Computer Fraud and Security
ISSN (online): 1873-7056

Extracting Abnormal Text Information from Asset Data under Natural Language Processing

Zhiqi Li[1,*], Yuening Wang[1], Hui Chen[2]
[1]*Planning and Finance Department, Yunnan Power Grid Co., Ltd, Kunming 650000, Yunnan, China*
[2]*Financial Shared Service Center, Yunnan Power Grid Co., Ltd., Kunming 650000, Yunnan, China*
**Corresponding author.*

Abstract:
Due to the large amount of information and complex content in text, efficient extraction and analysis of abnormal information in the text is currently a challenge. Therefore, this article utilized natural language processing (NLP) technology to extract and analyze abnormal text information from asset data in asset data. This article first took Chinese A-share listed companies from 2016 to 2022 as the research object, collected relevant asset data texts, and defined the main variables. Subsequently, this article utilized the idea of NLP to construct a corresponding model for extracting and analyzing abnormal text information from asset data for experimental testing. This article conducted experimental tests on five types of abnormal text in asset data, including revenue decline, profit decline, stock price volatility increase, debt ratio increase, and legal litigation risk, to verify the detection ability of NLP technology. The results showed that when faced with these five abnormalities, the detection rate was higher than 0.9. Experimental results have shown that NLP algorithms can effectively extract and analyze text anomalies in asset data, thereby improving the accuracy and credibility of asset data.

Keywords: natural language processing, asset data, abnormal text information, information extraction, outlier detection

INTRODUCTION

With the advent of the big data era, unstructured data represented by text is playing an increasingly important role in various fields. NLP is an emerging research direction in artificial intelligence (AI), aimed at equipping machines with the ability to understand and express human language. In finance, textual information (news, comments, social media posts, etc.) contains rich information and is key to understanding and predicting market dynamics. However, due to the complex structure of text data, it often contains a large amount of abnormal information. Therefore, how to effectively extract and analyze such data is of great practical significance for enterprise asset management and risk control.

Asset data is a type of data that contains a large amount of information, both numerical and textual. If there is incorrect information in the text information, it can have a significant impact on the asset management and risk control of the enterprise. Xie et al. proposed a digital twin based asset status monitoring and anomaly detection framework, and attempted to embed Bayesian change point detection methods to detect suspicious asset anomalies in real-time. The results indicate that with the support of digital twin data management capabilities, the framework has achieved continuous status monitoring and anomaly detection for individual assets [1]. Brintrup et al. discounted the use of data analytics in predicting Tier 1 supply chain disruptions and the use of historical data available to OEMs to detect anomalous information such as malicious texts [2]. Using a statistical cost accounting model, Owusu and Alhassan examined the relationship between profit and asset-liability management structure of 27 banks in Ghana from 2007 to 2015. This tested whether these asset data were abnormal and confirmed the central assumption of the cost accounting model [3]. Fujiki conducted an in-depth analysis of the anomalies in cross market asset data. The results indicated that asset prices interacted with each other in different markets, and abnormal fluctuations in one market might be transmitted to other markets [4]. Talafidaryani focused on asset data anomalies and risk management, and constructed a set of asset allocation models that could be dynamically adjusted with market changes, risk preferences, and other factors within the Bayesian theoretical framework. This model could provide timely warning and response to abnormal situations in asset data [5]. Hartmann and Henkel conducted in-depth research on the linkage between financial markets and macroeconomics, using macroeconomic indices and leading indicators. He predicted and detected abnormal fluctuations in the asset market, providing important references for investors and regulatory agencies in their decision-making [6]. Jiang et al. was committed to researching asset data in financial markets and established a new time series based asset pricing model based on this foundation [7]. These scholars have adopted anomaly recognition methods. Although it has certain advantages for complex financial data, in practice, there are problems such as high computational complexity and difficulty in determining anomaly thresholds, which reduce the real-time performance and accuracy of the algorithm. This article addresses the issues of high computational complexity and difficulty in determining anomaly thresholds in the application of anomaly recognition methods in financial data. By optimizing the algorithm model and introducing adaptive threshold mechanisms, the real-time performance and accuracy of the algorithm have been improved, effectively solving these problems.

NLP technology is an emerging discipline of AI that can efficiently target abnormal texts in asset data. Existing research mostly adopts two methods: rule-based and machine learning. Nti et al. conducted extensive research on the correlation between NLP and abnormal asset data, and found that abnormal fluctuations in some financial assets in the financial market could be effectively predicted by analyzing comments on social media [8]. Bag and Omrane were engaged in micro analysis of financial markets, particularly conducting in-depth research on abnormal phenomena in securities trading. He revealed the impact of investor behavior and market structure on asset prices, and proposed relevant anomaly detection and prediction methods [9]. Hens and Naebi’s main job was to model and detect anomalies in financial data in the financial market. He proposed a high-order statistical method based on NLP, which could effectively detect abnormal patterns in complex financial data [10]. Hao et al. proposed an ICPS (Industrial cyber physical system) asset protection priority optimization defense resource allocation scheme. This plan fully considered the balance between asset vulnerability, cost, and criticality, models defense resource allocation as a multi-objective optimization problem, and used Pareto optimal solution generation method for solution [11]. The above-mentioned scholars have achieved certain results in the research of asset data anomalies, but still need to comprehensively consider multiple factors to improve the real-time and accuracy of algorithms, and provide more effective solutions for the detection and prediction of asset data anomalies. This article proposes an efficient extraction and analysis method for asset data anomalies by optimizing NLP algorithms and combining them with the characteristics of asset data, which effectively improves the real-time and accuracy of the algorithm.

This article intends to use NLP techniques to study how to efficiently extract abnormal information from asset data and conduct in-depth analysis on it. With the advent of the big data era, unstructured text information plays an increasingly important role in enterprise resource management, and these data often contain various types of abnormal data states [12,13]. This article first elaborates on the basic theories and methods of NLP, and proposes an asset information mining method based on NLP technology. Furthermore, this article utilizes NLP methods such as particle embedding, named entity recognition, and semantic parsing to achieve accurate localization and extraction of massive text data.

DESIGN OF ASSET DATA ANOMALY UNDER NLP

Sample Selection and Data Sources

This article takes Chinese A-share listed companies from 2016 to 2022 as the object, and uses text mining methods to establish a specialized dictionary. It quantifies the frequency of data asset information disclosure in the annual reports of listed companies and explores the correlation between this frequency and the accuracy of analyst earnings forecasts. On the data source, this article integrates the semantic related word frequency, annual report total vocabulary, annual report reading difficulty index, and risk disclosure level from the “text structure database”. On this basis, this article integrates multi-dimensional data such as company performance forecast information from the Wind database and company characteristics, analyst behavior variables, and market benchmark indices obtained from the CSMAR database.

In the data preprocessing process, this article used the following methods for screening: 1) All samples of financial listed companies were excluded. 2) It can exclude ST (Special Treatment) listed companies. 3) This article removed observations of omitted variable records from the sample. Through this rigorous data cleaning program, a sample containing annual observation data from 24537 companies was finally obtained. At the same time, to reduce the impact of extreme values, this article performed tail trimming on all continuous variables at the 1% level to ensure the stability and reliability of the model.

Variable Definition

Although current data assets have not been included in the financial statements of enterprises, some listed companies have actively included relevant data assets in their annual reports [14,15]. This makes it possible for asset analysts to obtain and use this information. Therefore, this article first uses text mining methods to conduct in-depth analysis of the annual reports of Chinese listed companies, in order to achieve a quantitative evaluation of the frequency of data asset information disclosure for each company.

Based on core terms such as “data assets” and “data resources”, and using the Word2Vec neural network model, this article conducts deep learning and training on the massive financial text data publicly published by Chinese listed companies. Therefore, this article selects words from the WinGo text database that are more similar to “data assets” and “data resources” than 0.4, and establishes information disclosure that can better reflect enterprise data assets. The corresponding summary variables are shown in Table 1.

On this basis, this article constructs a specific model to calculate the frequency of data assets mentioned by companies in public information, and uses it as an important measurement standard. The calculation of data assets is as follows:

$$DataAsset_{in} = \frac{\sum_{j=1}^{n} GS_{inj}}{TotalGS_{in}} \times 100\% \tag{1}$$

Among them, $\mathrm{GS_{inj}}$ is the word frequency of the jth word in the set of similar words in the nth year’s annual report of enterprise i; $\mathrm{TotalGS_{in}}$ refers to the total number of words used by a company in its annual report.

Table 1. Summary of main variables

Variable symbols	Variable names	Definition description	Calculation method
DataAsset	Data assets	Data asset information in the company’s annual report	See Formula (1) for details.
Centralize	Customer concentration	The sum of the company’s sales to the top five customers as a proportion of total sales	Sum of top 5 customer sales/total sales
Size	Scale of enterprise	Ln (1+ total assets at year end)	Natural logarithm (1 + year-end total assets)
Lev	Financial leverage	Overall liabilities at the end of the year/overall assets in the year	Overall liabilities at the end of the year/overall assets in the year
ROA	Profitability	Annual net profit/end of year total assets	Annual net profit/end of year total assets
Indep	Independence of the Board	Independent directors/number of board of directors	Number of independent directors/number of board of directors
Dual	Two jobs in one	The value is 1 if the company’s chairman and general manager are unified, and 0 otherwise	The value is 1 if the two jobs are one, and 0 otherwise
Quick	Current ratio	Current assets/current liabilities	Current assets/current liabilities
BM	Book-to-market ratio	Total assets at year-end/Total market capitalization at year-end	Total assets at year-end/Total market capitalization at year-end
SOE	Nature of property rights	The value is 1 for SOEs and 0 otherwise	The value is 1 if it is a SOE, 0 otherwise

Building Models Based on NLP

This article is based on NLP technology, extracting and cleaning text information from NLP, including word segmentation, disappearing words, part of speech tagging, etc. Subsequently, through methods such as word frequency statistics and sentiment analysis, abnormal information in the text can be analyzed, and abnormal phenomena in the text can be identified [16,17]. Finally, based on the analysis results, this article corrects and removes abnormal information in the text. The operational process of the asset data abnormal text information extraction and analysis model under NLP technology constructed in this article is shown in Figure 1. The following are the specific steps for extracting and analyzing abnormal text information based on NLP:

(1) Text preprocessing: During the text preprocessing process, the text can be first segmented, dividing adjacent text strings into several separate words or labels [18,19]. Secondly, words that appear frequently in the text but have little impact on the meaning of the text can be removed during preprocessing. Finally, word stems can be extracted to restore each word to its basic form ("jumping" can be simplified to "jump"). These steps are beneficial for subsequent feature extraction and model construction [20,21]. Among them, stopping word filtering can be achieved through set operations, and the corresponding mathematical formula is as follows:

$$\mathrm{Pr}\,ocessedWords = \{word \in RawWords | word \notin stopwordsSet\} \tag{2}$$

In the formula, $\mathrm{RawWords}$ is the word list obtained by segmenting the original text, while $\mathrm{stopwordsSet}$ is a predefined stop word set.

(2) Feature extraction: next, the preprocessed text sequence is transformed into a word embedding matrix X, which is transformed into a vector expression representing one word in a row, that is, the row vector of each word [22]. If the length of a paragraph is m, its matrix expression is:

$$X \in R^{m \times d} \tag{3}$$

Here d is the dimension of word embedding. The text feature extraction processing using convolutional neural networks can be achieved using a set of mathematical operations. Convolutional kernels can be used to slide the input sequence and perform convolution operations. The convolutional kernel is generally a small matrix W of size (k, d), where k represents the width (window size) of the convolutional kernel. By adding a bias term b, the calculated local feature map can be expressed as:

$$C_j = f(\mathrm{Re}\,LU(W \cdot X_{j;j+k-1} + b)) \tag{4}$$

$X_{j;j+k-1}$ represents the word embedding submatrix from the jth word to the $j+k-1$th word; ReLU is an activation function; f is a pooling operation to obtain a fixed length feature representation.

Outlier detection: In text feature based outlier detection, the unsupervised learning method K-means clustering is used in this paper to discover abnormal patterns in the dataset. (3) For the vector representation set $X = \{x_1, x_2, \ldots, x_n\}$ extracted from text features, $x_i \in R^d$ is the d dimensional feature vector of the ith sample. An anomaly measurement is defined by the distance between a sampling point and its cluster center, and its degree of anomaly increases with the increase of distance. The corresponding anomaly detection formula is as follows:

$$AnomalyScore(X_i) = \left\| X_i - U_k \right\|^2 \tag{5}$$

Among them, μ_k is the centroid of the cluster to which sample x_i belongs. According to the similarity with other classes, when a class has a significant difference from other classes in that class, this class is considered an outlier point.

(4) Abnormal information extraction: Based on obtaining outlier points, deep analysis can be performed on the discovered abnormal text to extract important information. This can involve identifying specific keywords, phrases, or themes to understand why this paragraph is marked as abnormal. This includes identifying keywords, phrases, and topics in abnormal text, and exploring why this part of the text is marked as abnormal.

In addition, by combining contextual information of the text, industry background knowledge, and historical data, a deeper understanding of the essence of these abnormal phenomena can be obtained, providing valuable information for risk management, decision support, or further data investigation. Furthermore, by analyzing the context and other relevant information of the corpus, one can gain a deeper understanding of its meaning.

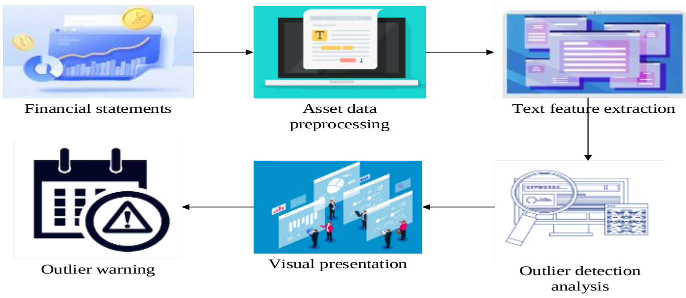

Figure 1. Process flow of asset data anomaly text information extraction and analysis model based on NLP

EXPERIMENT ON EXTRACTING AND ABNORMAL TEXT INFORMATION FROM ASSET DATA

Detection Capability for Different Types of Anomalies

This article conducts research on extracting and analyzing textual abnormal information from asset data to test the detection ability of the NLP technology used in this article for various abnormal phenomena. This article selects five common types of anomalies, including revenue decline, profit decline, stock price volatility increase, debt ratio increase, and legal litigation risk, as experimental objects, numbered 1-5 in Figure 2. The experimental results use an intuitive and efficient visualization method-heat map method, through a 5×5 matrix, the detection results of various anomalies are presented on a 5×5 matrix. This article can also conduct an in-depth analysis of the experimental results to provide a theoretical basis for subsequent research and application.

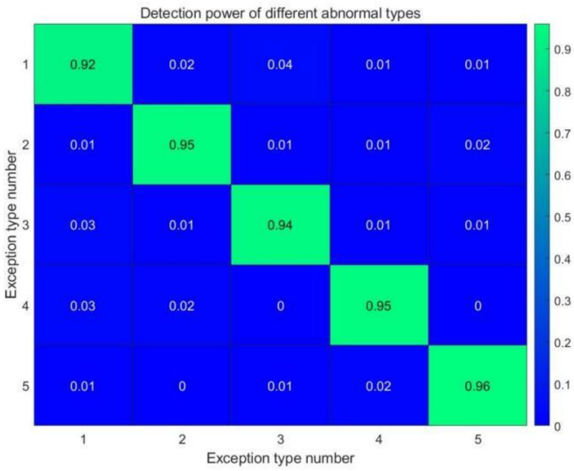

Figure 2. The detection capabilities of different abnormal types

From the heat map data in Figure 2, there is a 5×5 matrix, each element represents the detection ability for different abnormal types. By looking at the heat map data, the next conclusions can be made:

(1) The anomaly detection rates for numbers 1-5 are 0.92, 0.95, 0.94, 0.95, and 0.96, respectively, with the strongest detection ability for legal litigation risk.

(2) The larger the number in the main diagonal direction, the better the method proposed in this paper can detect corresponding anomalies. Correspondingly, the smaller the non diagonal values in the heat map, the better the false detection ability of these anomaly types can be controlled by the method proposed in this paper.

This article also analyzes the correlation between abnormal asset data and the text mentioned, and Table 2 further reveals their correlation. By comparing the size of the anomaly index and the number of times the text is cited, it can be seen that some abnormal changes occur frequently in the article, while others rarely occur. Specifically, the decline in company performance and changes in stock prices may not necessarily be directly proportional to the number of citations in the text. This means that investors and the media can be more concerned about these two abnormal changes. In addition, the frequency of abnormal texts in legal litigation is also high, which may be related to factors such as information disclosure and risk awareness within the industry. By studying the degree of correlation between arguments and anomalies, a deeper understanding of which abnormal changes have a stronger correlation with the content mentioned in the text can be gained, and the concerns and expectations of market entities can be thus better understood.

Table 2. Correlation analysis between abnormal asset data and textual mention

Anomaly indicator	Abnormal variation	Number of text references	Mention the correlation with anomalies
Revenue decline	-15%	37	0.75
Profit decline	-30%	62	0.82
Rise in stock price volatility	+200%	105	0.68
Increase in debt ratio	+11%	45	0.79
Risk of legal action	+27%	85	0.74

Comparison of Accuracy of Different Models

This article also conducts empirical research on asset data anomaly extraction models based on NLP through a series of detailed experiments. On the same standard library, this article quantitatively evaluates the recognition and classification ability of different models for abnormal text based on accuracy and other criteria. In Figure 3, the horizontal axis represents the number of 500 abnormal texts (Table 3), while the vertical axis displays the classification accuracy percentage of each model.

Table 3. Exception text types

Text type	Quantity	Text type	Quantity
Error description	95	Negative emotion	35
False information	85	Syntax error	55
Incomplete information	60	Technical problem	90
Other anomalies	40	Formatting error	40

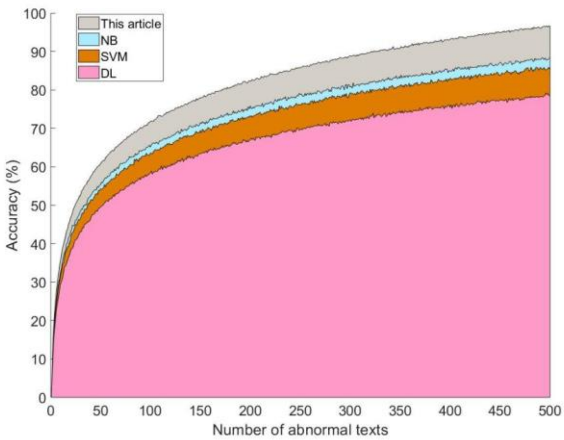

Figure 3. Comparison of accuracy of different models

The model in this paper maintained high prediction accuracy throughout the entire experiment, while the prediction accuracy of NB, SVM, and deep learning models decreased from high to low in Figure 3. Among them, the accuracy of the model constructed in this article ultimately remained at 96.6% after 500 training sessions, and the corresponding prediction accuracy of NB, SVM, and deep learning models were 88.4%, 85.9%, and 79.2%, respectively. Through such visual comparison, the performance level of NLP models in dealing with large-scale abnormal text problems is intuitively demonstrated, and the superiority of the new model in this scenario is highlighted.

Comparison of Text Processing Efficiency

This article compares the text processing efficiency of NB, SVM, and deep learning models. The experiment evaluated it from five aspects: processing speed, resource utilization, scalability, consistency, and ease of use. Finally, based on the characteristics of the five dimensional radar, this article obtains the scores of each model in different aspects and drew corresponding radar graphs.

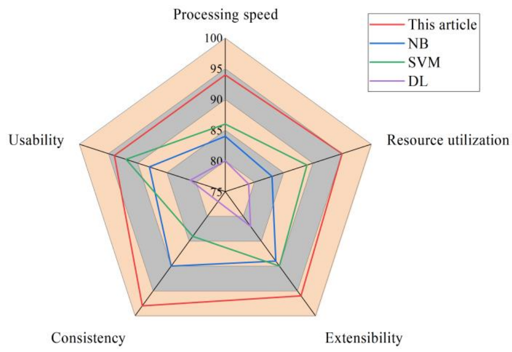

Figure 4. Text processing efficiency results

The model based on NLP technology adopted in this article has its advantages and characteristics in all five evaluation dimensions in Figure 4. In terms of computational speed, the model in this article outperforms NB, SVM, and DL models with a score of 94. In terms of resource utilization, the model in this article performs the best, followed by SVM, NB, and DL with the lowest score. In terms of scalability, consistency, and ease of use, the paper's model has also achieved high scores.

Satisfaction Comparison

This article also evaluates the satisfaction of the model among user groups through comparative experiments, and continues to compare it with the three models mentioned above. This experiment selects 20 real experience personnel with different backgrounds and levels of experience. This article adopts a 100 point scoring mechanism to test the four different abnormal text processing models from the perspectives of functionality, usability, response speed, and accuracy.

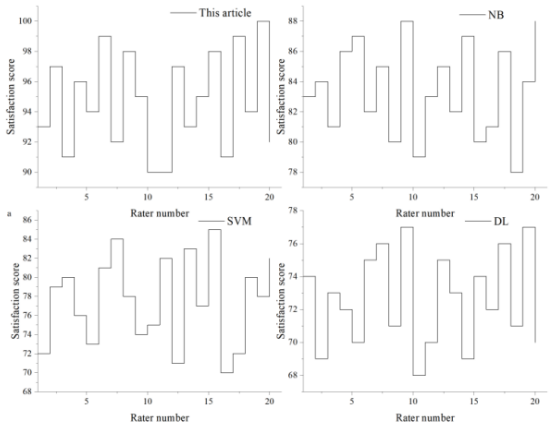

Figure 5. User satisfaction results

The 20 participants have different evaluations of these four models in Figure 5. The score range of the NLP model constructed in this article is between 90 and 100, with an average score of 94.7 points. The average satisfaction scores of NB, SVM, and DL models are 83.5, 77.6, and 72.6. Overall, the model constructed in this article achieved the highest level of satisfaction in all cases, with an average score significantly higher than the other three models.

CONCLUSIONS

The asset data anomaly extraction model based on NLP used in this article demonstrated strong detection capabilities, with a detection rate of over 92% for various abnormal phenomena, among which the ability to detect legal litigation risks was the strongest, reaching 96%. In terms of accuracy, the model achieved a high accuracy of 96.6% after 500 training sessions, far exceeding naive Bayes (88.4%), support vector machine (85.9%), and deep learning models (79.2%). At the same time, the model also performed well in terms of processing speed and resource utilization, with a calculation speed score of 94 points and the best performance in resource utilization. In summary, this model can provide strong support for the detection and prediction of asset data anomalies with its efficient and accurate characteristics.

REFERENCES

[1] Xie X, Lu Q, Parlikad A K. Digital twin enabled asset anomaly detection for building facility management. IFAC-PapersOnLine, 2020, 53(3): 380-385. DOI: https://doi.org/10.1016/j.ifacol.2020.11.061

[2] Brintrup A, Pak J, Ratiney D. Supply chain data analytics for predicting supplier disruptions: a case study in complex asset manufacturing. International Journal of Production Research, 2020, 58(11): 3330-3341. DOI: https://doi.org/10.1080/00207543.2019.1685705

[3] Owusu F B, Alhassan A L. Asset-Liability Management and bank profitability: Statistical cost accounting analysis from an emerging market. International Journal of Finance & Economics, 2021, 26(1): 1488-1502. DOI: https://doi.org/10.1002/ijfe.1860

[4] Fujiki H. Crypto asset ownership, financial literacy, and investment experience. Applied Economics, 2021, 53(39): 4560-4581. DOI: https://doi.org/10.1080/00036846.2021.1904125

[5] Talafidaryani M. A text mining-based review of the literature on dynamic capabilities perspective in information systems research. Management Research Review, 2021, 44(2): 236-267. DOI: doi/10.1108/MRR-03-2020-0139/full/html

[6] Hartmann P, Henkel J. The rise of corporate science in AI: Data as a strategic resource. Academy of Management Discoveries, 2020, 6(3): 359-381. DOI: https://doi.org/10.5465/amd.2019.0043

[7] Jiang L, Liu J, Peng L. Investor attention and asset pricing anomalies. Review of Finance, 2022, 26(3): 563-593. DOI: https://doi.org/10.1093/rof/rfab032

[8] Nti I K, Adekoya A F, Weyori B A. A systematic review of fundamental and technical analysis of stock market predictions. AI Review, 2020, 53(4): 3007-3057. DOI:https://doi.org/10.1007/s10462-019-09754-z

[9] Bag S, Omrane A. Corporate social responsibility and its overall effects on financial performance: Empirical evidence from Indian companies. Journal of African Business, 2022, 23(1): 264-280. DOI: https://doi.org/10.1080/15228916.2020.1826884

[10] Hens T, Naebi F. Behavioural heterogeneity in the capital asset pricing model with an application to the low-beta anomaly. Applied Economics Letters, 2021, 28(6): 501-507. DOI: https: //doi.org/10.1080/13504851.2020.1761529

[11] Hao W, Yao P, Yang T. Industrial cyber–physical system defense resource allocation using distributed anomaly detection. IEEE Internet of Things Journal, 2021, 9(22): 22304-22314. DOI: 10. 1109/JIOT.2021.3088337

[12] Maroufkhani P, Iranmanesh M, Ghobakhloo M. Determinants of big data analytics adoption in small and medium-sized enterprises (SMEs). Industrial Management & Data Systems, 2023, 123(1): 278-301. DOI:doi/10.1108/IMDS-11-2021-0695/full/html

[13] Mortaheb R, Jankowski P. Smart city re-imagined: City planning and GeoAI in the age of big data. Journal of Urban Management, 2023, 12(1): 4-15. DOI: https://doi.org/10.1016/j.jum.2022.08.001

[14] Van Doorn N, Badger A. Platform capitalism's hidden abode: producing data assets in the gig economy. Antipode, 2020, 52(5): 1475-1495. DOI: https://doi.org/10.1111/anti.12641

[15] Kehoe P J, Lopez P, Midrigan V. Asset prices and unemployment fluctuations: A resolution of the unemployment volatility puzzle. The Review of Economic Studies, 2023, 90(3): 1304-1357. DOI: https://doi.org/10.1093/restud/rdac048

[16] Desai R H, Choi W, Henderson J M. Word frequency effects in naturalistic reading. Language, cognition and neuroscience, 2020, 35(5): 583-594. DOI: https://doi.org/10.1080/23273798.2018.1527376

[17] Baek H, Lee Y, Choi W. Proficiency versus lexical processing efficiency as a measure of L2 lexical quality: Individual differences in word-frequency effects in L2 visual word recognition. Memory & Cognition, 2023, 51(8): 1858-1869. DOI: https://doi.org/10.3758/s13421-023-01436-0

[18] Hickman L, Thapa S, Tay L. Text preprocessing for text mining in organizational research: Review and recommendations. Organizational Research Methods, 2022, 25(1): 114-146. DOI: https://doi.org/10.1177/1094428120971683

[19] Mahilraj J, Tigistu G, Tumsa S. Text preprocessing method on Twitter sentiment analysis using machine learning. International Journal of Innovative Technology and Exploring Engineering, 2020, 9(12): 233-240. DOI: 10.35940/ijitee.K7771.0991120

[20] Krysiak Y, Maslyk M, Silva B N. The elusive structure of magadiite, solved by 3D electron diffraction and model building. Chemistry of Materials, 2021, 33(9): 3207-3219. DOI: https: //doi.org/10. 1021/acs.chemmater.1c00107

[21] Cascarano G L, Giacovazzo C. Towards the automatic crystal structure solution of nucleic acids: Automated model building using the new CAB program. Acta Crystallographica Section D: Structural Biology, 2021, 77(12): 1602-1613. DOI: https://doi.org/10.1107/S2059798321010937

[22] Zebari R, Abdulazeez A, Zeebaree D. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. Journal of Applied Science and Technology Trends, 2020, 1(2): 56-70. DOI: https://doi.org/10.38094/jastt1224