Woody International Publish Limited
An Multidisciplinary Academic Journal Publisher

Journal of Artificial Intelligence and Information, Volume 1, 2024
https://www.woodyinternational.com/

Current Applications of Deep Learning Models in Natural Language Processing Technology

Alexander[1*], Tatyana[2], Nikolai[3]

[1]Network Security, Compton University, UK
[2]Software Engineering, University of Montpellier, France
[3]Information Technology, Autonomous University of Madrid, Spain
alex2323Agmail.com
**Author to whom correspondence should be addressed.*

Abstract: ***This article traces the evolution of natural language processing (NLP) technologies from early statistical models like N-grams to contemporary deep learning frameworks such as recurrent neural networks (RNNs) and transformer-based architectures like BERT and GPT. It discusses how these advancements have revolutionized tasks like information extraction and machine translation, enabling more efficient linguistic data processing and reducing the dependency on manually labelled datasets. Despite challenges such as computational intensity, modern NLP continues to push boundaries in understanding and generating human language, driving significant advancements in various practical applications.***

Keywords: Natural Language Processing; Deep learning; GTP-3; Google; Language processing.

Cited as: Alexander, Tatyana, & Nikolai. (2024). Current Applications of Deep Learning Models in Natural Language Processing Technology. *Journal of Artificial Intelligence and Information*, *1*, 17–23. Retrieved from https://woodyinternational.com/index.php/jaii/article/view/32

1. Introduction

Long before the rise of deep learning, natural language processing technology was developed successfully and had many practical applications; at that time, it was mainly based on statistics and natural language processing [1]. The basic idea is to obtain statistical information on words and words in a large corpus, such as the extensive N-metalanguage model. Its principle can be understood as follows: suppose that in a sentence, the probability of the occurrence of the current word is related to the N-1 words in front of it [2]. The commonly used binary language model assumes that the probability of the occurrence of the current word is related to only one word in front of it. The formula is shown in Formula 1.

$$p(w_n) = p(w_1) \cdot p(w_2|w_1) \cdot p(w_3|w_2) \cdots p(w_n|w_{n-1}) \quad (1)$$

The probability of the current word wn is only related to the previous work w(n-1), and w(n-1) is only related to w(n-2), and so on. The model established in this way can be applied to scenarios such as machine translation and intelligent question answering [3]. Its advantage is that it avoids complex and clumsy grammar-based rule design, and it can achieve higher indexes in various tasks such as translation and information extraction by using only the binary model. However, this statistics-based language model also has disadvantages [4]. For example, in order to fully tap the features of the corpus, when N is obtained to a large extent, the number of parameters in the model increases exponentially, which will cause great difficulties in the deployment of the model.

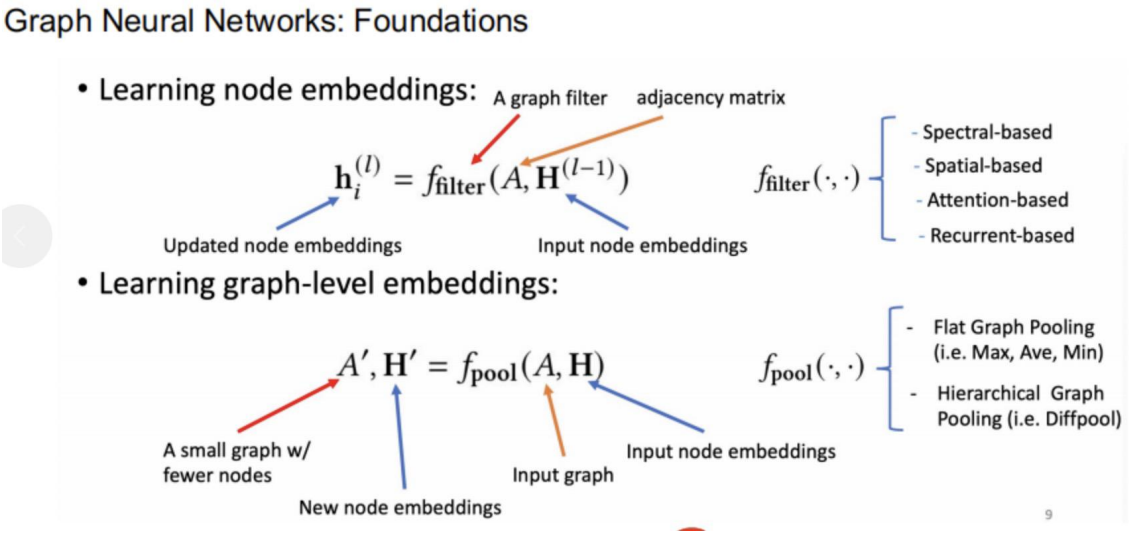
2. Natural Language Processing based on Deep Learning

At present, the mainstream ideas of natural language processing technology mainly include recurrent neural networks (RNN) [5], models based on pre-training and methods based on prompt words.

2.1 Recurrent Neural Network (RNN)

Recurrent neural network is very suitable for processing serialized data, its basic idea is similar to binary model, but the processing method is very different. First of all, all words will be encoded as word vectors, and this technology has existed before deep learning [6-8]. For example, word2vec represents each word as a

512-dimensional vector and carries out statistical learning in a large corpus. The obtained model has a great feature, that is, the cosine distance between synonyms is very close. This also makes perfect sense. Secondly, recurrent neural network uses artificial neural network to process the relationship between words. In the model, words participate in the calculation in the form of feature vectors [9]. Finally, the recurrent neural network uses parameter-sharing technology to compress the parameter number of the model, which significantly facilitates the deployment and application of the model.



The recurrent neural network can handle the word context well with a small number of parameters and at the same time, improve the effect of information extraction, machine translation, intelligent question answering and other tasks. However, it still does not fully mine the information in large corpora, and the use of corpus is limited to the generation of word vectors in the model [10]. If the model is to achieve better results, a large amount of data must be labelled for model training, and the cost of manual labelling becomes an urgent problem to be solved [11]. Another more serious problem is that the output of the current word features of the recurrent neural network depends on the hidden layer state vector of the output of the previous step, which leads to the network being output step by step during operation, and it cannot effectively use the parallelisation advantages of GPU and other accelerated computing devices. Compared with convolutional neural networks, its running speed becomes a disadvantage.

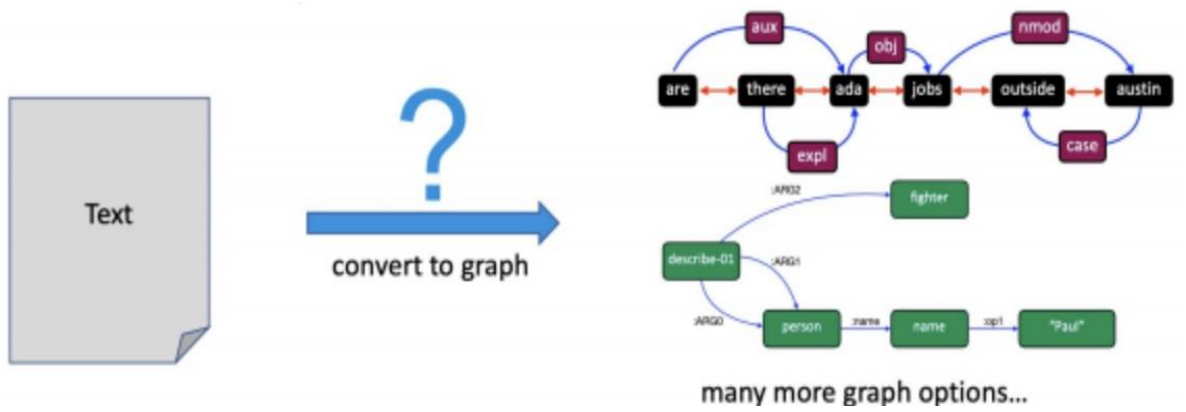
2.2 Model-based on Pre-training

In 2017, Google published a paper, "Attention is All You Need", that proposed a new network architecture - Transformer, which has reached a leading level in areas such as machine translation. The principle of the model is that by using only attention as the basic structure, the problem that the recurrent neural network cannot compute in parallel is effectively solved. The full exploitation of Transformer's advantages comes from [12] GPT [13], BERT and other methods proposed in 2018. The basic principle is to use Transformer's network structure for pre-training in a large corpus to excavate the information in the corpus fully [14]. The pre-trained model should be optimized more in terms of richness and the number of parameters. Applying the pre-trained model to downstream information extraction, machine translation and other tasks requires only a small amount of labeled data to achieve the effect of the previous recurrent neural network.

Since the introduction of the pre-training model, the pattern of pre-training + downstream task tuning has occupied a central position in natural language processing, and subsequent improvements to the model have also produced various variants of the pre-training model [15]. On the one hand, these variants make the pre-trained model perform better on downstream tasks. On the other hand, the running speed of the model is also improved. In application scenarios such as information extraction and machine translation, it is only necessary to select a suitable pre-trained model, which can be deployed to the actual use by annotating a small amount of data for tuning.

2.3 Advantages of Models based on Pre-training

Natural language processing based on the paradigm of pre-training model + downstream task tuning also has some disadvantages: First, the number of parameters of pre-training model is huge, for example, the number of parameters of BERT has reached hundreds of millions, and the direct result is that a pre-training needs to consume a lot of resources [16]. Google used 64 Tpus to run for four days when training the BERT-Large model, and OpenAI's GPT-3 [17] model released in 2020 has more than 170 billion participants, and the resources required for pre-training are surprisingly large.



Most companies are discouraged from pre-training on this scale, and the technology for future large-scale models may only be mastered by a few large companies. Another problem is the mismatch between the pre-trained model and the downstream task. Several tasks set in the pre-trained model are completely different from the downstream information extraction, machine translation, etc., which leads to the fact that the downstream task can not fully mine the information stored in the pre-trained model in the process of tuning [18].

2.4 Prompt Word-based Approach

To solve the problem of mismatch between the pre-trained model and the downstream task mentioned above, the prompt word-based approach is proposed, and the third normal form of natural language processing is formed. The basic principle of this method is that the situation of the downstream task is considered before the pre-training, and the downstream task is integrated into the model for pre-training during the pre-training stage [19]. This method can fully mine the information in the pre-trained model, and its advantage is that it only needs less data to achieve good results in the tuning of downstream tasks, so the goal of small sample learning or even zero sample learning for many downstream tasks can be realized.

Of course, this paradigm based on prompt words still has the problem of consuming large resources in the pre-training stage. In fact, since the concept of pre-training was proposed, the concept of large model has always existed in natural language processing [20]. In the development of recent years, this problem has not only not been alleviated but has become more and more serious. The parameters could reach 100 trillion.

3. Application Scenario based on Natural Language Processing Technology

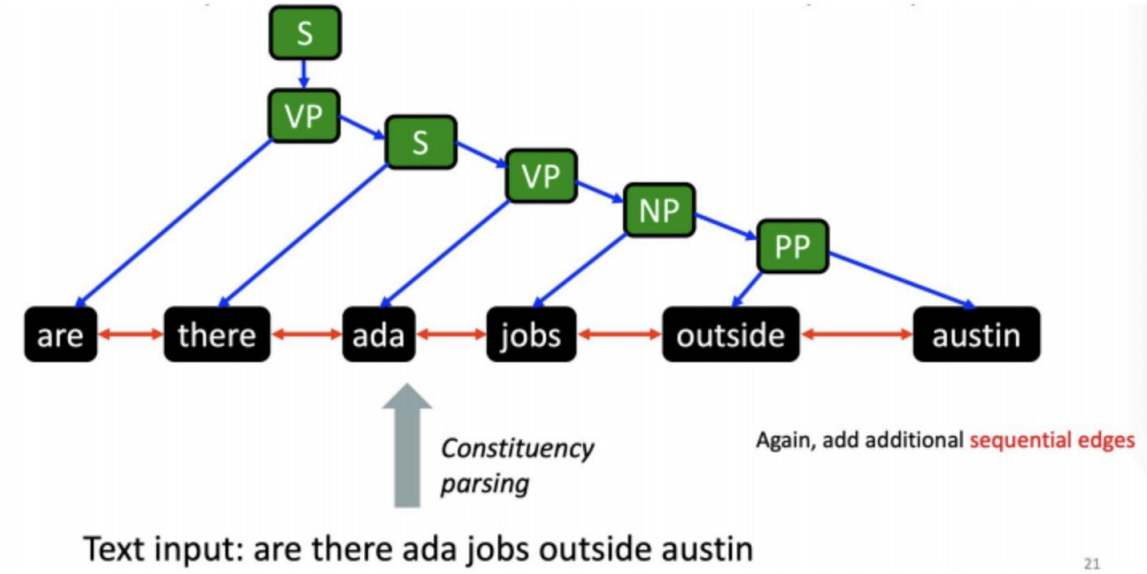
3.1 Information Extraction

Information extraction is one of the most successful scenarios of natural language processing applications based on deep learning [21]. The basic application scenario is to extract the desired information from a text, such as name, phone number, address, and so on. As far as the security industry is concerned, information extraction also has a very wide range of applications, such as extracting the address of the crime in the alarm phone, the time of the crime, etc. [22]; Extract the victim's name, ID card, suspect's name, ID card and so on from the case record; In the monitoring screen, according to the text watermark in the crime photo, the information such as the time and place of the crime is extracted.

Typically, information extraction is based on deep learning models such as named entity recognition (NER) and relational extraction to extract specific information from text, such as name, address, date, etc. In the security industry, the police can automatically extract key information of the case, such as the time, location and identity information of the relevant personnel [23], by analyzing alarm calls, case records and surveillance videos, thereby speeding up the case processing process.

3.2 Machine Translation

Machine translation is a well-defined application that translates text in one language into text in the target language. In the actual case handling situation, many case texts are not in Chinese; there are multiple languages and even small languages, which adds a lot of difficulties for the police to deal with the case. Machine translation can be timely and accurate to complete the translation of these languages to Chinese, which greatly improves the efficiency of case handling.



Typically, machine translation utilizes a neural machine translation model, such as a Transformer, to translate text from one language into another [24]. In case processing, police may need to deal with multilingual texts, and machine translation can quickly and accurately translate non-Chinese texts into Chinese, helping police understand and analyse case information from different languages.

3.3 Man-machine Dialogue

A human-machine conversation is a scenario of the standard Turing test, defined as multiple rounds of dialogue between a machine and a human, with the human being unable to distinguish whether the object of the conversation is a human or a machine [25]. At present, no model can officially pass the Turing test and be widely recognized. Still, the human-machine dialogue model established based on deep learning can complete several rounds of human-machine dialogue. Its effect is enough to solve problems in some practical scenarios, especially in some professional fields where the quality of human-machine dialogue has been relatively high [26-28]. For the security industry, for some simple legal advice, case searches, and other needs, voice dialogue robots, voice assistants, and other tools can be developed to solve the police effectively need to repeatedly answer questions, handle consultations, and handle other situations.

In general, human-machine conversation models based on deep learning, such as [29] BERT and [30] GPT, can achieve natural and smooth multi-round conversations. In the security field, the development of voice dialogue robots can provide police with legal advice, case inquiries and other services to improve work efficiency. Although it has yet to pass the Turing test, it has demonstrated good interaction in specific scenarios.

3.4 Other Applications

In recent years, the Imagen model released by Google and the [31-35] DALL-E model released by OpenAI has received great attention in the industry. Its purpose is to convert text into pictures; that is, only one needs to input a simple text description, and the model can generate the corresponding picture according to the text description. For example, Baidu released a knowledge-enhanced cross-modal large model - ERNIE-ViLG 2.0 [36], which made a breakthrough in AI painting and drew a work of "Wenxin Diao Dragon".

4. Conclusion

Based on the detailed exploration of natural language processing [37-39] (NLP) technology across different eras, this article delves into its evolution from statistical-based methods to contemporary deep learning approaches. Initially rooted in statistical models like [40] N-gram language models, which leverage word probabilities conditioned on preceding words, early NLP aimed at tasks such as machine translation and information extraction without the need for intricate grammar rules. However, limitations arose from scaling these models due to exponential parameter growth.

Deep learning ushered in transformative methods like recurrent neural networks (RNNs) and, more recently, transformer-based architectures. RNNs [41-44] encode words as vectors, facilitating sequential data processing while minimizing parameter count through parameter sharing. In contrast, transformers like BERT and GPT leverage attention mechanisms for parallel computation, addressing RNN's sequential processing limitations. Despite their resource-intensive nature, these pre-trained models have significantly advanced tasks such as information extraction and machine translation with reduced reliance on annotated data.

In conclusion, while each NLP [45] paradigm—from statistical methods to contemporary deep learning—has its advantages and drawbacks, the field continues to evolve towards more efficient and effective models capable of handling complex linguistic tasks across diverse applications.

Acknowledgments

Here, I would like to especially thank Xu Zhang for his research achievements in his article [1]“Analyzing Financial Market Trends in Cryptocurrency and Stock Prices Using CNN-LSTM Models”. His work provides valuable theoretical foundation and methodological support for my research. Xu Zhang's in-depth analysis and application of innovative CNN-LSTM model in cryptocurrency and stock price trend analysis have greatly inspired my thoughts on the construction and optimization of financial market prediction models.

Xu Zhang's research not only expanded my understanding of the application of deep learning models in the financial field, but also provided powerful practical guidance in empirical analysis and predictive model design. His work provides an important reference for me to explore the dynamics and trend analysis of the financial market, and has a profound influence on my research path in this field.

In addition, I want to especially thank Liu, Yingchia, Yang Xu, And Runze Song in their article [2]"Transforming User Experience (UX) through Artificial Intelligence (AI) in interactive media. Research results in design. They delve into how artificial intelligence can transform user experience (UX) in interactive media design. Their work has provided important theoretical frameworks and practical cases for my research, and has had a profound impact and inspiration on my approach and strategies when designing and developing interactive media products.

References

[1] Zhang, X. (2024). Analyzing Financial Market Trends in Cryptocurrency and Stock Prices Using CNN-LSTM Models.

[2] Liu, Yingchia, Yang Xu, and Runze Song. "Transforming User Experience (UX) through Artificial Intelligence (AI) in interactive media design." Engineering Science & Technology Journal 5.7 (2024): 2273-2283.

[3] Li, H., Wang, S. X., Shang, F., Niu, K., & Song, R. (2024). Applications of Large Language Models in Cloud Computing: An Empirical Study Using Real-world Data. International Journal of Innovative Research in Computer Science & Technology, 12(4), 59-69.

[4] Zhan, X., Shi, C., Li, L., Xu, K., & Zheng, H. (2024). Aspect category sentiment analysis based on multiple attention mechanisms and pre-trained models. Applied and Computational Engineering, 71, 21-26.

[5] Wu, B., Xu, J., Zhang, Y., Liu, B., Gong, Y., & Huang, J. (2024). Integration of computer networks and artificial neural networks for an AI-based network operator. arXiv preprint arXiv:2407.01541.

[6] Liang, P., Song, B., Zhan, X., Chen, Z., & Yuan, J. (2024). Automating the training and deployment of models in MLOps by integrating systems with machine learning. Applied and Computational Engineering, 67, 1-7.

[7] Li, A., Yang, T., Zhan, X., Shi, Y., & Li, H. (2024). Utilizing Data Science and AI for Customer Churn Prediction in Marketing. Journal of Theory and Practice of Engineering Science, 4(05), 72-79.
[8] Wu, B., Gong, Y., Zheng, H., Zhang, Y., Huang, J., & Xu, J. (2024). Enterprise cloud resource optimization and management based on cloud operations. Applied and Computational Engineering, 67, 8-14.
[9] Liu, B., Yu, L., Che, C., Lin, Q., Hu, H., & Zhao, X. (2024). Integration and performance analysis of artificial intelligence and computer vision based on deep learning algorithms. Applied and Computational Engineering, 64, 36-41.
[10] Li, A., Zhuang, S., Yang, T., Lu, W., & Xu, J. (2024). Optimization of Logistics Cargo Tracking and Transportation Efficiency based on Data Science Deep Learning Models.
[11] Xu, J., Yang, T., Zhuang, S., Li, H., & Lu, W. (2024). AI-Based Financial Transaction Monitoring and Fraud Prevention with Behaviour Prediction.
[12] Guo, L., Li, Z., Qian, K., Ding, W., & Chen, Z. (2024). Bank Credit Risk Early Warning Model Based on Machine Learning Decision Trees. Journal of Economic Theory and Business Management, 1(3), 24-30.
[13] Xu, Z., Guo, L., Zhou, S., Song, R., & Niu, K. (2024). Enterprise Supply Chain Risk Management and Decision Support Driven by Large Language Models. Applied Science and Engineering Journal for Advanced Research, 3(4), 1-7.
[14] Song, R., Wang, Z., Guo, L., Zhao, F., & Xu, Z. (2024). Deep Belief Networks (DBN) for Financial Time Series Analysis and Market Trends Prediction.
[15] Yuan, B. (2024). Design of an Intelligent Dialogue System Based on Natural Language Processing. Journal of Theory and Practice of Engineering Science, 4(01), 72-78.
[16] Yao, J., & Yuan, B. (2024). Optimization Strategies for Deep Learning Models in Natural Language Processing. Journal of Theory and Practice of Engineering Science, 4(05), 80-87.
[17] Bai, X., Zhuang, S., Xie, H., & Guo, L. (2024). Leveraging Generative Artificial Intelligence for Financial Market Trading Data Management and Prediction.
[18] Xu, X., Yuan, B., Song, T., & Li, S. (2023, November). Curriculum recommendations using transformer base model with infonce loss and language switching method. In 2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 389-393). IEEE.
[19] Zheng, H., Wu, J., Song, R., Guo, L., & Xu, Z. (2024). Predicting Financial Enterprise Stocks and Economic Data Trends Using Machine Learning Time Series Analysis.
[20] Guo, L., Song, R., Wu, J., Xu, Z., & Zhao, F. (2024). Integrating a Machine Learning-Driven Fraud Detection System Based on a Risk Management Framework.
[21] Gong, Y., Zhu, M., Huo, S., Xiang, Y., & Yu, H. (2024, March). Utilizing Deep Learning for Enhancing Network Resilience in Finance. In 2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE) (pp. 987-991). IEEE.
[22] Xu, J., Jiang, Y., Yuan, B., Li, S., & Song, T. (2023, November). Automated Scoring of Clinical Patient Notes using Advanced NLP and Pseudo Labeling. In 2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 384-388). IEEE.
[23] Tian, J., Li, H., Qi, Y., Wang, X., & Feng, Y. (2024). Intelligent medical detection and diagnosis assisted by deep learning. Applied and Computational Engineering, 64, 121-126.
[24] Xin, Q., Xu, Z., Guo, L., Zhao, F., & Wu, B. (2024). IoT Traffic Classification and Anomaly Detection Method based on Deep Autoencoders.
[25] Yang, T., Li, A., Xu, J., Su, G., & Wang, J. (2024). Deep Learning Model-Driven Financial Risk Prediction and Analysis.
[26] Yuan, B., & Song, T. (2023, November). Structural Resilience and Connectivity of the IPv6 Internet: An AS-level Topology Examination. In Proceedings of the 4th International Conference on Artificial Intelligence and Computer Engineering (pp. 853-856).
[27] Zhou, Y., Zhan, T., Wu, Y., Song, B., & Shi, C. (2024). RNA Secondary Structure Prediction Using Transformer-Based Deep Learning Models. arXiv preprint arXiv:2405.06655.
[28] Cui, Z., Lin, L., Zong, Y., Chen, Y., & Wang, S. (2024). Precision Gene Editing Using Deep Learning: A Case Study of the CRISPR-Cas9 Editor. Applied and Computational Engineering, 64, 134-141.
[29] Li, J., Wang, Y., Xu, C., Liu, S., Dai, J., & Lan, K. (2024). Bioplastic derived from corn stover: Life cycle assessment and artificial intelligence-based analysis of uncertainty and variability. Science of The Total Environment, 174349.
[30] Wang, B., He, Y., Shui, Z., Xin, Q., & Lei, H. (2024). Predictive Optimization of DDoS Attack Mitigation in Distributed Systems using Machine Learning. Applied and Computational Engineering, 64, 95-100.
[31] Zhang, X. (2024). Machine learning insights into digital payment behaviors and fraud prediction. Applied and Computational Engineering, 67, 61-67.

[32] Lu, W., Ni, C., Wang, H., Wu, J., & Zhang, C. (2024). Machine Learning-Based Automatic Fault Diagnosis Method for Operating Systems.
[33] Zhang, Y., Xie, H., Zhuang, S., & Zhan, X. (2024). Image Processing and Optimization Using Deep Learning-Based Generative Adversarial Networks (GANs). Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 5(1), 50-62.
[34] Xin, Q., Song, R., Wang, Z., Xu, Z., & Zhao, F. (2024). Enhancing Bank Credit Risk Management Using the C5. 0 Decision Tree Algorithm. Journal Environmental Sciences And Technology, 3(1), 960-967.
[35] Li, H., Wang, S. X., Shang, F., Niu, K., & Song, R. (2024). Applications of Large Language Models in Cloud Computing: An Empirical Study Using Real-world Data. International Journal of Innovative Research in Computer Science & Technology, 12(4), 59-69.
[36] Shi, Y., Yuan, J., Yang, P., Wang, Y., & Chen, Z. Implementing Intelligent Predictive Models for Patient Disease Risk in Cloud Data Warehousing.
[37] Zhan, T., Shi, C., Shi, Y., Li, H., & Lin, Y. (2024). Optimization Techniques for Sentiment Analysis Based on LLM (GPT-3). arXiv preprint arXiv:2405.09770.
[38] Lin, Y., Li, A., Li, H., Shi, Y., & Zhan, X. (2024). GPU-Optimized Image Processing and Generation Based on Deep Learning and Computer Vision. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 5(1), 39-49.
[39] Ping, Gang, et al. "Research on Optimizing Logistics Transportation Routes Using AI Large Models." Applied Science and Engineering Journal for Advanced Research 3.4 (2024): 14-27.
[40] Chen, Zhou, et al. "Application of Cloud-Driven Intelligent Medical Imaging Analysis in Disease Detection." Journal of Theory and Practice of Engineering Science 4.05 (2024): 64-71.
[41] Wang, B., Lei, H., Shui, Z., Chen, Z., & Yang, P. (2024). Current State of Autonomous Driving Applications Based on Distributed Perception and Decision-Making.
[42] Yang, P., Chen, Z., Su, G., Lei, H., & Wang, B. (2024). Enhancing traffic flow monitoring with machine learning integration on cloud data warehousing. Applied and Computational Engineering, 67, 15-21.
[43] Jiang, W., Qian, K., Fan, C., Ding, W., & Li, Z. (2024). Applications of generative AI-based financial robot advisors as investment consultants. Applied and Computational Engineering, 67, 28-33.
[44] Li, Zihan, et al. "Robot Navigation and Map Construction Based on SLAM Technology." (2024).
[45] Fan, C., Ding, W., Qian, K., Tan, H., & Li, Z. (2024). Cueing Flight Object Trajectory and Safety Prediction Based on SLAM Technology. Journal of Theory and Practice of Engineering Science, 4(05), 1-8.