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.13888549

Analyzing Financial News Sentiment with NLP to Forecast Market Trends

Ziwei Wang[1*], Qian Zhang[2], Tianzheng Liu[3] and Chao Li[4]
[1]The Chinese University of Hong Kong, Shenzhen, CHINA
[2]Tencent Inc, CHINA
[3]University of Rochester, USA
[4]Georgetown University, USA

[*]**Corresponding Author:** Ziwei Wang

Received: 29-08-2024 Revised: 14-09-2024 Accepted: 01-10-2024

ABSTRACT

In the era of information explosion, public sentiment, significantly influenced by various information sources including major portals and social media, plays a pivotal role in shaping financial markets. Investor and consumer emotions are highly susceptible to news, rumors, and comments, as exemplified by the "GameStop vs. Wall Street" event where social media influence led to a surge in GameStop's stock value, peaking at $30 billion—over 100 times higher than its value in August. Key social media figures, such as Elon Musk and Donald Trump, can also drastically affect markets with a single tweet, as seen with the Bitcoin and Dogecoin booms following Musk's endorsements.

Market sentiment can escalate in cycles of optimism, amplifying positive news and diminishing negative news, or vice versa in periods of pessimism, leading to market despair. This research employs a BERT model for sentiment analysis of financial emotions and integrates Ensemble Empirical Mode Decomposition (EEMD) with machine learning for financial market analysis, including the commodities and stock markets.

The study utilizes the latest data, including monthly (from January 2005 to September 2020) and weekly (from January 7, 2005, to October 2, 2020) gold price data points. The dataset is strategically divided into an 8:2 ratio for training and testing, with 151 monthly data points for training, 38 for testing, and 657 weekly data points for training, alongside 165 for testing. The Mean Absolute Percentage Error (MAPE) is used to gauge the forecasting model's accuracy, a standard metric for assessing predictive model performance.

By integrating EEMD with correlation analysis, this paper aims to elucidate the volatility of gold price closing values and identify the primary drivers of market fluctuations. The robust methodological framework presented enhances the precision of gold price predictions, offering valuable insights for investors and market analysts.

Keywords-- Financial Market Sentiment, Social Media Influence, Investor Emotion, BERT Sentiment Analysis, Ensemble Empirical Mode Decomposition (EEMD), Machine Learning, Gold Price Forecasting

I. INTRODUCTION

The analysis of sentiment has become a pivotal tool for understanding the underlying emotions and attitudes that drive human behavior. This paper delves into the realm of sentiment analysis, a subfield of natural language processing (NLP)[1] that has garnered significant attention due to its potential applications in various sectors, including but not limited to business, politics, and social sciences.

Sentiment analysis, also known as opinion mining, involves the computational study of people's opinions, sentiments, and emotions towards entities such as products, services, and organizations.[2] The rapid growth of social media platforms, with their vast repositories of user-generated content, has fueled the development of this field. From product reviews to political discourse, the digital footprint of human sentiment is now more accessible than ever, providing a rich dataset for analysis.

This paper aims to explore the theoretical underpinnings and technical aspects of sentiment analysis, with a particular focus on the methodologies that have been developed to extract and interpret emotional expressions from text. We will examine the evolution of the field, the challenges it faces, and the implications of its findings for various stakeholders[4,5,6,7,8,11,12].

The structure of this paper is as follows: The first section will provide a comprehensive overview of sentiment analysis, including its historical context and the driving forces behind its emergence as a critical area of study. The second section will delve into the technical aspects of sentiment analysis, highlighting the key algorithms and models that have been developed to process and analyze sentiment. This will include a discussion on topic-level and sentence-level sentiment analysis, as well as the role of machine learning in enhancing the accuracy and efficiency of sentiment detection.

Peer Reviewed & Refereed Journal
Volume-14, Issue-5 (October 2024)
https://doi.org/10.5281/zenodo.13888549

In the subsequent sections, we will explore case studies that demonstrate the practical applications of sentiment analysis in real-world scenarios. This will be followed by a critical evaluation of the current state of the field, including the limitations of existing methods and the potential for future advancements.

Finally, the conclusion will synthesize the key findings of the paper and propose directions for future research. By the end of this paper, readers will have a thorough understanding of the significance of sentiment analysis in today's data-driven world and the potential it holds for shaping our understanding of human sentiment and behavior.

II. METHODOLOGY

2.1 Traditional Sentiment Modeling

Before the advent of natural language processing technologies, traditional financial sentiment analysis relied on conventional methods. The Volatility Index (VIX)[7,9,10,14,15,16,17], also known as the "fear gauge," is formed due to the observation that the implied volatility of at-the-money options is generally lower than that of out-of-the-money options. Market participants exhibit a higher risk aversion during market downturns, leading to increased demand for put options and a subsequent rise in the implied volatility of deep out-of-the-money put options. The VIX reflects the expectations of options market participants regarding the future volatility of the market index and is often used as a contrarian indicator of market sentiment.

When the VIX is high, it indicates that market participants anticipate increased market volatility and reflect a state of unease; conversely, a low VIX suggests that market participants expect future market volatility to be moderate. The VIX is also considered an investor fear gauge. Typically, the VIX rises during market declines and falls during market advances. From another perspective, extreme highs or lows in the VIX signal that market participants are either in a state of panic, buying put options indiscriminately, or are overly optimistic, neglecting to hedge, which can often be a sign of an impending market reversal.[20,21,22,23,24,25,27,28]

We selected traditional modeling methods to quantify consumer sentiment, including:

Gold Index: The NYSE Arca Gold BUGS Index is a modified dollar-value weighted index of companies involved in gold mining. BUGS stands for a basket of unhedged gold stocks.

VIX: The VIX Index is the ticker symbol for the Chicago Board Options Exchange Volatility Index, which measures the implied volatility of S&P 500 index options. Commonly referred to as the "fear index" or "fear gauge."

FIN State: The Federal Reserve Bank of Chicago's National Financial Conditions Index and its impact on global financial markets.

GVZ (Gold Volatility Index): Gold prices tend to rise during inflation, when the currency (especially the US dollar) weakens, or during significant world events. Therefore, the volatility of gold has a more complex relationship with prices than the stock market.

Oil ETF: Oil ETFs, which have a strong correlation with commodities.

Google Index: The popularity of search terms.

Var	Y	ADF
Gold	D(Y)	(-3.44)***
goldindex	D(X2t)	(-3.44)"…
FINstate	D(X2x)	(-5.76)***
stateChicago FINSTATE	D(X3.t)	(-12.11)**
VIX	D(X4t)	(-3.967)**
GVZ	D(Xs.x)	(-3.001)
3VIX	D(X6t)	(-3.981)
NASDAK100	D(X7)	(-3.952)
oilETF	D(X8t)	(-3.719)***
Google index	D(X9.±)	(-3.719)***

We attempted to incorporate these elements into our model. These traditional sentiment indicators were found to be very noisy time series with correlation coefficients below 0.1. In Granger causality tests, we found that these traditional indicators exhibited significant lag, indicating that market behavior clearly precedes consumer behavior. Therefore, finding an indicator with lag that precedes market behavior is particularly important.

2.2 Google Index Sentiment Modeling

Google Trends is a search trend feature that shows how often a particular search term is entered relative to the total search volume of the website for a given time period. Google Trends can be used to compare keyword research and identify peaks in keyword search volumes triggered by events. Google Trends provides data related to keywords, including search volume indices and geographical information about search engine users.

We can observe that the popularity of different keywords in the Google engine [29,30,31,32,33,34,35,36,37,39]has varying impacts on the peaks and troughs of gold prices, with some keywords

Peer Reviewed & Refereed Journal
Volume-14, Issue-5 (October 2024)
https://doi.org/10.5281/zenodo.13888549

stimulating the gold price and others being stimulated by the gold price, causing flucuations. We used Granger causality tests to see if there is a causal relationship between the first-order differences of the two time series. Mathematical Formula Keywords Stable Granger p-value

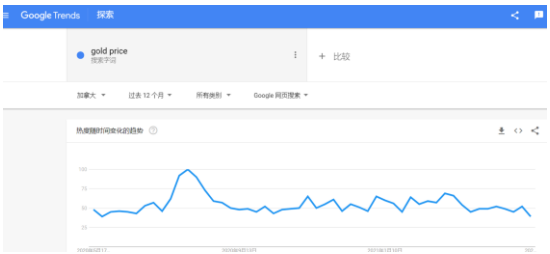


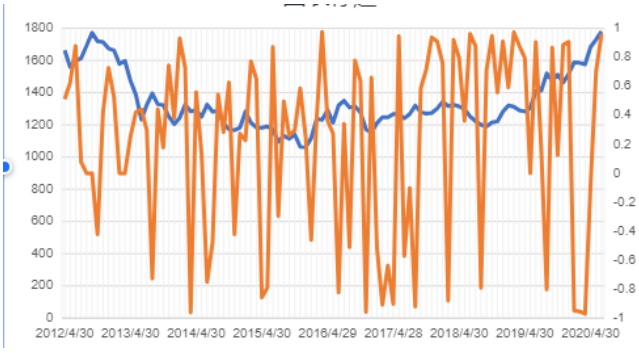
Figure 1: "Sell gold" and Gold Price

It can be seen that most Google Index keywords have a weak causal relationship with gold prices, with only "sell gold" having an impact and passing the Granger test, indicating a causal relationship with gold prices. We can also see from the graph that there is a lot of overlap in the peaks and troughs of the first-order differences of the two series, indicating that the popularity of the "sell gold" keyword has an impact on the gold market's fluctuations. When the popularity of the "sell gold" keyword is high, it leads to panic among gold investors, causing a drop in gold prices. However, through the construction of the Google Index, we also found that the interpretability of Google's keywords is relatively low, necessitating a more interpretable method that can pass time series and correlation coefficient tests. With the advent of natural language processing technology, we have adopted more advanced methods, starting from the latest and most authoritative textual information to quantify sentiment and predict stock prices.

2.3 Multimodal Sentiment Modeling

From our previous two sets of experiments, it is evident that traditional financial sentiment indicators based on VIX and Google Trends have significant limitations, mainly in two aspects. Firstly, the "noise" in the market is too high, with noise coming from traditional financial sentiment indicators that have many errors, especially Google Trends, which often incorrectly counts popularity, such as when counting the keyword "gold price," it often includes "gold price" from the game "Warcraft III," which is clearly unrelated, indicating that some Google Index noise comes from search engine statistical regression. Another part of the noise comes from the inherent noise of public opinion, with a large amount of public opinion information and extreme distribution of polarity leading to too much noise in traditional financial indicators. Another pain point is the lag in public opinion information, which usually follows market behavior. When the price of gold falls, it causes an uproar among investors. Therefore, the transmission is very slow. Only those sequences that precede price changes can be helpful for prediction tasks. Thus, we introduce a multimodal text sentiment modeling method to assist in predicting gold prices, hoping to predict gold prices from a public sentiment perspective. Combining news text for crude oil prediction includes three main parts:

News Title Mining: First, news titles are preprocessed, including tokenization, stopword filtering, and stemming. Then, we use GloVe to embed clean text into word vector matrices. Subsequently, topic modeling and sentiment analysis are used to calculate the topic intensity and



Lag Order Selection: First-order differencing is performed on non-stationary time series. We model the relationship between each exogenous sequence and the crude oil price sequence using the Vector Autoregression (VAR) model to obtain the optimal lag. Then we transform the multivariate time series prediction into a regression problem based on these lagged values.

$$y_t = \sum_{i=1}^{q} \alpha_i x_{t-i} + \sum_{j=1}^{q} \beta_j y_{t-j} + u_{1t}$$
$$x_t = \sum_{i=1}^{s} \lambda_i x_{t-i} + \sum_{j=1}^{s} \delta_j y_{t-j} + u_{2t}$$

Investing.com is a world-renowned financial website that provides real-time information and news on thousands of financial investment products, including global stocks, foreign exchange, futures, bonds, funds, and digital currencies, as well as various investment tools. We collected 28,220 news titles from the Investing.com futures news column as textual data for this study. We collected gold daily data from March 29, 2011, to March 22, 2019, from FRED Economic Data, and the collected news also covers this period. The base oil selected is West Texas Intermediate (WTI) crude oil, which is a common type in North America. Due to the United States' military and economic strength globally, WTI has become the benchmark for global crude oil pricing, and we used web crawling technology to crawl the commodity news from investing.com.

Peer Reviewed & Refereed Journal
Volume-14, Issue-5 (October 2024)
https://doi.org/10.5281/zenodo.13888549

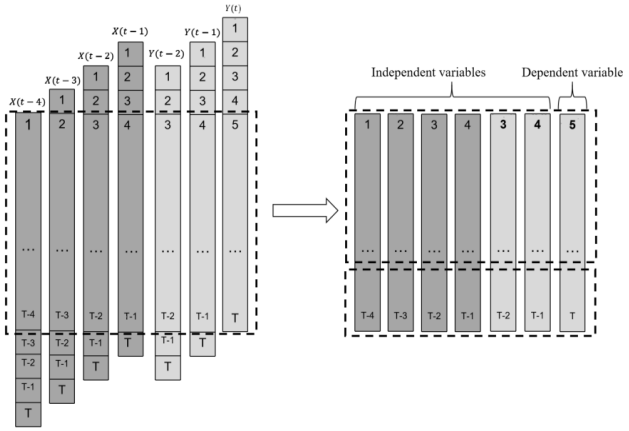
Preprocessing is a basic step in text mining, including word tokenization, stopword filtering, and word embedding. The first two steps aim to convert text into a set of words after removing unimportant words. In short, word embedding is a dimensionality reduction technique that maps high-dimensional words (unstructured information) to low-dimensional numerical vectors (structured information). In other words, word embedding aims to convert documents into mathematical representations that computers can read as input, making it an essential task in text analysis problems.

Stanford University proposed an unsupervised learning algorithm for word representation called GloVe (Pennington et al., 2014) [3], which is a new global log-bilinear regression model designed to combine the advantages of global matrix and local context window methods. In particular, explores the training of a word-word co-occurrence matrix, rather than the entire sparse matrix. Since uses global and local statistical information of words to generate language models and vectorized representations of words, it is a popular word vector representation in the field of natural language processing. This method makes up for some obvious shortcomings of traditional methods. Due to its superior structure, is difficult to use for word analogy tasks. Skip-gram is trained on individual local context windows, preventing it from capturing global information of the corpus.[41,42,43,44,45] considers the co-occurrence relationship of words to construct the embedding

IV. OUR APPROACH

Time series regression means that a target variable can be predicted by some regressors. A common method for predicting multivariate time series is to transform the prediction problem into a regression problem. Let's take a simple example to illustrate this approach. Given an endogenous variable with a lag of 2 and an exogenous variable with a lag of 4, we aim to use these lag values to make predictions. First, we get 4 and 2 copies of and Y, respectively. We then move the copy of the sum, as shown in the left section of the figure, deleting the row where the null value is, to get the data set for the regression model. Finally, the two lags are also included in the independent variable, and the regression equation for the independent variable can be written as the following formula, where these lags can be predicted

$$\hat{Y}_t = f(Y_{t-1}, Y_{t-2}, X_{t-1}, X_{t-2}, X_{t-3}, X_{t-4})$$



data.t1 <-data.frame（topic1，dprice）
data.t2 <-data.frame（topic2，dprice）
data.t3 <-data.frame（topic3，dprice）
data.t4 <-data.frame（topic4，dprice）
data.po <-data.frame（polarity，dprice）
VARselect（data.t1，lag.max = 10，type =“ const”）
VARselect（data.t2，lag.max = 10，type =“ const”）
VARselect（data.t3，lag.max = 10，type =“ const”）
VARselect（data.t4，lag.max = 10，type =“ const”）
VARselect（data.po，lag.max = 10，type =“ const”）
VARselect（dprice，lag.max = 10，type =“ const”）

V. CONCLUSION

The results illustrate the predictive performance of our proposed method compared to other benchmarks at layers 1, 2, and 3. In the table, the model named "no-text" incorporates no textual features; the model named "textblob" utilizes the traditional sentiment strength from the NLTK library; the model named "our method" includes the short-text topic and sentiment features we proposed. The general ARIMA and ARIMAX models encompass three parameters: represents the number of autoregressive terms, □q represents the number of moving average terms, and □d is the number of differences required to render the series stationary. We applied ARIMA (p,d,q) to the no-text model and ARIMAX(p,d,q) to the textblob and our method models. The parameters were set as ARIMA (4, 0, 3), ARIMAX (4, 1, 3), and ARIMAX (4, 1, 3) for the no-text, textblob, and our method, respectively. Appendix 7 lists the textual features selected by our method, from which we can discern the preferences for features among different models. From the results, we can conclude:

Peer Reviewed & Refereed Journal
Volume-14, Issue-5 (October 2024)
https://doi.org/10.5281/zenodo.13888549

REFERENCES

[1] 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)*. IEEE Computer Society.

[2] Bo, Shi & Minheng Xiao. (2022). *Dynamic risk measurement by evt based on stochastic volatility models via MCMC*. arXiv preprint arXiv:2201.09434.

[3] Wang, Randi & Morad Behandish. (2022). *Surrogate modeling for physical systems with preserved properties and adjustable tradeoffs*. arXiv preprint arXiv:2202.01139.

[4] 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.

[5] Li, Zhenglin, et al. (2023). Stock market analysis and prediction using LSTM: A case study on technology stocks. *Innovations in Applied Engineering and Technology*, 1-6.

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

[7] 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.

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

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

[10] Chen, M. (2021, December). (2021). 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] Mo, Yuhong, et al. (2024). Large language model (llm) ai text generation detection based on transformer deep learning algorithm. *International Journal of Engineering and Management Research, 14*(2), 154-159.

[12] 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.

[13] Xiao, Minheng, Shi Bo & Zhizhong Wu. (2024). *Multiple greedy quasi-newton methods for saddle point problems*. arXiv preprint arXiv:2408.00241.

[14] 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.

[15] Bo, Shi. (2022). *Application of K-means clustering algorithm in evaluation and statistical analysis of internet financial transaction data*. arXiv preprint arXiv:2202.03146.

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

[17] Han, Yi & Thomas CM Lee. (2024). Structural break detection in non-stationary network vector autoregression models. *IEEE Transactions on Network Science and Engineering*.

[18] Xiao, Minheng & Shi Bo. (2024). *Electroencephalogram emotion recognition via auc maximization*. arXiv preprint arXiv:2408.08979.

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

[20] 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*, 01023. EDP Sciences.

[21] Mo, Yuhong, et al. (2024). Password complexity prediction based on roberta algorithm. *Applied Science and Engineering Journal for Advanced Research, 3*(3), 1-5.

[22] Bo, Shi & Minheng Xiao. (2024). *Root cause attribution of delivery risks via causal discovery with reinforcement learning*. arXiv preprint arXiv:2408.05860.

[23] 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).

[24] 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.

[25] Yang, Yahe, et al. (2024). *Research on large scene adaptive feature extraction based on deep learning*.

[26] Liu, Tianrui, et al. (2024). Spam detection and classification based on distilbert deep learning algorithm. *Applied Science and Engineering Journal for Advanced Research*, *3*(3), 6-10

[27] 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.

[28] 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.

[29] Dai, Shuying, et al. (2024). The cloud-based design of unmanned constant temperature food delivery trolley in the context of artificial intelligence. *Journal of Computer Technology and Applied Mathematics*, *1*(1), 6-12.

[30] He, Shuyao, et al. (2024). Lidar and monocular sensor fusion depth estimation. *Applied Science and Engineering Journal for Advanced Research*, *3*(3), 20-26

[31] 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.

[32] Li, Shaojie, Yuhong Mo & Zhenglin Li. (2022). Automated pneumonia detection in chest x-ray images using deep learning model. *Innovations in Applied Engineering and Technology*, 1-6.

[33] 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.

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

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

[36] Song, Jintong, et al. (2024). A comprehensive evaluation and comparison of enhanced learning methods. *Academic Journal of Science and Technology*, *10*(3), 167-171.

[37] Mo, Yuhong, et al. (2024). Make scale invariant feature transform "fly" with CUDA. *International Journal of Engineering and Management Research*, *14*(3), 38-45.

[38] Liu, Jihang, et al. (2024). Unraveling large language models: From evolution to ethical implications-introduction to large language models. *World Scientific Research Journal*, *10*(5), 97-102.

[39] Yan, H., Wang, Z., Bo, S., Zhao, Y., Zhang, Y. & Lyu, R. (2024). *Research on image generation optimization based deep learning*.

[40] Tang, X., Wang, Z., Cai, X., Su, H. & Wei, C. (2024). *Research on heterogeneous computation resource allocation based on data-driven method*. arXiv preprint arXiv:2408.05671.

[41] Li, W., Li, H., Gong, A., Ou, Y. & Li, M. (2018, August). An intelligent electronic lock for remote-control system based on the internet of things. In: *Journal of Physics: Conference Series*, *1069*(1), pp. 012134. IOP Publishing.

[42] Wu, Z., Wang, X., Huang, S., Yang, H. & Ma, D. (2024). Research on prediction recommendation system based on improved markov model. *Advances in Computer, Signals and Systems*, *8*(5), 87-97.

[43] Huang, S., Yang, H., Yao, Y., Lin, X. & Tu, Y. (2024). *Deep adaptive interest network: personalized recommendation with context-aware learning*. arXiv preprint arXiv:2409.02425.

[44] Lu, Q., Guo, X., Yang, H., Wu, Z. & Mao, C. (2024). Research on adaptive algorithm recommendation system based on parallel data mining platform. *Advances in Computer, Signals and Systems*, *8*(5), 23-33.

[45] Tan, C., Wang, C., Lin, Z., He, S. & Li, C. (2024). Editable neural radiance fields convert 2d to 3d furniture texture. *International Journal of Engineering and Management Research*, *14*(3), 62-65.

[46] Qi, Z., Ma, D., Xu, J., Xiang, A. & Qu, H. (2024). *Improved YOLOv5 based on attention mechanism and fasternet for foreign object detection on railway and airway tracks*. arXiv preprint arXiv:2403.08499.

[47] Xiang, A., Huang, B., Guo, X., Yang, H. & Zheng, T. (2024). *A neural matrix decomposition recommender system model based on the multimodal large language model*. arXiv preprint arXiv:2407.08942.

[48] Xiang, A., Qi, Z., Wang, H., Yang, Q. & Ma, D. (2024). *A multimodal fusion network for student emotion recognition based on transformer and tensor product*. arXiv preprint arXiv:2403.08511.