Application of Natural Language Processing in Stock Forecasting

Subjects: Mathematics, Interdisciplinary Applications | Business, Finance

Contributor: Wai Khuen Cheng , Khean Thye Bea , Steven Mun Hong Leow , Jireh Yi-Le Chan , Zeng-Wei Hong , Yen-Lin Chen

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.

natural language processing | deep learning | stock forecasting | sentiment analysis | event extraction | multicollinearity

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 [1], 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 [2] 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 [3] 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 [4], 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, their 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 [5].

In [5], 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, [6] 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 [7]. 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 [8] 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 [9] 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 [10] 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 [11]. Other than English context, the study [12] 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 [13] 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 [14] and applying paragraph vector to embed the online financial news into a vector before using the Deep Neural Generative (DNG) model for prediction [15].

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 [16] 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, their 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 [10][13][15] analyzed the news title instead of the content to reduce the associated noise. The study [17] 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 [17].

In the study [18], 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 [19] 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 [20], 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 [21], Stock2Vec embedding was applied in predicting the stock movement. Specifically, their 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 [22] trained the Stock2Vec based on the stock news and sentiment dictionaries. Different from the study [21], 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.

2. Sentiment-Based Stock Forecasting Approach

Sentiment analysis (SA) is regarded as one of the Natural Language Processing (NLP) applications to study people's opinions, emotions, and attitudes toward a specific event, individual, or topic [23]. 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 [24] 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 [25], 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 [26] 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 [27], ConceptNet [28], SentiWordNet [29], Harvard General Inquirer, Henry Wordlist [30], Opinion Finder [31], 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 [32] 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, [33] 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, [34] 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 [35] mapped the words of the financial news onto the emotional spaces of two different dictionaries, Harvard IV-4 Dictionary and Loughran–McDonald Financial Dictionary. Their study found that the dictionaries-based sentiment features yielded a better result than the bag-of-words model. In addition, the study [36] 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 [37] 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 [38] 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 [38] 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). Their 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 [39] 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 [40] proposed to adapt the existence and intensity of emotion words as a base to classify the sentiment of the financial news. Their 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 [41], 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 [42] 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 [43] 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 [44] 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 [45] 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 [46] 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 [47] 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 [48] 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 [49][50] and DeepMoving [51] to process the Tweet text into sentiment polarity or scoring before composing a sentiment index to forecast the market. However, the study [46] 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) [52] is adopted as a topic model to compute the topic distribution over words. The study [53] 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 [54][55] 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, [56] 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 [57] applied the BERT model in analyzing the Chinese stock reviews and achieved a higher precision in sentiment analysis result. Moreover, the recent study [58] proposed the FinALBERT model to leverage and fine-tune the pretrained FinBERT models for stock movement prediction.

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 [59]. 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 [60].

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 [61] 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. [62].

In the literature, the ViewerPro system is employed in [63] 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 [61] 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, [64] 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 [61] 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 [65] to forecast the index movement of S&P 500 based on the financial news of Reuters and Bloomberg. However, there is another study [66] pointed the limitations [65] where the events of small companies were neglected since the relevant reporting was limited. The authors of [67] combined the merit of event embedding [64] based on news headline of “Reuters”, “Reddit”, and “Intrinio” with the same set of technical indicators of [10] and historical prices to improve the predictability of Index movement of S&P500 and DJIA.

On the other hand, the shortcoming of event embedding [64] was pointed out in their later study [68], 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 [68]. Therefore, the study [68] improved the quality of event embeddings by incorporating the knowledge graph into the training phase to encode the background information.

Moreover, the work [66] pointed out that the previous studies neglected event characteristics which may seriously degrade the predictive performance. Their 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 [66] 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 [69] criticized that the coarse-grained event structure such as (S, P, O) [61][70] 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). Their 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 [71], 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 [72] 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 [73] 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 [74] 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, their 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 [75] 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 [75] 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.

References

1. 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.
2. Luss, R.; d’Aspremont, A. Predicting abnormal returns from news using text classification. Quant. Financ. 2015, 15, 999–1012.
3. 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.

4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.

15. 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.
16. 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.
17. 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.
18. 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.
19. Liu, G.; Wang, X. A numerical-based attention method for stock market prediction with dual information. IEEE Access 2018, 7, 7357–7367.
20. 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.
21. 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.
22. Lu, R.; Lu, M. Stock trend prediction algorithm based on deep recurrent neural network. Wirel. Commun. Mob. Comput. 2021, 2021, 5694975.
23. Liu, B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions; Cambridge University Press: Cambridge, UK, 2020.
24. Tetlock, P.C. Giving content to investor sentiment: The role of media in the stock market. J. Financ. 2007, 62, 1139–1168.
25. Yadav, A.; Vishwakarma, D.K. Sentiment analysis using deep learning architectures: A review. Artif. Intell. Rev. 2020, 53, 4335–4385.
26. 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.

27. Fellbaum, C. A Semantic Network of English Verbs. WordNet: An Electronic Lexical Database; MIT Press: Cambridge, MA, USA, 1998; Volume 3, pp. 153–178.
28. Liu, H.; Singh, P. Conceptnet—A practical commonsense reasoning tool-kit. BT Technol. J. 2004, 22, 211–226.
29. 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.
30. Henry, E. Are investors influenced by how earnings press releases are written? J. Bus. Commun. 2008, 45, 363–407.
31. 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.
32. Yekrangi, M.; Abdolvand, N. Financial markets sentiment analysis: Developing a specialized lexicon. J. Intell. Inf. Syst. 2021, 57, 127–146.
33. Loughran, T.; McDonald, B. When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. J. Financ. 2011, 66, 35–65.
34. 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.
35. 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.
36. 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.
37. 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.
38. Oliveira, N.; Cortez, P.; Areal, N. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. Decis. Support Syst. 2016, 85, 62–73.
39. 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.
40. 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.

41. 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.
42. Bollen, J.; Mao, H.; Zeng, X. Twitter mood predicts the stock market. J. Comput. Sci. 2011, 2, 1–8.
43. 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.
44. 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.
45. 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.
46. 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.
47. 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.
48. Jacobs, G.; Hoste, V. Fine-Grained Implicit Sentiment in Financial News: Uncovering Hidden Bulls and Bears. Electronics 2021, 10, 2554.
49. 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.
50. 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.
51. Boukhers, Z.; Bouabdallah, A.; Lohr, M.; Jürjens, J. Ensemble and Multimodal Approach for Forecasting Cryptocurrency Price. arXiv 2022, arXiv:2202.08967.
52. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. J. Mach. Learn. Res. 2013, 3, 993–1022.
53. 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.

54. Nguyen, T.H.; Shirai, K.; Velcin, J. Sentiment analysis on social media for stock movement prediction. Expert Syst. Appl. 2015, 42, 9603–9611.
55. 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.
56. Alaparthi, S.; Mishra, M. Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey. arXiv 2020, arXiv:2007.01127.
57. Li, M.; Chen, L.; Zhao, J.; Li, Q. Sentiment analysis of Chinese stock reviews based on BERT model. Appl. Intell. 2021, 51, 5016–5024.
58. 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.
59. Xiang, W.; Wang, B. A survey of event extraction from text. IEEE Access 2019, 7, 173111–173137.
60. 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.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.
66. 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.

67. 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.
68. Ding, B.; Wang, Q.; Wang, B.; Guo, L. Improving knowledge graph embedding using simple constraints. arXiv 2018, arXiv:1805.02408.
69. 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.
70. 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.
71. 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.
72. 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.
73. 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.
74. 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.
75. Cheng, D.; Yang, F.; Xiang, S.; Liu, J. Financial time series forecasting with multi-modality graph neural network. Pattern Recognit. 2022, 121, 108218.