

# Crypto, Transformers e Baleias

USP

FGV

ripple

University Blockchain  
Research Initiative

Competidores:  
Heitor Gama  
Asaffe Duarte  
Solano Omar

Uma análise  
Inovadora dos  
Movimentos de  
Baleias em  
Criptomoedas

By Team HAS



# Roadmap



# Visão Geral

- Objetivo: Prever os movimentos de Crypto Baleias por meio de modelos estado da arte de Inteligência Artificial
- Perguntas-chave
  - Como as transações e movimentos de baleias são definidos no contexto do mercado de criptomoedas?
  - Quais são os principais desafios na previsão de atividades de baleias e como esses desafios foram abordados no desenvolvimento do modelo?
  - Quais são os resultados obtidos pelo modelo preditivo em termos de precisão e eficácia na previsão dos movimentos das baleias?
  - Como o modelo se comporta em condições de mercado voláteis e como a robustez foi avaliada?
  - De que maneira a análise preditiva de baleias pode contribuir para a compreensão mais ampla do comportamento do mercado de criptomoedas e suas dinâmicas?
  - Quais são as limitações do modelo proposto e quais áreas representam oportunidades para pesquisas futuras?
- Metodologia
  - Aquisição e pré-processamento da base de dados
  - Separação inteligente em treino+validação e teste
  - Calibração de hiperparâmetros usando o subset de validação
  - Escolha do melhor modelo
  - Verificação do desempenho no subset de teste
  - Análise dos resultados

**Principal Inovação: uso de Transformers num contexto de séries temporais de criptomoedas**



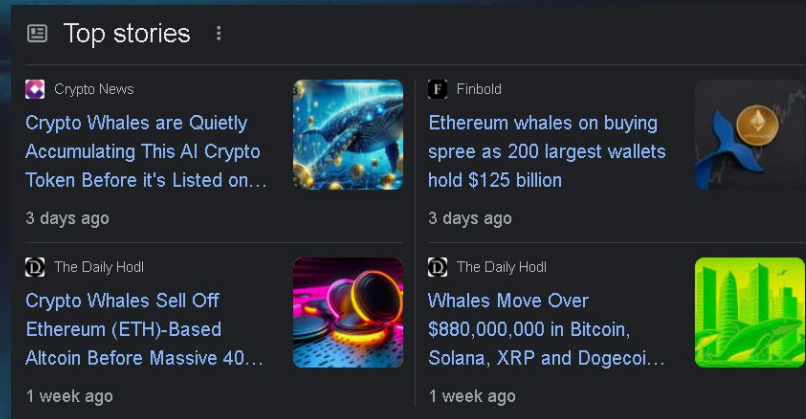


# Motivação

## Por que estudar Baleias?

- **O que é uma baleia?**
  - Uma "crypto whale" é um endereço de carteira que mantém uma quantidade substancial de criptomoeda.
- **Uma baleia tem grande impacto no mercado e o estudo de seus movimentos permite:**
  - **Antecipar Tendências de Mercado**
    - Analisar os padrões de transações das baleias pode proporcionar insights valiosos sobre as tendências futuras do mercado de criptomoedas.
  - **Identificar Potenciais Pontos de Volatilidade**
    - A compreensão dos movimentos das baleias ajuda na identificação de eventos que podem desencadear aumentos significativos na volatilidade do mercado.
  - **Mitigar Riscos e Otimizar Estratégias de Investimento**
    - Ao entender as ações das baleias, os investidores podem ajustar suas estratégias para mitigar riscos e otimizar suas posições no mercado.
  - **Antever Manipulações de Mercado:**
    - O estudo dos movimentos das baleias é crucial para detectar possíveis manipulações de mercado, contribuindo para a integridade e transparência do ecossistema cripto.
  - **Compreender o Comportamento do Investidor Institucional**
    - As baleias muitas vezes representam investidores institucionais, e o estudo de seus movimentos oferece insights valiosos sobre o comportamento desse segmento vital do mercado.
  - **Adaptar-se a Mudanças nas Dinâmicas do Mercado**
    - A análise contínua dos movimentos das baleias permite que os participantes do mercado se adaptem proativamente a mudanças nas dinâmicas do mercado, mantendo-se à frente de potenciais movimentos massivos.

## As ondas geradas por baleias



O estudo dos movimentos das baleias não apenas fornece uma compreensão mais profunda do comportamento do mercado de criptomoedas, mas também capacita os investidores a tomar decisões mais informadas e estratégicas.



# Motivação

## O que são Transformers?

- Arquitetura de modelo de aprendizado profundo.
- Destaque em processamento de linguagem natural (NLP).
- Introduzida em 2017 por Vaswani et al.
- Mecanismo chave: Atenção Multihead.

## Por quê usar Transformers nesse contexto?

### **Processamento Não Sequencial:**

Não depende de processamento sequencial, permitindo paralelização eficiente visando a sustentabilidade.

### **Captura de Relacionamentos Complexos:**

A atenção multihead captura relacionamentos de longo alcance, vital para séries temporais complexas.

### **Transferência de Aprendizado:**

Capacidade de pré-treinamento em grandes conjuntos de dados, facilitando adaptação a contextos específicos, como detecção de baleias.

### **Versatilidade Além do NLP**

- Tem grande potencial descoberto nas diversas áreas de ciência de dados a todo momento (Carion et al., 2020; Dosovitskiy et al., 2021; Liu et al., 2021b)
- Não encontramos na literatura estudos suficientemente desenvolvidos dessa poderosa ferramenta no contexto de criptomoedas e acreditamos que isso se deve apenas pelo fato de que os avanços dela são muito recentes até mesmo no contexto geral de Machine Learning (ML)
- Sua eficácia foi recentemente comprovada no contexto da geologia (Jordao et al., 2023), mostrando-se superior a todos os métodos tradicionais de ML.

N. Carion et al. [End-to-end object detection with transformers](#)  
Dosovitskiy et al., [An image is worth 16x16 words: Transformers for image recognition at scale](#)  
Liu et al. [Efficient training of visual transformers with small datasets](#)  
Jordao. et al. [Towards automatic and accurate core-log processing](#)





# Metodologia

## Aquisição e pré-processamento da base de dados

- Base de dados: múltiplos datasets de séries temporal de transações de maio 2023, 1 para cada moeda.  
<https://whale-alert.io/sample-data/>

```
21:48:45 19,999 USDT (20,466 USD) TRANSFERRED from binance to unknown wallet
21:48:47 152,565 USDC (152,280 USD) TRANSFERRED from gateio to kraken
21:48:50 11,613 XRP (7,141 USD) TRANSFERRED from unknown wallet to unknown wallet
```

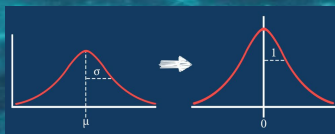
### Moedas escolhidas para análise e motivação

- Bitcoin: baseline para comparação, tem histórico
- Dogecoin e Litecoin:

Altcoins são as mais afetadas por baleias devido à baixa liquidez (Yekatsiaryna, 2023)

### Limpeza

- Drop de features sem correlação
- One-hot encoding
- Standard Scaling



## Detalhes do pré-processamento

- Sort das amostras por timestamp
- Drop de features por simplificação:  
`['hash', 'timestamp', 'raw_amount']`
- Extração das informações mais importantes dentro de cada sub-transação: o alvo é o montante da transação (somatório da quantidade de moedas em cada sub-transação multiplicada pela cotação atual dela)
- Conversão de objects para floats
- Drop NaN values
- 70% Treino, Validação = 10% Treino, 30% Teste

## Tratamento de séries temporais

- Foram criadas lag\_features: 50 novas features do histórico de cada transação anterior sequencialmente. Dessa forma usamos a memória para prever o futuro
- Criamos 9 categorias para classificar transações baleias:
  - 0 - Tiny Fish < 100m USD
  - 1 - Small Fish < 200m USD
  - 2 - Medium Fish < 500m USD
  - 3 - Big Fish < 1M USD
  - 4 - Small Whale < 5M USD
  - 5 - Medium Whale < 10M USD
  - 6 - Big Whale < 20M USD
  - 7 - Huge Whale < 50M USD
  - 8 - Mega Whale > 50M

Yekatsiaryna, [Are crypto whales good or bad?](#)

Para BTC multiplicamos os intervalos por um fator 8, pois os peixes costumam ser maiores



# Metodologia - rodando o Transformer

## O que exatamente prevemos?

- Criamos a feature target `max_whale_class_next_hour`, que é o maior peixe da próxima hora. Esse tempo poderia ser diminuído para menos caso houvesse maior poder computacional
- Prevemos um número de 0 a 8 indicando o tamanho do maior peixe

## Calibração de hiperparâmetros usando o subset de validação

- O pré-processamento termina gerando um 4 numpy-arrays comprimidos num arquivo .npz (`X_train`, `X_test`, `y_train` e `y_test`).
- Nosso programa `Transformer.py` carrega o npz, e primeiro calibra o Transformer com um subset de validação de 10% do treino.

## Escolha do melhor modelo

- Testamos a acurácia na validação para diferentes hiperparâmetros e escolhemos o conjunto que der a maior acurácia na validação
- Por simplicidade e economia no tempo de execução, rodamos por 10 épocas cada dataset, mas esse número poderia ser elevado para melhores resultados a custo de maior tempo de treino
- O modelo de melhor performance é usado para prever o `y_test` apartir do `X_test`

## Métricas de avaliação

- Avaliamos o desempenho pela acurácia e a loss usando `categorical_crossentropy` na previsão do maior peixe na próxima hora
- Salvamos os logs dos resultados num arquivo `'sigla_da_moeda_logs.txt'`





# Resultados

## Verificação do desempenho no subset de teste



Bitcoin

Excuting on BTC dataset...

```
Epoch 1/10
2856/2856 - 152s - loss: 0.3059 - accuracy: 0.8745 - 152s/epoch - 53ms/step
Epoch 2/10
2856/2856 - 142s - loss: 0.1031 - accuracy: 0.9599 - 142s/epoch - 50ms/step
Epoch 3/10
2856/2856 - 138s - loss: 0.0815 - accuracy: 0.9684 - 138s/epoch - 48ms/step
Epoch 4/10
2856/2856 - 142s - loss: 0.0672 - accuracy: 0.9743 - 142s/epoch - 50ms/step
Epoch 5/10
2856/2856 - 140s - loss: 0.0562 - accuracy: 0.9788 - 140s/epoch - 49ms/step
Epoch 6/10
2856/2856 - 142s - loss: 0.0479 - accuracy: 0.9815 - 142s/epoch - 50ms/step
Epoch 7/10
2856/2856 - 140s - loss: 0.0451 - accuracy: 0.9829 - 140s/epoch - 49ms/step
Epoch 8/10
2856/2856 - 149s - loss: 0.0395 - accuracy: 0.9852 - 149s/epoch - 52ms/step
Epoch 9/10
2856/2856 - 139s - loss: 0.0363 - accuracy: 0.9863 - 139s/epoch - 49ms/step
Epoch 10/10
2856/2856 - 140s - loss: 0.0366 - accuracy: 0.9864 - 140s/epoch - 49ms/step
2549/2549 [=====] - 23s 9ms/step
```

Testing Accuracy on BTC: 0.7240

```
max_whale_class_next_hour
4 113737
5 74928
7 72327
6 10824
Name: count, dtype: int64
whale_class
0 269576
1 1119
2 618
3 273
4 216
5 16
7 6
6 2
Name: count, dtype: int64
```

Acurácia:  
72,40%



Dogecoin

Excuting on DOGE dataset...

```
Epoch 1/10
227/227 - 16s - loss: 0.9065 - accuracy: 0.5941 - 16s/epoch - 72ms/step
Epoch 2/10
227/227 - 11s - loss: 0.6521 - accuracy: 0.6923 - 11s/epoch - 49ms/step
Epoch 3/10
227/227 - 11s - loss: 0.5337 - accuracy: 0.7443 - 11s/epoch - 49ms/step
Epoch 4/10
227/227 - 11s - loss: 0.4662 - accuracy: 0.7834 - 11s/epoch - 48ms/step
Epoch 5/10
227/227 - 10s - loss: 0.3833 - accuracy: 0.8171 - 10s/epoch - 45ms/step
Epoch 6/10
227/227 - 10s - loss: 0.3563 - accuracy: 0.8351 - 10s/epoch - 44ms/step
Epoch 7/10
227/227 - 11s - loss: 0.2916 - accuracy: 0.8700 - 11s/epoch - 49ms/step
Epoch 8/10
227/227 - 12s - loss: 0.2892 - accuracy: 0.8699 - 12s/epoch - 51ms/step
Epoch 9/10
227/227 - 11s - loss: 0.2815 - accuracy: 0.8773 - 11s/epoch - 49ms/step
Epoch 10/10
227/227 - 11s - loss: 0.2180 - accuracy: 0.9094 - 11s/epoch - 50ms/step
195/195 [=====] - 2s 9ms/step
```

Testing Accuracy on DOGE: 0.8756

```
max_whale_class_next_hour
8 53211
4 23929
3 15637
5 3051
Name: count, dtype: int64
whale_class
0 94518
2 668
1 427
3 155
8 33
4 25
5 2
```

Acurácia:  
87,56%



LiteCoin

Excuting on LTC dataset...

```
Epoch 1/10
1049/1049 - 54s - loss: 0.4821 - accuracy: 0.7966 - 54s/epoch - 51ms/step
Epoch 2/10
1049/1049 - 45s - loss: 0.2093 - accuracy: 0.9124 - 45s/epoch - 43ms/step
Epoch 3/10
1049/1049 - 49s - loss: 0.1138 - accuracy: 0.9578 - 49s/epoch - 47ms/step
Epoch 4/10
1049/1049 - 49s - loss: 0.1055 - accuracy: 0.9636 - 49s/epoch - 47ms/step
Epoch 5/10
1049/1049 - 49s - loss: 0.0694 - accuracy: 0.9780 - 49s/epoch - 47ms/step
Epoch 6/10
1049/1049 - 48s - loss: 0.0433 - accuracy: 0.9863 - 48s/epoch - 45ms/step
Epoch 7/10
1049/1049 - 49s - loss: 0.0622 - accuracy: 0.9818 - 49s/epoch - 47ms/step
Epoch 8/10
1049/1049 - 51s - loss: 0.0573 - accuracy: 0.9834 - 51s/epoch - 49ms/step
Epoch 9/10
1049/1049 - 49s - loss: 0.0428 - accuracy: 0.9867 - 49s/epoch - 46ms/step
Epoch 10/10
1049/1049 - 47s - loss: 0.0389 - accuracy: 0.9879 - 47s/epoch - 45ms/step
899/899 [=====] - 8s 8ms/step
```

Testing Accuracy on LTC: 0.9976

```
max_whale_class_next_hour
5 10218
4 8660
7 966
3 871
Name: count, dtype: int64
whale_class
0 20037
3 241
2 182
1 120
4 117
5 17
7 1
```

Acurácia:  
99,76%





# Resultados

## 1. Diferenças nas Características das Criptomoedas:

A acurácia da previsão pode variar devido às características únicas de cada criptomoeda (Bitcoin, Dogecoin, Litecoin).

Hipótese: As transações e padrões associados ao valor das criptomoedas são distintos, influenciando a eficácia do modelo.

## 4. Diferenças nas Comunidades e Usos:

As comunidades e os casos de uso das criptomoedas variam.

Hipótese: O comportamento dos usuários e o propósito das transações podem ser fatores importantes na previsão.

## 7. Sensibilidade ao Período de Tempo:

A escolha do período de tempo para treinamento e teste pode ser crítica.

Hipótese: Modelos treinados em períodos mais recentes podem ter melhor desempenho na previsão futura.

## 2. Volume e Atividade de Transação:

O volume total de transações pode impactar a precisão do modelo.

Hipótese: Criptomoedas com maior volume de transações podem ter padrões mais previsíveis, resultando em maior acurácia.

## 5. Sensibilidade ao Tamanho da Amostra:

Tamanhos diferentes de conjuntos de dados podem levar a variações na acurácia.

Hipótese: Tamanhos de conjuntos de dados menores podem levar a maior variabilidade nos resultados.

## 8. Impacto de Eventos Externos:

Eventos externos, como atualizações de software ou notícias, podem afetar os resultados.

Hipótese: Modelos podem ser sensíveis a eventos externos que não foram considerados.

## 3. Variações de Preço e Volatilidade:

A volatilidade nos preços das criptomoedas pode influenciar a precisão do modelo.

Hipótese: Criptomoedas mais estáveis ou voláteis podem apresentar desafios diferentes na previsão de grandes transações.

## 6. Feature Engineering e Seleção:

As características escolhidas e como são manipuladas podem influenciar os resultados.

Hipótese: A engenharia de recursos e a seleção de recursos podem ser otimizadas para melhorar a precisão.

## 9. Complexidade e Ajuste do Modelo:

A escolha do modelo e seus hiperparâmetros pode impactar a acurácia.

Hipótese: Modelos mais complexos podem se ajustar demais aos dados de treinamento, resultando em menor generalização.



Muito mais dados, muito maior frequência, as maiores baleias em USD, mas que concentram menor percentual



Muito menos dados, menor frequência, menores baleias em USD, mas maiores em percentual

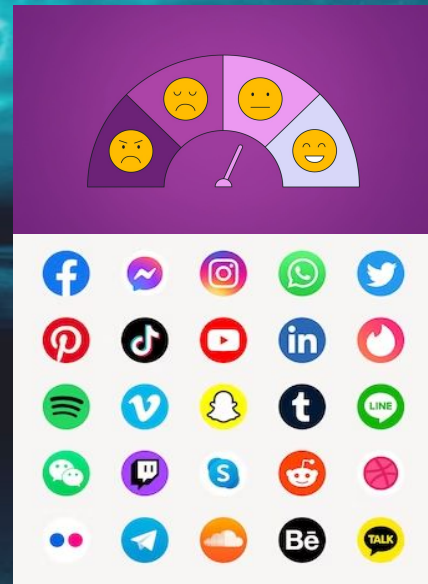


Quantidade intermediária de dados, frequência e baleias tanto percentualmente quanto em valor absoluto



# Limitações e Melhorias

- Inclusão de Análise Sentimental: dados de redes sociais são ótimos indicadores de movimentos de baleias (Raheman et al., 2022)
- Identificação e acompanhamento das baleias
- Maior tempo de treinamento = melhor modelo preditivo
- Com maior robustez conseguimos prever baleias cada vez mais próximas do presente
- Usar um modelo de regressão: ex: qual será a maior transação nos próximos 30 min
- Aumentar o número de classes de baleias



**Repositório  
disponível em:**





# Conclusão

## 1. Definição de Transações de Baleias:

- No mercado de criptomoedas, transações de baleias referem-se a movimentações significativas de ativos, geralmente envolvendo grandes volumes financeiros. Essas transações podem indicar atividades estratégicas de investidores de grande porte, influenciando o comportamento do mercado.

## 4. Comportamento em Condições de Mercado Voláteis:

- A resposta do modelo a condições de mercado voláteis foi avaliada, considerando a sensibilidade do modelo a mudanças rápidas nos padrões de transações. A robustez foi testada em cenários simulados de alta volatilidade para garantir que o modelo mantivesse desempenho estável.

## 2. Desafios na Previsão de Atividades de Baleias:

- Desafios principais incluem a volatilidade do mercado, a complexidade das transações e a necessidade de identificar padrões sutis. Para abordar esses desafios, o modelo foi desenvolvido considerando a escolha apropriada de algoritmos e treinamento com conjuntos de dados representativos.

## 5. Contribuição para a Compreensão do Mercado:

- A análise preditiva de baleias oferece insights valiosos sobre o comportamento do mercado de criptomoedas. Ao antecipar movimentos significativos, os investidores podem tomar decisões mais informadas. Além disso, a identificação de padrões de baleias pode contribuir para a compreensão mais ampla das dinâmicas do mercado.

## 3. Resultados do Modelo Preditivo:

- Os resultados obtidos indicam uma acurácia de previsão específica para diferentes criptomoedas. Por exemplo, para Bitcoin, a acurácia foi de 72.4%, para Dogecoin foi de 87.56%, e para Litecoin atingiu 99.76%. Esses resultados demonstram a eficácia do modelo em prever transações de baleias.

## 6. Limitações do Modelo e Oportunidades para Pesquisas Futuras:

- Limitações podem incluir a dependência de condições específicas do mercado e a sensibilidade a eventos externos. O modelo pode ser refinado para lidar com diferentes cenários e considerar dados mais recentes. Oportunidades futuras incluem explorar técnicas avançadas de aprendizado de máquina e integrar dados externos para melhorar a precisão e a generalização.

