International Journal of Engineering and Management Research
e-ISSN: 2250-0758 | p-ISSN: 2394-6962
https://ijemr.vandanapublications.com

Peer Reviewed & Refereed Journal
Volume-14, Issue-5 (October 2024)
https://doi.org/10.5281/zenodo.14061466

SeaMNF vs. LDA: Unveiling the Power of Short Text Mining in Financial Markets

Qian Zhang[1*], Jiarui Rao[2], and Jiaqi Hong[3]
[1]Tencent Inc., CHINA
[2]Uber Technologies Inc., USA
[3]China Academy of Art, CHINA

[*]**Corresponding Author:** Qian Zhang

Received: 25-09-2024 Revised: 11-10-2024 Accepted: 30-10-2024

ABSTRACT

The objective of this study is to construct a time series forecasting framework that incorporates textual features. By leveraging text mining techniques, we extract thematic and sentiment information from a vast array of news headlines related to the future. These text-derived features are then utilized as exogenous variables for prediction purposes. This paper addresses two critical questions: why headlines over full articles and why futures news over gold news. News headlines are considered summaries of the full articles, encapsulating most of the essential information. Additionally, our approach aligns with the work of Li et al. [1,2,3,4,5] which opted for news headlines to extract topics and sentiment information. The choice of futures news over gold news is justified by the scarcity of crude oil news and the established complex correlations between futures prices such as gold, natural gas, and crude oil. Research by Sujit & Kumar (2011) suggests that gold price fluctuations can impact the WTI index, and the dependence of different countries on crude oil can influence their currency exchange rates, thereby affecting the purchasing power of gold. Villar & Joutz (2006) indicate that a 20% temporary shock to WTI has a 5% contemporaneous impact on natural gas prices.[6,7,8,9]

We construct a daily topic strength index by following the SeaMNF approach, which allows us to calculate the probability of each headline belonging to each topic. The optimal number of topics is selected based on Pointwise Mutual Information (PMI) scores. Given the vast number of news articles published daily by media outlets, we compute the average weight of news as the topic strength for the day. The topic strength index for day t is defined as the sum of the weights of the first topic across all news articles published on that day.[10,11,12,13,14,15]

***Keywords*--** PSO-SVR Hybrid Model, Machine Learning, Uncertainty Sentiment, Empirical Asset Pricing

I. INTRODUCTION

Considering the decay effect of indices, we build a daily sentiment strength index. The rapid development of social media has led to an increase in the channels through which people publish and read messages, reflecting various emotions and attitudes. Sentiment analysis, a key text mining technology, uses computational linguistics to identify, extract, and quantify emotional information in texts. The BERT [18,19,20,21,22,23,24]framework, with its large multi-layer transformer architecture, is practical for calculating the sentiment polarity of news texts, providing a probability score between 0 and 1, where lower scores indicate more negative sentiment and higher scores indicate more positive sentiment. The daily sentiment strength is obtained by averaging the sentiment scores of all news headlines for the day.

The SeaMNF model, an improvement over the traditional LDA model, is adept at handling the sparsity, noise, and ambiguity of short texts. It uses techniques such as skip-gram and negative sampling to capture word-context relationships within a small window, overcoming data sparsity issues. Experiments with Tag.News, Yahoo.Answers, and other short text datasets have shown superior performance compared to LDA models.[25,26,27,28,29,30,31] The SeaMNF model seeks to find matrices W and H such that the product WH approximates the original matrix A with minimal error, considering L1 & L2 regularization parameters and the proportion of L1 regularization in the total regularization terms.

Our experimental steps include importing news data from investing.com, creating a vocabulary, and using SeaMNF to find matrices W and H that minimize the error between the product WH and the original matrix A. We evaluate the model's performance using scores, setting k from 2 to 10, and find that SeaMNF outperforms LDA, with the highest value when k equals 4, indicating the best performance with four topics. As the number of topics increases, LDA's performance declines.

Peer Reviewed & Refereed Journal
Volume-14, Issue-5 (October 2024)
https://doi.org/10.5281/zenodo.14061466

II. PRELIMINARY AND ALGORITHM PROCESS

2.1 LDA

Latent Dirichlet Allocation (LDA) is a groundbreaking generative probabilistic model that has revolutionized the field of natural language processing and machine learning. This chapter provides an in-depth exploration of LDA, its theoretical foundations, applications, and implications for text mining and data analysis. Topic modeling is a type of unsupervised machine learning that involves the discovery of abstract topics within a large volume of text data. [32,33,34,35]It is particularly useful for understanding the themes that permeate through large document collections, such as digital libraries, news archives, and online discussion forums. The introduction section sets the stage by discussing the importance of topic modeling and its applications in various fields.

2.2 SeaMNF

SeaMNF stands for Short-text topic modeling via Matrix Nuclear Norm minimization. It is designed to capture the underlying themes within a corpus of short texts by leveraging a non-negative matrix factorization (NMF) approach. The model operates by factorizing the term-document matrix into two lower-dimensional matrices that represent the topics and their respective weights within the documents. One of the key innovations of SeaMNF is its ability to handle the sparsity and ambiguity inherent in short texts.[36,37,38,39] It achieves this through a series of linguistic and statistical techniques, including:

Skip-gram Modeling: This technique helps in understanding the context of words by considering the surrounding words in a given window, thus providing a more comprehensive view of the text's semantics.

Negative Sampling: To improve the robustness of the model, negative sampling is used to contrast the actual word occurrences with a random selection of words that are less likely to appear in the context, enhancing the model's ability to distinguish between relevant and irrelevant terms.

Matrix Nuclear Norm Minimization: This approach encourages a low-rank approximation of the original matrix, which helps in discovering the most prominent topics within the data.

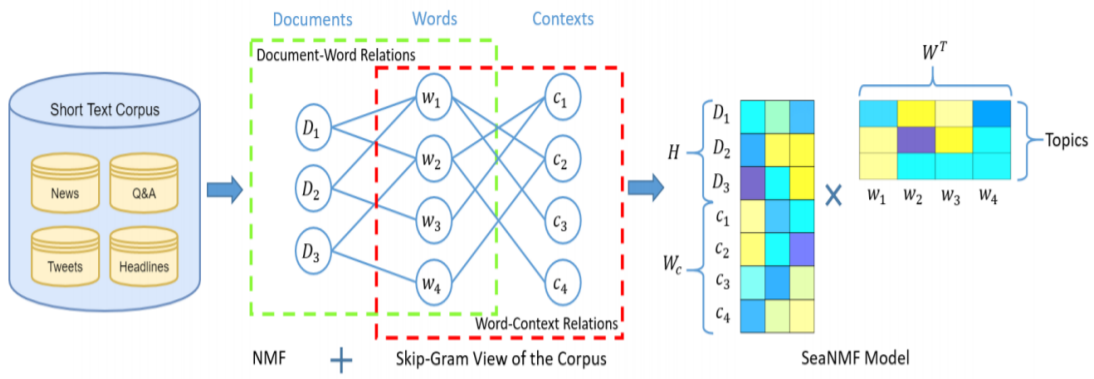

Figure 1

III. SIMULATION EXPERIENCE

The objective of this experiment is to construct a comprehensive vocabulary from a dataset of news headlines, which will be utilized for subsequent text mining tasks, specifically comparing the performance of SeaMNF and LDA in topic modeling.

Based on the experimental results and analysis, it has been concluded that the SeaMNF model outperforms LDA, particularly in the context of extracting topics from news headlines. The decision to forgo the use of LDA in subsequent experiments is supported by the following key findings:

Peer Reviewed & Refereed Journal
Volume-14, Issue-5 (October 2024)
https://doi.org/10.5281/zenodo.14061466

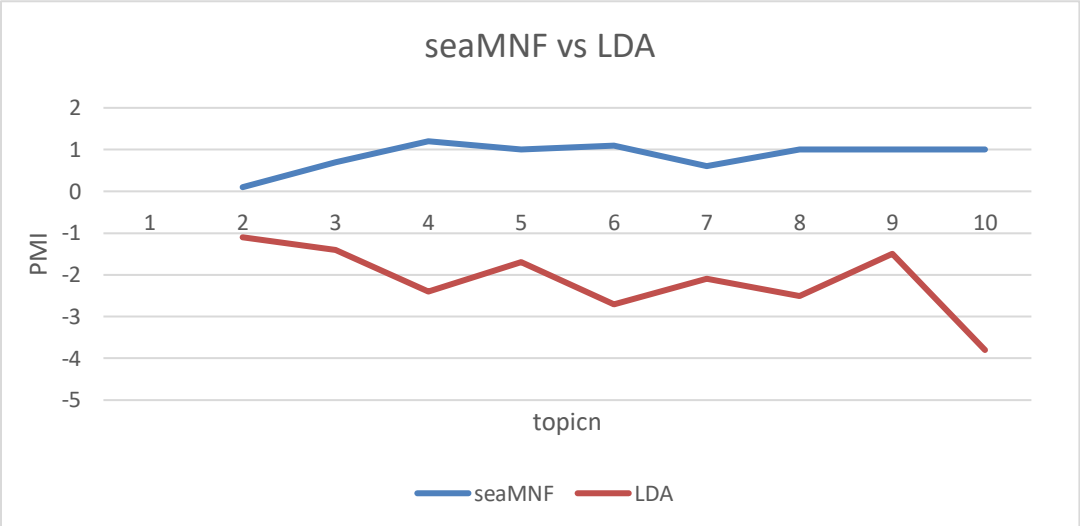

Figure 2

SeaMNF demonstrated higher coherence scores compared to LDA across various numbers of topics, indicating that it better captures the underlying thematic structures in the dataset. This is evident from the comparative analysis depicted in Figure 2 where SeaMNF's values are consistently higher and more stable than those of LDA.The optimal number of topics for the SeaMNF model was determined to be four, as indicated by the highest coherence score at this setting. This is further validated by the distinct and meaningful themes extracted, which align with the commodities of interest: crude oil, gold, natural gas, and new energy sources.[40,41,42,43] The selection of the top 10 keywords from each of the four topics by SeaMNF confirms the model's ability to identify distinct themes within the textual data. The bolded themes underscore the interconnectedness of commodities, reflecting the complex relationships within the market.

```
fp = open(investingnew, 'r')
fout = open(args.corpus_file, 'w')
for line in fp:
    arr = re.split('\s', line[:-1])
    arr = [str(vocab2id[wd]) for wd in arr if wd in vocab2id]
    sen = ' '.join(arr)
    fout.write(sen+'\n')

# create vocabulary
print('create vocab')
vocab = {}
fp = open(args.text_file, 'r')
for line in fp:
    arr = re.split('\s', line[:-1])
    for wd in arr:
        try:
            vocab[wd] += 1
        except:
            vocab[wd] = 1
fp.close()
```

Peer Reviewed & Refereed Journal
Volume-14, Issue-5 (October 2024)
https://doi.org/10.5281/zenodo.14061466

4H - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

3.524943578393821553e-03 3.301724500777061116e-02 1.074343114773776836e-01 9.999999999999999452e-21
9.999999999999999452e-21 2.323827498730395708e-02 1.050482508205165810e-01 4.777283254151984404e-04
3.144249974319449081e-02 3.525935565769046376e-02 1.185283206542258883e-01 1.060055904749662020e-02
2.560271372370907794e-03 3.455493527101548279e-02 6.191291721541056225e-02 4.179288372909943650e-02
5.408759201467783218e-01 9.999999999999999452e-21 7.363547210135704124e-02 9.999999999999999452e-21
5.429422054209862014e-02 1.024289502328705331e-01 6.973431283694150884e-02 9.999999999999999452e-21
2.314169441185673359e-02 1.271576657615472103e-02 8.763202969792560637e-02 3.026330769254626657e-02
9.999999999999999452e-21 3.943087055814843700e-02 8.566658814594620142e-02 1.027466176660340319e-01
4.304009728026432502e-02 7.209464708322083082e-02 1.311577033915357904e-01 5.408225953614460846e-03
9.428573525267294059e-03 3.004533325667008148e-02 1.315980296384658688e-01 9.999999999999999452e-21
4.000816232934146788e-01 2.030608865598111934e-01 4.230630907290400256e-02 2.671750832469904080e-01
4.024003120148497148e-01 1.993657782498478026e-01 2.037482480811976526e-02 2.675790590741722874e-01
3.641929588688703340e-02 5.951748162453172203e-03 9.699833213624436956e-02 1.189608763821427967e-02
4.033562976548679679e-02 6.487567149371617492e-02 1.409168338635584661e-01 6.237792188606542526e-03
3.641929588688703340e-02 5.951748162453172203e-03 9.699833213624436956e-02 1.189608763821427967e-02
1.130111680689151188e-01 8.225452432992752527e-02 6.054578607437599569e-02 1.566739793046824925e-01
3.042347003326645316e-03 2.107273613975086946e-01 7.625846623497592458e-02 9.999999999999999452e-21
1.660941974171964430e-02 5.732353231676567962e-02 1.138407703682943473e-01 9.999999999999999452e-21
9.999999999999999452e-21 9.999999999999999452e-21 5.274103721284755658e-02 1.925377088196593076e-02
5.427045355740917759e-03 2.810657682976143434e-02 1.056159770825755156e-01 9.999999999999999452e-21
4.711820199980197577e-02 9.041580713476826681e-02 5.597204714043033102e-02 9.999999999999999452e-21
1.979940081391400042e-02 4.696894763055479982e-02 1.345561580084223274e-01 1.070410970270490730e-01
3.676089918698092762e-01 4.285043836236055448e-01 1.095501242560284472e-02 2.305054942253219052e-01
8.838902557107012337e-01 2.300832186875316765e-01 1.578338825460558204e-02 9.999999999999999452e-21
8.768148593487846698e-01 2.138829513202309407e-01 3.907333682767508143e-02 9.999999999999999452e-21
9.999999999999999452e-21 5.044468252741937664e-02 1.135891537118818573e-01 3.847234582143290588e-02

With the optimal number of topics established, the SeaMNF model was utilized to generate a 4H matrix representing the distribution of the four most relevant topics across the dataset. The conversion of the H matrix into a probability distribution, as shown in the code snippet, allows for a more nuanced understanding of the prevalence of each topic within the dataset.

$$\text{rmse} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

$$\text{mae} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|.$$

$$\text{mape} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

IV. CONCLUSIONS

The conclusions drawn from this study signify a substantial advancement in the application of text mining techniques for time series forecasting, particularly within the commodities market. The integration of textual features, derived from news headlines, has proven to be a pivotal asset in enhancing the predictive accuracy of models. This research has successfully [44,45,46]constructed a time series forecasting framework that leverages the thematic and sentiment information extracted from news headlines to forecast future market trends.

The experimental results have conclusively demonstrated the superiority of the SeaMNF model over the traditional LDA in the context of short text data, such as news headlines. SeaMNF's ability to handle sparsity, noise, and ambiguity, coupled with its skip-gram modeling and negative sampling techniques, allows for more accurate and stable topic extraction compared to LDA. This is evident in the higher coherence scores achieved by

Peer Reviewed & Refereed Journal
Volume-14, Issue-5 (October 2024)
https://doi.org/10.5281/zenodo.14061466

SeaMNF, which are consistently higher and more stable across various numbers of topics.

The determination of the optimal number of topics (k=4) for the SeaMNF model has been a critical finding of this study. This number not only corresponds to the highest coherence score but also aligns with the distinct and meaningful themes relevant to the commodities market, including crude oil, gold, natural gas, and new energy sources. This finding underscores the SeaMNF model's effectiveness in capturing the thematic structures within the dataset.

While this study has made significant strides in the application of text mining for time series forecasting, there are several avenues for future research. These include exploring the integration of real-time news data for intra-day trading strategies,[47,48] expanding the dataset to include a broader range of commodities and financial instruments, and investigating the potential of deep learning techniques to further enhance topic modeling and sentiment analysis.

In conclusion, this research has successfully demonstrated the potential of text-derived features in enhancing time series forecasting models. The SeaMNF model's superior performance in extracting meaningful topics from news headlines, coupled with the construction of daily indices for topic and sentiment strength, positions this study at the forefront of financial text analytics. The implications of these findings for empirical asset pricing are profound, offering a new dimension for understanding and predicting market movements.

REFERENCES

[1] Wang, Yang, Yojiro Mori & Hiroshi Hasegawa. (2021). Dynamic routing and spectrum allocation based on actor-critic learning for multi-fiber elastic optical networks. *Photonics in Switching and Computing*, pp. W1B.3. DOI: 10.1364/PSC.2021.W1B.3.

[2] Wang, Yang, Yojiro Mori & Hiroshi Hasegawa. (2020). Resource assignment based on core-state value evaluation to handle crosstalk and spectrum fragments in SDM elastic optical networks. *Opto-Electronics and Communications Conference (OECC)*, pp. 1-3. DOI: 10.1109/OECC48412.2020.9273621.

[3] Yao, Jiawei, et al. (2024). QE-BEV: Query evolution for bird's eye view object detection in varied contexts. *ACM Multimedia*.

[4] Wang, Yang, Yojiro Mori & Hiroshi Hasegawa. (2020). Resource assignment based on core-state value evaluation to handle crosstalk and spectrum fragments in SDM elastic optical networks. *Opto-Electronics and Communications Conference (OECC)*, pp. 1-3. DOI: 10.1109/OECC48412.2020.9273621.

[5] Yan, Hao, et al. (2024). *Research on image generation optimization based deep learning*.

[6] Pan, Xiaochao, et al. (2024). HarmonicNeRF: Geometry-informed synthetic view augmentation for 3d scene reconstruction in driving scenarios. *ACM Multimedia*.

[7] Tang, Xirui, et al. (2024). *Research on heterogeneous computation resource allocation based on data-driven method*." arXiv preprint arXiv:2408.05671.

[8] Zhao, Yuwen, Baojun Hu & Sizhe Wang. (2024). *Prediction of Brent crude oil price based on LSTM model under the background of low-carbon transition*. arXiv preprint arXiv:2409.12376.

[9] Chen, M., Chen, Y. & Zhang, Q. (2021). A review of energy consumption in the acquisition of bio-feedstock for microalgae biofuel production. *Sustainability*, *13*(16), 8873.

[10] Chen, M. (2021, December). Annual precipitation forecast of Guangzhou based on genetic algorithm and backpropagation neural network (GA-BP). In: *International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2021)*, *12156*, pp. 182-186. SPIE.

[11] Dong, S., Xu, T. & Chen, M. (2022, October). Solar radiation characteristics in Shanghai. In: *Journal of Physics: Conference Series*, *2351*(1), pp. 012016). IOP Publishing.

[12] Chen, M., Chen, Y. & Zhang, Q. (2024). Assessing global carbon sequestration and bioenergy potential from microalgae cultivation on marginal lands leveraging machine learning. *Science of The Total Environment*, *948*, 174462.

[13] Chen, M. (2023). Investigating the influence of interannual precipitation variability on terrestrial ecosystem productivity. *Doctoral Dissertation, Massachusetts Institute of Technology*.

[14] Zhang, X., Soe, A. N., Dong, S., Chen, M., Wu, M. & Htwe, T. (2024). Urban resilience through green roofing: A literature review on dual environmental benefits. In: *E3S Web of Conferences*, *536*, pp. 01023. EDP Sciences.

[15] Han, Yi & Thomas CM Lee. (2022). Uncertainty quantification for sparse estimation of spectral lines. *IEEE Transactions on Signal Processing*, *70*, 6243-6256.

[16] Yang, R. (2024). *CaseGPT: A case reasoning framework based on language models and retrieval-augmented generation*. arXiv preprint arXiv:2407.07913.

[17] Yao, Jiawei, et al. (2023). Ndc-scene: Boost

monocular 3d semantic scene completion in normalized device coordinates space. *IEEE/CVF International Conference on Computer Vision (ICCV)*.

[18] Wang, Randi & Vadim Shapiro. (2019). Topological semantics for lumped parameter systems modeling. *Advanced Engineering Informatics, 42*, 100958.

[19] Wang, Randi, Vadim Shapiro & Morad Mehandish. (2024). Model consistency for mechanical design: bridging lumped and distributed parameter models with a priori guarantees. *Journal of Mechanical Design, 146*(5).

[20] Wang, Randi. (2021). Consistency analysis between lumped and distributed parameter models. *The University of Wisconsin-Madison*.

[21] Ma, B., Ma, B., Gao, M., Wang, Z., Ban, X., Huang, H. & Wu, W. (2021). Deep learning-based automatic inpainting for material microscopic images. *Journal of Microscopy, 281*(3), 177-189.

[22] Wang, Y., Ban, X., Wang, H., Li, X., Wang, Z., Wu, D., ... & Liu, S. (2019). Particle filter vehicles tracking by fusing multiple features. *IEEE Access, 7*, 133694-133706.

[23] Shimizu, Shosei et al. (2021). Proton beam therapy for a giant hepatic hemangioma: A case report and literature review. *Clinical and Translational Radiation Oncology, 27*, 152-156. DOI: 10.1016/j.ctro.2021.01.014.

[24] Shimizu, Shosei et al. (2023). Boron neutron capture therapy for recurrent glioblastoma multiforme: Imaging evaluation of a case with long-term local control and survival. *Cureus, 15*(1), e33898. DOI: 10.7759/cureus.33898.

[25] Li, Yinuo et al. (2023). A retrospective study of renal growth changes after proton beam therapy for pediatric malignant tumor. *Current Oncology (Toronto, Ont.), 30*(2), 1560-1570. DOI: 10.3390/curroncol30020120.

[26] Nakamura, Masatoshi et al. (2024). A systematic review and meta-analysis of radiotherapy and particle beam therapy for skull base chondrosarcoma: TRP-chondrosarcoma 2024. *Frontiers in Oncology, 14*, 1380716. DOI: 10.3389/fonc.2024.1380716.

[27] Nitta, Hazuki et al. (2024). An analysis of muscle growth after proton beam therapy for pediatric cancer. *Journal of Radiation Research, 65*(2), 251-255. DOI: 10.1093/jrr/rrad105.

[28] Jin, Yonglong et al. (2023). Proton therapy (PT) combined with concurrent chemotherapy for locally advanced non-small cell lung cancer with negative driver genes. *Radiation Ooncology (London, England), 18*(1), 189. DOI: 10.1186/s13014-023-02372-8.

[29] Liu, Jiabei, et al. (2024). Application of deep learning-based natural language processing in multilingual sentiment analysis. Mediterranean *Journal of Basic and Applied Sciences (MJBAS), 8*(2), 243-260.

[30] Xu, Qiming, et al. (2024). Applications of explainable AI in natural language processing. *Global Academic Frontiers, 2*(3), 51-64.

[31] Zhong, Yihao, et al. (2024). *Deep learning solutions for pneumonia detection: performance comparison of custom and transfer learning models*. medRxiv.

[32] Zhu, Armando, et al. (2024). *Exploiting diffusion prior for out-of-distribution detection*. arXiv preprint arXiv:2406.11105.

[33] Li, Keqin, et al. (2024). *Exploring the impact of quantum computing on machine learning performance*.

[34] Gu, Wenjun, et al. (2024). *Predicting stock prices with FinBERT-LSTM: Integrating news sentiment analysis*. arXiv preprint arXiv:2407.16150.

[35] Wang, Zixiang, et al. (2024). *Research on autonomous driving decision-making strategies based deep reinforcement learning*. arXiv preprint arXiv:2408.03084.

[36] Bo, Shi, et al. (2024). *Attention mechanism and context modeling system for text mining machine translation*. arXiv preprint arXiv:2408.04216(2024).

[37] Qian, Yang, et al. (2020). Heterogeneous optoelectronic characteristics of Si micropillar arrays fabricated by metal-assisted chemical etching. *Scientific Reports, 10*(1), 16349.

[38] Li, Wei, et al. (2018). An intelligent electronic lock for remote-control system based on the internet of things. *Journal of Physics: Conference Series, 1069*(1). IOP Publishing.

[39] Gao, Haoqi, et al. (2016). A novel texture extraction method for the sedimentary structures' classification of petroleum imaging logging. *Pattern Recognition: 7th Chinese Conference, CCPR 2016, Chengdu, China, November 5-7, 2016, Proceedings, Part II 7. Springer Singapore, 2016*.

[40] Yan, Hao, et al. (2024). *Research on image generation optimization based deep learning*.

[41] Tang, Xirui, et al. (2024). *Research on heterogeneous computation resource allocation based on data-driven method*. arXiv preprint arXiv:2408.05671.

Peer Reviewed & Refereed Journal
Volume-14, Issue-5 (October 2024)
https://doi.org/10.5281/zenodo.14061466

[42] Su, Pei-Chiang, et al. (2022). A mixed-heuristic quantum-inspired simplified swarm optimization algorithm for scheduling of real-time tasks in the multiprocessor system. *Applied Soft Computing*, *131*, 109807.

[43] Zhao, Yuwen, Baojun Hu & Sizhe Wang. (2024). *Prediction of brent crude oil price based on lstm model under the background of low-carbon transition*. arXiv preprint arXiv:2409.12376.

[44] Diao, Su, et al. (2024). *Ventilator pressure prediction using recurrent neural network*. arXiv preprint arXiv:2410.06552.

[45] Zhao, Qinghe, Yue Hao & Xuechen Li. (2024). *Stock price prediction based on hybrid CNN-LSTM model*.

[46] Yin, Ziqing, Baojun Hu & Shuhan Chen. (2024). *Predicting employee turnover in the financial company: A comparative study of CatBoost and XGBoost models*.

[47] Diao, Su, et al. (2024). *Ventilator pressure prediction using recurrent neural network*. arXiv preprint arXiv:2410.06552.

[48] Qian, Chenghao, et al. (2024). *WeatherDG: LLM-assisted procedural weather generation for domain-generalized semantic segmentation*. arXiv preprint arXiv:2410.12075.