3MEthTaskforce: Multi-source Multi-level Multi-token Ethereum Data Platform

Anonymous Author(s)

Abstract

This paper introduces 3MEthTaskforce (**3meth.github.io**), a comprehensive multi-source platform that integrates token transactions, token-specific information, global market indices, and sentiment data from the Ethereum network. It addresses key challenges in cryptocurrency research, particularly the need for a more holistic approach to understanding market risks and user behavior. 3MEthTaskforce highlights three critical tasks: User Behavior Prediction, Token Price Prediction, and User Behavior Marking, utilizing machine learning models such as dynamic GNNs to tackle these tasks. By offering benchmarks for these tasks, the platform facilitates deeper insights into the behavioral risks associated with cryptocurrency investments, enabling stakeholders to anticipate sharp market movements and mitigate risks.

CCS Concepts

• **Information systems → Data mining**; • **Computing methodologies →** *Neural networks*; Multi-agent systems; • **Applied computing → Electronic data interchange**.

Keywords

cryptocurrency, Ethereum, risk, benchmark, sentiment

ACM Reference Format:

Anonymous Author(s). 2018. 3MEthTaskforce: Multi-source Multi-level Multi-token Ethereum Data Platform. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 15 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The cryptocurrency market has experienced remarkable growth over the past decade, driven by the rapid development of Ethereum and the widespread adoption of the ERC20 token standard. As of 2024, the total market capitalization of cryptocurrencies stands at approximately 1.5 trillion USD, with over 300 million people holding cryptocurrencies and around 10,025 active tokens in circulation[1,2]. Ethereum alone handles an average of 1.2 million transactions per day[3], underscoring the vast scale and complexity of the cryptocurrency ecosystem.

However, this market is notoriously volatile, influenced by a range of factors including token prices, global market indicators, user behavior, and sentiment. The UST de-pegging event in May 2022 [9, 47], which triggered widespread panic and the collapse of LUNA (now LUNC), highlighted how rapidly sentiment can shift and impact related assets. As a result, researchers are increasingly focused on using machine learning to predict trends, detect anomalies, and assess risks in cryptocurrency markets [2, 10, 11, 25].

Despite the growing interest in machine learning for cryptocurrency analysis, current datasets are often fragmented, focusing on narrow aspects of the market. For instance, repositories like Zenodo [36] and Chartalist [42] provide transaction-level data, while others, like Kaggle and Huggingface[4], offer textual data from news and social media. These datasets typically isolate specific components, missing the broader picture of how factors like transaction activity, sentiment, and global market trends interact. A comprehensive dataset that integrates multiple sources of data to model these complex interactions is currently lacking.

To address this gap, this paper introduces the **M**ulti-source **M**ulti-level **M**ulti-token **Eth**ereum **Taskforce** (3MEthTaskforce), a platform designed to support machine learning research on cryptocurrency risk analysis. The platform provides a rich dataset consisting of 303 million transaction records from 35 million users across **3,880 tokens**, as well as text-based sentiment data from Reddit (2014–2024) and key market indicators such as 24-hour trading volumes. By integrating these diverse data sources, 3MEthTaskforce enables comprehensive modeling of the relationships between user behavior, market sentiment, and token performance.

In addition to offering this dataset, 3MEthTaskforce defines three core tasks – *User Behavior Prediction*, *Token Price Prediction*, and *User Behavior Marking* – to advance research in cryptocurrency risk analysis. The platform includes benchmarks for each task, featuring 6 dynamic GNN-based models for user behavior prediction and 2 RNN-based models for token price forecasting, with systematic experimental results validating the performance of these models.

Furthermore, the platform introduces a novel method for assessing the *behavioral risk* of cryptocurrency investments, validated through real-world events such as the LUNA incident. This method offers a new framework for evaluating user behavior and improving risk management strategies in decentralized finance (DeFi). The key contributions of this paper are summarized as follows:

- Introducing 3MEthTaskforce, the first multi-source, multi-token platform that integrates transaction data, token information, textual sentiment data, and market indicators, to address gaps in existing single-source datasets.
- Defining three key tasks: User Behavior Prediction, Token Price Prediction, and User Behavior Marking, and supporting research into these tasks by providing comprehensive data and evaluation frameworks.

[1]https://etherscan.io/chart/marketcap
[2]https://www.demandsage.com/blockchain-statistics/
[3]https://ycharts.com/indicators/ethereum_transactions_per_day
[4]https://huggingface.co/datasets?sort=trending&search=blockchain

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

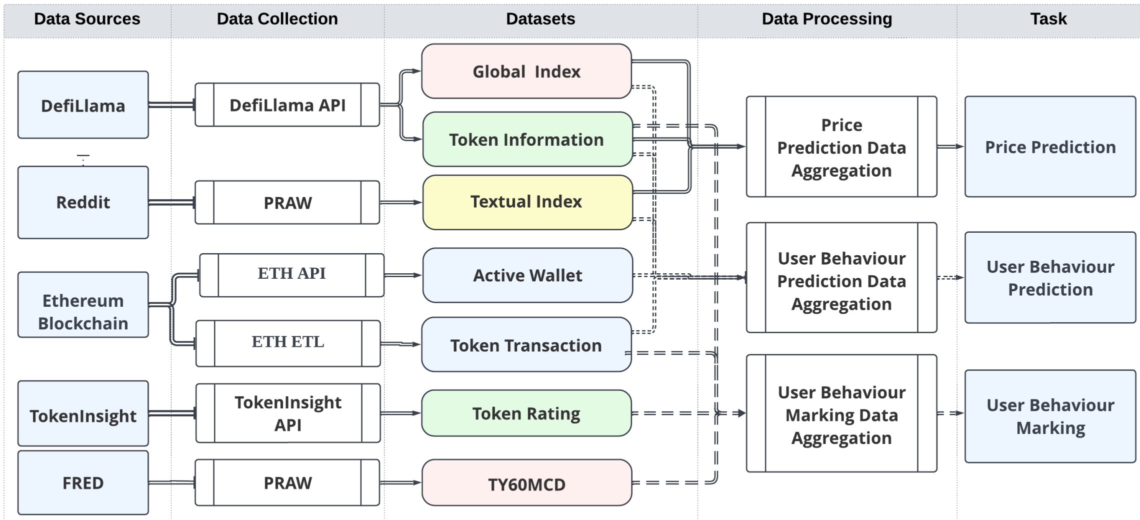

Figure 1: 3MEthTaskforce Data Pipeline. The third column shows datasets from Transaction Record (blue), Token Information (green), Global Market Indices (red), Textual Indices (yellow)

- Offering a comprehensive set of benchmarks for the proposed tasks, with systematic experimental results validating the performance of these models.
- Introducing a novel method for assessing behavioral risk in cryptocurrency investments, validated using real-world events such as the LUNA incident.

2 Related Work

Cryptocurrency-related data repositories. Most existing blockchain data repositories focus on transaction networks. For example, the Ethereum dataset on Zenodo (2021) covers 40 ERC20 tokens but lacks market indices or user sentiment data [36]. Similarly, Chartalist (2022) offers graph learning benchmarks but is limited to a few cryptocurrencies and omits sentiment analysis, global market indices, and user behavior [42]. Public Ethereum datasets on BigQuery, Kaggle, and Athena provide comprehensive transaction data, including complete blockchain records, but lack token prices, global market indices, and social media posts, restricting their use in tasks requiring multi-source data integration. More recent datasets, such as EX-GRAPH (2024), link Ethereum transactions with user social media profiles, enhancing tasks like link prediction and fraud detection, though the text data in this set cannot be used as a market sentiment indicator [49].

Some datasets focus on sentiment in cryptocurrency-related textual data. For example, the dataset by Mohamad et al. (2024) analyzes sentiment in cryptocurrency news and tweets [34], while other similar datasets on Huggingface and Kaggle focus on natural language processing tasks. However, none integrate textual data with price trends, market indices, or user behavior.

In contrast, the 3MEth dataset integrates multi-token transaction data, long-term community sentiment, and global market indices, offering a comprehensive multimodal dataset for node- and edge-level tasks in blockchain graphs. Its inclusion of community sentiment as a global signal enhances tasks like user behavior analysis and market prediction.

Application of ML in link prediction and price prediction. Graph Neural Networks (GNNs) have proven effective for blockchain-related link-based tasks [39], including anomaly detection [18, 23, 37], user identity inference, and transaction prediction [24, 27, 31]. However, these studies often overlook key features like token price fluctuations, community sentiment, and market indices. Recent benchmarks such as the Temporal Graph Benchmark and Live Graph Lab [22, 55] further showcase GNNs' capabilities but highlight the lack of multi-modal data integration in tasks like user classification and link prediction for NFT networks.

Price prediction is typically framed as a time series problem, with models like LSTM, GRU, and random forests commonly used [10, 17, 26, 56, 57]. While some studies incorporate social media sentiment [19], they often focus on a limited set of tokens. In contrast, our dataset enables time series predictions across thousands of cryptocurrencies, offering broader opportunities for analysis.

Risk assessment in cryptocurrency markets. Most studies on cryptocurrency user behavior adopt a financial perspective, often focusing on market-level price fluctuations to analyze risk [6, 12, 33, 52]. For instance, under conditions of extreme volatility, Bitcoin has been found to be more stable than Ethereum, aligning with predictions from extreme value theory [4]. Some research categorizes market participants into types, such as optimists, pessimists, active, and passive investors [5], while other studies employ machine learning techniques like SVM and k-means clustering to classify investor types and identify behavioral patterns [30]. While progress has been made in understanding broader cryptocurrency investment scenarios, these studies generally offer macro-level assessments and fall short of quantifying specific user behaviors. Although some research labels investment behaviors as "high risk" or "low risk" [29], they lack a detailed, data-driven method for scoring individual investment behaviors.

3 The 3MEthTaskforce Dataset

The 3MEthTaskforce dataset is designed to integrate *multi-source*, *multi-level*, and *multi-token* data. Appendix A provides some examples of raw data. The dataset offers a rich and diverse collection of data spanning transactions, token information, global market indicators, and social media texts. Below is a detailed description of each section of the datasets. Figure 1 is a flow chart showing the 3MEthTaskforce pipeline from raw data, to datasets, to tasks.

Section 1: Token Transactions. This section provides **303 million transaction records** from **3,880 tokens** and **35 million users** on the Ethereum blockchain, stored in **3,880 CSV files**, each representing a specific token. Each transaction includes:

- Sender and receiver wallet addresses: Enables network analysis and user behavior studies.
- Token address: Links transactions to tokens for token-specific analysis.
- Transaction value: Reflects the number of tokens transferred, essential for liquidity studies.
- Blockchain timestamp: Captures transaction timing for temporal analysis.

See Table 8 for examples. Apart from the large csv file, we also provide a smaller csv file containing 267, 242 **transaction records** of 29, 164 **wallet addresses**. This smaller dataset involves a total of 1, 194 tokens, covering the time period T_s = [Sep 2016, Nov 2023]. This detailed transaction data is critical for studying user behavior, liquidity patterns, and tasks like link prediction and fraud detection.

Section 2: Token Information. This section offers metadata for **3,880 tokens**, stored in corresponding CSV files, each with:

- Timestamp: Marks the time of data update.
- Token price: Useful for price prediction and volatility studies.
- Market capitalization: Reflects the token's market size and dominance.
- 24-hour trading volume: Indicates liquidity and trading activity.

Additionally, two CSV files provide *rating data* for **269 tokens** from TokenInsight.com, with metrics such as performance and team strength. This data supports token price prediction, risk analysis, and token ranking.

Section 3: Global Market Indices. This section provides *macro-level data* to contextualize token transactions, stored in separate CSV files. Key indicators include:

- Bitcoin dominance: Tracks Bitcoin's share of the cryptocurrency market.
- Total market capitalization: Measures the overall market's value, with breakdowns by token type.
- Stablecoin market capitalization: Highlights stablecoin liquidity and stability.
- 24-hour trading volume: A key measure of market activity.
- Treasury Yield for 60-month Certificates of Deposit (TY60MCD): A macroeconomic benchmark for risk-free returns.

These indices are essential for integrating global market trends into predictive models for volatility and risk-adjusted returns.

Section 4: Textual Indices. This section contains sentiment data from Reddit's Ethereum community, covering **7,800 top posts from 2014 to 2024**. Each post includes:

- Comment text: For sentiment analysis and NLP tasks.
- Post score (net upvotes): Reflects engagement and sentiment strength.
- Timestamp: Aligns sentiment with price movements.
- Number of comments: Gauges sentiment intensity.
- Sentiment indices: Sentiment scores computed using methods presented in Section 5.1.

This data is valuable for understanding social dynamics in the market and enhancing sentiment analysis models that can explain market movements and improve behavioral predictions.

3.1 Data Collection

We employed a multi-source strategy, integrating data from five providers to ensure diversity and representativeness.

Ethereum Public ETL Tool for Token Transactions: We used the Ethereum Public ETL tool [33] to collect token transaction data for the 3MEth dataset. This tool efficiently extracts, transforms, and loads transaction data from the Ethereum blockchain, providing 3, 880 token transaction records. Additionally, we used a free Ethereum API (etherscan.io/) to gather 5, 855 active wallet addresses and their transactions. This results in 267, 242 transaction records among 29, 164 wallet addresses and 1, 194 tokens which form our smaller dataset.

DefiLlama for General Token Information: DefiLlama[5] was used to gather historical prices for 3, 880 tokens and global market indices (excluding TY60MCD), accessing data from multiple exchanges.

TokenInsight API for Token Ratings: We obtained token ratings for 269 Ethereum tokens using the TokenInsight API[6], which provides evaluations based on market performance, technical strength, and other key factors.

Reddit API (PRAW) for Textual Data: The Python Reddit API Wrapper (PRAW)[7] was used to extract approximately 7, 800 timestamped posts from eight popular cryptocurrency subreddits (e.g., r/Ethereum, r/CryptoCurrency).

TY60MCD from Federal Reserve Economic Data: Treasury Yield for 60-month Certificates of Deposit (TY60MCD) data was collected from the Federal Reserve Economic Data (FRED)[8], providing essential macroeconomic indicators.

3.2 Ethics and Privacy

This research adheres to ethical guidelines to minimize privacy risks. Ethereum's public blockchain and publicly sourced Reddit data ensure transparency, with no private content accessed. The dataset maintains pseudonymity, anonymizing transactions and Reddit posts by masking identifiers and personal information. Strict protocols prevent re-identification or unauthorized data use. The

[5]https://defillama.com/
[6]https://tokeninsight-api.readme.io/reference/get_rating-coins
[7]https://defillama.com/
[8]https://fred.stlouisfed.org/series/TY60MCD

study aims to provide valuable insights into cryptocurrency trading and decentralized finance, with potential benefits outweighing minimal risks. See more details in Appendix B.

4 Tasks

The 3MEthTaskforce platform introduces three key tasks: *User Behavior Prediction*, *Token Price Prediction*, and *User Behavior Marking*. These tasks utilize the platform's diverse dataset to model and predict interactions between users and tokens, providing a framework to systematically study and mitigate behavioral risks driven by both individual decisions and broader market factors. To formally define these tasks, we introduce the following key terminologies:

- **Transaction History**: Captures all user buy and sell activities across various tokens over a time period $(1, 2, \ldots, \tau)$, provided by the Token Transactions section (Section 1). For a user $i \in U$, token $j \in C$, and time step $t \leq \tau$, let $a_{i,j,t} \in \{0, 1\}$ denote a transaction, where $a_{i,j,t} = 1$ indicates a buy/sell transaction. The transaction matrix is:
$$\boldsymbol{A} = (a_{i,j,t})_{i \in U, j \in C, t \leq \tau}.$$
- **Token-Level Features**: Provided by the Token Information section (Section 2), these include time-dependent features for each token $j \in C$:
  - *Price*: $p_{j,t}$, the market value of token j at time t.
  - *Market Capitalization*: $m_{j,t}$, calculated as the product of the token price and circulating supply.
  - *Volume*: $v_{j,t}$, the trading volume at time t.

  These features are represented as time-series vectors over $1, \ldots, \tau$:
$$\boldsymbol{p}_j = (p_{j,1}, \ldots, p_{j,\tau}), \quad \boldsymbol{m}_j = (m_{j,1}, \ldots, m_{j,\tau}), \quad \boldsymbol{v}_j = (v_{j,1}, \ldots, v_{j,\tau}).$$
- **Global Market Indices**: Provided by the Global Market Indices section (Section 3), these reflect broader market trends, such as Bitcoin dominance and total market capitalization. The time-series vector is:
$$\boldsymbol{g} = (g_1, g_2, \ldots, g_\tau).$$
- **Market Sentiment**: Derived from the Textual Indices section of the dataset, market sentiment is captured by a sentiment index s_t, representing collective emotions and opinions. The sentiment is represented as a time-series vector:
$$\boldsymbol{s} = (s_1, s_2, \ldots, s_\tau).$$

This provides the essential terminologies for modeling user behavior and market conditions in our tasks.

4.1 User behavior prediction

This task aims to forecast users' purchasing and selling behaviors in the cryptocurrency market. By predicting which users are likely to buy or sell tokens and identifying the specific tokens they are most likely to transact with, stakeholders – such as traders, investors, and institutions – can anticipate sharp market movements and respond proactively. For this task, we consider the interplay of multiple factors influencing user behavior.

Task definition. The goal of the *user behavior prediction* task is to forecast future transactions $\boldsymbol{A}_{:,:,t}$ at a future time $t > \tau$, based on historical data. The task can be formally defined as follows:

User behavior prediction task	
Input:	a transaction matrix $\boldsymbol{A}$, price vectors $(\boldsymbol{p}_j)_{j \in C}$, market capitalization vectors $(\boldsymbol{m}_j)_{j \in C}$ and volume vectors $(\boldsymbol{v}_j)_{j \in C}$ of all tokens, global market vector $\boldsymbol{g}$ and market sentiment vector $\boldsymbol{s}$.
Output:	predicted transactions $\boldsymbol{A}_{:,:,t}$ at time $t > \tau$

4.2 Token Price Prediction

This task involves forecasting the future price of a cryptocurrency token, providing insights into price fluctuations and supporting informed investment decisions. We approach this task in a broader market context by incorporating historical token data, global market indices, and sentiment analysis from the textual indices. The task is formulated as:

Token price prediction task	
Input:	price vectors $(\boldsymbol{p}_j)_{j \in C}$ of all tokens, global market vector $\boldsymbol{g}$ and market sentiment vector $\boldsymbol{s}$.
Output:	predicted prices of all tokens $(p_{j,t})_{j \in C}$ at time $t > \tau$

4.3 User behavior marking

This task is designed to evaluate the investment risk associated with users' cryptocurrency trading activities. While risk has many interpretations, in this work, we specifically focus on the risk stemming from user investment behaviors. Unlike traditional financial markets, cryptocurrency markets introduce additional uncertainties driven by extreme price fluctuations, speculative behavior, and the nascent nature of digital assets. Evaluating how users' buying and selling decisions respond to such market volatility is a vital aspect of understanding investment risk for traders, investors, and institutions alike [32].

To complete this task, one may utilize insights from two primary sources, *user selling behavior* and *predicted token prices*, in conjunction with *historical price vectors* to generate a *risk score*. The risk score reflects how a user's trading behavior, aligns with the price movements of the traded tokens. The objective is to assess the degree of risk involved in an individual's recent investment decisions, accounting for market volatility and the potential for significant financial loss.

To define the task formally, given an event where user $i \in U$ sells token $j \in C$ at time t_2, the first step is to identify the time t_1 when the user initially purchased token j. We then construct a price vector for token j over the time period from t_1 to t_2, denoted as $\boldsymbol{p}_{j,t_1:t_2} = (p_{j,t_1}, p_{j,t_1+1}, \ldots, p_{j,t_2})$. This price vector may be derived from historical prices, predicted prices, or a combination of both, capturing the market fluctuations that occurred during the user's holding period.

The User Behavior Marking task takes this constructed price vector as input and returns a risk score $r_{i,j}$ that indicates the level of risk associated with user i's investment in token j. The formulation of the task is presented as follows:

Limitations: Some users, in order to mitigate risk and protect privacy, use a wallet to trade only a single token or a small number

User behavior marking task	
Input:	user i, token j, price vector $\boldsymbol{p}_{j,t_1:t_2}$
Output:	risk score $r_{i,j}$

of tokens [53], which implies that a single user may possess multiple addresses. However, our user behavior marking strategy is based on the assumption of a one-to-one relationship between users and addresses. For instance, when a user has multiple addresses, transactions across different addresses are treated as independent activities, making it difficult to effectively consolidate them and thus failing to capture the user's overall trading patterns. Future work could address this limitation by developing methods to capture the behavior associated with multiple addresses belonging to the same user.

5 Baselines and Experiments

3MEthTaskforce implementes and evaluates several baseline methods for the tasks defined above.

5.1 User behavior prediction

5.1.1 Baseline methods. We treat this task as a *link prediction task* on an *edge-labelled temporal bipartite graph*. The set of vertices V of the graph is $U \cup C$ containing both the set of users and tokens. A *temporal edge* is of the form (u, c_j, t) where $u \in U$, $c_j \in C$, and $1 \le t \le \tau$ is the timestamp of the edge. This edge denotes a transaction where user u purchases some amount of token c_j at timestamp t. For each temporal edge (u, c_j, t), associate q and $\text{trans}_{j,t}$ to form a *labelled temporal edge* $(u, c, q, \text{trans}_{j,t})$, where q represents the amount of token c_j purchased during this transaction, and $\text{trans}_{j,t}$ is a *transaction label*, discussed below.

For the user behavior prediction task, we construct a graph (V, E) where E contains the set of all such labelled temporal edges up to a certain timestamp τ. This graph can be constructed using the dataset in 3MEthTaskforce. Our goal is to train a GNN that predicts, for any $u \in U$ and $c \in C$ and a future timestamp $t' > \tau$, whether a temporal edge (u, c, t') will appear.

To describe our baseline methods in detail, we need to elaborate on the following issues.

Textual Sentiment Index Extraction. To convert textual data into sentiment indices, we concatenate the top posts at time t into a vector $\textbf{txt}_t = \text{txt}_{t,1} \mid \text{txt}_{t,2} \mid \cdots \mid \text{txt}_{t,\ell}$. We then apply two methods to extract sentiment indices from $\textbf{txt}_t$:

(1) Following [7], we use a large language model (DeepSeek) to generate an overall sentiment score s_t^{overall} on a scale of 0 to 10, where 5 represents neutral sentiment, below 5 indicates negative sentiment, and above 5 indicates positive sentiment.

(2) The second method refines the first by incorporating the post timestamps to account for cryptocurrency trends and generating two separate scores: s_t^{pos} for positive sentiment and s_t^{neg} for negative sentiment.

The prompts used for these methods are provided in Appendix D. The sentiment index aggregation methods are provided in Appendix F

Transaction Labels. To comprehensively analyze performance of models for this link prediction task, we define several different transaction label vectors:

- $\text{trans\_record}_{j,t} = (p_{j,t}, m_{j,t}, v_{j,t})$ includes only token information.
- $\text{trans\_global}_{j,t} = g_t$ includes only global market index.
- $\text{trans\_text}_{j,t} = s_t^{\text{overall}}$ includes the only overall sentiment score.
- $\text{trans\_text\_llm}_{j,t} = (s_t^{\text{pos}}, s_t^{\text{neg}})$ contains both positive and negative sentiment scores.
- $\text{trans\_all}_{j,t} = (p_{j,t}, m_{j,t}, v_{j,t}, g_t, s_t^{\text{overall}}, s_t^{\text{pos}}, s_t^{\text{neg}})$ includes all features: token information, global market index, and sentiment scores.

3MEthTaskforce implements **six baseline dynamic GNN models** to the constructed graph G: DyGFormer [54], JODIE [28], DyRep [46], TGAT [51], TGN [41], and TCL [48]. These models were chosen as they were recently proposed for this task and have demonstrated promising performance in various link prediction tasks. For each of these models, we measure the performance of link prediction using all six transaction label vectors above. A detailed description of these models can be found in Appendix C.

5.1.2 Experiment Setup. We aim to compare different GNN models, input features, and the impact of two types of sentiment features on performance.

Dataset. We use the smaller dataset from the Transaction Record section, containing approximately 260,000 transactions and 29,164 active wallet addresses. From this, we extract the transaction label vectors as defined above. The datasets are split chronologically into train/validation/test sets with a 70%/15%/15% ratio.

Parameter Settings. We train models using Adam [15] with binary cross-entropy loss. All models are trained for 100 epochs, with early stopping after 20 epochs of no improvement. The learning rate is set to 0.0001, and the batch size is 200.

Performance Metrics. Following [38, 40, 50, 51], we evaluate user behavior prediction using test set average precision (TAP) and new node average precision (NAP). The inductive negative sampling strategy is described in [38]. All results are averaged over three runs.

5.1.3 Experiment Results. Table 1 reports the performance of six models across datasets using TAP and NAP as metrics.

Impact of LLM-enhanced sentiment. Models incorporating LLM-derived sentiment scores (`trans_text_llm`) outperform those using only overall sentiment (`trans_text`). For example, TCL improves from 0.737 NAP on `trans_text` to 0.786 on `trans_text_llm`, demonstrating the benefit of incorporating background knowledge from LLMs for sentiment annotation.

Effectiveness of additional features. Adding features like token information and sentiment improves performance across models. For instance, DyRep achieves a TAP of 0.926 on `trans_text_llm`, compared to 0.914 using only transaction data. Similarly, DyGFormer sees a TAP increase from 0.928 on transactions to 0.938 on `trans_text_llm`, indicating that these additional features help predict user-token interactions more effectively.

Table 1: Comparison of six dynamic GNN models on six datasets for the User Prediction task. TAP is the average precision on the test set, and NAP is the average precision on the new node set. An underline indicates the best result in each column, and bold font highlights the best result in each row.

Dataset	DyRep		TCL		TGN		TGAT		JODIE		DyGFormer	
	tap	nap	tap	nap	tap	nap	tap	nap	tap	nap	tap	nap
trans	0.914	0.870	0.859	0.778	0.917	0.860	0.878	0.794	**0.935**	**0.898**	0.928	0.879
trans_record	0.918	0.876	0.870	0.794	0.899	0.834	0.861	0.773	**0.930**	**0.890**	0.925	0.879
trans_global	0.909	0.864	0.853	0.767	0.881	0.810	0.834	0.750	0.931	0.889	**0.939**	**0.905**
trans_text	0.923	0.876	0.835	0.737	0.917	0.861	0.826	0.705	**0.942**	**0.900**	0.919	0.870
trans_text_llm	0.926	0.881	0.867	0.786	0.926	0.873	0.841	0.733	**0.955**	**0.920**	0.938	0.896
trans_all	0.922	0.869	0.846	0.750	0.887	0.806	0.845	0.743	**0.941**	**0.898**	0.913	0.860

Performance of different GNN models. JODIE and DyGFormer consistently perform best, with JODIE reaching a TAP of 0.941 on `trans_all` and 0.955 on `trans_textual_llm`. DyRep shows stable results, with a high of 0.926 on `trans_text_llm`, while TCL performs weaker, reaching only 0.750 TAP on `trans_all`.

5.2 Price Prediction

5.2.1 Baseline Method. We approach the price prediction task as a *time series forecasting* problem using two baseline models: LSTM and GRU [14, 16]. See Appendix E. We define four input vectors for evaluation:

- $\text{price}_{j,t} = p_{j,t}$
- $\text{price\_global}_{j,t} = (p_{j,t}, g_{j,t})$
- $\text{price\_text}_{j,t} = (p_{j,t}, s_t^{\text{pos}}, s_t^{\text{neg}})$
- $\text{price\_all}_{j,t} = (p_{j,t}, g_{j,t}, s_t^{\text{pos}}, s_t^{\text{neg}})$

5.2.2 Experiment Setup. The experiment aims to: (1) Compare LSTM and GRU across multiple tokens. (2) Test the impact of additional features (global index and sentiment). (3) Assess prediction performance based on how long tokens have been on the market.

We classify all tokens by the time in which they appear on the market. This can be extracted from the Token Information Section of the datasets. In particular, fix $x \in [0, 1]$, and call a token c_j *x-recent* if the first timestamp x_j in which c_j has a non-zero price record is less than or equal to x, i.e., c_j appears on the market before x of the total timestamps have elapsed. For our experiment, we set $x \in \{0, 0.4, 0.8\}$. Intuitively, the 0-recent, 0.4-recent, and 0.8-recent tokens represent tokens with the longest, longest-to-medium, and longest-to-shortest market presence:

- 0-recent tokens: 4 tokens (Peercoin, MaidSafeCoin, Swarm Network, OKcash)
- 0.4-recent tokens: 211 tokens (e.g., Obyte, Blox, Augur, AppCoins)
- 0.8-recent tokens: 1592 tokens (e.g., Chimpion, GogolCoin, PoolTogether)

We use *mean squared error* (MSE) as the performance metric and report the average performance across all tokens.

5.2.3 Results. In Table 2, we report the performance of the above four datasets on both LSTM and GRU models.

Effect of adding sentiment information. Incorporating sentiment information slightly improves model performance for tokens with shorter market history. Specifically, when $x = 0.8$, the MSE of LSTM on the price is 0.532, whereas that of LSTM on the price_text decreases to 0.511. MSE of GRU over price is 1.077, while that of GRU over price_text decreases to 1.019.

Effect of value of x. As x increases (i.e., adding more new tokens), both models show a decreasing trend in prediction performance, but their performance significantly improves on datasets that include global and sentiment features. For example, in both LSTM and GRU models, the MSE on price increases as x moves from 0 to 0.8, with the LSTM's MSE increasing from 0.001 to 0.532, and GRU's MSE increasing from 0.001 to 1.077. However, on price_global and price_all, the MSE significantly decreases as x increases. Specifically, in LSTM, the MSE on the price_global drops from 83.910 to 2.242, while in the GRU model, the MSE decreases from 45.330 to 2.259.

Performance of different models. Overall, LSTM model performs slightly better than GRU on most datasets, especially on those that include additional features. For instance, on price_all, the MSE of LSTM is 1.667, compared to 2.271 for GRU. Similarly, on price_sentiment, LSTM achieves an MSE of 0.511, whereas GRU's MSE is 1.019.

5.3 User Behavior Marking

5.3.1 Baseline methodology. 3MEthTaskforce deploys a method proposed by [3] to calculate the users' risk score. From CAPM and Tobin's Separation Theorem [43, 44], [3] derives the relationship between the market portfolio and user risk preferences. This approach links traditional financial theories to the cryptocurrency markets, where user behavior can be highly speculative and driven by volatility.

Suppose user i purchases token j at time t_1 and price p_{j,t_1}, and sells it at time t_2 and price p_{j,t_2}. The percentage of the market portfolio is calculated by

$$w = \frac{E_{RM} - R_f}{r\sigma_M^2}.$$

where r is the investor's risk aversion coefficient, R_f is the risk-free rate, $E_{RM} = (p_{j,t_2} - p_{j,t_1})/p_{j,t_1}$ is the expected return of the market portfolio, and σ_M^2 represents its variance, calculated by

$$\sigma_M^2 = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2-1} (\rho_j(t) - \overline{\rho_j})^2$$

Table 2: Comparison of LSTM and GRU models with different x values. The metric is MSE (Mean Squared Error). An underline indicates the best result in each column, and bold font highlights the best result in each row

Dataset	LSTM			GRU		
	$x = 0$	$x = 0.4$	$x = 0.8$	$x = 0$	$x = 0.4$	$x = 0.8$
price	**<u>0.001</u>**	0.006	0.532	**<u>0.001</u>**	0.020	1.077
price_global	83.910	5.645	**<u>2.242</u>**	45.330	7.686	2.259
price_text	0.004	**<u>0.021</u>**	0.511	0.003	0.103	<u>1.019</u>
price_all	40.440	5.069	1.667	34.405	**<u>1.127</u>**	2.271

where $\rho_{j,t} = (p_{j,t+1} - p_{j,t})/p_{j,t}$ is the return at time t, and $\overline{\rho_j}$ represents the average return over the time interval $[t_1, t_2]$.

Let R_f be the real-world risk-free rate (e.g., TY60MCD). Due to its relatively low volatility and limited impact on the cryptocurrency market [1, 8, 13], we treat R_f as a constant, calculated as the average yield of treasury over the past 5 years.

In the context of our dataset, we consider only the user's returns from token volatility, meaning the user's entire portfolio is in the market, implying $w = 1$. In this case, we obtain:

$$r = \frac{E_{RM} - R_f}{\sigma_M^2}.$$

For simplicity, we treat each user's investment as an independent market portfolio. By calculating the market portfolio's expected return E_{RM} and its volatility σ_M^2, we can determine how the user's risk preferences and decisions align with the overall market's performance. A higher risk aversion coefficient r would indicate that the user is more conservative, taking on less risk, while lower r values suggest a higher tolerance for risk. Therefore, we compute the user's risk preference for the current portfolio as

$$Risk(i, j, \boldsymbol{p}_j, t_1, t_2) = r.$$

5.3.2 Experiment Setup. Our experiment aims to validate the risk score computed above using the 3MEthTaskforce dataset. We conduct two primary evaluations:

(1) *TokenInsight rating evaluation:* We analyze the relationship between our calculated risk scores and the token ratings of 269 tokens ranked by the website TokenInsight.com. The rating are AAA, AA, A, BBB, BB, B, CCC, CC, C, D, arranged in descending order, with AAA being the highest rating, followed by AA and A, while D represents the lowest rating.
(2) *Time period risk evaluation:* We investigate whether the general investment principle—short-term trading tends to carry higher risks, while long-term investments tend to carry lower risks—applies to our risk score [21, 35].

For this, we select two sets of tokens for these experiments:

- *TokenInsight rating evaluation:* We use a set of 269 tokens $C = \{c_1, c_2, \ldots, c_{269}\}$, each with a rating R_j) from TokenInsight. Ratings are categorized into 10 levels, from AAA (highest) to D (lowest).
- *Time period risk evaluation:* We use 1,559 tokens with at least two years of price history to assess risk over various investment periods.

For each token c_j, we calculate the risk score for user i over a time periods of d days. The risk score $Risk(i, j, \boldsymbol{p}_j, t_1, t_2)$ is based on the price sequence $\boldsymbol{p}_j[t_1 : t_2]$ between the start time t_1 and end time t_2 where $t_2 - t_1 + 1 = d$.

- For rating evaluation, we calculate the average risk score among all tokens with rating R, where R is one of the ratings from TokenInsight. Let C_R denote the set of tokens with rating R. For each $d \in \{30, 90, 180, 365\}$, we examine how the average risk score among all tokens in C_R for a period of d days varies across different token ratings to validate the correlation between token risk and token rating.
- For time period evaluation, we randomly select two weeks as the end time t_2, with the period ranging from 1 to 26 weeks. We then compute the risk for each user i, where the investment period $d \in \{7, 14, \ldots, 182\}$, and analyze the risk distribution across different investment periods. We assess whether the risk score follows the general investment principle that shorter investment periods carry higher risks, while longer periods carry lower risks, by analyzing risk values across various investment durations.

5.3.3 Result. Our experimental results indicate that all calculated risk values are negative, which suggests that the cryptocurrency market inherently carries high risks. To facilitate analysis and presentation of the results, we reported the absolute values of the risk scores, meaning that higher absolute values correspond to greater risks.

The results shown in Figure 2 demonstrate a significant correlation between users' risk preferences and the token rating provided by TokenInsight. Specifically, investing in tokens with higher ratings indicates a higher risk preference among users, and this trend becomes more pronounced as the investment period increases.

Moreover, the experimental findings shown in Figure 3 reveal a relationship between investment duration and risk. We observed that longer investment horizons are generally associated with more stable and lower risks.

From these analyses, we can conclude that Token Ratings not only reflect the intrinsic ratings of tokens but also serve as an effective measure of users' risk preferences, especially in long-term investments. Additionally, the increase in risk during the Luna event validates the sensitivity and accuracy of our model.

5.4 Case Study: LUNA Incident

LUNC's price began to decline on May 5th [9]. During this period, the cryptocurrency market experienced significant turbulence, with many cryptocurrency prices exhibiting strong fluctuations due to

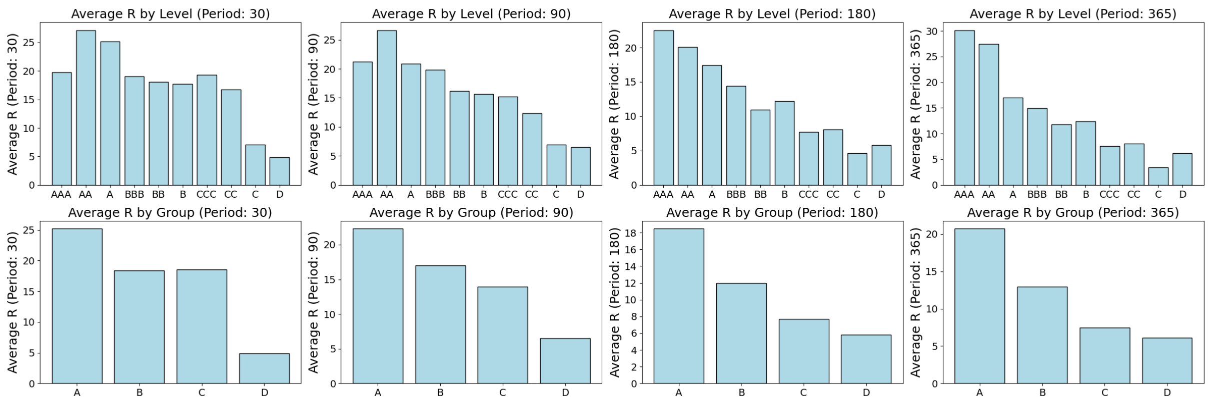

Figure 2: This figure presents the performance of tokens included in the Token Rating in terms of their calculated risk across different trading periods. The top four panels show the average risk of tokens at each rating level during the respective period, while the bottom four panels group tokens with ratings containing A, B, and C, respectively, and display the average risk for each group during the same periods.

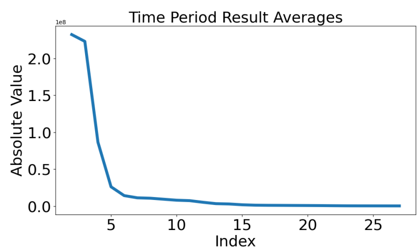

Figure 3: This figure reports the changes in the average risk of 1,599 tokens over an investment period ranging from 1 to 26 weeks.

the impact of the LUNA incident [47]. To compare the investment risk in the cryptocurrency market during the 7 days of the event and the 7 days prior, we randomly selected 200 tokens with price records during this period and calculated their sell transactions from April 28th to May 4th (before the incident) and from May 5th to May 11th (LUNA incident). For simplicity, we assume that the transaction period is no longer than 90 days.

We define the set of transactions during the LUNA event (May 5th–11th) as $\boldsymbol{A}_{\text{luna}}$, where each transaction has a defined buy time t_{1_a} and sell time t_{2_a}. The price change for the token involved in each transaction is noted over this period. The set of users who made these transactions is represented as $\boldsymbol{U}_{\text{luna}}$.

For each transaction during the LUNA event, we calculate a risk score based on the user's behavior and the token's price change between t_{1_a} and t_{2_a}. The average risk for all transactions during this event is then computed by taking the mean of all individual risk scores.

Similarly, for the week prior to the LUNA event (April 28th–May 4th), we define the set of transactions as $\boldsymbol{A}_{\text{common}}$, with corresponding buy and sell times for each transaction. The users involved in these transactions are denoted by $\boldsymbol{U}_{\text{common}}$, and the risk scores are calculated in the same way as for the LUNA event. The average risk for these transactions is also computed by averaging all individual risk scores.

By comparing the average risk during the LUNA event with the average risk in the prior week, we can assess how the market turbulence caused by the LUNA incident impacted investment risk in the cryptocurrency market.

Indeed, during the Luna incident, our computed risk score reaches **24,658,527.37**, significantly higher than the **21,353,272.87** observed in the week prior, further supporting this conclusion. This result aligns with our intuitive expectations.

6 Broader Impact

The rapid growth of blockchain in decentralized finance (DeFi) has introduced both new opportunities and more complex risk management challenges. Incidents like the Luna de-pegging have exposed the limitations of traditional risk assessment methods in addressing the high-frequency, dynamic nature of cryptocurrency transactions. To tackle these issues, 3MEthTaskforce employs an improved financial model to deliver more precise risk calculations for cryptocurrency trading activities. This framework not only enhances traditional risk management but also sets the stage for machine-learning-driven, proactive risk monitoring systems.

The 3MEthTaskforce dataset, by integrating diverse data sources, creates a robust foundation for tasks like user behavior prediction and price forecasting. By combining these machine learning models with financial formulas, we can anticipate market trends, enhance predictive capabilities, and offer a more forward-looking approach to risk management.

As DeFi continues to evolve, the tools provided by 3MEthTaskforce will be crucial in advancing real-time risk monitoring, user behavior analysis, and the detection of malicious activities. The benchmarks introduced in this platform serves to empower researchers and stakeholders to address the complexities of market fluctuations and emerging risks, contributing to automated decision-making in dynamic market conditions.

References

[1] Halvor Aarhus Aalborg, Peter Molnár, and Jon Erik de Vries. What can explain the price, volatility and trading volume of bitcoin? *Finance Research Letters*, 29:255–265, 2019.

[2] V^1 Akila, MVS Nitin, I Prasanth, Sandeep Reddy, and Akash Kumar. A cryptocurrency price prediction model using deep learning. In *E3S Web of Conferences*, volume 391, page 01112. EDP Sciences, 2023.

[3] Evan Albers, Mohammad T Irfan, and Matthew J Botsch. Beliefs, shocks, and the emergence of roles in asset markets: An agent-based modeling approach. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 40–48, 2024.

[4] Muhamad Anuar Danial Hairul Anuar and Saiful Izzuan Hussain. Behaviour of extreme volatility in cryptocurrency: Bitcoin vs ethereum. In *AIP Conference Proceedings*, volume 2905. AIP Publishing, 2024.

[5] Ayana T Aspembitova, Ling Feng, and Lock Yue Chew. Behavioral structure of users in cryptocurrency market. *Plos one*, 16(1):e0242600, 2021.

[6] S Baranidharan. Investor behaviour, market efficiency, and regulatory challenges in digital currency investments: A comprehensive literature review. *Digital Currency Assets and Challenges to Financial System Stability*, pages 144–177, 2024.

[7] Saachin Bhatt, Mustansar Ghazanfar, and M Amirhosseini. Machine learning based cryptocurrency price prediction using historical data and social media sentiment. *Computer Science & Information Technology (CS & IT)*, 13(10):1–11, 2023.

[8] Elie Bouri, Peter Molnár, Georges Azzi, David Roubaud, and Lars Ivar Hagfors. On the hedge and safe haven properties of bitcoin: Is it really more than a diversifier? *Finance Research Letters*, 20:192–198, 2017.

[9] Antonio Briola, David Vidal-Tomás, Yuanrong Wang, and Tomaso Aste. Anatomy of a stablecoin's failure: The terra-luna case. *Finance Research Letters*, 51:103358, 2023.

[10] Junwei Chen. Analysis of bitcoin price prediction using machine learning. *Journal of Risk and Financial Management*, 16(1):51, 2023.

[11] Zheshi Chen, Chunhong Li, and Wenjun Sun. Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, 365:112395, 2020.

[12] Christian Conrad, Anessa Custovic, and Eric Ghysels. Long-and short-term cryptocurrency volatility components: A garch-midas analysis. *Journal of Risk and Financial Management*, 11(2):23, 2018.

[13] Shaen Corbet, Brian Lucey, Andrew Urquhart, and Larisa Yarovaya. Cryptocurrencies as a financial asset: A systematic analysis. *International Review of Financial Analysis*, 62:182–199, 2019.

[14] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE, 2017.

[15] P Kingma Diederik. Adam: A method for stochastic optimization. *(No Title)*, 2014.

[16] Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[17] Mohammad J Hamayel and Amani Yousef Owda. A novel cryptocurrency price prediction model using gru, lstm and bi-lstm machine learning algorithms. *Ai*, 2(4):477–496, 2021.

[18] Beibei Han, Yingmei Wei, Qingyong Wang, Francesco Maria De Collibus, and Claudio J Tessone. Mt 2 ad: multi-layer temporal transaction anomaly detection in ethereum networks with gnn. *Complex & Intelligent Systems*, 10(1):613–626, 2024.

[19] G Haritha and N Sahana. Cryptocurrency price prediction using twitter sentiment analysis. In *CS & IT conference proceedings*, volume 13. CS & IT Conference Proceedings, 2023.

[20] Anne Hove Henriksen Heise, Soraj Hongladarom, Anna Jobin, Katharina Kinder-Kurlanda, Sun Sun, Elisabetta Locatelli Lim, Annette Markham, Paul J Reilly, Katrin Tiidenberg, and Carsten Wilhelm. Internet research: Ethical guidelines 3.0. 2019.

[21] Darien Huang, Christian Schlag, Ivan Shaliastovich, and Julian Thimme. Volatility-of-volatility risk. *Journal of Financial and Quantitative Analysis*, 54(6):2423–2452, 2019.

[22] Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. Temporal graph benchmark for machine learning on temporal graphs. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Woochang Hyun, Jaehong Lee, and Bongwon Suh. Anti-money laundering in cryptocurrency via multi-relational graph neural network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 118–130. Springer, 2023.

[24] Arijit Khan. Graph analysis of the ethereum blockchain data: A survey of datasets, methods, and future work. In *2022 IEEE International Conference on Blockchain (Blockchain)*, pages 250–257. IEEE, 2022.

[25] Ahmed M Khedr, Ifra Arif, Magdi El-Bannany, Saadat M Alhashmi, and Meenu Sreedharan. Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. *Intelligent Systems in Accounting, Finance and Management*, 28(1):3–34, 2021.

[26] Gyeongho Kim, Dong-Hyun Shin, Jae Gyeong Choi, and Sunghoon Lim. A deep learning-based cryptocurrency price prediction model that uses on-chain data. *IEEE Access*, 10:56232–56248, 2022.

[27] Jaehyeon Kim, Sejong Lee, Yushin Kim, and Sunghyun Cho. A graph embedding-based identity inference attack on blockchain systems. In *2022 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–3. IEEE, 2022.

[28] Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1269–1278, 2019.

[29] Lin Li. Investigating risk assessment in post-pandemic household cryptocurrency investments: an explainable machine learning approach. *Journal of Asset Management*, 24(4):255–267, 2023.

[30] Xingyi Li, Yujun Liu, Zhuang Liu, and Shushang Zhu. Cryptocurrency return prediction: A machine learning analysis. *Available at SSRN 4703167*, 2024.

[31] Xiao Liu, Zaiyang Tang, Peng Li, Song Guo, Xuepeng Fan, and Jinbo Zhang. A graph learning based approach for identity inference in dapp platform blockchain. *IEEE Transactions on Emerging Topics in Computing*, 10(1):438–449, 2020.

[32] Yukun Liu, Aleh Tsyvinski, and Xi Wu. Common risk factors in cryptocurrency. *The Journal of Finance*, 77(2):1133–1177, 2022.

[33] Evgeny Medvedev. the d5 team. *Ethereum ETL. Available online: https://github.com/blockchain-etl/ethereum-etl (accessed on 25 February 2021)*, 2018.

[34] Nur Azmina Mohamad Zamani, Norhaslinda Kamaruddin, and Ahmad Muhyiddin B Yusof. Dataset on sentiment-based cryptocurrency-related news and tweets in english and malay language. *Language Resources and Evaluation*, pages 1–36, 2024.

[35] Alan Moreira and Tyler Muir. Should long-term investors time volatility? *Journal of Financial Economics*, 131(3):507–527, 2019.

[36] C Ozturan, A Sen, and B Kilic. Transaction graph dataset for the ethereum blockchain. *Zenodo*, page 4718440, 2021.

[37] Vatsal Patel, Lei Pan, and Sutharshan Rajasegarar. Graph deep learning based anomaly detection in ethereum blockchain network. In *International conference on network and system security*, pages 132–148. Springer, 2020.

[38] Farimah Poursafaei, Shenyang Huang, Kellin Pelrine, and Reihaneh Rabbany. Towards better evaluation for dynamic link prediction. *Advances in Neural Information Processing Systems*, 35:32928–32941, 2022.

[39] Yuxin Qi, Jun Wu, Hansong Xu, and Mohsen Guizani. Blockchain data mining with graph learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[40] Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. arxiv 2020. *arXiv preprint arXiv:2006.10637*.

[41] Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*, 2020.

[42] Kiarash Shamsi, Friedhelm Victor, Murat Kantarcioglu, Yulia Gel, and Cuneyt G Akcora. Chartalist: Labeled graph datasets for utxo and account-based blockchains. *Advances in Neural Information Processing Systems*, 35:34926–34939, 2022.

[43] William F Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442, 1964.

[44] James Tobin. Liquidity preference as behavior towards risk. *The review of economic studies*, 25(2):65–86, 1958.

[45] Leanne Townsend and Claire Wallace. Social media research: A guide to ethics. *University of Aberdeen*, 1(16):1–16, 2016.

[46] Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. Dyrep: Learning representations over dynamic graphs. In *International conference on learning representations*, 2019.

[47] Harald Uhlig. A luna-tic stablecoin crash. Technical report, National Bureau of Economic Research, 2022.

[48] Lu Wang, Xiaofu Chang, Shuang Li, Yunfei Chu, Hui Li, Wei Zhang, Xiaofeng He, Le Song, Jingren Zhou, and Hongxia Yang. Tcl: Transformer-based dynamic graph modelling via contrastive learning. *arXiv preprint arXiv:2105.07944*, 2021.

[49] Qian Wang, Zhen Zhang, Zemin Liu, Shengliang Lu, Bingqiao Luo, and Bingsheng He. Ex-graph: A pioneering dataset bridging ethereum and x. In *The Twelfth International Conference on Learning Representations*.

[50] Yanbang Wang, Yen-Yu Chang, Yunyu Liu, Jure Leskovec, and Pan Li. Inductive representation learning in temporal networks via causal anonymous walks. *arXiv preprint arXiv:2101.05974*, 2021.

[51] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. *arXiv preprint arXiv:2002.07962*, 2020.

[52] Libo Yin, Jing Nie, and Liyan Han. Understanding cryptocurrency volatility: The role of oil market shocks. *International Review of Economics & Finance*, 72:233–253, 2021.

[53] Xin Yin, Zhen Liu, Guomin Yang, Guoxing Chen, and Haojin Zhu. Secure hierarchical deterministic wallet supporting stealth address. In *European Symposium on Research in Computer Security*, pages 89–109. Springer, 2022.
[54] Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. Towards better dynamic graph learning: New architecture and unified library. *Advances in Neural Information Processing Systems*, 36:67686–67700, 2023.
[55] Zhen Zhang, Bingqiao Luo, Shengliang Lu, and Bingsheng He. Live graph lab: Towards open, dynamic and real transaction graphs with nft. *Advances in Neural Information Processing Systems*, 36:18769–18793, 2023.
[56] Wang Zhengyang, Li Xingzhou, Ruan Jinjin, and Kou Jiaqing. Prediction of cryptocurrency price dynamics with multiple machine learning techniques. In *Proceedings of the 2019 4th International Conference on Machine Learning Technologies*, pages 15–19, 2019.
[57] Thanasis Zoumpekas, Elias Houstis, and Manolis Vavalis. Eth analysis and predictions utilizing deep learning. *Expert Systems with Applications*, 162:113866, 2020.

Appendix

A Raw Data

Table 3 is an example of the Token Rating in the Token Info Section. It presents the names, symbols, rating levels, and scores of five cryptocurrencies, including Ethereum, Bitcoin, BNB, Dai, and Uniswap. Each cryptocurrency is rated based on its performance (either AAA or AA), with a corresponding numerical score.

Table 3: Examples of token_rating.cav Table

Name	Symbol	Rating Level	Rating Score
Ethereum	ETH	AAA	81.34
Bitcoin	BTC	AA	78.47
BNB	BNB	AA	77.83
Dai	DAI	AA	76.17
Uniswap	UNI	AA	73.66

Table 4 is an example of the Market Cap in the Global Index Section. It shows the time-series data of the total market capitalization of cryptocurrencies, covering five dates from April 29, 2013, to May 7, 2013.

Table 4: Examples of Total Market Capitalization Data

DateTime	Market Cap
2013/4/29 12:00	1583440000
2013/5/1 12:00	1637389952
2013/5/3 12:00	1275410048
2013/5/5 12:00	1335379968
2013/5/7 12:00	1313900032

Table 5 is an example of the 24h Volume Data in the Global Index Section. This table presents the 24-hour trading volume from February 25, 2014, to March 5, 2014. The volume ranges from 7,047,3200 to 11,784,0000, indicating fluctuations in market activity during this period.

Table 5: Examples of 24h Volume Data

DateTime	Volume (24h)
2014/2/25 13:00	70473200
2014/2/27 13:00	84957000
2014/3/1 13:00	41190400
2014/3/3 13:00	17991000
2014/3/5 13:00	117840000

Table 6 is an example of the Stablecoin Market Capitalization Data in the Global Index Section. It shows the stablecoin market capitalization data from March 10, 2016, to March 18, 2016.

Table 6: Examples of Stablecoin Market Capitalization Data

DateTime	Stablecoin Market Cap
2016/3/10 13:00	1451448.067
2016/3/12 13:00	1451593.424
2016/3/14 13:00	1451479.734
2016/3/16 13:00	1451602.250
2016/3/18 13:00	1451575.947

Table 7 is an example of the TY60MCD.csv in the Global Index Section. It lists the TY60MCD data from April 2021 to August 2021.

Table 7: Examples in TY60MCD.csv

DateTime	TY60MCD
2021/4/1	0.75
2021/5/1	0.92
2021/6/1	0.86
2021/7/1	0.87
2021/8/1	0.87

Table 8 is an example of Token Transaction Data for token AAVE. This table records the transaction information for AAVE, including token address, sender and receiver addresses, transaction value, transaction hash, log index, block timestamp, and block number. The transaction times range from the early hours of July 26, 2024, to later that night, showcasing multiple large transactions of AAVE tokens.

Table 9 is an example of the token_general Table in the Token Info Section. This table lists information about five tokens, including ID, symbol, name, Ethereum address, and decimals.

Table 10 is an example of Token Recording BNB in the Token Info Section. It records the historical price, market cap, and total volume for BNB tokens. The data covers timestamps, prices, market capitalizations, and total volumes across several time intervals.

Table 11 is an example of the Token Sentiment Data using background knowledge from an LLM for sentiment scoring. The table provides sentiment analysis of textual data, including score, timestamp, number of comments, text content, and positive/negative sentiment scores. For instance, a comment on November 19, 2013, received 3 positive and 7 negative scores, with the text partially masked.

Table 12 is an example of Bitcoin Dominance Data in the Global Index Section. It includes the market capitalizations of Bitcoin (BTC), Ethereum (ETH), Tether (USDT), BNB, Solana (SOL), and other cryptocurrencies.

B Ethics and Privacy

This research adheres to the ethical guidelines outlined by the Association of Internet Researchers [20] and Townsend & Wallace [45]. In conducting this study, we have carefully considered several factors to minimize potential ethical and privacy risks.

Ethereum is a public and permissionless blockchain, which means its entire network is open for anyone to join, participate in consensus, execute transactions, and view the ledger. This openness

Table 8: Examples in Token Transaction Data for AAVE

Token Address	From Address	To Address	Value	Transaction Hash	Log Index	Block Timestamp	Block Number
0x...dae9	0x...7fad	0x...7e1c	8.40279E+16	0x...276e3	101	2024-07-26 01:28:59	20387439
0x...dae9	0x...5622	0x...9a81	8.51435E+18	0x...90c4c	484	2024-07-26 01:14:11	20387365
0x...dae9	0x...f2c8	0x...4bee	1.15726E+20	0x...13aac	171	2024-07-26 01:12:11	20387356
0x...dae9	0x...5145	0x...0703	1.14393E+21	0x...ac8cf	367	2024-07-25 20:47:23	20386044
0x...dae9	0x...699c	0x...5fb6	3.40423E+18	0x...84f0	204	2024-07-25 20:10:47	20385863

Table 9: Examples in token_general_3880.csv Table

ID	Symbol	Name	ETH Address	Decimal
0	zcn	Zus	...38f3b78	10
1	0kn	0 Knowledge Network	...7d29036	18
2	ome	O-MEE	...826977e	18
3	zrx	0x Protocol	...699f498	18
4	0x0	0x0.ai: AI Smart Contract	...0811ad5	9

Table 10: Examples in Token History Information for BNB ...1bdd52.csv

Timestamp	Price	Market Caps	Total Volumes
1.50552E+12	0.107250624	10725062.44	1.051223307
1.50561E+12	0.154041291	15404129.09	14.67858722
1.50569E+12	0.173491239	17349123.91	6.001766938
1.50578E+12	0.168334191	16833419.06	3.878927407
1.50587E+12	0.166627925	16662792.49	40.6876186

Table 11: Examples in Text Data with Sentiment Analysis

Score	Timestamp	Number of Comments	Text (masked)	Positive	Negative
3191	2013/11/19 19:15	472	I'm one of the Senators attending...	3	7
3193	2013/11/25 1:38	282	I was bored so I animated the...	1	0
3524	2014/2/13 23:49	470	on r/bitcoin right now...	0	7
3055	2014/2/18 20:15	463	Bitcoin takes a walk with Dogecoin...	2	1
3455	2014/2/26 16:41	489	Open Letter to Michael Casey - WSJ reporter...	5	7
3954	2014/2/28 8:17	416	We've gotta be able to laugh at ourselves...	1	0

underscores its public nature, with no central authority controlling access to the network or its data. Similarly, the Reddit data used in this study is sourced exclusively from public subreddits and posts. In line with the emphasis on publicness, no private messages or content from restricted communities have been accessed or included.

In maintaining the pseudonymity of transactions, our dataset preserves the same level of anonymity inherent to the Ethereum network. Reddit users typically operate under pseudonyms, and to further minimize privacy risks, we have anonymized the data by replacing post identifiers with randomly generated values. In addition, we applied data processing techniques to mask any personally identifiable information (PII) and sensitive information that might have been present in the raw data.

We have implemented strict protocols governing the use and distribution of this dataset. Researchers seeking access will be required to agree to terms that prohibit any attempts to re-identify individuals or use the data for purposes other than approved research.

The insights gained from this dataset have the potential to contribute significantly to our understanding of cryptocurrency trading and its broader ecosystems. These findings can help advance the development of more secure, equitable, and robust blockchain and decentralized finance (DeFi) systems. We believe the potential benefits of this research outweigh the minimal risks associated with using publicly available data.

Table 12: Examples in Cryptocurrency Bitcoin Dominance Data

DateTime	BTC Cap	ETH Cap	USDT Cap	BNB Cap	SOL Cap	Others Cap
2023/9/3 12:00	5.04E+11	1.97E+11	82901682430	32993907647	7967342850	2.18E+11
2023/9/10 12:00	5.04E+11	1.97E+11	82992946191	32957692204	7984825883	2.19E+11
2023/9/17 12:00	5.18E+11	1.97E+11	83065027653	33069547650	7871412204	2.22E+11
2023/9/24 13:00	5.18E+11	1.92E+11	83206758665	32396394279	8035147376	2.21E+11
2023/10/1 13:00	5.26E+11	2.01E+11	83260095938	33046306303	8834373543	2.25E+11

C GNN Model

DyGFormer. lies in introducing the self-attention mechanism of Transformers to capture long-term dependencies in dynamic graphs. By incorporating temporal encoding within the Transformer, DyGFormer effectively captures complex dynamic changes in graphs. Compared to other models that utilize short-term temporal windows, DyGFormer excels in tasks involving long-term dependencies.

JODIE. focuses on continuous time-series interaction modeling, tracking the long-term interaction dynamics between users and items. This bidirectional embedding update mechanism makes it particularly effective in scenarios with frequent user-item interactions (e.g., recommendation systems). JODIE leverages RNNs to capture the historical dependencies of nodes (users and items), making it well-suited for long-term behavior prediction.

DyRep. lies in its distinction between two types of events: interaction events between nodes (such as transactions or communications) and structural events (such as the creation or deletion of nodes/edges). DyRep divides dynamic graph tasks into these two categories of events and models them using event time sequences, making it ideal for tasks involving constantly changing nodes and edges.

TGAT. is the combination of Graph Convolutional Networks (GCNs) with temporal encoding and the use of graph attention mechanisms to capture dynamic relationships in the graph. Unlike simple temporal updates, TGAT focuses on important neighboring nodes through attention mechanisms, making it well-suited for dynamic networks with frequent interactions.

TGN. introduces a message-passing mechanism to update node states, making it particularly suitable for learning node embeddings in large-scale dynamic graphs. Unlike TGAT, TGN is not restricted to attention mechanisms but combines message-passing and temporal updates, allowing it to scale to large graphs and handle complex dynamic environments.

TCL. utilizes contrastive learning to capture the temporal evolution of node embeddings in dynamic graphs by comparing node representations at different time points. Unlike other models that rely on supervised learning, TCL constructs positive and negative sample pairs (e.g., the same node at different time points) to capture temporal changes. This unsupervised learning approach enables TCL to perform well in dynamic graph tasks without explicit labels.

D Prompt

Prompt 1: You are a useful cryptocurrency social media post sentiment analysis expert. Now I give you the text of the top reddit posts on cryptocurrency scope, then you need to understand and analyze the sentiment score of this post in the context of cryptocurrency (on a scale of 10, 0-4 is negative, higher is positive, lower is negative, 6-10 is positive, 5 is neutral).Below is the text of reddit's top crypto-scoped posts that you should analyze as above, just give me the final score, just give a number like "6" :

Prompt 2: You are a helpful cryptocurrency social media posts sentiment analysis expert. Now I give you the cryptocurrency scope top posts' text of reddit with its timestamp, and then you need understand and analyse the following content in context of cryptocurrency with considering the trends and news about cryptocurrencies at the time of the timestamp and give me the negative score and the positive score. The sentiment score may take into account not only the sentiment of the post regarding cryptocurrencies, but also whether the time of the post and the context (e.g., big events in the cryptocurrency space) seemed to have a positive or negative impact on cryptocurrencies at the time. For example, if the text say: Bitcoin is a kind of cryptocurrency, this is a totally neutral, so the negative score and the positive score are 0; if the text say: I like Bitcoin, while bitcoin has high volatility; this somewhat negative and somewhat positive, so the negative score is about 5(up to 10) and positive score is about 5(up to 10); if the text say, although bitcoin has high volatility, I like it, it should be higher positive score than negative score; these examples are make you understand how marking. A forementioned content is sample cryptocurrency sentiment analysis, the actual senario should be more complicated. For example the first text said: 1384888505 I'm one of the Senators attending today's U.S. Senate Banking Committee hearing related to bitcoin. What would you like me to know? The timestamp of this post is 1384888505, and then the content is related to the hearing, indicating that at that time, Bitcoin may have attracted a lot of attention, but at that time, people do not know whether this thing is good or not, and may have to hold a hearing. You need to consider the information at this level, and then assign positive and negative scores separately. The following is the cryptocurrency scope top posts' text of reddit with its timestamp, you should analyze follow the aforementioned requirements and only give me the final score like "positivate score: 2, negative score: 3":

E LSTM and GRU

LSTM (Long Short-Term Memory) in Time Series Problems. LSTM is a type of recurrent neural network (RNN) specifically designed to overcome the limitations of traditional RNNs in handling long-term dependencies. It addresses the vanishing gradient problem, which commonly arises in training standard RNNs, making LSTM particularly well-suited for time series forecasting. In LSTM, memory cells maintain information over long periods through three key gates: the input gate, forget gate, and output gate. These gates regulate the flow of information, allowing the network to selectively retain or forget information based on its relevance to the prediction task. As a result, LSTM can capture both short-term and long-term patterns in time series data, making it a powerful tool for tasks such as stock price prediction, weather forecasting, and anomaly detection in time-dependent data.

GRU (Gated Recurrent Unit) in Time Series Problems. GRU is another variant of recurrent neural networks, similar to LSTM but with a simpler architecture. Unlike LSTM, GRU has only two gates: the update gate and the reset gate. These gates control how much past information is retained and how new input is incorporated into the network's state. GRUs are often preferred for time series problems when computational efficiency is a concern, as they tend to train faster and require fewer parameters than LSTMs while maintaining competitive performance. GRU models are particularly effective for shorter time series or when the data does not exhibit complex long-term dependencies. Despite their simpler design, GRUs have proven successful in tasks such as traffic flow prediction, energy consumption forecasting, and speech recognition, where the ability to capture temporal patterns is essential.

F LLM Knowledge Base Marked Sentiment Data Algorithm

Algorithm 1, named LLM Knowledge Base Marked Sentiment Data, processes sentiment data from a CSV file to generate a DataFrame with updated values for scores, timestamps, number of comments, and sentiment attributes (positive and negative). The algorithm begins by reading and cleaning the data, then it processes each row by applying a decay factor $k = 0.5$ to reduce the scores, comments, and sentiment values for certain time intervals. It also handles missing timestamps by filling in empty entries. The final output is a clean, processed DataFrame that includes adjusted sentiment information over a specified period.

Algorithm 1 LLM Knowledge Base Marked Sentiment Data

Input: Path to the CSV file, decay factor $k = 0.5$
Output: Processed DataFrame with updated score, timestamp, number of comments, positive and negative attributes
Read the CSV file and clean the data
Initialize an empty list *processed_data*
for each row in data frame **do**
  Extract $score$, $timestamp$, $number\_of\_comment$, $positive$, $negative$ from the row
  for $j \leftarrow 0$ to 2 **do**
    Set $new\_timestamp = timestamp + j \cdot 1$ day
    Append $[score, new\_timestamp]$ to $processed\_data$
    Append $[number\_of\_comment]$ to $processed\_data$
    Append $[positive, negative]$to $processed\_data$
  end for
  for $j \leftarrow 3$ to 6 **do**
    $score \leftarrow score \cdot k$
    $number\_of\_comment \leftarrow number\_of\_comment \cdot k$
    $positive \leftarrow positive \cdot k$
    $negative \leftarrow negative \cdot k$
    Set $new\_timestamp = timestamp + j \cdot 1$ day
    Append $[score, new\_timestamp]$ to $processed\_data$
    Append $[number_of_comment]$to $processed\_data$
    Append $[positive, negative]$to $processed\_data$
  end for
  if next post timestamp > $timestamp + 7$ days **then**
    while last post day < next post timestamp **do**
      Append $[0, last\_post\_day, 0, 0, 0]$ to $processed\_data$
      Increment last_post_day by 1 day
    end while
  end if
end for
Check for invalid timestamps and clean processed data
Return *processed_data* as a DataFrame

Algorithm 2, named Common LLM Marked Sentiment Data Processing, processes sentiment data from a CSV file to generate a DataFrame with updated attributes such as scores, timestamps, comment counts, and overall sentiment. The procedure starts by cleaning the data (converting timestamps, ensuring numeric columns, and removing missing values). It then processes each row, applying a decay factor k to adjust the scores and comments over time. The algorithm also fills in missing timestamps when necessary. Finally, it aggregates the data by summing scores and comments and averaging sentiment before returning the cleaned and sorted DataFrame.

Algorithm 2 Common LLM Marked Sentiment Data Processing

Input: Path to CSV, decay factor k
Output: Processed DataFrame with updated score, timestamp, number of comments, and sentiment
Step 1: Data Cleaning
Convert 'timestamp' to datetime, ensure columns are numeric, remove NaNs
Step 2: Data Processing
for each row in the data **do**
  Extract features: $score$, $timestamp$, $number\_of\_comment$, $sentiment$
  for $j = 0$ to 2 **do**
    Update and append $score$ and $new\_timestamp$ to $processed\_data$
  end for
  for $j = 3$ to 6 **do**
    Apply decay to $score$ and $number\_of\_comment$, append to $processed\_data$
  end for
  if next post timestamp exceeds threshold **then**
    Append filler rows for missing days
  end if
end for
Step 3: Aggregation and Averaging
Group by 'timestamp', sum 'score' and 'number_of_comment', average 'sentiment'
Step 4: Final Clean-up
Sort by 'timestamp', remove invalid entries, return $processed\_data$

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009