predicting stock movement of publicly listed companies. *Inf. Syst.* **2017**, *69*, 81–92. [CrossRef]
140. Shi, Y.; Liu, W.M.; Ho, K.Y. Public news arrival and the idiosyncratic volatility puzzle. *J. Empir. Financ.* **2016**, *37*, 159–172. [CrossRef]
141. Zhang, G.; Xu, L.; Xue, Y. Model and forecast stock market behavior integrating investor sentiment analysis and transaction data. *Clust. Comput.* **2017**, *20*, 789–803. [CrossRef]
142. Bouktif, S.; Fiaz, A.; Awad, M. Augmented Textual Features-Based Stock Market Prediction. *IEEE Access* **2020**, *8*, 40269–40282. [CrossRef]
143. Shi, Y.; Ho, K.Y.; Liu, W.M. Public information arrival and stock return volatility: Evidence from news sentiment and Markov Regime-Switching Approach. *Int. Rev. Econ. Financ.* **2016**, *42*, 291–312. [CrossRef]
144. Araci, D. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv* **2019**, arXiv:1908.10063.
145. Vora, V.; Shah, M.; Chouhan, A.; Tawde, P. Stock Market Prices and Returns Forecasting Using Deep Learning Based on Technical and Fundamental Analysis. In *Proceedings of the Information and Communication Technology for Competitive Strategies (ICTCS 2021)*; Kaiser, M.S., Xie, J., Rathore, V.S., Eds.; Springer Nature Singapore: Singapore, 2022; pp. 717–728. [CrossRef]
146. Sun, C.; Huang, L.; Qiu, X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv* **2019**, arXiv:1903.09588.
147. Huang, J.; Xing, R.; Li, Q. Asset pricing via deep graph learning to incorporate heterogeneous predictors. *Int. J. Intell. Syst.* **2022.** *37*, 8462–8489. [CrossRef]
148. Yadav, A.; Jha, C.; Sharan, A.; Vaish, V. Sentiment analysis of financial news using unsupervised approach. *Procedia Comput. Sci.* **2020**, *167*, 589–598. [CrossRef]
149. Yadav, R.; Kumar, A.V.; Kumar, A. News-based supervised sentiment analysis for prediction of futures buying behaviour. *IIMB Manag. Rev.* **2019**, *31*, 157–166. [CrossRef]
150. Vilas, A.F.; Redondo, R.P.D.; Crockett, K.; Owda, M.; Evans, L. Twitter permeability to financial events: An experiment towards a model for sensing irregularities. *Multimed. Tools Appl.* **2019**, *78*, 9217–9245. [CrossRef]
151. Kruiper, R.; Vincent, J.F.; Chen-Burger, J.; Desmulliez, M.P.; Konstas, I. In Layman's Terms: Semi-Open Relation Extraction from Scientific Texts. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020.
152. Zhou, S.; Yu, B.; Sun, A.; Long, C.; Li, J.; Yu, H.; Sun, J.; Li, Y. A Survey on Neural Open Information Extraction: Current Status and Future Directions. *arXiv* **2022**, arXiv:2205.117252022.
153. Gurin, Y.; Szymanski, T.; Keane, M.T. Discovering news events that move markets. In Proceedings of the 2017 Intelligent Systems Conference (IntelliSys), London, UK, 21–22 September 2016; IEEE: Piscataway, NJ, USA, 2017; pp. 452–461.
154. Li, Q.; Chen, Y.; Jiang, L.L.; Li, P.; Chen, H. A tensor-based information framework for predicting the stock market. *ACM Trans. Inf. Syst. (TOIS)* **2016**, *34*, 1–30. [CrossRef]
155. Checkley, M.; Higón, D.A.; Alles, H. The hasty wisdom of the mob: How market sentiment predicts stock market behavior. *Expert Syst. Appl.* **2017**, *77*, 256–263. [CrossRef]
156. Granger, C. Testing for causality: A personal viewpoint. *J. Econ. Dyn. Control* **1980**, *2*, 329–352. [CrossRef]

157. Wei, Y.C.; Lu, Y.C.; Chen, J.N.; Hsu, Y.J. Informativeness of the market news sentiment in the Taiwan stock market. *North Am. J. Econ. Financ.* **2017**, *39*, 158–181. [CrossRef]
158. de Araújo, J.G.; Marinho, L.B. Using Online Economic News to Predict Trends in Brazilian Stock Market Sectors. In Proceedings of the 24th Brazilian Symposium on Multimedia and the Web, Salvador, Brazil, 16–19 October 2018; WebMedia '18; Association for Computing Machinery: New York, NY, USA, 2018; pp. 37–44. [CrossRef]
159. Romanov, V.; Naletova, O.; Panteleeva, E.; Federyakov, A. Fractal model of estimating news and insider influence on market volatility. *Autom. Doc. Math. Linguist.* **2007**, *41*, 141–149. [CrossRef]
160. Zhou, W.X. Multifractal detrended cross-correlation analysis for two nonstationary signals. *Phys. Rev. E* **2008**, *77*, 066211. [CrossRef]
161. Alamatian, Z.; Vafaei Jahan, M.; Milani Fard, A. Using Market Indicators to Eliminate Local Trends for Financial Time Series Cross-Correlation Analysis. In Proceedings of the 34th Canadian Conference on Artificial Intelligence (Canadian AI), Vancouver, BC, Canada, 25–28 May 2021. [CrossRef]
162. Chen, K.; Luo, P.; Liu, L.; Zhang, W. News, search and stock co-movement: Investigating information diffusion in the financial market. *Electron. Commer. Res. Appl.* **2018**, *28*, 159–171. [CrossRef]
163. Omrane, W.B.; Hussain, S.M. Foreign news and the structure of co-movement in European equity markets: An intraday analysis. *Res. Int. Bus. Financ.* **2016**, *37*, 572–582. [CrossRef]
164. Omrane, W.B.; Tao, Y.; Welch, R. Scheduled macro-news effects on a Euro/US dollar limit order book around the 2008 financial crisis. *Res. Int. Bus. Financ.* **2017**, *42*, 9–30. [CrossRef]
165. Birz, G. Stale economic news, media and the stock market. *J. Econ. Psychol.* **2017**, *61*, 87–102. [CrossRef]
166. Fang, L.; Yu, H.; Huang, Y. The role of investor sentiment in the long-term correlation between US stock and bond markets. *Int. Rev. Econ. Financ.* **2018**, *58*, 127–139. [CrossRef]
167. El Akraoui, B.; Daoui, C. Deep Reinforcement Learning for Bitcoin Trading. In *Proceedings of the Business Intelligence*; Fakir, M., Baslam, M., El Ayachi, R., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 82–93. [CrossRef]
168. Oliveira, N.; Cortez, P.; Areal, N. The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Syst. Appl.* **2017**, *73*, 125–144. [CrossRef]
169. Ghasemaghaei, M. The role of positive and negative valence factors on the impact of bigness of data on big data analytics usage. *Int. J. Inf. Manag.* **2020**, *50*, 395–404. [CrossRef]
170. Kiymaz, H. The effects of stock market rumors on stock prices: Evidence from an emerging market. *J. Multinatl. Financ. Manag.* **2001**, *11*, 105–115. [CrossRef]
171. Ranjan, S.; Sood, S. Investor community sentiment analysis for predicting stock price trends. *Int. J. Manag. Technol. Eng.* **2019**, *9*, *6012–6020.*
172. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
173. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.14165.
174. Zhang, J.; Zhao, Y.; Saleh, M.; Liu, P.J. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv* **2019**, arXiv:1912.08777.