mathematics

MDPI

Review

A Review of Sentiment, Semantic and Event-Extraction-Based Approaches in Stock Forecasting

Wai Khuen Cheng [1,*], Khean Thye Bea [2], Steven Mun Hong Leow [2], Jireh Yi-Le Chan [2], Zeng-Wei Hong [3] and Yen-Lin Chen [4,*]

[1] Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar 31900, Malaysia
[2] Faculty of Business and Finance, Universiti Tunku Abdul Rahman, Kampar 31900, Malaysia; beakheanthye@1utar.my (K.T.B.); steven.utar@1utar.my (S.M.H.L.); jirehchan@utar.edu.my (J.Y.-L.C.)
[3] Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, Taiwan; zwhong@fcu.edu.tw
[4] Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei 10608, Taiwan
* Correspondence: chengwk@utar.edu.my (W.K.C.); ylchen@mail.ntut.edu.tw (Y.-L.C.)

Abstract: Stock forecasting is a significant and challenging task. The recent development of web technologies has transformed the communication channel to allow the public to share information over the web such as news, social media contents, etc., thus causing exponential growth of web data. The massively available information might be the key to revealing the financial market's unexplained variability and facilitating forecasting accuracy. However, this information is usually in unstructured natural language and consists of different inherent meanings. Although a human can easily interpret the inherent messages, it is still complicated to manually process such a massive amount of textual data due to the constraint of time, ability, energy, etc. Due to the different properties of text sources, it is crucial to understand various text processing approaches to optimize forecasting performance. This study attempted to summarize and discuss the current text-based financial forecasting approaches in the aspect of semantic-based, sentiment-based, event-extraction-based, and hybrid approaches. Afterward, the study discussed the strength and weakness of each approach, followed with their comparison and suitable application scenarios. Moreover, this study also highlighted the future research direction in text-based stock forecasting, where the overall discussion is expected to provide insightful analysis for future reference.

Keywords: natural language processing; deep learning; stock forecasting; sentiment analysis; event extraction

MSC: 68T50; 68T07

check for updates

Citation: Cheng, W.K.; Bea, K.T.; Leow, S.M.H.; Chan, J.Y.-L.; Hong, Z.-W.; Chen, Y.-L. A Review of Sentiment, Semantic and Event-Extraction-Based Approaches in Stock Forecasting. *Mathematics* **2022**, *10*, 2437. https://doi.org/10.3390/math10142437

Academic Editors: Peter Schwendner, Mark James Thompson, Jan-Alexander Posth, Per Bjarte Solibakke and Kristina Šutienė

Received: 2 June 2022
Accepted: 9 July 2022
Published: 13 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Stock is a financial instrument to represent the business owner where the demand-supply determines the prices in the market. Supply factors usually refer to the internal corporate policy, such as issuing new shares or share buybacks. In contrast, the demand factors are diverse across the business factors, industry prospects, macro-environment, market sentiment, etc. In the demand perspective, the stock prices reflect the investor's expectation toward the future value of a particular business. In other words, the trading activities are information driven, where the investors are regularly reacting to new information and revising their expectations toward the future business status.

According to the theory of Efficient Market Hypothesis (EMH) [1], rational investors should fully respond to all available information instantly when participating in trading activities under a perfectly efficient market. Under such conditions, the arbitrage action

eliminates the price deviation, and the market prices should be fair and neither overvalued nor undervalued. To this, the price movement should obey the Random Walk Hypotheses [2] and negate the forecasting attempt to achieve more than 50% accuracy. However, the following question remained: "Is perfect market condition exists in practice?"

In the existing literature, extensive studies [3–7] continuously realized excess return in the stock market. Those findings contradicted EMH and raised the fundamental question: "Where does the abnormal return come from?" To this, the Adaptive Market Hypothesis (AMH) [8] is proposed to explain the market anomalies with the theories of behavioral finance. In practice, the ability of informational mining and investors' behavior toward the information can differ across the investors. The information asymmetry is the primary reason to distort the perfect market condition and imply the likelihood to realize the excess return in the stock market [9].

Although excess return can be realized in practice, the nonlinear and dynamic attributes of financial markets are the biggest challenges to be addressed for financial forecasting [10]. Financial time series usually come with a high degree of variability and extensive noise to disrupt the forecasting performance [11,12]. The study [13] claimed that the financial market might have a complex correlation across different markets, companies, or countries. Based on this, later studies attempted to address the forecasting problem by modelling the correlation between the variables of different market [14]. Recently, the advancement of web technologies is driving the growth of textual data, and the massively available text data could be used to reveal the unexplained variability in the financial market and improve forecasting performance.

Moreover, the natural language in the text is usually unstructured and implied with different inherent meanings. Those messages can be easily interpreted by humans but are difficult to be translated by a machine. Furthermore, such a massive amount of text data is difficult to be processed manually by a human due to the constraint of time, ability, energy, etc. The invention of Natural Language Processing (NLP) has provided a solution to develop computational models that enable the machine to understand human languages and automatically solve practical problems. Therefore, the application of NLP is becoming an important tool to reveal the investor behavioral information to explain the market variability and improve the stock prediction performance [15].

Previous survey studies mostly focused on discussing the deep learning approaches in financial forecasting. Some studies reviewed the application of deep learning approaches in different financial aspects such as algorithmic trading, risk assessment, portfolio management, etc. [16], and studied the different approaches in stock forecasting [17]. Some studies reviewed the stock forecasting studies in seven different aspects [18,19] and provided a comparative study in applying the deep learning approaches in stock forecasting [20]. The most related study [9] discussed the application of the NLP technique for financial forecasting. Their study summarized the finding in three points: types of text sources employed, analysis algorithms, and research results. However, the rapid development in the field, to some extent, makes those surveys outdated in presenting state-of-the-art approaches. Different from these surveys, this review focuses on discussing the application of NLP in stock forecasting and covers studies up to 2021.

In terms of the methodology, the study conducted the first search in December 2021, using the google scholar search engine based on the keywords (search terms) that related to text-based stock forecasting techniques in the title. The primary search terms are "text", "news", "sentiment", "event", "stock", "forecast", "predict", "neural network", "machine learning", "financial", etc., along with the Boolean operators ('AND', 'OR', 'NOT', etc.). The study browses the first 15 pages of search results for each of the recent five years and manually screens the appropriate article title before briefing the article's abstract to filter the irrelevant articles. Due to previous related review studies available in covering the earlier literature, this study mainly focusses on the last five years studies to discuss the recent development in the field. The study is further tracing the earlier significant studies from the literature of recent studies to review the development in a chronological manner.

For the selection criteria, the study attempted to address the relevant studies with a high citation score, followed by their journal quality which are reflected from the Web of Science (WoS) ranking, impact factors, etc.

Overall, this research summarizes and discusses the current text-based financial forecasting approach in the aspect of semantical-based, sentiment-based, event-extraction-based, and hybrid approaches in Section 2. Moreover, this study evaluates the advantage and disadvantages of each approach in Section 3 and suggests the future research direction in text-based financial forecasting in Section 4.

2. Application of NLP in Stock Forecasting

According to [9], NLP techniques majorly focus on processing the textual information into numerical features representation to enable the machine to understand the semantic meaning of text. In the financial aspect, the text sources can be the (1) corporate disclosure, (2) financial reports, (3) professional periodicals, (4) aggregated news, (5) message boards, and (6) social media. Each of the financial text sources might possess different properties based on their variety of length, subjectivity, frequency of updates, informative contents, etc. Therefore, the textual analytical techniques to transform the textual data into suitable meaningful representations are relatively important as the result could affect the entire outcome of the prediction task.

In the literature, the text-based stock forecasting approaches can be generally classified into three different methods which are semantical-based, sentiment-based, and event-extraction-based approaches. The categorizations are mainly based on the different textual analytical techniques applied in the financial text. In the earlier literature, researchers usually relied on the statical method to represent the financial text such as bag-of-words, n-grams, TF-IDF, etc. However, the traditional methods failed to take the dependency of text into account and suffered from the "curse of dimensionality" issue. The development of neural network provides solutions to overcome the problem above by introducing the word embedding technique, to semantically represent the text and facilitating the semantic-based approach.

However, it is obvious that the difference in financial text sources might have different properties and it is hard to ensure that the same method is effective for all text sources [7]. For example, the investors are likely to be passively or actively influenced by the text information during the investment decision-making process. The opinion review on social media might directly and actively influence the investor since the user comments are usually inherent with a strong opinion and stance. In such a way, the researcher might prefer another text analytical technique such as sentiment-based approaches to analyze the emotional state for stock movement forecasting. The sentiment-based approaches can be derived from the semantic-based approach to perform a further task training of sentiment analysis to transform the word vector into sentiment polarities or scoring.

Moreover, the semantic-based approach is limited in processing a long-text information due to the "long-range dependency" issue. In other words, a fixed size word vector is hardly to encode every information from a long text without any information loss. Based on this, the event extraction-based approach is derived to extract important information from text to produce a better financial text representation. Based on this, Section 2 attempts to discuss and summarize the approaches in text-based stock forecasting studies.

2.1. Semantic-Based Stock Forecasting Approach

In earlier stages, text-based stock forecasting studies usually rely on the bag-of-words (BOW) approach to quantify financial texts. This approach breaks up the text into a list of words where each word and the frequency of word appearance serve as features to represent the semantics of the text. In [21], the NewsCATS model was proposed to adopt the BoW approach to represent the press releases in vector form and applied KNN and SVM models to examine the news impacts on stock markets. The study [22] applied the BoW

approach and frequency-inverse document frequency weighting (tf-IDF) as textual features to represent the press releases before adapting multiple kernels learning for prediction.

Other than the Bow approach, later studies increasingly applied different textual features such as noun phrases, named entities, and tf-IDF to represent the text. The study [23] examined the breaking news effect on stock prices based on the SVM model with three separated textual features, which are BoW, Named entities, and Noun phrases. Their experimental results suggested that the noun phrases model yields a better result than BoW and Named entities models regarding directional accuracy, simulated trading, and closeness. In [24], TF-IDF was applied to represent the news headline of BBC and 20 newsgroup datasets before serving as input to SVM to predict the stock movement.

However, using the BoW approaches, it is difficult to distinguish the semantical relation between words instead of treating each of the words independently. This problem is critical to the BoW approach as the natural language is usually contextual-dependent. Based on this, the n-gram feature is introduced to track the problem above. Instead of extracting each word, n-gram takes a contiguous sequence of n items of words from a text sequence which might better capture the syntactic information between the words. In short, the n-gram method is developed based on the concept of lexical co-occurrence statistics. For example, the unigram where n = 1 indicated similar features with BoW. For bigram or trigram, two and three contiguous words are regarded as an entity [25].

In [25], the effectiveness of different n-gram features was examined using ad hoc announcement data. Their result found that the bi-gram features yield a better result than trigram and unigram. On the other hand, [26] criticized that the n-gram approach might suffer from the "Curse of Dimensionality" issue. It implied the scenario where the dimension of the co-occurrence matrix is booming along with the increasing vocabulary size. Consequently, it might sparser the data distribution and weaken the model's robustness. The n-gram feature is also limited to addressing the language-dependency problem. Based on this, the word embedding technique emerged to overcome the issues of the "curse of dimensionality" by learning the low-dimensional dense word vector directly [27]. Each dimension of the word embedding vector is regarded as a latent word feature, and the linguistic patterns are also encoded within the word vector. The word vector can be semantically composited, such as the vec('France') − vec('Paris') + vec('Japan') is close to vec('Tokyo'), and the similarity between the word vector can be measured via the cosine similarity.

The later study by [28] applied n-gram, and word2Vec approaches to represent the text of the tweet and adopted random forest algorithms to measure the correlation between the sentiment and stock price movement. Another similar study [29] attempted to examine the effect of the tweets on stock market trends based on the textual representation of BoW and word embedding. Although there is an improvement made on the BoW vocabulary size, the tweet vector in 300 dimensions generated by Word2Vec embedding is yielding a better result than BoW representation. Based on this, the strong capability of word embedding in measuring the word semantic was shown. This has induced the subsequence study to adopt this approach widely. For example, the study [30] applied theWord2Vec model to embed the financial news title of Reuter into sentences embedding before combining with seven technical indicators to predict the Index movement of S&P 500. Similar word2vec embedding was also applied to a Bidirectional gated recurrent unit (GRU) to predict the movement of S&P500 [31]. Other than English context, the study [32] applied the skipgrams model to transform Korean sentences into a vector before being adopted to a CNN model as input to forecast the stock price after five trading days.

Instead of assigning each word into a distinct vector based on word embedding, the study [33] adopted the character level embedding to retain the sub-word information of news for S&P500 movement prediction. It is mentioned that the character level embedding may capture additional word's morphology information and mitigate the limitation of word embedding where the ability is limited to deal with the unknown word missing in the training corpus which is known as the out-of-vocabulary (OOV) issue. Moreover, some

studies attempted to embed the entire sentences or paragraph into a vector for prediction such as applying the paragraph vector for newspaper articles embedding [34] and applying paragraph vector to embed the online financial news into a vector before using the Deep Neural Generative (DNG) model for prediction [35].

Furthermore, it is not surprising that the quality of articles might significantly affect the performance of text-based financial forecasting. However, the problems of the online contents in practice are the unstable in quality, trustworthiness, and comprehensiveness. Based on this, the study [5] proposed the way to process such chaotic online news with the three principles of (1) Sequential Context Dependency where the news is interdependent, (2) Diverse Influence where the news impact varies across each other, and (3) Effective and efficient learning. In short, the study proposed the Hybrid Attention network (HAN) that consists of two different attention modules of (1) news level attention to effectively capture important news information and (2) temporal attention to better capture the time-varying effect to address the issue above.

However, there is another issue associated with the word embedding approach, which is the long-range dependency problem. It is hard to compress the entire information of a long text, such as a document, into a fixed vector without any information loss. The noise of irrelevant text might also affect the quality of the word vector. Based on this, numerous studies [30,33,35] analyzed the news title instead of the content to reduce the associated noise. The study [36] claimed that the news title usually is a short description and ignores extensive essential information. Based on this, the news abstract is employed as the target to weigh the sentence in news content and generate a target-specific abstract-guided news document representation to reduce noise while preserving important information in news content [36].

In the study [37], the hierarchical complementary attention network (HCAN) was proposed which composed of two attention mechanisms: (1) word-level attention to capture the significance words in news title and content, and (2) sentence-level attention to capture the significance of each sentence in content. The title representation and content representation are concatenated to predict the stock price movement. Moreover, the study [38] attempted to improve the interconnection between the market and news data by proposing the numerical-based attention (NBA) method to align the news embedding with the stock vector. Ideally, the model enables to assign more weight of news impact to its relevant stock.

On the other hand, there is another research direction to adapt the Word2Vec algorithm in sector embedding to predict the company stock properties, namely, Stock2Vec. In the study of [39], the idea of Stock2Vec is to vectorize the stock information in a low dimension vector. Based on this, the stocks with similar properties are likely to appear in a same neighborhood in the vector space. Their study claimed that the modelling of the intrinsic relationship among stocks could improve the predictive performance. In the study of [40], Stock2Vec embedding was applied in predicting the stock movement. Specifically, the study preprocessed the financial news and labeled the news and stock prices with their corresponding polarities, either positive or negative. Afterward, a Bidirectional gated recurrent unit (BGRU) was applied to learn the Stock2Vec embedding based on the labeled news, Harvard IV-4 dictionary and daily S&P 500 Index. Their result emphasized the effectiveness of Stock2Vec as compared to Glove and Word2Vec since the sentiment value of words was taken into consideration. Similarly, the study [41] trained the Stock2Vec based on the stock news and sentiment dictionaries. Different from the study [40], they employed additional political news and stock forum speech for sentiment measure and trained the Stock2Vec on CSI 300 stock data, using the LSTM-based model. Table 1 summarizes the findings of Semantic-based stock forecasting studies.

Table 1. Summary table of semantic-based stock forecasting approach.

Ref	Year	Model	Method
[34]	2016	Long Short-Term Memory (LSTM)	Applied paragraph vector method to embed the newspaper articles
[30]	2017	Recurrent Convolutional Neural Network (RCNN)	Applied word2vec to generate word vectors of the news title and combine with technical indicator vector for prediction
[5]	2018	HAN	Applied two attention modules (1) news level attention and (2) temporal attention to better capture asymmetric news and time dependency effect
[38]	2018	NBA	Proposed (NBA) method to improve the dual sources interaction between market and news data for stock market prediction
[36]	2018	Document embedding model	Using news abstract as target model to weight the sentence in news content and generate a target specific abstract guided news documents representation
[37]	2018	HCAN	Proposed HCAN model with two attention mechanism (1) word-level attention and (2) sentence-level attention to retain important text information and mitigate the noise of irrelevant text
[32]	2019	BGRU	Applied word2vec to obtain the word vectors of the financial news before adapting (BGRU) for prediction
[41]	2021	Stock2Vec	Vectorized the stock information in conjunction of news sentiment as stock representation

2.2. Sentiment-Based Stock Forecasting Approach

Sentiment analysis (SA) is regarded as one of the NLP applications to study people's opinions, emotions, and attitudes toward a specific event, individual, or topic [42]. In the financial aspect, sentiment analysis plays a vital role in transforming the unstructured financial text into the sentiment signal to reflect the inner thoughts of the investors. The study [43] observed that the market price is having a downtrend after a high appearance of pessimism-scored reports. The finding above supports that the emotional text information is key in revealing investor expectations when reacting to different available text information. Therefore, sentiment analysis has become one of the famous research directions in the financial area.

In the past literature, sentiment-based financial forecasting usually involves a two-step procedure in which a sentiment analyzer is first applied to extract the sentiment features of the text sources and subsequently adapts the sentiment features to a prediction model for stock forecasting. According to the study [44], existing sentiment classification techniques usually fall into three categories: (1) lexicon-based approach to compound a sentiment dictionary to determine the sentiment polarity of each word, (2) Machine Learning approach to train a sentiment classifier based on the statistical semantic features such as n-gram, Term frequency-inverse document frequency (TF-IDF), bag-of-word, etc., and (3) Deep Learning approach to automatically extract the feature representation via the neural network for sentiment measure. In addition, the study [45] reviewed the sequential transfer learning approaches in analyzing different sentiment-oriented tasks.

Due to simplicity and efficiency, the lexicon-based approach is the most famous approach in sentiment-based financial forecasting. Once the sentiment wordlist is compiled, the researcher may easily measure the corresponding text sentiment. The existing approaches to compile the sentiment wordlist fall into two categories: (1) Dictionary-based and (2) Corpus-based approaches. The former was to construct the sentiment word list based on the synonyms and antonyms of predetermined sentiment words in the general dictionary such as WordNet [46], ConceptNet [47], SentiWordNet [48], Harvard General Inquirer, Henry Wordlist [49], Opinion Finder [50], etc. In contrast, the latter was to exploit the syntactic pattern of co-occurrence words in the corpus to compile the sentiment wordlist. Other than this, the study [51] attempted to combine both approaches to develop a specialized financial lexicon statistically and semantically.

However, the general sentiment dictionary is criticized for performing weakly in domain-specific sentiment. For instance, [52] found that around 73.8% of the negative words in the Harvard general inquirer was not considered as negative in the financial domain. Thus, the Loughran and McDonald's wordlist (Financial lexicon) was developed

based on the corpus-based approach to exploit the financial sentiment wordlist from the U.S. Securities and Exchange Commission report. Based on this, [53] applied the McDonald dictionary and AffectiveSpace2 to extract the sentiment embedding of the summarized financial news article that related to the twenty most capitalized companies listed in the NASDAQ 100 index. The sentiment embedding is adopted together with technical indicators for market analysis.

The study [15] mapped the words of the financial news onto the emotional spaces of two different dictionaries, Harvard IV-4 Dictionary and Loughran–McDonald Financial Dictionary. The study found that the dictionaries-based sentiment features yielded a better result than the bag-of-words model. In addition, the study [54] pre-processed the news headline with the approaches of Relevant Gloss Retrieval, Similarity Threshold, Verb Nominalization before applying the SentiWordNet for sentiment score measuring. Their study showed that identifying a proper sense of significant words in news headlines is useful to improve the sentiment-based market forecasting. Other than the open-source dictionary above, numerous studies attempted to develop the financial lexicon by themselves to enhance the quality of sentiment measures. For instance, the study [55] manually constructed a specific sentiment wordlist in the pharmaceutical domain to identify the sentiment polarity of each word. The sentiment score was computed based on the sentiment word count and directly applied for trading decisions. Another study [56] developed an alternative financial lexicon using financial microblogging data. The corresponding lexicon comprises 7000 unigrams, 13,000 bigrams of words with their respective sentiment scores, and considers the affirmative and negated contexts.

Afterward, the subsequent study [56] applied the proposed lexicon to measure the sentiment of each Twitter message, and the results are aggregately used to compute various Daily Twitter Sentiment Indicators, including Bullish Ratio, Bullishness Index (BI), Agreement (AG), and Variation (VA). The study applied the Kalman Filter (KF) procedure to combine the Daily Twitter Sentiment Indicators with weekly American Association of Individual Investors (AAII), Investors Intelligence (II) values, monthly University of Michigan Surveys of Consumers (UMSC), and Sentix values. The aggregated sentiment index used to predict the daily returns, trading volume, volatility of various indices, such as Standard & Poor's 500 (SP500), Russell 2000 (RSL), Dow Jones Industrial Average (DJIA), Nasdaq 100 (NDQ), and portfolios (e.g., formed on size and industries).

The study [57] combined the sentiment features measured by four different lexicons of NTUSD, HowNet-VSA, NTgUFSD, and iMFinanceSD with the statistical word information to measure the relationship between the financial news and stock price trend. The result emphasized that the sources of financial news do affect the accuracy of stock forecasting. Another study [58] proposed to adapt the existence and intensity of emotion words as a base to classify the sentiment of the financial news. The study employed the context entropy model to measure the semantical similarity between two words by comparing their contextual distributions based on the entropy measure. In the study [59], the sentiment on several major events in four different countries were examined and the result indicated that the event sentiment is improving the predictive result.

In a similar study, the authors of [3] applied two different mood tracking methods: (1) Opinion Finder for binary moods examining (Positive and Negative), and (2) Google Profile of Mood States (GPOMS) for fine-grained mood examination such as (Happy, Kind, Vital, Sure, Alert, Clam) to analyze daily twitter content. The psychological features are then to be proven to improve the predictive performance for the Dow Jones Industrial Average (DJIA) index. They trained a SOFNN (Self-Organizing Fuzzy Neural Network) and showed that one of the six mood dimensions called "Calm" was a statistically significant mood predictor for the DJIA daily price up and down change. The study [60] applied the sentiment analysis on the financial web news, forum discussions, and tweets with google trends to predict the Ghana stock market movement. The combined dataset achieved the highest predictive accuracy ranging from (70.66–77.12%) in a different time window.

In the aspect of Chinese microblogging data, there is a study [61] that selectively filtered the Weibo posts related to three influential topics based on the relevant keyboard. After that, Chinese Emotion Word Ontology (CEWO) was applied to measure the sentiment score in 7 categories (Happiness, Sadness, Surprise, Fear, Disgust, Anger, and Good). The discrete sentiment score was aggregated to construct the daily emotional time series for each category. The finding indicated that the public mood states of "Happiness" and "Disgust" has cause an obvious change in China stock price. Moreover, the Tsinghua Sentiment Dictionary was expanded by [62] to add specific financial terms and adjectives to analyze the sentiment of the Sina Weibo (Chinese social network) post. The sentiment score and technical indicators then served as the input to a novel "Deep Random Subspace Ensembles" (DRSE) model for market forecasting.

The study [63] constructed the "aggregate news sentiment index" (ANSI) based on the term frequencies of the optimism and pessimism characteristic terms in Chinese financial news to study the relationship between financial news and the Taiwan stock market. Their results suggested that the sentiment level of the financial news has a significant effect in constructing the financial portfolio. In addition, the study by [64] applied the Word2Vec model to transform the user comments from (www.eastmoney.com, accessed on 1 December 2021) into a textual representation before employing CNN to measure the user's bullish-bearish tendencies. It is highlighted that the user's bearish tendencies implied a higher market volatility and reflect a possible higher market return. Moreover, there is a SENTiVENT corpus presented by [65] consisting of the token-level annotations for target spans, polar spans and polarity, and training the model based on corresponding annotations for stock movement prediction.

Furthermore, some studies applied the open-source sentiment tools such as TextBlob API [66,67] and DeepMoving [68] to process the Tweet text into sentiment polarity or scoring before composing a sentiment index to forecast the market. However, the study [63] pointed out that the measure of sentiment polarities is not sufficient as the sentiment polarities are dynamic across different topics or domains. For example, the word "low" in the phrase between "low tax" and "low profit" is having an opposite sentiment meaning. Therefore, the Latent Dirichlet Allocation (LDA) [69] is adopted as a topic model to compute the topic distribution over words. The study [70] adopted the LDA to filter the unrelated topic from financial microblogs ("Weibo") and then applied the financial lexicon to obtain the sentiment polarities to forecast the market index. Similar studies by [71,72] performed topic-based sentiment analyses to predict the stock market.

Recently, the pretrained model has achieved the state-of-art results across different NLP downstream tasks, with no exception of sentiment analysis. Thanks to the advanced computing power and massive textual data available on the web, researchers enabled to self-supervised training a model from corpus to learn the common knowledge that can be transferred and benefit different NLP downstream tasks. In sentiment-based stock forecasting, [73] examined the relative effectiveness of four different sentiment models named SentWordNet, logistic regression, LSTM, and BERT. Their result indicated that the BERT model is outperforming the others in sentimental sequential forecasting. The authors of [74] applied the BERT model in analyzing the Chinese stock reviews and achieved a higher precision in sentiment analysis result. Moreover, the recent study [75] proposed the FinALBERT model to leverage and fine-tune the pretrained FinBERT models for stock movement prediction. Table 2 summarized the findings of sentiment-based stock forecasting studies.

Table 2. Summary table of sentiment-based stock forecasting approach.

Ref	Year	Model	Method
[3]	2011	Self-Organizing Fuzzy Neural Network	Applied two mood tracking methods (1) Opinion Finder and (2) Google Profile of Mood States (GPOMS) to analyze daily twitter content for prediction
[61]	2016	Neural Network	Applied Chinese Emotion Word Ontology (CEWO) to measure sentiment score in 7 categories and aggregated the score into a daily emotional time series
[70]	2016	user-group model	Applied Latent Dirichlet Allocation (LDA) to filer irrelevant topic of financial microblogs from "Weibo" and then applied the financial lexicon to obtain the sentiment polarities to forecast the market index.
[53]	2019	Feedforward Neural Network	Applied McDonald dictionary and AffectiveSpace2 to extract sentiment embedding
[63]	2017	Vector AutoRegressive (VAR)	Measure the market optimism and pessimism to construct the aggregate news sentiment index
[64]	2020	CNN	Applied CNN to measure the user's bullish-bearish tendencies for prediction
[74]	2021	BERT	Applied BERT to measure the sentiment of Chinese stock reviews
[75]	2021	FinALBERT + BiSTM + Attention	Fine-tuned the FinBERT model to extract the sentiment information to predict the stock prices
[68]	2022	Multimodal AdaBoost-LSTM ensemble approach	Model the multimodalities data such as media sentiment, trading data, volumes, blockchain information to predict the Bitcoin price

2.3. Event Extraction-Based Stock Forecasting Approach

The event extraction approach is another NLP application in financial forecasting. Unlike the previous semantic-based approach to process the entire sentence, paragraph, and document of texts, event extraction focuses on retrieving essential event information from the text and representing the event in a structural form. The core objective is to distill vital information to reduce the noise of irrelevant text. The definition of an event is defined as a specific occurrence of something that happens in a particular time, a particular place that involves one or more people, which can be described as a change of state [76]. The task of event extraction is to detect the event-mentioned sentences from the text based on the event triggers (keyword in identifying the occurrence of a specific type of event) and identify the event type as well as its event arguments [77].

In other words, event extraction summarized the unstructured natural languages into a structural set of linked relations which can describe the "5W1H" questions of "Who, when, what, where, why, and how" of a real-world event. The structural event representation can be further adapted for logical reasoning or inference. Thus, the study by [78] claimed that the news event might change the state of investor's mind, trigger their trading action, and influence the stock movement. The applications of event extraction are penetrating the business and financial domain, such as helping the companies to rapidly discover the market responses, inferencing signals for the trading suggestion, risk analysis, etc. [79].

In the literature, the ViewerPro system is employed in [80] to extract companies' events from Reuters news articles that related to the FTSE 50 stock index. The ViewerPro system filtered irrelevant news and identified events through pattern matching on a domain-specific knowledge repository. Subsequently, the study [78] was the first to represent the structured news events in the tuple of (Actor-Action-Object-Time) based on the Open-IE approach to predict the stock movement of S&P 500. However, the predictive performance based on the structured events tuple represented by one-hot feature vectors is limited by the sparsity issues of discrete features in statistical models, thus, [81] improved their study by representing the structured events into the dense vector event embeddings. Specifically, the three components of word embeddings for "Agent", "Predicate", and "Object" are extracted from the raw text and combined to produce the structured event embedding of (A, P, O). The goal of event representation learning is to ensure that similar events should be embedded close to each other in the same vector space, and different events should be far from each other.

As a result, the event embeddings led to better stock market prediction than the formal discrete event-based approach [78] since the structural relations can be captured via the semantic compositionality. In addition, they found that the CNN model is better in capturing the long-term event effects. In later studies, the technique of structured-event embedding is widely applied such as the study [82] to forecast the index movement of S&P 500 based on the financial news of Reuters and Bloomberg. However, there is another study [83] pointed the limitations [82] where the events of small companies were neglected since the relevant reporting was limited. The authors of [84] combined the merit of event embedding [81] based on news headline of "Reuters", "Reddit", and "Intrinio" with the same set of technical indicators of [30] and historical prices to improve the predictability of Index movement of S&P500 and DJIA.

On the other hand, the shortcoming of event embedding [81] was pointed out in their later study [85], where the relationship between the similar semantic and syntactic events are not captured if they do not have similar word embedding. Moreover, the approach is constrained by assuming that the event with similar word embedding will have similar semantics. For example, the event embedding of "Peter quits Apple" and "Steve Jobs leaves Apple" are reflecting a huge semantic difference since "Peter" is the customer where "Steve Jobs" is the CEO of Apple, but the event embedding does not capture the semantic differences [85]. Therefore, the study [85] improved the quality of event embeddings by incorporating the knowledge graph into the training phase to encode the background information.

Moreover, the work [83] pointed out that the previous studies neglected event characteristics which may seriously degrade the predictive performance. The study mentioned three different event properties which are the "Imbalanced distribution of events", "Inconsistent effect of the event", "Distinct important of events", and the "Temporal effect of the event". The property of the imbalanced distribution of events suggested that the financial news usually tends to report for big enterprises rather than small enterprises. It may cause an imbalanced reporting volume where the event dictionary will be either too sparse or too dense for different stocks. Secondly, the event effect is typically inconsistent and diverse across the industries. For example, the news of "Rebound of COVID-19 in Malaysia" will be positive against the healthcare industry but negative against the tourism industry. The reasons behind are due to the increasing need of medical supplements, while reducing of public contact. Thirdly, the magnitude of events' impact is diverse across the news, indicating the need to distinguish the significance of events. Lastly, the event usually has a different long-lasting effect, indicating the causality relationship or dependency between the events.

Thus, the study [83] attempted to address the event characteristic above by proposing the event attention network (EAN) to exploit the sentimental event embedding in which the event effect with the sentiment properties is simultaneously captured to improve the prediction toward 20 various companies' stock trend in Hong Kong and Shenzhen Market. Specifically, the event information is extracted from "Finet", "Tencent News", "Sina News" and structurally embedded into (Time-Location-Name-Action). The attention mechanism is adopted to distinguish the particular importance of events where the Bi-directional LSM with CNN layer is designed to capture the sequential dependency of events and extract the stock-driven feature representation. Meanwhile, the sentiments are analyzed from the social media platform of "East Money", "Facebook", "Twitter" and to be classified into six dimensions (Happy, Vital, Kind, Sure, Clam, Alert). The inclusion of additional sentiment information does have a significant improvement on the predictive result.

On the other hand, the study [4] criticized that the coarse-grained event structure such as (S, P, O) [7,78] may omit specific semantic information of different types of events, thus proposing the Japanese financial event dictionary (TFED) to extract the fine-grained events automatically from financial news. Generally, the TFED specified the type of financial events and their corresponding trigger words and event structures. For example, trigger words such as (acquisition, merge, acquire) are used to detect the M&A event, where (fund,

funding) indicates the Funding event. The event details will be further extracted in their corresponding structure (Firm-Time-Method) and (Who-Action-Target-Time-Location). The study introduced the Multi-task Structured Stock Prediction model (MSSPM) to jointly learn the event extraction and stock prediction since both tasks are highly correlated. In a later study [86], the structured events were extracted from the text before they were transformed into an event vector to learn the correlation between the events.

Moreover, recent studies [87] attempted to enhance the predictive model, such as proposing the CapTE (Capsule network based on Transformer) to learn the deep semantic information and structural relation of tweets. In contrast, the study [88] introduced the dilated causal convolution networks with attention (Att-DCNN) to produce the event knowledge embedding to learn the direct and inverse relationship among the events and to take the financial indicators into account for index prediction based on S&P500. Both studies outperformed the previous baseline model with an accuracy of 64.22% by CapTE and 72.23% Att-DCNN on different datasets, respectively.

Furthermore, recent studies attempted to exploit the potential of graph neural network in modelling the interrelation between the events and stocks. For example, the study [89] proposed a relational event-driven stock trend forecasting (REST) framework to capture the stock-dependent influence and Cross-stock influence. Their study observed that the effect of an event is having a couple of properties in practice. Firstly, the event effect is varying on stocks connected with different relations. Secondly, the event effect is having a dynamic propagation strength between two stocks. Thirdly, the event effect could take a multi-hop propagation across the stocks. Based on this, the study proposed REST with three distinct components namely (1) Event information encoder to compute the event representation, (2) Stock context Encoder to model the stock context information, and (3) Graph convolutional network to capture the stock-dependent and cross-stock influence for stock trend prediction.

On the other hand, the study [90] highlighted that the event impact might have different speeds across different stocks. This might cause co-movement, yet the movement is asynchronous over time which is known as lead-lag effect. For instance, "Qualcomn suites against Apple" may have a direct impact on both companies but will also influence the upstream and downstream-related companies. Based on this, the study [90] proposed a multi-modality graph neural network (MAGNN) with inter-modality sources attention and inner-modality graph attention to capture the complex relation between multimodality sources. The multi-modality sources such as historical prices, News, and events, along with a knowledge graph are employed as input to predict the financial time series. The overall findings of event extraction-based stock forecasting studies are summarized in Table 3.

2.4. Hybird Approach in Text Based Financial Forecasting

In Section 2.4, the past studies involving more than one category of NLP approaches are discussed. These studies attempted to process each different source with different methods. In [91], a tensor-based stock information analyzer named (TeSIA) was proposed to improve the stock forecasting result. The study categorized the investor information sources into three types of firm-specific, event-specific, and sentiment modes. Specifically, the study represented the firm-specific mode with fundamental market data (price, trading volume, P/E ratio, P/B ratio) and used the BoW approach to process the news articles into the term vector from noun and sentiment words to represent the event-specific modes. The study followed the sentiment analysis approach [92] in processing the posts from two financial discussion boards, "Sina.com" and "eastmoney.com", into three sentiment features of market sentiment intensity, optimistic and pessimistic mood. The information above was further composed and processed by a tensor model to ensure the interconnection between different information. In the base of the TeSIA model, the study [93] expanded the study by enhancing the capability of the predictive model with a tensor-based event-driven LSTM model. In addition to technical indicators and media sentiment, the study incorporated the correlations among companies to capture the indirect influence from the relevant company.

Experimental results showed that the proposed approach outperformed the baseline model by at least 22.8% over AZFinText, eMAQT, and Tesia.

Table 3. Summary table of event extraction-based stock forecasting approach.

Ref	Year	Model	Contribution
[78]	2014	Event-Neural Network	First attempt to represented news events in a structured tuple of (Actor-Action-Object-Time) based on Open-IE approach
[81]	2015	Event Embedding-CNN	Improve the discrete structure events into dense vector structured event embeddings (Agent, predicate, object)
[85]	2018	Knowledge Graph Event Embdding CNN	Incorporated knowledge graph into training phase to encode background information to improve the representation of event embedding
[83]	2019	EAN	Incorporated sentiment information and adapted the attention mechanism to model the asymmetric event impact and time dependency effect to improve the prediction
[4]	2019	SSPM	Proposed a novel Japanese financial event dictionary (TFED) to extract fine-grained structured events information and fuse with news text to generate structure-aware text representation. SSPM model employed the structured events as distant supervised label to further training a multi-task framework for both event extraction and stock prediction
[88]	2020	Att-DCNN	Proposed (Att-DCNN) model to better learning the direct and inverse relationship between events to improve the event-knowledge embedding for prediction
[89]	2021	REST	Applied Graph convolutional network to model the stock-dependent influence and cross-stock influence in particular events
[90]	2022	MAGAN	Applied multi-modality graph neural network to integrate the event impact, news, stock prices for financial prediction

The study [7] claimed that the previous studies that are merely focusing on a single data source might lose important information and degrade the predictive performance of the stock movement. Thereby, the "multi-source integration model" is proposed to integrate heterogeneous information in the aspect of market data, public sentiment, and structured events into a comprehensive system to predict the movement of the Shanghai Composite Index. In their study, the LDA-S method was adopted to extract the topic sentiment, while the HanLP method was used to extract the main structure event's information from sentences before applying the Restricted Boltzmann Machine (RBM) approach to generate the pretrained vector as the input to sentence2vec for event embedding. The study found that the news event is the most influential predictive factor among each other.

On the other hand, there is another research direction to address the stock correlation measure based on the graph neural network. For example, the study [94] proposed the LSTM Relational Graph Convolutional Network (LSTM-RGCN) to embed the news headline as the node features in a graph structure and learn the interconnectivity between the node (news correlation). Afterward, the node vectors are employed to forecast the overnight stock price movement. Another study by [95] proposed the multi-GCGRU model by integrating the GCN with gated recurrent unit (GRU) to model the cross-correlation features for temporal stock dependency measure. The authors of [96] applied the graph network to encode the three different sources (trading data, stock news, and graphic indicators) into graph data before aggregating them for prediction.

3. Summary

In Section 2, the overview text-based stock forecasting has been discussed in three different approaches of semantical-based, sentiment-based and event-extraction-based methods. Section 3 attempts to discuss the strengths and weaknesses of each approach and analyze their application in different scenarios.

In earlier studies, text-based financial forecasting studies mostly relied on the word statistical information for representation such as bag-of-words, n-gram, TF-IDF, proper nouns, etc. However, those approaches failed to take the word ordering into account instead of treating each word independently. Along with the increasing vocabulary size, the word statistical matrix will be significantly expanded and ultimately result in the "curse of dimensionally" issue. Based on this, the semantical-based approach provided the way to

overcome the issue above by introducing the word embedding method to represent the text in a low-dimensional dense vector. The better textual representation is the key to enhance the performance of text-based stock forecasting.

Recently, the development of sequential transfer learning has benefited the semantical-based approach where a large language model is first pretrained in a large corpus before further fine-tuned to better extracting the contextual semantic of a text. This may lower the application difficulty and reduce the computational overhead to lower the computational cost since the model can be initialized by the parameter of the pretrained model instead of training from scratch. Although the semantic-based approach may understand the word semantic via the text encoding process, it is still difficult to encode a long text such as a long paragraph or document into a fixed-size vector without information loss [30,33,35]. In other words, the fixed-size vector is limited to represent the entire long text information and the poor feature representation is likely to degrade the predictive performance. Thus, the long-range dependency issue is remaining a major challenge in this approach.

Moreover, the long text content probably consists of irrelevancy which might disrupt the information encoding process. The redundancy of text might result in noise, affect the quality of text vector representation, and deteriorate the final predictive performance [36]. As compared to sentiment or event-extraction-based approaches, the noise effect of long text can be reduced by feature transformation or the information extracting process. Although the accuracy of the sentiment measure can be affected by the long-text noise, the simplicity of feature transformation into sentiment polarities or scoring can reduce the redundancy of the predictive model and mitigate the underlying effect. For event extraction, the event information is retrieved to neglect the redundancy of information which is helpful in long-text scenario.

However, the sentiment-based approach failed to identify the interconnectivity effect between different time-steps [63,64]. The sentiment features are likely to be independent between each other. For example, the comments from individual A and B against a specific event should be independent in practice. While converting into sentiment features, the opinion impact between different sources is neglected. The opinion impact from the US president and a common individual are obvious different; however, the approach is limited to distinguish the impact above and treat every opinion standardly. As compared to the Semantic and event-extraction-based approach, additional information can be retrained in a high-dimensional vector which allows the approach to distinguish the asymmetric impact and capture the difference between different contents. Both approaches also have the strength to capture the interconnectivity between different information across timesteps.

Furthermore, the encoded information in vector form is unable to be interpreted by a human. The neural operation is a black box which might hinder the process of reasoning. Although several good predictive results were achieved by [5,37,38], an investor might still be confused with the result and doubt the reason behind the prediction. In the financial aspect, logical reasoning is relatively important since the financial decisions are interrelated between each other for planning the investment strategies, reducing the investment risk, and avoiding financial losses. As compared to the sentiment-based approach, the word vectors are further transformed into the sentiment scoring or polarities as input features to a predictive model which has a better transparency in contributing the final predictive result. However, neither semantic nor sentiment-based approaches are limited in conducting the logical reasoning process. Both approaches are unable to draw the causal relationship between different events and scenarios. For event extraction-based approaches, it is possible to apply the knowledge graph technique to link the associated events and model the causal relationship for explanation [85].

For the semantic-based approach, it is highlighted that the domain similarity between the pretraining model and downstream application is relatively significant. The large discrepancy between the two domains might result in a situation of negative transfer and deteriorate the performance of the sentiment measure. For example, "underestimate" is negative sentiment but indicates a favorable investment opportunity to investors in the

financial domain. Therefore, the researcher is required to identify the appropriateness of pretrained embedding or increase the computational cost to refine the embedding approach with a financial text. For the event extraction-based approach, it remained a challenging task since it involved a series of NLP subtasks such as event detection, document classification, argument identification, etc., thereby contributing the difficulty for application.

Moreover, there is a limitation in recent sentiment-based stock forecasting studies where those studies mostly focus on classifying the general financial text such as sentence level, paragraphs level, or document level. The neglection of fine-grained information is critical and might cause of misleading. For example, "A is good, but B is bad". Both positive or negative classification results for the entire sentences are not accurate and the general classification will loss of more information. As compared to semantic-based or even-extraction based approaches, the information of noun can be encoded within a vector and the vectorized information is valuable for subsequent stock prediction.

Lastly, [7] pointed out that the online textual contents from distinct sources may function well in a different way due to the different properties of text sources. The social media content usually comes with a strong sentiment sense and lexically incomplete which is more suitable for sentiment analysis. For example, the social media comment of "Stock A is good because of yesterday announcement". The event extraction approaches are unable to identify any factual information from the example above and critically affect their performance. On the contrary, the expert financial news usually with the property of nature of professionalism (e.g., neutral statement) and the comprehended contents are more suited than social media for event extraction. Based on the discussion above, the researchers are recommended to consider the text properties such as subjectivity, length of text, and content comprehensiveness or the strength and weakness of each approach in deciding the text-based stock forecasting approaches. The overall methodology discussion is summarized in Table 4.

Table 4. Summary of discussion in comparing different approaches.

	Semantic-Based	Sentiment-Based	Event-Extraction-Based
Logical reasoning from predictive result	No	No	Yes
Identify the asymmetric content impact	Yes	No	Yes
Identify the interconnectivity impact between different time steps	Yes	No	Yes
Consideration of fine-grained information in text	Yes	Yet to be reached	Yes
Content Subjectivity	Applicable in any subjectivity content	More suitable for subjective opinion contents	More suitable for objective fact contents
Content detail and comprehensiveness	Applicable in both general or detail contents	Applicable in both general or detail contents	Perform poorly when content details are lacking
Length of text	Perform poorly in long text scenario	Perform poorly in long text scenario due to accuracy affection in sentiment measure	Perform better in long-text scenario to reduce redundancy in text
Example of applicable Financial text sources	News, Financial report, corporate disclosures, etc.	Social Media, discussion board, etc.	News, Financial reports, etc.

4. Discussion on Future Prospect

Financial time series with the properties of a high degree of variability, nonlinear nature, and the presence of noise are significant challenges to stock forecasting. Although the NLP techniques enable the model to capture insight from a financial text, it still has limitations. In Section 4, this study attempts to discuss the future stock forecasting research based on the weakness of each approach presented in Section 3.

4.1. Aspect 1: Aspect-Based Sentiment Stock Forecasting

In Section 3, it is mentioned that the opinion text is usually composed with the characteristic of diversity to express a varied opinion on different aspects. Due to computational efficiency, current sentiment-based stock forecasting studies mainly focus on measuring text sentiment at either document or sentence level. Although the emerged word embedding and deep learning approaches provide a way to globally analyze the semantical and syntactical text information, ignoring such fine-grained sentiment information is critical and might cause of misleading. For example, the statement of "I love stock A but hate stock B". The general sentiment model is limited to distinguish the correct opinion on its corresponding aspect. Thereby, the general evaluation is a constraint and might not be accurately analyzed. Based on this, a future sentiment-based forecasting study is suggested to looking into aspect-based sentiment analysis.

4.2. Aspect 2: Capability of Contextualized Text Encoder

From Section 3, it is obvious that the semantic-based and sentiment-based approaches might suffer from the long-range dependencies issue where the model is limited to encode a long text into a fixed vector without information loss. In the long text scenario, the content is possible to contain extensive irrelevancy and noise. The high redundancy in the text might result in the low quality of textual representation and affect the text-based predictive performance. In a recent study, the pretrained model (BERT) based on the transformer architecture was able to process a limit of 512 tokens. In practice, the word count of a financial report or a corporate disclosure report is overwhelmingly exceeding the processing limit. Based on this, the following research question is raised: "how to effectively identify important information while eliminate the text noise to generate a better textual representation".

4.3. Aspect 3: Stock Correlation Measure Based on Diverse Information

The chaos of financial time series is the biggest problem to be addressed in stock forecasting. Companies' business relations are incredibly complex, involving supply–customer relationships, creditor–debtor relationships, business cooperation, shareholding, or even a multinational business. Such multiple correlations between companies formulate a vast and complex enterprise network. Based on this, the correlation between stocks is not deniable to become the critical factor in financial forecasting since the correlated stocks might cause co-movement. However, earlier forecasting studies measured the correlation between stocks that were merely based on the similarity between corresponding stock price data.

The stock correlation might also be a challenge in text-based financial forecasting. Although news and opinion on social media can be quantified and further processed, the company background information and its business interrelation might be ignored. It is not surprising that the current mainstream research primarily focuses on processing the concurrent semantical text information for stock forecasting and ignoring the interrelation between business. For example, "The incoming iPhone 14 with unique features is dominative". Such a piece of news might also affect its competitor's stock prices in addition to the origin beneficial company of "Apple".

Another news headline example is "A (poor performance) resign from the company". Commonly, the sentiment of news is negative. However, it consists of the implicit meaning to refresh the adverse situation of the company. It might cause of misleading when the background information is omitted. Thus, future study is encouraged to explore the knowledge graph in adapting the common knowledge to enhance the application. In the recent study, the graph neural network has achieved a great result in learning the correlation between stocks. The method enables to embed the textual data, historical data or chart data into a graph structure and learn the interconnectivity between the nodes of graph to facilitate the forecasting performance. Future study is encouraged to fuse different sources information for prediction based on a graph neural network.

5. Conclusions

This review surveyed noteworthy research on stock forecasting based on natural language processing approaches. First, the study reviews different NLP approaches in stock forecasting and categorizes them into semantic-based, sentiment-based, and event-extraction-based, and hybrid-based methods. Afterwards, the study discussed the strength and weakness of each approach, compared their differences and discussed their suitable application scenario. Finally, the study highlighted the future research direction in applying the aspect-based sentiment stock forecasting, enhancing the capability of a contextualized text encoder, and fusing diverse information for stock correlation measure. The review attempted to assist people from different backgrounds to easily understand the corresponding approaches and laid a foundation to advance the field of stock forecasting.

Author Contributions: W.K.C. and K.T.B. investigated the ideas, reviewed the approaches, and wrote the manuscript; S.M.H.L. provided the survey studies and methods; J.Y.-L.C. conceived of the presented ideas and wrote the manuscript with support from Y.-L.C.; Z.-W.H. provided the suggestions on analytical writing; W.K.C. and Y.-L.C. both provided funding supports. Both authors W.K.C. and K.T.B. contributed equally to the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Fundamental Research Grant Scheme provided by the Ministry of Higher Education of Malaysia under grant number FRGS/1/2019/STG06/UTAR/03/1. This work was also supported by Ministry of Science and Technology in Taiwan, grant MOST-109-2628-E-027-004-MY3, MOST-110-2218-E-027-004, and MOST-110-2622-E-027-002.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: There are no data applicable in this study.

Conflicts of Interest: No potential conflict of interest were reported by the authors.

References

1. Fama, E.F. Efficient capital markets: A review of theory and empirical work. *J. Financ.* **1970**, *25*, 383–417. [CrossRef]
2. Malkiel, B.G. Efficient market hypothesis. In *Finance*; Springer: Berlin, Germany, 1989; pp. 127–134.
3. Bollen, J.; Mao, H.; Zeng, X. Twitter mood predicts the stock market. *J. Comput. Sci.* **2011**, *2*, 1–8. [CrossRef]
4. Chen, D.; Zou, Y.; Harimoto, K.; Bao, R.; Ren, X.; Sun, X. Incorporating fine grained events in stock movement prediction. *arXiv* **2019**, arXiv:1910.05078.
5. Hu, Z.; Liu, W.; Bian, J.; Liu, X.; Liu, T.-Y. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, Marina del Rey, CA, USA, 5–9 February 2018; pp. 261–269.
6. 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]
7. Zhang, X.; Qu, S.; Huang, J.; Fang, B.; Yu, P. Stock market prediction via multi-source multiple instance learning. *IEEE Access* **2018**, *6*, 50720–50728. [CrossRef]
8. Lo, A.W. Reconciling efficient markets with behavioral finance: The adaptive markets hypothesis. *J. Invest. Consult.* **2005**, *7*, 21–44.
9. Xing, F.Z.; Cambria, E.; Welsch, R.E. Natural language based financial forecasting: A survey. *Artif. Intell. Rev.* **2018**, *50*, 49–73. [CrossRef]
10. Schwert, G.W. Why does stock market volatility change over time? *J. Financ.* **1989**, *44*, 1115–1153. [CrossRef]
11. De Long, J.B.; Shleifer, A.; Summers, L.H.; Waldmann, R.J. Noise trader risk in financial markets. *J. Political Econ.* **1990**, *98*, 703–738. [CrossRef]
12. Verma, R.; Verma, P. Noise trading and stock market volatility. *J. Multinatl. Financ. Manag.* **2007**, *17*, 231–243. [CrossRef]
13. Chan, J.Y.L.; Leow, S.M.H.; Bea, K.T.; Cheng, W.K.; Phoong, S.W.; Hong, Z.W.; Chen, Y.L. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics* **2022**, *10*, 1283. [CrossRef]
14. Chan, J.Y.L.; Leow, S.M.H.; Bea, K.T.; Cheng, W.K.; Phoong, S.W.; Hong, Z.W.; Chen, Y.L. A Correlation-Embedded Attention Module to Mitigate Multicollinearity: An Algorithmic Trading Application. *Mathematics* **2022**, *10*, 1231. [CrossRef]
15. Li, Q.; Wang, T.; Li, P.; Liu, L.; Gong, Q.; Chen, Y. The effect of news and public mood on stock movements. *Inf. Sci.* **2014**, *278*, 826–840. [CrossRef]
16. Jiang, W. Applications of deep learning in stock market prediction: Recent progress. *Expert Syst. Appl.* **2021**, *184*, 115537. [CrossRef]

17. Ozbayoglu, A.M.; Gudelek, M.U.; Sezer, O.B. Deep learning for financial applications: A survey. *Appl. Soft Comput.* **2020**, *93*, 106384. [CrossRef]
18. Chopra, R.; Sharma, G.D. Application of Artificial Intelligence in Stock Market Forecasting: A Critique, Review, and Research Agenda. *J. Risk Financ. Manag.* **2021**, *14*, 526. [CrossRef]
19. Shah, D.; Isah, H.; Zulkernine, F. Stock market analysis: A review and taxonomy of prediction techniques. *Int. J. Financ. Stud.* **2019**, *7*, 26. [CrossRef]
20. Shahi, T.B.; Shrestha, A.; Neupane, A.; Guo, W. Stock price forecasting with deep learning: A comparative study. *Mathematics* **2020**, *8*, 1441. [CrossRef]
21. Mittermayer, M.-A.; Knolmayer, G.F. Newscats: A news categorization and trading system. In Proceedings of the Sixth International Conference on Data Mining (ICDM'06), Hong Kong, China, 18–22 December 2006; IEEE: Piscataway, NJ, USA, 2006; pp. 1002–1007.
22. Luss, R.; d'Aspremont, A. Predicting abnormal returns from news using text classification. *Quant. Financ.* **2015**, *15*, 999–1012. [CrossRef]
23. Schumaker, R.P.; Chen, H. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst. (TOIS)* **2009**, *27*, 1–19. [CrossRef]
24. Dadgar, S.M.H.; Araghi, M.S.; Farahani, M.M. A novel text mining approach based on tf-idf and support vector machine for news classification. In Proceedings of the 2016 IEEE International Conference on Engineering and Technology (ICETECH), Coimbatore, India, 17–18 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 112–116.
25. 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]
26. Kiros, R.; Salakhutdinov, R.; Zemel, R. Multimodal neural language models. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21 June–26 June 2014; pp. 595–603.
27. Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21 June–26 June 2014; pp. 1188–1196.
28. Pagolu, V.S.; Reddy, K.N.; Panda, G.; Majhi, B. Sentiment analysis of twitter data for predicting stock market movements. In Proceedings of the 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Online, 3–5 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1345–1350.
29. Garcia-Lopez, F.J.; Batyrshin, I.; Gelbukh, A. Analysis of relationships between tweets and stock market trends. *J. Intell. Fuzzy Syst.* **2018**, *34*, 3337–3347. [CrossRef]
30. Vargas, M.R.; De Lima, B.S.; Evsukoff, A.G. Deep learning for stock market prediction from financial news articles. In Proceedings of the 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Annecy, France, 26–28 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 60–65.
31. 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.
32. Yun, H.; Sim, G.; Seok, J. Stock prices prediction using the title of newspaper articles with korean natural language processing. In Proceedings of the 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Okinawa, Japan, 11–13 February 2019; IEEE: Piscataway, NJ, USA, 2017; pp. 19–21.
33. dos Santos Pinheiro, L.; Dras, M. Stock market prediction with deep learning: A character-based neural language model for event-based trading. In Proceedings of the Australasian Language Technology Association Workshop, Brisbane, Australia, 1 December 2017; pp. 6–15.
34. Akita, R.; Yoshihara, A.; Matsubara, T.; Uehara, K. Deep learning for stock prediction using numerical and textual information. In Proceedings of the 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, Japan, 26–29 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–6.
35. Matsubara, T.; Akita, R.; Uehara, K. Stock price prediction by deep neural generative model of news articles. *IEICE Trans. Inf. Syst.* **2018**, *101*, 901–908. [CrossRef]
36. Duan, J.; Zhang, Y.; Ding, X.; Chang, C.Y.; Liu, T. Learning target-specific representations of financial news documents for cumulative abnormal return prediction. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 2823–2833.
37. 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, Turin, Italy, 22–26 October 2018; pp. 1603–1606.
38. Liu, G.; Wang, X. A numerical-based attention method for stock market prediction with dual information. *IEEE Access* **2018**, *7*, 7357–7367. [CrossRef]
39. Wang, X.; Wang, Y.; Weng, B.; Vinel, A. Stock2Vec: A hybrid deep learning framework for stock market prediction with representation learning and temporal convolutional network. *arXiv* **2021**, arXiv:2010.01197.
40. Minh, D.L.; Sadeghi-Niaraki, A.; Huy, H.D.; Min, K.; Moon, H. Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *IEEE Access* **2018**, *6*, 55392–55404. [CrossRef]

41. Lu, R.; Lu, M. Stock trend prediction algorithm based on deep recurrent neural network. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 5694975. [CrossRef]
42. Liu, B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*; Cambridge University Press: Cambridge, UK, 2020.
43. Tetlock, P.C. Giving content to investor sentiment: The role of media in the stock market. *J. Financ.* **2007**, *62*, 1139–1168. [CrossRef]
44. Yadav, A.; Vishwakarma, D.K. Sentiment analysis using deep learning architectures: A review. *Artif. Intell. Rev.* **2020**, *53*, 4335–4385. [CrossRef]
45. Chan, J.Y.L.; Bea, K.T.; Leow, S.M.H.; Phoong, S.W.; Cheng, W.K. State of the art: A review of sentiment analysis based on sequential transfer learning. *Artif. Intell. Rev.* **2022**, 1–32. [CrossRef]
46. Fellbaum, C. *A Semantic Network of English Verbs. WordNet: An Electronic Lexical Database*; MIT Press: Cambridge, MA, USA, 1998; Volume 3, pp. 153–178.
47. Liu, H.; Singh, P. Conceptnet—A practical commonsense reasoning tool-kit. *BT Technol. J.* **2004**, *22*, 211–226. [CrossRef]
48. 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), Marseille, France, 11–16 May 2020.
49. Henry, E. Are investors influenced by how earnings press releases are written? *J. Bus. Commun.* **2008**, *45*, 363–407. [CrossRef]
50. Wilson, T.; Hoffmann, P.; Somasundaran, S.; Kessler, J.; Wiebe, J.; Choi, Y.; Cardie, C.; Riloff, E.; Patwardhan, S. Opinionfinder: A system for subjectivity analysis. In Proceedings of the HLT/EMNLP 2005 Interactive Demonstrations, Vancouver, BC, Canada, 7 October 2005; pp. 34–35.
51. Yekrangi, M.; Abdolvand, N. Financial markets sentiment analysis: Developing a specialized lexicon. *J. Intell. Inf. Syst.* **2021**, *57*, 127–146. [CrossRef]
52. Loughran, T.; McDonald, B. When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *J. Financ.* **2011**, *66*, 35–65. [CrossRef]
53. Picasso, A.; Merello, S.; Ma, Y.; Oneto, L.; Cambria, E. Technical analysis and sentiment embeddings for market trend prediction. *Expert Syst. Appl.* **2019**, *135*, 60–70. [CrossRef]
54. 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]
55. Shah, D.; Isah, H.; Zulkernine, F. Predicting the effects of news sentiments on the stock market. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 4705–4708.
56. 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]
57. Day, M.-Y.; Lee, C.-C. Deep learning for financial sentiment analysis on finance news providers. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 18–21 August 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1127–1134.
58. Yu, L.-C.; Wu, J.-L.; Chang, P.-C.; Chu, H.-S. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowl.-Based Syst.* **2013**, *41*, 89–97. [CrossRef]
59. Maqsood, H.; Mehmood, I.; Maqsood, M.; Yasir, M.; Afzal, S.; Aadil, F.; Muhammad, K. A local and global event sentiment based efficient stock exchange forecasting using deep learning. *Int. J. Inf. Manag.* **2020**, *50*, 432–451. [CrossRef]
60. Nti, I.K.; Adekoya, A.F.; Weyori, B.A. Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence from Ghana. *Appl. Comput. Syst.* **2020**, *25*, 33–42. [CrossRef]
61. Chen, W.; Cai, Y.; Lai, K.; Xie, H. A topic-based sentiment analysis model to predict stock market price movement using weibo mood. In *Web Intelligence*; IOS Press: Amsterdam, The Netherlands, 2016; Volume 14, pp. 287–300.
62. Wang, Q.; Xu, W.; Zheng, H. Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles. *Neurocomputing* **2018**, *299*, 51–61. [CrossRef]
63. 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]
64. Qian, Y.; Li, Z.; Yuan, H. On exploring the impact of users' bullish-bearish tendencies in online community on the stock market. *Inf. Process. Manag.* **2020**, *57*, 102209. [CrossRef]
65. Jacobs, G.; Hoste, V. Fine-Grained Implicit Sentiment in Financial News: Uncovering Hidden Bulls and Bears. *Electronics* **2021**, *10*, 2554. [CrossRef]
66. Gupta, I.; Madan, T.K.; Singh, S.; Singh, A.K. HiSA-SMFM: Historical and Sentiment Analysis based Stock Market Forecasting Model. *arXiv* **2022**, arXiv:2203.08143.
67. Korivi, N.; Naveen, K.S.; Keerthi, G.C.; Manikandan, V.M. A Novel Stock Price Prediction Scheme from Twitter Data by using Weighted Sentiment Analysis. In Proceedings of the 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Virtual, 27–28 January 2022; pp. 623–628.
68. Boukhers, Z.; Bouabdallah, A.; Lohr, M.; Jürjens, J. Ensemble and Multimodal Approach for Forecasting Cryptocurrency Price. *arXiv* **2022**, arXiv:2202.08967.
69. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2013**, *3*, 993–1022.

70. Zhao, B.; He, Y.; Yuan, C.; Huang, Y. Stock market prediction exploiting microblog sentiment analysis. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 4482–4488.
71. Nguyen, T.H.; Shirai, K.; Velcin, J. Sentiment analysis on social media for stock movement prediction. *Expert Syst. Appl.* **2015**, *42*, 9603–9611. [CrossRef]
72. Si, J.; Mukherjee, A.; Liu, B.; Li, Q.; Li, H.; Deng, X. Exploiting topic-based twitter sentiment for stock prediction. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; Volume 2, pp. 24–29.
73. Alaparthi, S.; Mishra, M. Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey. *arXiv* **2020**, arXiv:2007.01127.
74. Li, M.; Chen, L.; Zhao, J.; Li, Q. Sentiment analysis of Chinese stock reviews based on BERT model. *Appl. Intell.* **2021**, *51*, 5016–5024. [CrossRef]
75. Jaggi, M.; Mandal, P.; Narang, S.; Naseem, U.; Khushi, M. Text mining of stocktwits data for predicting stock prices. *Appl. Syst. Innov.* **2021**, *4*, 13. [CrossRef]
76. Xiang, W.; Wang, B. A survey of event extraction from text. *IEEE Access* **2019**, *7*, 173111–173137. [CrossRef]
77. Zheng, S.; Cao, W.; Xu, W.; Bian, J. Doc2edag: An end-to-end document-level framework for chinese financial event extraction. *arXiv* **2019**, arXiv:1904.07535.
78. 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; pp. 1415–1425.
79. Yang, H.; Chen, Y.; Liu, K.; Xiao, Y.; Zhao, J. Dcfee: A document-level Chinese financial event extraction system based on automatically labeled training data. In Proceedings of the ACL 2018, System Demonstrations, Melbourne, Australia, 15–20 July 2018; pp. 50–55.
80. Nuij, W.; Milea, V.; Hogenboom, F.; Frasincar, F.; Kaymak, U. An automated framework for incorporating news into stock trading strategies. *IEEE Trans. Knowl. Data Eng.* **2013**, *26*, 823–835. [CrossRef]
81. 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.
82. Nascimento, J.B.; Cristo, M. The impact of structured event embeddings on scalable stock forecasting models. In Proceedings of the 21st Brazilian Symposium on Multimedia and the Web, Manaus, Brazil, 27–30 October 2015; pp. 121–124.
83. Wang, Y.; Li, Q.; Huang, Z.; Li, J. Ean: Event attention network for stock price trend prediction based on sentimental embedding. In Proceedings of the 10th ACM Conference on Web Science, Amsterdam, The Netherlands, 27–30 May 2018; pp. 311–320.
84. Oncharoen, P.; Vateekul, P. Deep learning for stock market prediction using event embedding and technical indicators. In Proceedings of the 2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA), Krabi, Thailand, 14–17 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 19–24.
85. Ding, B.; Wang, Q.; Wang, B.; Guo, L. Improving knowledge graph embedding using simple constraints. *arXiv* **2018**, arXiv:1805.02408.
86. Wu, J.; Wang, Y. A Text Correlation Algorithm for Stock Market News Event Extraction. In Proceedings of the International Conference of Pioneering Computer Scientists, Engineers and Educators, Taiyuan, China, 17–20 September 2021; Springer: Singapore, 2021; pp. 55–68.
87. Liu, J.; Lin, H.; Liu, X.; Xu, B.; Ren, Y.; Diao, Y.; Yang, L. Transformer-based capsule network for stock movement prediction. In Proceedings of the First Workshop on Financial Technology and Natural Language Processing, Macao, China, 12 August 2019; pp. 66–73.
88. Daiya, D.; Wu, M.-S.; Lin, C. Stock movement prediction that integrates heterogeneous data sources using dilated causal convolution networks with attention. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 8359–8363.
89. Xu, W.; Liu, W.; Xu, C.; Bian, J.; Yin, J.; Liu, T.Y. REST: Relational Event-driven Stock Trend Forecasting. In Proceedings of the Web Conference, Ljubljana, Slovenia, 19–23 April 2021; pp. 1–10.
90. Cheng, D.; Yang, F.; Xiang, S.; Liu, J. Financial time series forecasting with multi-modality graph neural network. *Pattern Recognit.* **2022**, *121*, 108218. [CrossRef]
91. 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]
92. Li, Q.; Jiang, L.; Li, P.; Chen, H. Tensor-based learning for predicting stock movements. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29.
93. Li, Q.; Tan, J.; Wang, J.; Chen, H. A multimodal event-driven lstm model for stock prediction using online news. *IEEE Trans. Knowl. Data Eng.* **2020**, *33*, 3323–3337. [CrossRef]
94. Li, W.; Bao, R.; Harimoto, K.; Chen, D.; Xu, J.; Su, Q. Modeling the Stock Relation with Graph Network for Overnight Stock Movement Prediction. *IJCAI* **2020**, *20*, 4541–4547.

95. Ye, J.; Zhao, J.; Ye, K.; Xu, C. Multi-graph convolutional network for relationship-driven stock movement prediction. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 6702–6709.
96. Li, X.; Wang, J.; Tan, J.; Ji, S.; Jia, H. A graph neural network-based stock forecasting method utilizing multi-source heterogeneous data fusion. *Multimed. Tools Appl.* **2022**, 1–23. [CrossRef]