Abstract
Maintaining financial system stability is critical to economic development, and early identification of risks and opportunities is essential. The financial industry contains a wide variety of data, such as financial statements, customer information, stock trading data, news, etc. Massive heterogeneous data calls for intelligent algorithms for machines to process and understand. This paper mainly focuses on the stock trading data and news about China A-share companies. We present a financial data analysis application, Financial Quotient Porter, designed to combine textual and numerical data by using a multi-strategy data mining approach. Additionally, we present our efforts and plans in deep learning financial text processing application scenarios using natural language processing (NLP) and knowledge graph (KG) technologies. Based on KG technology, risks and opportunities can be identified from heterogeneous data. NLP technology can be used to extract entities, relations, and events from unstructured text, and analyze market sentiment. Experimental results show market sentiments towards a company and an industry, as well as news-level associations between companies.

Introduction
Maintaining financial system stability is critical to economic development. To meet this challenge, perceptual and cognitive technologies are necessary to identify risks and opportunities in advance (Xiao et al. 2021). Artificial intelligence (AI) can help humans analyze large amounts of financial data in real-time and output conclusions and decisions. Intelligent algorithms are a fundamental part of financial technology (fintech) and are widely used in many application scenarios, e.g., auditing (Fisher, Garnsey, and Hughes 2016), intelligent early warning (Gao et al. 2021; Oro, Ruffolo, and Pupo 2020), public opinion monitoring (Liu et al. 2019; 2020a), quantitative investment (Sorensen 2019), etc. The financial industry contains a wide variety of data such as company financial statements, customer information, news, stock trading data, industry research reports, etc. In terms of document structure, there are structured, semi-structured, and unstructured data that require different technical solutions.

This paper mainly focuses on applying NLP and KG technologies to textual and numerical data for comprehensive analysis in the financial industry. This paper mainly focuses on the processing of stock trading data and news about China A-share companies. This paper is divided into two parts: the first part introduces an application based on multi-strategy financial data mining, and the second part presents the efforts and plans of our application scenarios of deep learning financial text processing in detail.

Existing financial NLP has a wide range of applications, such as customer service chatbot (Okuda and Shoda 2018; Quah and Chua 2019), auditing, financial sentiment analysis (Araci 2019; Wang et al. 2021), public opinion monitoring, intelligent early warning, and stock behavior prediction (Mehtab and Sen 2019; Ris and Sjöberg 2021; Khedr, Yaseen, and others 2017). However, existing approaches still face many challenges, such as leveraging knowledge graphs to improve language understanding, and using NLP technology to enrich the content of knowledge graphs. Furthermore, the relationship between textual data processing and numerical data mining needs to be further explored. For example, stock trading data, e.g., stock prices, trading volume, Stochastic oscillator (KDJ), moving average convergence/divergence (MACD) and other indicators are purely numerical, while financial news, industry reports, company announcement are textual data. While the gap between numerical and textual data is obvious, they influence each other. On the one hand, news has a clear impact on the stability of the financial system, and on the other hand, stock trading data heralds upcoming news content. Sometimes breaking news events, such as wars, natural disasters, etc., may invalidate data mining techniques for stock trading data. Combining these two types of information, multi-strategy data mining is able to simultaneously consider stock trading data and news from around the world to identify opportunities and risks for conclusions and decisions.

Combining trading data and financial news to give a comprehensive analysis is a valuable task. The first challenge is the construction of the China A-share companies knowledge graph (CAKG). To resolve this problem, we employ a knowledge graph construction pipeline (Zhu 2022). The second challenge is the heterogeneous data mining of financial data. We propose a multi-strategy data mining pipeline to consider both the trading-level and news-level data.
Application scenarios of NLP and KG in fintech are extensive. For deep learning financial text processing, KG, NLP, and computer vision (CV) technologies make the machine better understand multi-modal data. KG technology allows the machine to reorganize, link, and understand multi-source heterogeneous data. For financial text processing, entity linking and typing can be used to identify an infinite number of concepts (labels) of entities. Information provenance (Hartig 2009; Lu et al. 2018) and complex network theory (Shu et al. 2017; Tang 2009) can be used to verify the reliability or correctness of news. KG reasoning can be used for conflict verification and to discover new opportunities and risks. Users can obtain answers to complex financial questions through KG question answering (QA).

NLP technology allows machines directly understand the unstructured text or extract structured key information. For financial text processing, NLP can be used to classify company announcements, news, and user comments into different sentiment categories (positive, neutral, negative, etc.) and different levels of credibility. Extracting entities, relations, and company events of interest from the plain text will enable machines to analyze key information. Machine reading comprehension returns satisfactory answers from massive documents. Text clustering can be used to compare the data across industries. Abstractive summarization helps investors and analysts read documents more efficiently. The main contributions of this paper are as follows.

1. We propose a multi-strategy financial data mining pipeline to analyze numerical and textual data.
2. We develop an application to process stock trading data and news and evaluate company performance from multiple dimensions.
3. We present our efforts and plans in deep learning financial text processing application scenarios, including challenges and methods for the wider use of NLP and KG in financial data analysis.

Related Work

For financial NLP, Araci (Araci 2019) pre-train a FinBERT on a large financial corpus and a small task-specific corpus for sentiment analysis which help to classify how the markets will react to the information presented in the text. Liu et al. (Liu et al. 2020b) pre-train a FinBERT to improve the performance on financial text understanding. This model is trained on 6 self-supervised tasks on financial corpora. Yang et al. (Yang, Uy, and Huang 2020) pre-train a FinBERT on a large financial communication corpus and achieve performance improvements on sentiment analysis. Khan et al. (Khan and Rabbani 2021) propose a chatbot architecture to deal with the business in the finance and banking industry. Yildirim et al. (Yildirim et al. 2018) propose a machine learning pipeline to resolve the problem of classifying the financial news as significant and nonsignificant.

For financial KG, Alam et al. (Alam and Ali 2022) propose to combine KG and machine learning to improve the prediction performance of loan default risk. They build a KG to train the KG embeddings to use as input to the model. Huai et al. (Elhammadi et al. 2020) propose a high-precision knowledge extraction pipeline for extracting key information in financial news, in which they use semantic role labels and a conditional random field (CRF) model to identify semantic relationships between entities in noisy text. Cheng et al. (Cheng et al. 2020) propose to use KG-based event embeddings for quantitative investment by jointly training FinKGS, events, and relations and feeding them into models to derive investment strategies. Zehra et al. (Zehra et al. 2021) build a financial report query system to query annual financial reports to help investors in the banking sector. They extract information from annual reports to build KGS and use ontologies to help understand user queries. Liao et al. (Liao 2020) construct a financial event graph by extracting causal relationships between events and use this technique for stock market forecasting.

For financial data mining, Zhang et al. (Zhang et al. 2020) propose a hierarchical evolutionary algorithm to locate promising search spaces and mine alpha factors in quantitative investment (Sorensen, Chen, and Mussalli 2021). Cheng et al. (Cheng et al. 2021) present the way they combine decision tree and Apriori algorithm (Agrawal et al. 1996; Han, Pei, and Yin 2000) into investment decision models. Chang et al. (Chang, Wang, and Chuang 2021) present a data mining pipeline based on neural network, support vector machine, mixed data sampling, etc. for stock price prediction. Kim (Kim 2021) proposes a financial market data mining framework, including the whole process of data preprocessing, feature selection, model, evaluation, and reporting. Li et al. (Li et al. 2021) improve clustering algorithms by proposing a criterion based on spectral graph theory to evaluate the cluster quality. Researchers (Al-Hashedi and Magalingam 2021; Sanad and Al-Sartawi 2021) review the data mining technology for financial fraud detection.

Approach

In this section, we first introduce our application in multi-strategy financial data mining. Then we present the application scenarios of KG and NLP techniques in deep learning financial text mining.

Multi-strategy Financial Data Mining

Multi-strategy financial data mining aims to use multi-strategy data analysis algorithms to comprehensively analyze numerical data and textual data to obtain market sentiments of companies and industries, such as food and beverage, semiconductor industry, lithium battery industry, etc. Finally, we can infer whether company performance or target industry performance is in line with expectations.

We present a mobile app “financial quotient porter” (财商数据) which aims to analyze the stock trading data from multiple perspectives. Share prices typically reflect a company’s current quarter performance or future growth potential. To estimate company performance, we first construct a China A-share companies knowledge graph to support the analysis of stock trading data. We propose a multi-dimensional scoring (MDS) pipeline which is designed to analyze a company from multiple dimensions. We assign different weights to different dimensions, for example, market sentiment towards a company, sentiment comparison
among different investment groups, joint analysis of sentiment among multiple investment groups, etc. We further analyze the sentiment of different industries and estimate whether the development of these industries is in line with expectations.

News is a non-negligible factor in maintaining financial stability because the stock market is sensitive to news. Many stocks go up by the limit when there is good news, and down by the limit when there is bad news. Some stocks react to the news too quickly for the market to identify opportunities or avoid risks. If investors want to discover indirect opportunities or avoid indirect risks, the news-level association can help find closely related companies. News-level associations are time-sensitive, meaning two companies weren’t connected a week ago, but now they’re connected through an event, and their connection may disappear a month later. We can discover news-level related companies and conduct a comprehensive analysis of their trading data. To find news-level related companies, we employ information extraction techniques (Zhu 2022). We construct a weighted graph based on the extracted structured information to analyze the degree of news-level association between companies. Different from static relations such as (company A, holding, company B), (person X, directorOf, company A) & (person X, directorOf, company B), we focus on the real-time news-level associations rather than relations that have always existed in the past.

Deep Learning Financial Text Processing

We present our application scenarios of deep learning financial text processing with our efforts and plans to identify financial opportunities and risks using NLP and KG techniques. Note that this subsection mainly describes the application scenarios, but deep learning financial text processing is not limited to this. Many technologies not mentioned are worth exploring further application scenarios.

Financial KG Applications  News or intelligence has a great impact on the stability of financial markets, and early identification or intelligence inference is crucial for finding opportunities and avoiding risks. The presence of undetected entities in news may cause problems, requiring us to identify complete entities as much as possible. For example, given the sentence “Due to covid-19, 1664 will reduce sales by x% in 2022.”, we humans can simply identify the entity “1664”, i.e., the beer brand, craft beer, and many people may not know which company this brand belongs to. Entity linking (EL) (Sevgili et al. 2020; Shen, Wang, and Han 2014), also known as named entity disambiguation (NED), includes both named entity recognition and disambiguation processes. The challenges lie in irregular entity mentions, long-tail entities, and entity ambiguity. Although named entity recognition (NER) models achieve state-of-the-art performance on some datasets, they can only recognize limited entities and require large amounts of high-quality training data. If the word “1664” was filtered out as a meaningless number, the system would not find news about this brand. With a knowledge graph containing rich entities, we can identify complete investment-related entities from plain text using an entity linking pipeline.

Entity disambiguation is important for accurately targeting the market segments (subdivision), which is the second challenge of entity linking. Many entities have the name “1664”, such as E1:/year (/年份,. E2:/food_and_beverages/beer (/食品饮料/啤酒), E3:/brand/beer/brand (/品牌/啤酒/品牌), etc. Using a context-aware entity disambiguation algorithm, we can link “1664” to the correct entity E3:/brand/beer/brand. Obscure entities are more challenging. For example, “The data shows that the net profit of Great Wall fell by x% year-on-year, which may be related to the upgrade of the energy structure” where “Great Wall” maybe E4:/auto (/汽车), E5:/auto/new_energy (/汽车/新能源), E6:/auto/traditional_energy (/汽车/传统能源), E7:/company/China_Great_Wall (/公司/中国长城), E8:/company/Great_Wall_Motors (/公司/长城汽车), E9:/travel/tourist_attraction/the_Great_Wall (/旅游/景点/长城), etc. With the help of KG, this model can infer the correct entity by reading the “energy structure” in the context, and conclude this “Great Wall” should be the E6:/auto/traditional_energy, /company/Great_Wall_Motors.

Assigning sub-industry concepts to entities facilitates reasoning for better decision-making. Let’s still take the above example, “Due to covid-19, 1664 will reduce sales by x% in 2022.”, we humans know that the fine-grained labels for “1664” are /company/Chongqing_Beer, /alcoholic_beverages/beer, /alcoholic_beverages/beer/craft_beer, /brand/beer/brand, /food_and_beverages, etc. From these labels, we can infer that this is risky news for the /company/Chongqing_Beer (重美). At the same time, this is risky news for other products belonging to /alcohol_beverages/beers only. At such as “Corona Beer”. Nevertheless, that does not mean it is risky news for /alcoholic_beverages/liquor or /company/Kweichow_Moutai. Fine-grained entity-typing (Murty et al. 2018; Onoe and Durrett 2020; Liu et al. 2021) aims to assign fine-grained concepts to entities. The challenge lies in the large number of concepts and the concept hierarchy. Prompt-learning (Ding et al. 2021) achieves state-of-the-art performance on some datasets. For financial application scenarios, there will be more concepts beyond the capacity of a single model, and the use of knowledge graphs and ontologies can help address this problem, allowing machines to more accurately infer information about opportunities and risks.

News spreads like wildfire, but its reliability and correctness are difficult to evaluate. For example, “Pork prices will increase by x% in 2 months”. Whether this outgoing news is reliable is difficult to assess, but investment opportunities often come with risks. Knowledge validation from heterogeneous data is an important technology to verify the reliability of the information. The challenge lies in the way to find evidence and track news sources (provenance (Hartig 2009)). It is a good choice to verify the information from
the perspective of graph structure (complex network) (Zhou and Zafarani 2019; Allassad, Hussain, and Agarwal 2019).

When dealing with large amounts of company performance data, analysts need to ensure data consistency. For example, extracting from different files we got the performance of Q1 result > Q2 result > Q3 result > Q1 result (第一季度>第二季度>第三季度>第一季度) which is contradictory. The out-of-range attribute value is also an error, and it is difficult to manually check for implicit errors.

Knowledge validation aims to improve knowledge quality (Owoc, Ochmanska, and Gladysz 1999). Identifying conflicts in data is an important aspect of KG reasoning, and there are some knowledge reasoning tools (Vermesan and Coenen 2013). Whether it is an obvious data error or an implicit property range error, the location of the error can be tracked down.

Investors need to infer new opportunities and risks based on existing data. For example, (Company A, business partner, Company B), (Company B, suffer, huge fine) → (Company A, stock price, falling). (Company A, holding, Company B), (Company B, performance, +120%) → (Company A, stock price, rising). Knowledge reasoning (Chen, Jia, and Xiang 2020; Zhang and Yao 2021) aims to identify errors and infer new facts from existing data. The challenge lies in inference rule mining and interpretability of inference results. Inference algorithms can be divided into rule mining (Galárraga et al. 2013; Lao, Mitchell, and Cohen 2011), reinforcement learning (Xiong, Hoang, and Wang 2017), knowledge representation learning (Saxena, Tripathi, and Talukdar 2020; Bordes et al. 2013), etc. By mining rules or having experts design them, machines can infer opportunities and risks through knowledge graphs.

Querying company data based on complex questions is a common task. For example, “List the top 5 companies with the fastest fourth quarter growth of all companies producing premium beers” (列出所有生产高端啤酒的公司中第四个季度业绩增速最高的前5名的公司). While investors can obtain this data from search engines or stock brokers websites (such as 10jqka.com.cn, eastmoney.com, etc.), managing this data is time-consuming. Knowledge graph question answering (KGQA) (Huang et al. 2019; Hao et al. 2017) aims to find answers to natural language questions over a knowledge graph. Investors can ask complex questions to the Knowledge Graph. The challenges lie in the completeness of the knowledge graph, the semantic parsing of questions, and the reasoning of answers. While the query language for interacting with knowledge graphs is SPARQL, the system should have the ability to parse natural language questions into SPARQL (Yih et al. 2016; Cao et al. 2020). Deep learning models have achieved good performance in translating natural language question into SPARQL (Yin, Grommann, and Rudolph 2021; Luz and Finger 2018).

Financial NLP Applications

News can be classified into different degrees of good news or bad news according to sentiment, or news can be divided into credible news, medium credible news, untrustworthy news, and junk news according to credibility. It is important to analyze the sentiment of financial news or the sentiment of user comments. Text classification (Kowsari et al. 2019; Kim 2014; Yang et al. 2016) aims to assigns a set of predefined categories to the documents. The challenge lies in that the same content can be expressed in different ways, and different types of documents express sentiment in different ways. In terms of sentiment classification (Zhang, Wang, and Liu 2018; Yadav and Vishwakarma 2020), documents of different genres have different target categories and their expressions are different. For example, news expresses sentiment objectively, user comments express sentiment implicitly, and financial reports express sentiment implicitly. Different models can be trained on different types of documents to predict the sentiment distribution of events, companies, industries, etc.

Structured text only accounts for a small proportion, and the content in massive unstructured text is underutilized. Extracting structured knowledge from the unstructured text can be used to complete knowledge graphs for better services. Information extraction (IE) (Deng and Liu 2018; Miwa and Bansal 2016) includes two subtasks, named entity recognition (NER) (Ma and Hovy 2016; Lample et al. 2016) and relation extraction (RE) (Han et al. 2019; Yao et al. 2019). NER is used to identify entities in unstructured text, and RE is used to extract relations between entities. NER faces many challenges, such as nested entities, high-frequency irregular entity abbreviations, long-tail entities, etc. RE is divided into supervised RE, distant-supervision RE and open IE (Mausam 2016). Supervised RE involves the definition of pre-defined relations, data annotation, model training, and evaluation. Distant-supervision (DS) RE uses relations in KGs to automatically annotate corpora. This method can conduct relation extraction and performance evaluation without manual annotation of the corpus. However, DSRE faces the problem of noisy samples in the training data. Open IE does not require pre-defined relations and extracts entity relations based on various strategies such as statistics, dictionaries, rules, syntactic parsing, word vectors, etc. The challenges lie in quality control, entity relation normalization, etc. Graph search algorithms can be used on the extracted graphs to identify market opportunities and risks.

Various events occur every day around the world, some of which affect the current and future prices of the stocks of related companies. For example, “Due to covid-19, 1664 will decrease sales by x% in 2022”, the event is “decrease sales”, and the event arguments are (company: Chongqing beer), (amount: decrease x%) and (cause: covid-19). Extracting events (Wang et al. 2020) from documents can build an event graph to analyze the opportunities and risks. Event extraction (Xiang and Wang 2019) aims to extract structured event types and arguments from unstructured text. The challenge lies in that we encounter many types of events in the financial market and the event arguments have various expressions contexts. Furthermore, manual labeling of training
data requires summarizing and annotating events in unstructured text, which is labor-intensive. Pipeline or joint extraction (Sha et al. 2018) models can be used to extract event triggers and arguments.

“After the price increase of Kweichow Moutai, what happened to the prices of other brands of liquor?” (贵州茅台涨价的时候其他品牌的白酒的价格怎么变化的?) When we want to know something about financial markets, we usually use a search engine, which returns web pages from open domains and requires manual sifting of the web pages. Information retrieval (IR) based question answering (Abbasiantaeb and Momtazi 2021; Zhu et al. 2021) aims at finding short passages for users’ questions. The answers are directly searched from unstructured documents. This task consists of two stages, coarse ranking, and reranking. The challenge lies in the semantic matching of question answering pairs. Pre-trained language models achieve state-of-the-art performance on this task. Question answering systems allow users to get concrete answers directly.

With the boom of news, reports, social media, self-media, etc., massive documents need to be classified and organized based on their key information such as policies and companies in each cluster. Dividing documents into different groups for different market segments, policies, and industries can be used for fine-grained comparison and analysis. Text clustering aims to group a set of unlabeled texts with high similarity in the same group and low similarity in different groups. Different clusters can also be compared to each other. The challenge lies in the representation of the document and the clustering algorithms. Pre-trained language model achieves state-of-the-art performance to encode documents. Clustering algorithms (Bishop 2006) use hierarchical clustering, k-means, DBSCAN (Ester et al. 1996; Schubert et al. 2017), etc. to group documents into hierarchical or single-level clusters.

Massive unstructured reports are not adequately read. Investors need machines to compress text length to improve reading efficiency. Text summarization aims to condense long documents into a shorter version while preserving the key information and meaning of the content. The challenge lies in the summary generation and evaluation. The deep learning model (El-Kassas et al. 2021) has achieved good performance in generating concise and accurate abstracts.

Many natural language technologies are not used as separate application scenarios, but are used in different stages of KG and NLP applications, e.g., word vectors (Řehůrek and Sojka 2010), pre-trained language models (Devlin et al. 2019), word segmentation, part-of-speech tagging, syntax parsing (Manning et al. 2014; Bird and Loper 2004), etc.

**Experiment**

In this section, we start by analyzing the sentiments of different groups towards the company. We then analyze the sentiments of different groups towards an industry. Finally, we show news-level correlations between companies.

**Setup**

We run this app on an iPhone XR. The cloud services run on an AMD Ryzen 5 1500X Quad-Core Processor @ 3.5GHz (Mem: 16G) and 1 Tesla T4 GPU (16G).

**Results of Company Analysis**

Figure 1 and Figure 2 show the market sentiment analysis of Kweichow Moutai (贵州茅台) from February 2022 to April 2022, where the blue and red lines denote the sentiment of different groups in the market and stock prices respectively. Figure 1 shows the sentiment of group-1 towards Kweichow Moutai. Their views on Kweichow Moutai are relatively stable, with only noticeable changes in sentiment on specific dates such as February 23, 2022, and April 19, 2022. Figure 2 shows the sentiment of Group-2 towards Kweichow Moutai. We can see that their sentiment are more sensitive. For example, on March 11, 2022, their sentiment reached a minimum, and on March 18, 2022, their sentiment reached a maximum. Overall, the sentiment of Kweichow Moutai is still divided, but not too bad.

**Results of Industry Analysis**

Figure 3 and Figure 4 show the market sentiment analysis for the metaverse industry from February 2022 to April 2022, where the blue line denotes the sentiment of different groups in the market and the red line represents a comprehensive indicator, taking into account market sentiment and the actual performance of the industry. As shown in Figure 3, from March 18, 2022, to April 20, 2022, positive sentiment continued to rise. However, the red line is very unstable. For example, a maximum value was reached on March 2,
the market sentiment retreated immediately, and a minimum value was reached on March 18. This means the actual performance of the metaverse industry is not ideal. As shown in Figure 4, the sentiment of Group-2 towards the metaverse continued to grow from March 23rd. Unlike Group-1, the red line of Group-2 continued to rise. This is because the positive sentiment level of Group-2 is higher than that of Group-1, so the upward trend of the red line is relatively stable.

Results of News-level Associations

We take Contemporary Amperex Technology (宁德时代, CATL) as an example to analyze the relationships between companies on April 21, 2022. Figure 5 shows a graph of news-level associations between companies centered on CATL, where CATL is marked with a red circle. Red nodes represent industry concepts and green nodes represent companies. Edges represent relationships between nodes, and edge thicknesses represent connection weights. We can see that concepts directly related to CATL include new energy (新能源), lithium batteries (锂电池), and power exchange (换电). The concept pandemic (疫情) is indirectly related to the CATL through new energy, indicating that the pandemic has an impact on the CATL through the new energy industry. Companies directly related to CATL, in descending order of weight, are BYD (比亚迪), TSLA (特斯拉), Ningbo Shanshan (杉杉股份), China Molybdenum (洛阳钼业), Shanghai Putailai New Energy Technology (璞泰来), GanFeng Lithium (赣锋锂业), Yunnan Energy New Material (恩捷股份), Hangzhou First PV Material (福斯特), China Zhenhua (Group) Science & Technology (振华科技), AVIC Jonhon Optronic Technology (中航光电), etc.
Conclusion and Future Work

This paper introduces a financial data analysis app that combines stock trading data and news for comprehensive analysis and decision making. We propose a multi-strategy financial data mining pipeline and news-level associations. Besides, we present our deep learning financial text processing application scenarios. We describe our efforts and plans to uncover financial opportunities and risks using NLP and KG technologies. In the future, we hope to resolve the problem of unified representation of numerical and textual data and apply it to quantitative investment, while expanding the application scenarios of NLP and KG in financial data.

References

Abbasiantaeb, Z., and Momtazi, S. 2021. Text-based question answering from information retrieval and deep neural network perspectives: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11(6):e1412.

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; Verkamo, A. I.; et al. 1996. Fast discovery of association rules. Advances in knowledge discovery and data mining 12(1):307–328.

Al-Hashedi, K. G., and Magalingam, P. 2021. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. Computer Science Review 40:100402.

Alam, M. N., and Ali, M. M. 2022. Loan default risk prediction using knowledge graph. In International Conference on Knowledge and Smart Technology (KST), 34–39. IEEE.

Allassad, M.; Hussain, M. N.; and Agarwal, N. 2019. Finding fake news key spreaders in complex social networks by using bi-level decomposition optimization method. In International Conference on Modelling and Simulation of Social-Behavioural Phenomena in Creative Societies, 41–54. Springer.

Araci, D. 2019. Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

Bird, S., and Loper, E. 2004. Nltk: the natural language toolkit. Association for Computational Linguistics.

Bishop, C. M. 2006. Pattern Recognition and Machine Learning. Springer.

Bordes, A.; Usunier, N.; García-Durán, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In Proceedings of NIPS, 2787–2795.

Cao, S.; Shi, J.; Pan, L.; Nie, L.; Xiang, Y.; Hou, L.; Li, J.; He, B.; and Zhang, H. 2020. Kqa pro: A dataset with explicit compositional programs for complex question answering over knowledge base. arXiv e-prints arXiv–2007.

Chang, T.-H.; Wang, N.; and Chuang, W.-B. 2021. Stock price prediction based on data mining combination model. Journal of Global Information Management (JGIM) 30(7):1–19.

Chen, X.; Jia, S.; and Xiang, Y. 2020. A review: Knowledge reasoning over knowledge graph. Expert Systems with Applications 141:112948.

Cheng, D.; Yang, F.; Wang, X.; Zhang, Y.; and Zhang, L. 2020. Knowledge graph-based event embedding framework for financial quantitative investments. In Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, 2221–2230.

Cheng, K.-C.; Huang, M.-J.; Fu, C.-K.; Wang, K.-H.; Wang, H.-M.; and Lin, L.-H. 2021. Establishing a multiple-criteria decision-making model for stock investment decisions using data mining techniques. Sustainability 13(6):3100.

Deng, L., and Liu, Y. 2018. Deep learning in natural language processing. Springer.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, 4171–4186. Association for Computational Linguistics.

Ding, N.; Chen, Y.; Han, X.; Xu, G.; Xie, P.; Zheng, H.; Liu, Z.; Li, J.; and Kim, H. 2021. Prompt-learning for fine-grained entity typing. CoRR abs/2108.10604.

El-Kassas, W. S.; Salama, C. R.; Rafea, A. A.; and Mohamed, H. K. 2021. Automatic text summarization: A comprehensive survey. Expert Systems with Applications 165:113679.

Elhammadi, S.; Lakshmanan, L. V.; Ng, R.; Simpson, M.; Huai, B.; Wang, Z.; and Wang, L. 2020. A high precision pipeline for financial knowledge graph construction. In Proceedings of International Conference on Computational Linguistics, 967–977.

Estler, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of KDD, volume 96, 226–231.

Fisher, I. E.; Garnsey, M. R.; and Hughes, M. E. 2016. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. Intelligent Systems in Accounting, Finance and Management 23(3):157–214.
Galárraga, L. A.; Teflioudi, C.; Hose, K.; and Suchanek, F. 2013. Amie: association rule mining under incomplete evidence in ontological knowledge bases. In Proceedings of World Wide Web, 413–422.

Gao, R.; Zhang, Z.; Shi, Z.; Xu, D.; Zhang, W.; and Zhu, D. 2021. A review of natural language processing for financial technology. In International Symposium on Artificial Intelligence and Robotics, volume 11884, 262–277. SPIE.

Han, X.; Gao, T.; Yao, Y.; Ye, D.; Liu, Z.; and Sun, M. 2019. Opennre: An open and extensible toolkit for neural relation extraction. In Proceedings of EMNLP-IJCNLP, 169–174. Association for Computational Linguistics.

Han, J.; Pei, J.; and Yin, Y. 2000. Mining frequent patterns without candidate generation. ACM sigmod record 29(2):1–12.

Hao, Y.; Zhang, Y.; Liu, K.; He, S.; Liu, Z.; Wu, H.; and Zhao, J. 2017. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In Proceedings of IJCAI, 221–231.

Hartig, O. 2009. Provenance information in the web of data. In LDOW.

Huang, X.; Zhang, J.; Li, D.; and Li, P. 2019. Knowledge graph embedding based question answering. In Proceedings of ACM international conference on web search and data mining, 105–113.

Khan, S., and Rabbani, M. R. 2021. Artificial intelligence and nlp-based chatbot for islamic banking and finance. International Journal of Information Retrieval Research 11(3):65–77.

Kheder, A. E.; Yaseen, N.; et al. 2017. Predicting stock market behavior using data mining technique and news sentiment analysis. International Journal of Intelligent Systems and Applications 9(7):22.

Kim, Y. 2014. Convolutional neural networks for sentence classification. In Proceedings of EMNLP, 1746–1751. ACL.

Kim, M. 2021. A data mining framework for financial prediction. Expert Systems with Applications 173:114651.

Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; and Brown, D. 2019. Text classification algorithms: A survey. Information 10(4):150.

Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural architectures for named entity recognition. In Proceedings of NAACL-HLT, 260–270. The Association for Computational Linguistics.

Lao, N.; Mitchell, T.; and Cohen, W. 2011. Random walk inference and learning in a large scale knowledge base. In Proceedings of EMNLP, 529–539.

Li, T.; Kou, G.; Peng, Y.; and Philip, S. Y. 2021. An integrated cluster detection, optimization, and interpretation approach for financial data. IEEE Transactions on Cybernetics.

Liao, K. 2020. Research on key technologies of financial domain-oriented eventic graph construction. Master’s thesis, Harbin Institute of Technology.

Liu, K.; Ergu, D.; Cai, Y.; Gong, B.; and Sheng, J. 2019. A new approach to process the unknown words in financial public opinion. Procedia Computer Science 162:523–531.

Liu, D.; Zhang, H.; Yu, H.; Zhao, X.; Wang, W.; Liu, X.; and Ma, L. 2020a. Research on network public opinion analysis and monitor method based on big data technology. In IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC), 195–199. IEEE.

Liu, Z.; Huang, D.; Huang, K.; Li, Z.; and Zhao, J. 2020b. Finbert: A pre-trained financial language representation model for financial text mining. In Proceedings of IJCAI.

Liu, Q.; Lin, H.; Xiao, X.; Han, X.; Sun, L.; and Wu, H. 2021. Fine-grained entity typing via label reasoning. In Proceedings of EMNLP, 4611–4622. Association for Computational Linguistics.

Lu, R.; Jin, X.; Zhang, S.; Qiu, M.; and Wu, X. 2018. A study on big knowledge and its engineering issues. IEEE Transactions on Knowledge and Data Engineering 31(9):1630–1644.

Luz, F. F., and Finger, M. 2018. Semantic parsing natural language into SPARQL: improving target language representation with neural attention. CoRR abs/1803.04329.

Ma, X., and Hovy, E. H. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In Proceedings of ACL. The Association for Computer Linguistics.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In Proceedings of ACL, 55–60.

Mausam, M. 2016. Open information extraction systems and downstream applications. In Proceedings of IJCAI, 4074–4077.

Mehtab, S., and Sen, J. 2019. A robust predictive model for stock price prediction using deep learning and natural language processing. CoRR abs/1912.07700.

Miwa, M., and Bansal, M. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In Proceedings of ACL, 1105–1116. Berlin, Germany: Association for Computational Linguistics.

Murty, S.; Verga, P.; Vilnais, L.; Radovanovic, I.; and McCallum, A. 2018. Hierarchical losses and new resources for fine-grained entity typing and linking. In Proceedings of the ACL, 97–109.

Okuda, T., and Shoda, S. 2018. Ai-based chatbot service for financial industry. Fujitsu Scientific and Technical Journal 54(2):4–8.

Onoe, Y., and Durrett, G. 2020. Fine-grained entity typing for domain independent entity linking. In Proceedings of AAAI, volume 34, 8576–8583.

Oro, E.; Ruffolo, M.; and Pupo, F. 2020. A cognitive automation approach for a smart lending and early warning application. In EDBT/ICDT Workshops.

Owoc, M. L.; Ochmanska, M.; and Gladysz, T. 1999. On principles of knowledge validation. In Validation and Verification of Knowledge Based Systems. Springer. 25–35.
Quah, J. T., and Chua, Y. 2019. Chatbot assisted marketing in financial service industry. In *International Conference on Services Computing*, 107–114. Springer.

Řehůřek, R., and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA. http://is.muni.cz/publication/884893/en.

Ris, E., and Sjöberg, A. 2021. Index prediction on the swedish stock market using natural language processing methods on swedish news. *Master's thesis in Matematical Sciences*.

Sanad, Z., and Al-Sartawi, A. 2021. Financial statements fraud and data mining: a review. In *European, Asian, Middle Eastern, North African Conference on Management & Information Systems*, 407–414. Springer.

Saxena, A.; Tripathi, A.; and Talukdar, P. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of ACL*, 4498–4507.

Schubert, E.; Sander, J.; Ester, M.; Kriegel, H. P.; and Xu, X. 2017. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)* 42(3):1–21.

Sevgili, O.; Shelmanov, A.; Arkhipov, M.; Panchenko, A.; and Biemann, C. 2020. Neural entity linking: A survey of models based on deep learning. *arXiv preprint arXiv:2006.00575*.

Sha, L.; Qian, F.; Chang, B.; and Sui, Z. 2018. Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction. In *Proceedings of the AAAI*, volume 32.

Shen, W.; Wang, J.; and Han, J. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering* 27(2):443–460.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19(1):22–36.

Sørensen, E.; Chen, M.; and Mussalli, G. 2021. The quantitative approach for sustainable investing. *The Journal of Portfolio Management* 47(8):38–49.

Sørensen, E. H. 2019. The golden age of quant. *The Journal of Portfolio Management* 46(1):12–24.

Tang, X. 2009. Qualitative meta-synthesis techniques for analysis of public opinions for in-depth study. In *International Conference on Complex Sciences*, 2338–2353. Springer.

Vermeeren, A., and Coenen, F. 2013. Validation and verification of knowledge based systems: Theory, tools and practice. Springer Science & Business Media.

Wang, Y.; Guo, Y.; Zeng, D.; and Yang, X. 2021. Application of emotion recognition technology in the field of futures pricing. *Bankers* (5):125–126.

Xiang, W., and Wang, B. 2019. A survey of event extraction from text. *IEEE Access* 7:173111–173137.

Xiao, J.; Wang, L.; Yang, Y.; Li, N.; Zhao, M.; Chen, Y.; and Tan, T. 2021. Research on the application of perceptual cognitive technology in financial risk early warning. *Chinese Journal of Intelligent Science and Technology* 16(5):940–961.

Xiong, W.; Hoang, T.; and Wang, W. Y. 2017. Deeppath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of EMNLP*, 564–573. Association for Computational Linguistics.

Yadav, A., and Vishwakarma, D. K. 2020. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review* 53(6):4335–4385.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT*, 1480–1489.

Yang, Y.; Uy, M. C. S.; and Huang, A. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Yao, Y.; Ye, D.; Li, P.; Han, X.; Lin, Y.; Liu, Z.; Liu, Z.; Huang, L.; Zhou, J.; and Sun, M. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of ACL*, 764–777. Association for Computational Linguistics.

Yih, W.-t.; Richardson, M.; Meek, C.; Chang, M.-W.; and Suh, J. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of ACL*, 201–206.

Yıldırım, S.; Jothimani, D.; Kavaklioğlu, C.; and Başar, A. 2018. Classification of “hot news” for financial forecast using nlp techniques. In *IEEE International Conference on Big Data*, 4719–4722. IEEE.

Yin, X.; Gromann, D.; and Rudolph, S. 2021. Neural machine translating from natural language to sparql. *Future Generation Computer Systems* 117:510–519.

Zehra, S.; Mohsin, S. F. M.; Wasi, S.; Jami, S. I.; Siddiqui, M. S.; and Syed, M. K.-U.-R. R. 2021. Financial knowledge graph based financial report query system. *IEEE Access* 9:69766–69782.

Zhang, Y.; and Yao, Q. 2021. Knowledge graph reasoning with relational directed graph. *CoRR* abs/2108.06040.

Zhang, T.; Li, Y.; Jin, Y.; and Li, J. 2020. Autoalpha: an efficient hierarchical evolutionary algorithm for mining alpha factors in quantitative investment. *arXiv preprint arXiv:2002.08245*.

Zhang, L.; Wang, S.; and Liu, B. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8(4):e1253.

Zhou, X., and Zafarani, R. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD explorations newsletter* 21(2):48–60.
Zhu, H.; Tiwari, P.; Ghoneim, A.; and Hossain, M. S. 2021. A collaborative ai-enabled pretrained language model for aiot domain question answering. *IEEE Transactions on Industrial Informatics*.

Zhu, H. 2022. Metaaid: A flexible framework for developing metaverse applications via ai technology and human editing. *arXiv preprint arXiv:2204.01614*. 