**Modelo para negociação de criptomoedas usando detecção de anomalias e análise de sentimentos de rede social**

Daniel Benicio de Freitas Alves ¹\*;Juliano Domingues da Silva2

1 Universidade de São Paulo. Discente do MBA em Data Science e Analytics. Rua Alexandre Herculano, 120 – Vila Monteiro; 13418-445 Piracicaba, SP, Brasil

2 Universidade Estadual de Maringá. Professor Adjunto do Departamento de Administração. Av. Colombo, 5790 - Zona 7, 87020-900, Maringá - PR, Brasil.

\*autor correspondente: dalves@usp.br

**Modelo para negociação de criptomoedas usando detecção de anomalias e análise de sentimentos de rede social (TCC)**

**Resumo**

O mercado financeiro é hoje uma das maiores demandas de DSA e a predição de tendências de altas e baixas do mercado de negociação de títulos e outros índices financeiros é um tema frequentemente estudado, ainda mais com os mecanismos existentes de automação da compra e venda de ativos por meio de algoritmos. Contudo, é um problema também bastante desafiador por se tratar de um sistema complexo, cuja previsão das tendências ideais para as operações é caracterizada por relacionamentos ocultos, ruído e alto grau de incerteza. Por esse motivo, se faz necessário o aperfeiçoamento da eficiência do Trading, e é apresentada aqui neste estudo uma estratégia em razão do monitoramento do mercado e de tendências de alta e baixa de criptomoedas utilizando a Análise de Sentimentos de rede social conjugada com a Detecção de Anomalias sobre dados do Twitter e do mercado financeiro para identificar pontos onde a atividade intensa e anormal, e que poderiam caracterizar ótimas oportunidades de Trading diante de tendências de altas e baixas dos preços das criptomoedas. No presente estudo foi possível avaliar com vinte criptomoedas como Bitcoin, Ethereal e outras a performance dos modelos de Sentiment Analysis AFINN, TextBlob e Vader Sentiment, que conjugados com as tecnicas estatísticas de Cruzamento de Médias Móveis Exponenciais e Desvio da Média Absoluta resultou em uma melhoria de performance significativa de retorno oito vezes maior que o investimento de longo prazo.

**Palavras-chave:** Algorithmic Trading, Stock, NLP, Sentiment Analysis.

**Cryptocurrency trading model using Anomaly Detection and Sentiment Analysis of social network**

**Abstract**

The financial market is today one of the biggest demands of DSA and the prediction of up and down trends of the bond trading market and other financial indices is a topic often studied, even more with the existing mechanisms of automation of the purchase and sale of assets. through algorithms. However, it is also a very challenging problem because it is a complex system, whose prediction of ideal trends for operations is characterized by hidden relationships, noise and a high degree of uncertainty. For this reason, it is necessary to improve the efficiency of Trading, and a strategy is presented here in order to monitor the market and the up and down trends of cryptocurrencies using the Sentiment Analysis of a social network combined with the Detection of Anomalies on Twitter and financial market data to identify points where intense and abnormal activity, which could characterize great trading opportunities in the face of up and down trends in cryptocurrency prices. In the present study it was possible to evaluate with twenty cryptocurrencies such as Bitcoin, Ethereal and others the performance of the Sentiment Analysis models AFINN, TextBlob and Vader Sentiment, which combined with the statistical techniques of Crossing Exponential Moving Averages and Deviation from the Absolute Average resulted in an improvement of significant performance with a return eight times greater than the long-term investment.

**Keywords:** Algorithmic Trading, Stock, NLP, Sentiment Analysis

**Introdução**

O mercado financeiro é hoje uma das maiores demandas de DSA (Data Science and Analytics) e a predição de tendências de altas e baixas do mercado de negociação de títulos e outros índices financeiros é um tema frequentemente estudado (Valle-Cruz et al., 2021; Attigeri et al., 2015), ainda mais com os mecanismos existentes de automação da compra e venda de ativos por meio de algoritmos (em inglês, Algorithmic Trading, ou A.T, quando feito por robôs, ou simplesmente Trading no contexto geral de compra e venda de ativos automaticamente ou de maneira manual). Contudo, é um problema também bastante desafiador por se tratar de um sistema complexo, cuja previsão das tendências ideais para as operações é caracterizada por relacionamentos ocultos, ruído e alto grau de incerteza (Attigeri et al., 2015). Por esse motivo, se faz necessário o aperfeiçoamento da eficiência do Trading, e é apresentada aqui neste estudo uma estratégia em razão do monitoramento do mercado e de tendências de alta e baixa de criptomoedas. Este grupo de ativos é relativamente novo, apresentando ainda maior nível de incerteza e volatilidade que os demais investimentos, o que resulta em elevado risco para o investimento (Royal, 2022), tendo surgido a pouco mais de uma década, e estando atualmente disponíveis a serem negociados através da Internet por serem definidas como um número finito e valioso de recursos obtidos originalmente através da mineração de novos blocos do Blockchain (Frankenfield, 2022), têm semelhança com a negociação de divisas ou câmbio de moedas estrangeiras (ou Foreign Exchange – FOREX, como costuma ser referenciado em inglês), embora normalmente irrastreável e sem lastro, que são alguns dos motivos para as criptomoedas terem sido banidas em alguns países por também representarem algum risco a economia ou segurança nacional. Contudo, justamente devida a maior incerteza e volatilidade, e portanto com maiores oportunidades de guinadas para tendências de altas e baixas, as criptomoedas foram definidas como alvo de estudo à fim de encontrarmos aqui resultados de real significância para as áreas de DSA e A.T.

Além disso, o ponto basilar da estratégia de A.T. aqui proposta é utilizar a Análise de Sentimentos (Sentiment Analysis em inglês) de rede social conjugada com a Detecção de Anomalias do Aprendizado de Máquinas Não-Supervisionado sobre dados do Twitter e do mercado financeiro para identificar pontos onde a atividade intensa e anormal, e que poderiam caracterizar ótimas oportunidades de Trading diante de tendências de altas e baixas dos preços das criptomoedas. No presente estudo foi possível avaliar a performance dos modelos de Processamento de Linguagem Natural (em inglês, Natural Language Processing ou NLP) das versões estáveis de frameworks pré-treinados popularmente usados de Sentiment Analysis baseados em valência: AFINN (Finn, 2011), TextBlob (Loria, 2018) e Vader Sentiment (Hutto & Gilbert, 2014), e também no framework com melhor desempenho entre os três ainda foram avaliados os impactos de atualizações do léxicon de palavras e símbolos mais usados durante o período com pertinência à área de Trading de criptomoedas.

Para fins de repetição e continuidade futura de estudos sobre o tema bem como maior detalhamento dos dados aqui apresentados, o presente artigo para leitura em sua última revisão e alguns arquivos de apoio na análise dos dados ficarão disponíveis no repositório de dados no link abaixo:

<https://bit.ly/tcc-usp-mbadsa-dalves>

**Trabalhos relacionados**

O da Análise de Sentimentos tem sido bastante abordado ultimamente, grandes referências a área foram os estudos de (Abraham et al., 2018; Pano et al.,2020; Pillai et al.,2021), que trouxeram contribuições significativas para a área financeira, bem como (Hansun et al., 2017) com sua análise compartiva dentre diversas médias móveis distintas e que hoje são a página inicial deste imenso mundo de oportunidades que é o Trading.

Além disso, (Ahmed et al.,2017) trouxeram uma análise de desempenho muito valiosa com seu experimento também a esse domínio, no campo do aprendizado de máquina não supervisionado.

**Fundamentação teórica**

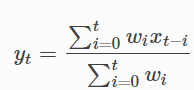
**Trading, HODL e Cryptocurrency Trading**

Usualmente Trading é uma operação de curto prazo – e com criptomoedas envolve comprar e vendê-las (entrar e sair de uma posição, respectivamente) usando uma plataforma de trading de criptomoedas (Cryptocurrency Trading, em inglês, ou simplesmente Crypto Trading como costuma se falar); difere de um investimento que é quando se detém um ativo por um longo período com o objetivo de conseguir comprar recorrentemente mais ativos a longo prazo e aumentar o patrimônio através da valorização do preço inicial da(s) posição(ões) que algum dia serão liquidadas, ou seja, terão seus ativos vendidos pelo novo preço, rendendo de acordo com a flutuação do mercado ao longo do tempo. Esse tipo de estratégia de investimento é chamado Buy and Hold, em inglês, ou HODL, que é um termo que surgiu de um erro de digitação que viralizou no meio financeiro). Já no Trading, existem diversas estratégias diferentes para obter lucro no curto ou médio prazo, mas resumidamente a ideia é comprar um ativo no início de uma tendência de alta, e após a sua valorização momentânea, vendê-lo o quanto antes diante de uma tendência de baixa (Abrol, 2022), e prosseguir assim repetidamente, até obter o maior rendimento possível com a especulação do recurso financeiro.

Aqui neste estudo optou-se por analisar a performance do Crypto Trading usando o modelo proposto em comparação com HODL, por ser este um padrão de mercado, usado aqui portanto como grupo controle, primeiramente numa simples comparação dentre abordagens com lógica distinta, e então com o aumento do dicionário de palavras conhecidas da abordagem que melhor performou no primeiro caso, comparando com sua atualização. Durante as simulações, para cada criptomoeda negociada foi realizada uma única operação de compra ao valor do mercado na primeira tendência de alta detectada com emprego de médias móveis, e realizada a liquidação ao fim do período analisado. Com seu rendimento medido a partir da diferença percentual entre o valor final e o valor inicial do investimento, ou seja o seu retorno percentual.

**Cruzamento de Médias Móveis Exponenciais**

O presente estudo consiste basicamente na simulação de operações de Crypto Trading de acordo com o cruzamento de duas Médias Móveis. As médias móveis são abordagens estatísticas empregadas com frequência na Análise Técnica do mercado de financeiro, que é um ramo analítico de investimentos que procura compreender os padrões de movimento dos preços de títulos e índices (Jason, 2022), ao mesmo tempo que são alvo de estudo na análise de séries temporais em DSA, por traçar suavizações das tendências e servirem de apoio a compreensão do comportamento do dado observado. ‎Tratam-se de conjuntos com um comprimento predefinido para o número de valores para serem calculados à média, e este conjunto de valores avança à medida que mais dados são adicionados com o tempo‎, deslocando-se a diante ao longo das horas, dias ou qualquer que seja a cadência temporal analisada, e excluindo o primeiro elemento do subconjunto anterior, e adicionando o elemento imediatamente após o subconjunto anterior ao novo subconjunto de forma a manter o comprimento fixo. Quando se trabalha com ao menos duas médias móveis, cada uma em períodos distintos e a média móvel mais curta cruza uma média móvel mais longa, indica potencial para comprar ou vender um ativo (Pratik, 2022).

 Médias Móveis Simples ou Simple Moving Average (SMA) são calculadas pela média da soma de todos os valores dentro da janela móvel onde cada um dos pontos valorados têm o mesmo peso, indiferente de sua posição na sequência. Ainda no pensamento linear, existem as Médias Móveis Ponderadas ou Weighted Moving Average (WMA) onde algum tipo de progressão não exponencial é aplicada na diferenciação dos valores mais recentes. Já Médias Móveis Exponenciais, ou Exponential Moving Average (EMA) e Médias Móveis Exponenciais Ponderadas, ou Exponential Weighted Moving Average (EWMA) adotam alguma função exponencial como base para formar um peso de ponderação, de forma que os valores mais recentes (ou seja, à direita da sequência), sejam tratados com maior peso do que os valores mais antigos da janela móvel. Para que não ocorra confusão entre WMA e EWMA, e considerando que na prática essa última é a usada com o cruzamento de duas ou mais médias móveis, aqui neste estudo a EWMA foi empregada nomeada como CEWMA daqui em diante, e implementada usando a biblioteca Pandas do Python cujo método

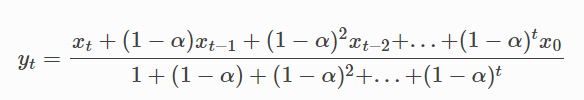
ewma( ) é nativo e calculado conforme a fórmula

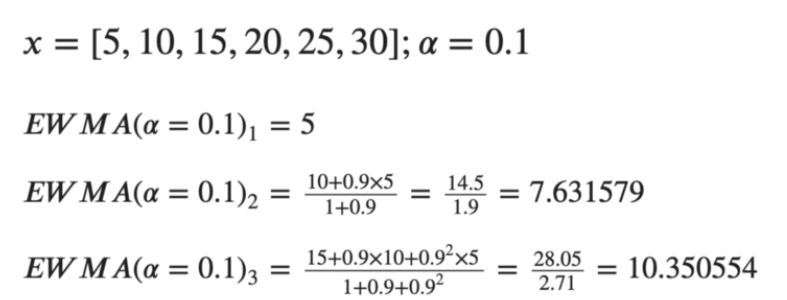


Onde é o valor de entrada e é o resultado, e por default usa o peso



O que dá:



Conforme pode ser observado no exemplo abaixo:

Nas simulações realizadas no decorrer desta pesquisa, as avaliações de diferença percentual do retorno financeiro além da métrica HODL usada para controle, também possuem a título de informação a diferença percentual do retorno financeiro para o Cruzamento de Média Móvel Exponencial Ponderada, ou CEWMA (do acrônimo para Crossing Exponential Weighted Moving Average, em inglês), além da diferença percentual com relação ao grupo controle.

**Algoritmos de Sentiment Analysis com base na valência das palavras**

Outro condicional da estratégia de negociação simulada nesta pesquisa, foi a ocorrência de sentimento positivo ou negativo quanto ao ativo, em torno do momento onde ocorrem os cruzamentos das médias móveis sobre as variações das criptomoedas obtida pela análise de sentimentos detectados nos dados extraídos da rede social, comparando três modelos de algoritmos pré-treinados: AFINN, TextBlob e Vader Sentiment, inicialmente nas versões disponíveis na Internet (conforme especificação no anexo), e posteriormente um deles, que melhor performou, foi selecionado para atualização do léxicon com os termos da área de Crypto Trading mais presentes no dataset mas que ainda não eram compreendidos pelo algoritmo.

Existem outros métodos de análise de sentimentos, mas aqui selecionamos aqueles que usam a estratégia de valência das palavras por razões de familiaridade com esta abordagem popular.

**Median Absolute Deviation**

Além das médias móveis, outro indicador de compra e venda base da presente simulação foi a identificação de anomalias de base estatística. Neste estudo foi implementado a detecção de anomalias usando Median Absolute Deviation (MAD), da biblioteca statsmodels (Seabold et al., 2010), que atua em séries temporais quantitativas para identificação de anomalias na variação dos dados, quando um ou mais pontos se distanciam dos demais em um valor denominado *threshold*, maior que N vezes a média absoluta da janela móvel de X períodos, sendo essa uma técnica estatística bastante empregada na identificação de anomalias da aprendizagem de máquina não supervisionada (MARJUNI et al., 2019; Fitriyah et al., 2019; Ahmad et al., 2011).

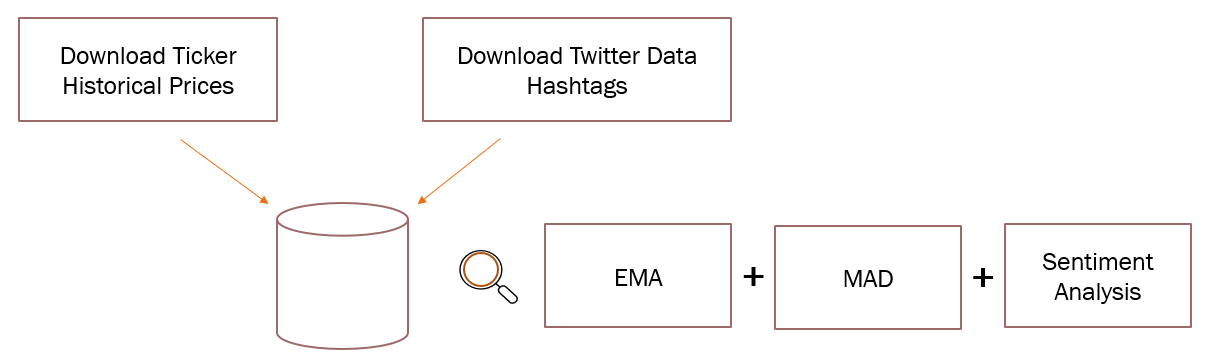
Neste caso, optamos por usar um valor usual de threshold de 3 vezes a média, numa janela móvel de 10 períodos, e monitorando possíveis pontos de atenção sobre os dados:

1. Quando ocorre volume de investimento atípico em torno do momento sendo avaliado.
2. Quando ocorre repercussão atípica da rede social relacionada com aquela criptomoeda em torno do momento sendo avaliado, sendo essa repercussão momentânea da moeda quantificada à média das repercussões individuais das postagens relacionadas com a hashtag naquela janela de tempo, que é calculada pela soma de todos os hits sobre daquela mensagem (1 post + X retweets + Y likes + Z replies).

**Material e Métodos**

A pesquisa descritiva aqui desenvolvida tem como objetivo avaliar o comportamento de algoritmos populares de análise de sentimentos baseados em valência e sua relação com a melhora ou piora do retorno financeiro simulado em operações de Crypto Trading, conforme o diagrama. Além disso, com a atualização do léxicon do algoritmo que melhor performou, espera-se:

1. Identificar qual algoritmo de Sentiment Analysis dentre os selecionados performou melhor, e o quanto melhor ele foi que os demais, sem quaisquer customizações do dicionário de palavras e símbolos conhecidos.
2. Identificar a viabilidade de especialização do idioma conhecido na forma de um jargão técnico.
3. Avaliar se há diferença perceptível na performance de NLP entre as linguagens verbal e não verbal acerca de Crypto Trading.



**Dados analisados**

Os dados analisados consistem em uma base histórica com informações agregadas de várias séries temporais entre 2021-01-01 e 2022-07-31 com observações de 60 em 60 minutos.

A base contém dados de cotação de criptomoedas disponíveis para Trading, dentre as quais foram selecionadas arbitrariamente 20 moedas populares em 2022 sem qualquer relação intencional entre si. O único requisito era terem cotação em USD disponível em intervalos de 60 min na API do Yahoo Finance. Dessa forma, foram analisadas neste estudo as criptomoedas listadas abaixo com seus *tickers* (código de identificação de ativos do mercado financeiro) e nomes populares se tiverem entre parênteses:

1. ATOM (Cosmos)
2. ADA (Cardano)
3. BCH (Bitcoin Cash)
4. BNB (Binance Coin)
5. BTC (Bitcoin)
6. DASH
7. DCR (Decred)
8. DOGE (Dogecoin)
9. DOT (Polkadot)
10. EOS
11. ETH (Ethereal)
12. ETC (Ethereum Classic)
13. HEX
14. LTC (Litecoin)
15. LUNA (Terra)
16. MIOTA (Iota)
17. SHIB (Shiba Inu)
18. SOL (Solana)
19. XMR (Monero)
20. XRP (Ripple)

A base também contém todo o histórico de tweets identificados pela hashtag com nome do ticker em questão, que além do post e códigos de identificação também possuem informações a respeito de likes, retweets e replies. Estes dados foram obtidos usando o web scraper de código aberto snscrape.

**Pré-processamento dos dados**

Foi realizado o Data Wrangling da base, compreendendo este processo de limpeza e uniformização dos dados em um único dataframe para ser analisado, cobrindo as seguintes operações:

• Remoção de URLs e figuras (emoticons foram deixados, uma vez que o algoritmo Vader Sentiment consegue interpretá-los, embora os outros dois algoritmos usados para análise de sentimentos não tenham essa feature no momento).

• Remoção de menções de perfis de usuários

• Remoção de hashtags e cashtags

• Remoção de “Stop Words”

• Alteração da capitalização dos textos para minúsculas. .

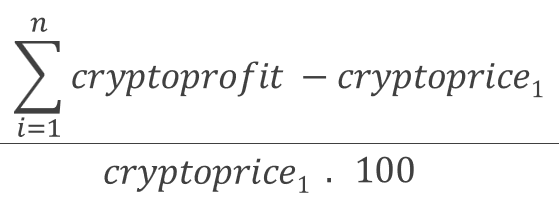
**Avaliações de performance**

Necessário nesse ponto identificar os indicadores que serão comparados para avaliar dentre as estratégias aquela de mais alta performance. Neste caso, adotamos dois indicadores principais:

1. HPP (HODL Profit Percentage)):

Dado pelas fórmulas abaixo

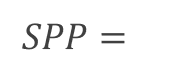
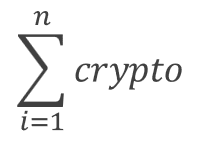
1. Cada Cryptocurrency analisada tem a soma de todos os retornos obtidos nas operações de Trading (nomeado aqui simplesmente cryptoprofit pra simplificar) decrescido do preço da sua primeira posição, dividido pelo preço da primeira posição vezes 100. Isso resulta no retorno percentual da Cryptocurrency, ou simplesmente crypto, ou seja a percentagem de retorno de cada criptomoeda analisada.

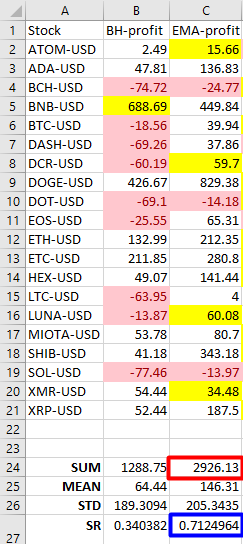


Onde



1. O SPP (abreviação de Strategy Profit Percentage) é o somatório de todos os retornos percentuais obtidos para aquela estratégia.





No exemplo ao lado, a segunda coluna lista o retorno percentual HODL para cada cryptomoeda na primeira coluna, enquanto a terceira coluna é o retorno percentual do CEWMA. Cujo SPP daria 2926.13% conforme o exemplo.

iii. O HPP é a diferença percentual do SPP CEWMA com o HODL SPP, que daria portanto 127% conforme o exemplo.

*(2926.13 – 1288.75) . 100 = 127*

*1288.75*

1. SR (Sharpe Ratio):

É uma medida estatística muito usada no mercado financeiro e principalmente na indústria de fundos e investimentos (Bittencourt, 2022). Na prática, quanto mais alto melhor, demonstrando quais técnicas chegam a resultados mais uniformes uma vez que seu desvio padrão é o menor possível.

Segundo (Treleaven et al, 2013), se denotarmos por Rt os retornos diários, ou seja, as variações diárias relativas (%) do valor total, então o índice de Sharpe associado (SR) é dado por:



Onde SR é igual a média(Rt) pelo desvio padrão(Rt) são, respectivamente, a média e o desvio padrão dos retornos calculados ao longo de todo o período de treinamento da estratégia de investimento, que compreende o período da amostra dos dados.

Na figura acima,

*146.31 = 0.7124964*

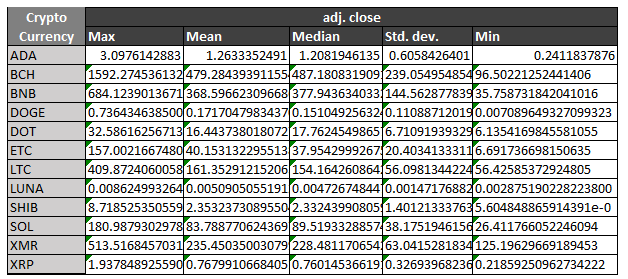
*205.3435*

**Resultados e Discussão**

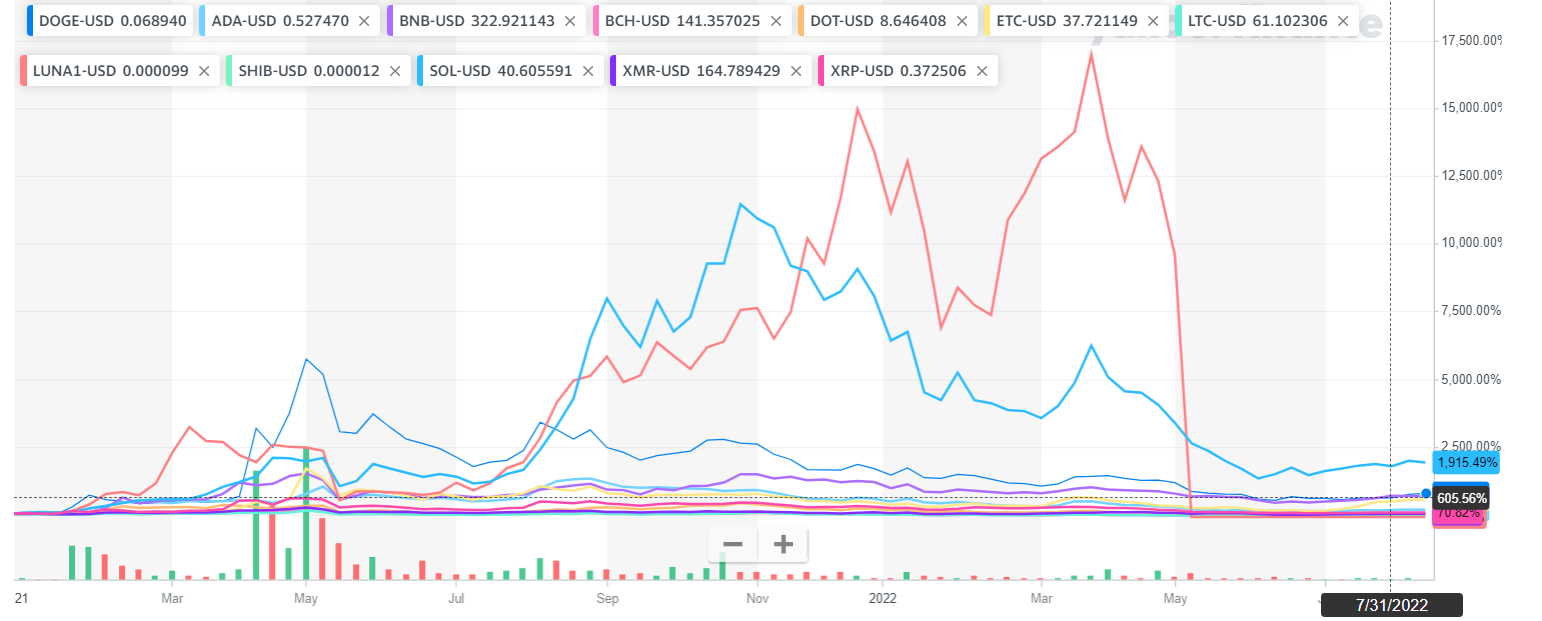
**Detalhamento da base de dados**

Os dados de interesse do dataset pós Wrangling, seguem listados abaixo. Para maiores detalhes, observar o diretório <https://bit.ly/tcc-usp-mbadsa-dalves> com todos os dados desta pesquisa.

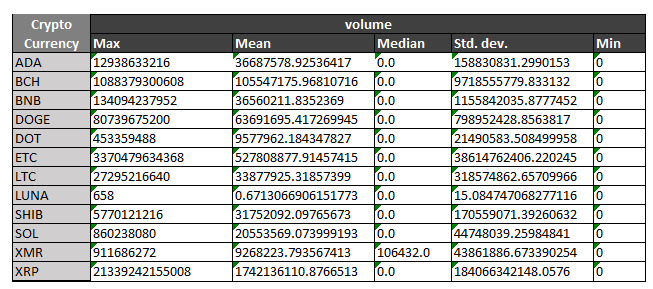
Adj. Close é campo com preço ajustado pelo mercado em que o ativo encerrou a negociação naquele dado momento. Abaixo a figura resume estatisticamente alguns preços observados:



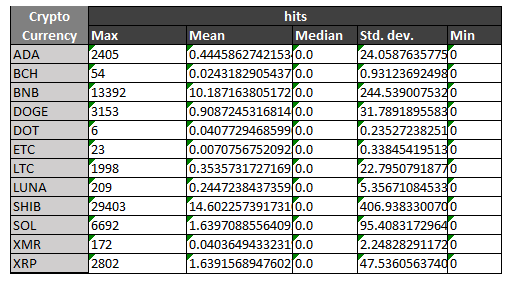
Ao longo do tempo, as variações do preço das criptomoedas analisadas com relação ao USD apresentaram comportamentos distintos com suas respectivas séries temporais, conforme o exemplo demonstrado no gráfico abaixo:



Volume é campo com o total negociado das operações Trading do ativo naquela janela de tempo, conforme exemplificado abaixo.



Hits são a repercussão momentânea da moeda quantificada à média das repercussões individuais das postagens relacionadas com a hashtag naquela janela de tempo, calculada pela soma de todos os hits sobre daquela mensagem (1 post + X retweets + Y likes + Z replies), conforme a amostra de dados a seguir:

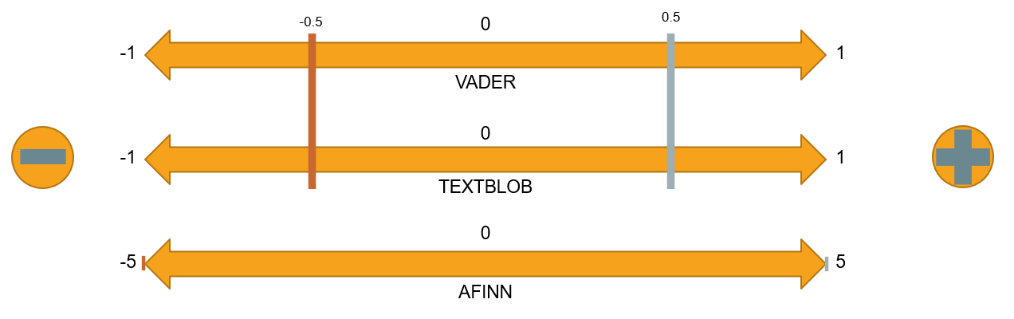


Quanto as nuvens de palavras observada, abaixo seguem algumas que chamaram mais atenção por conta dos termos e símbolos utilizados. A totalidade das nuvens para cada ativo pode ser observada no repositório de dados.

|  |  |
| --- | --- |
| DASH | Dogecoin |
| Litecoin | Iota |
| Shiba Inu | Solana |

**Limitações encontradas**

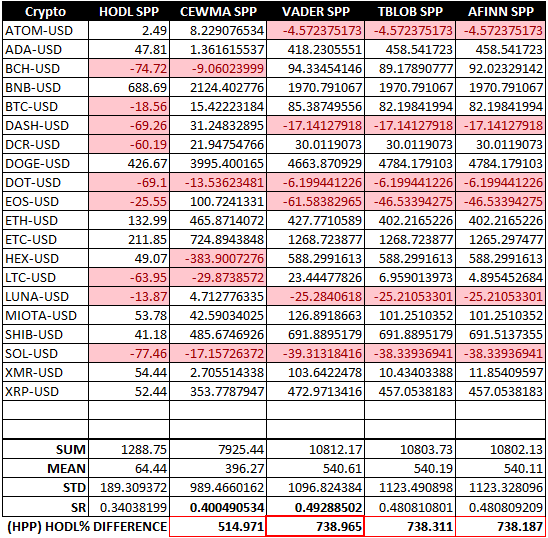
Conforme especificação a ser detalhada adiante, dentre os algoritmos escolhidos para a análise de sentimentos deste estudo, alguns não possuem a documentação tão detalhada quanto o desejável, o suporte a símbolos e outras linguagens diferentes do inglês. Ainda, é possível identificar que a lógica entre os algoritmos também difere, pois por exemplo, conforme ilustrado na imagem abaixo, é possível ver que todos trabalham com output trinários, ou sejam, suportam três valores distintos com resultado do seu processamento: negativo, neutro e positivo. Contudo, tanto Vader quanto Textblob têm uma normalização do resultado, de forma que sempre de -1 a -0.5 tem sentimento negativo, e de 0.5 a 1 tem sentimento positivo, ou de outra forma é neutro. O Afinn, por outro lado, não possui essa normalização, e conforme sua especificação qualquer valor abaixo de -5 é negativo, e acima de 5 é positivo, embora seu output não tenha um limite claro e não é raro seu valor orbitar entre -8 em um momento e quase 19 num momento seguinte, e essa divergência poderia levar a erros caso o mesmo threshold dos primeiros algoritmos lhe fosse atribuída.



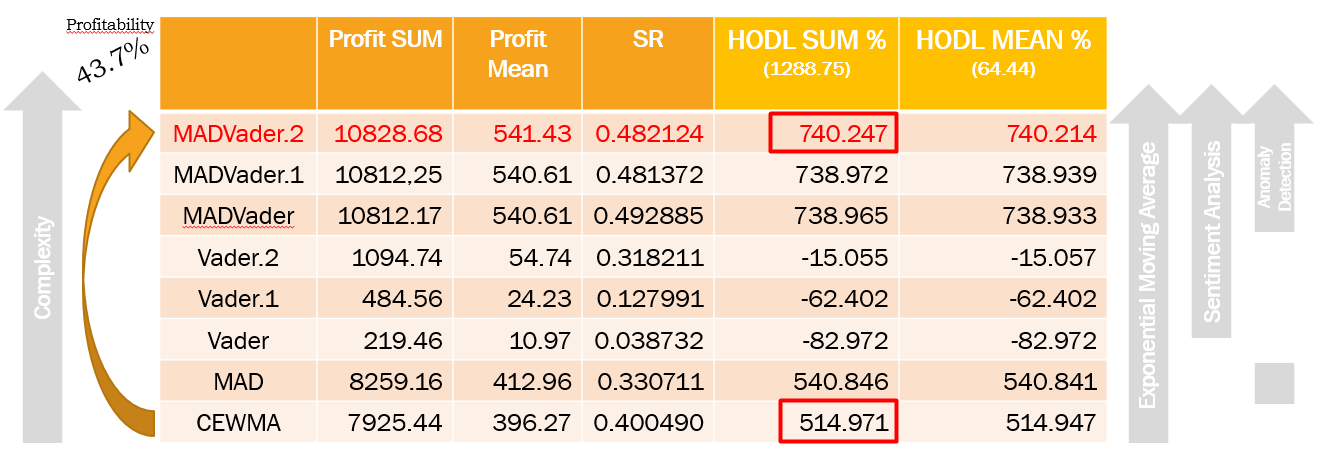
Outro fator de limitação encontrado nas estratégias de Sentiment Analysis foi, por exemplo com o Vader, a valoração de palavras de seu dicionário é realizada seguindo uma média de pontuações feitas por dez indivíduos sem relação entre si, o que requer em uma primeira análise tempo e engajamento da comunidade para tal. Dito isso, e admitindo que esse é o algoritmo que mais performou na análise em primeira mão sem melhoria do dicionário de palavras, conforme detalhamento adiante, ainda que uma issue foi registrada para que um jargão técnico de Crypto Trading fosse construído para atender o prazo de apresentação da pesquisa à banca foi possível seguir com o presente estudo simplesmente usando o próprio algoritmo para computar as novas palavras de interesse extraídas das nuvens de palavras com o simples processamento de termos extraídos de um glossário popular do nicho de criptomoedas (Hooson, 2022).

**Avaliação dos resultados**

Após processados nos respectivos algoritmos de análise de sentimentos, sem quaisquer modificações dos dicionários de palavras das versões originais dos três algoritmos de Sentiment Analysis avaliados, foi observada a maior performance da estratégia com Vader Sentiment nas duas métricas, SR e HPP, indicando respectivamente os retornos menos dispersos e o maior lucro percentual, em torno de 10812.17% em comparação com o HODL cujo lucro percentual foi 1288.75%, ou seja, nesse domínio da pesquisa ela performou em torno de oito vezes melhor.



Prosseguindo então com os devidos updates do dicionário de palavras no Vader Sentiment, seguindo a estratégia original de atualizar primeiro o léxicon com as palavras ainda desconhecidas pertinentes a Crypto Trading (nomeados \*Vader.1 nos resultados abaixo) e então os símbolos (nomeados \*Vader.2), e ainda isolando também quanto a utilização da detecção de anomalias (MAD) ou do simples cruzamento da média móvel exponencial (CEWMA), foi possível observar a diferença percentual da estratégia deste estudo. E em comparação com o HODL, ela se manteve nesta razão de oito para um com um discreto aumento de menos de menos de 2%, de forma que permanece incerto até o momento qualquer afirmação sobre a determinância da linguagem não verbal neste domínio e mais estudos poderiam seguir esta linha de investigação no futuro.



Ainda, comparando a abordagem mais simples de A.T. com uso apenas de médias móveis exponenciais contra o novo modelo aqui proposto, foi possível observar uma lucratividade 43.7% maior (em vermelho) embora o SR tenha demonstrado uma leve queda na dispersão dos resultados com relação ao modelo com Vader “saído da caixa”, ou seja, sem atualizações do seu léxicon (MADVader x MADVader.2).

Tais resultados corroboram a relevância do Vader na análise de sentimentos apresentada por (Abraham et al., 2018; Pano et al.,2020; Pillai et al.,2021) e das Médias Móveis Exponenciais por (Hansun et al., 2017) em operações no mercado financeiro. Entretanto, a técnica estatística MAD para detecção de anomalias mostrou alguma eficiência sozinha, e bastante eficiência quando conjugada com Vader e CEWMA, contradizendo parcialmente (Ahmed et al.,2017) em sua afirmação sobre abordagens como k-NN (Nearest Neigbour) e CMGOS (Clustering-based Multivariate Gaussian Outlier Score) serem mais adequadas ao invés das abordagens estatísticas no domínio de mercado financeiro. Entretanto MAD não foi uma técnica analisada na ocasião, e isso pode ter sido um dos fatores para chegar a tal conclusão. Novos estudos são necessários para melhor compreensão deste cenário, e talvez mesmo o k-NN e CMGOS possam performar ainda melhor em conjunto com as técnicas Vader e CEWMA empregadas no estudo atual.

**Conclusões**

Para abordar eficientemente o problema de previsão das tendências ideais para as operações de Trading do mercado financeiro em meio aos ruídos e incerteza naturais desse cenário, foi apresentado neste artigo uma estratégia inovadora usando Detecção de Anomalias e Análise de Sentimentos, além do Cruzamento de duas Médias Móveis Exponenciais Ponderadas, o que já costuma ser empregado com alguma frequência entre outras técnicas mais populares. Observamos num primeiro momento com os dados trabalhados que dentre as versões estáveis dos algoritmos de Análise de Sentimentos estudados o Vader Sentiment demonstrou um discreto retorno percentual maior em comparação com investimentos realizados em Buy and Hold do que os demais algoritmos Afinn e TextBlob, e no segundo momento que se manteve com melhorias discretas quando experimentamos atualizar seu dicionário de palavras conhecidas para as palavras e símbolos do domínio mais frequentes no dataset que compilamos. Ainda, a performance do método aqui apresentado foi muito superior quando comparada ao simples investimento a longo prazo, na proporção de 8 vezes maior, o que consideramos um resultado significativo para a área.

Além disso, o presente estudo abre oportunidades para estudos futuros usando outros métodos, como por exemplo k-NN e CMGOS, entre outras abordagens clusterizadas no lugar da Detecção de Anomalias aqui empregada, outras abordagens de Sentiment Analysis que não sejam baseadas em valência das palavras, ou que tenham um dicionário de palavras específico do mercado financeiro e de Crypto Trading, a comparação da performance desta mesma técnica quando empregada em outros ativos do mercado financeiro como por exemplo Fundos Imobiliários, Ações, FOREX, etc., e a ampliação do dataset analisado incluindo feeds de notícias e fóruns especializados.

**Agradecimento**

Agradeço à minha esposa Layla Alves por todo apoio e compreensão com meu engajamento acadêmico e na execução deste trabalho, e aos meus filhos Helena e Erik, que me incentivam diariamente a ir cada vez mais longe para quem sabe, um dia, inspirá-los a uma grandeza ainda maior. E ao criador do universo, tenha o nome que tiver em culturas e filosofias diferentes, e todos os seus agentes pela inspiração necessária na condução deste trabalho.

**Referências**

Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. SMU Data Science Review, 1(3), 1.

Abrol, Ayushi, “Complete Guide To Cryptocurrency Trading For Beginners”, <https://www.blockchain-council.org/cryptocurrency/complete-guide-to-cryptocurrency-trading-for-beginners/> , acessado em 16 de Outubro de 2022.

Ahmad, M., Haq, D. I. U., Mushtaq, Q., & Sohaib, M. (2011). A new statistical approach for band clustering and band selection using K-means clustering. Int. J. Eng. Technol, 3(6), 606-614.

Ahmed, M., Choudhury, N., & Uddin, S. (2017, July). Anomaly detection on big data in financial markets. In 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 998-1001). IEEE.

Akbik, Alan and Bergmann, Tanja and Blythe, Duncan and Rasul, Kashif and Schweter, Stefan and Vollgraf, Roland (2019), 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), p. 54-59

Attigeri G V, MM MP, Pai RM, Nayak A. Stock market prediction: a big data approach. In: TENCON 2015–2015 IEEE Region 10 Conference. 2015. p. 1–5.

Bittencourt, Andre. “Sharpe Ratio e o Teste de Hipóteses”, <https://www.linkedin.com/pulse/sharpe-ratio-e-o-teste-de-hip%C3%B3teses-andr%C3%A9-bittencourt-cqf> , acessado em 15 de Agosto de 2022.

Finn Årup Nielsen, "A new ANEW: evaluation of a word list for sentiment analysis in microblogs", Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages. Volume 718 in CEUR Workshop Proceedings: 93-98. 2011.

Fitriyah, H., & Budi, A. S. (2019). Outlier Detection in Object Counting based on Hue and Distance Transform using Median Absolute Deviation (MAD). 2019 International Conference on Sustainable Information Engineering and Technology (SIET). doi:10.1109/siet48054.2019.8985993

Frankenfield, James, “What is Bitcoin? How to Mine, Buy and Use It”, <https://www.investopedia.com/terms/b/bitcoin.asp> , acessado em 16 de Outubro de 2022.

Hansun, S., & Kristanda, M. B. (2017, November). Performance analysis of conventional moving average methods in forex forecasting. In 2017 International Conference on Smart Cities, Automation & Intelligent Computing Systems (ICON-SONICS) (pp. 11-17). IEEE.

Hooson, Mark, “Cryptocurrency Glossary Of Terms & Acronyms”, <https://www.forbes.com/advisor/investing/cryptocurrency/crypto-glossary/> , acessado em 16 de Outubro de 2022.

Jason Fernando., “Moving averages” , [http://www.investopedia.com/university/moving-](mailto:http://www.investopedia.com/university/movingaverage/)

[average/](mailto:http://www.investopedia.com/university/movingaverage/) , acessado em 20 de Março de 2022.

Loria, S. (2018). textblob Documentation. *Release 0.16*.

Marjuni, A., Adji, T. B., & Ferdiana, R. (2019). Unsupervised software defect prediction using median absolute deviation threshold based spectral classifier on signed Laplacian matrix. Journal of Big Data, 6(1). doi:10.1186/s40537-019-0250-z

Pano, T., & Kashef, R. (2020). A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the era of COVID-19. Big Data and Cognitive Computing, 4(4), 33.

Pillai, S., Biyani, D., Motghare, R., & Karia, D. (2021, June). Price prediction and notification system for cryptocurrency share market trading. In 2021 International Conference on Communication information and Computing Technology (ICCICT) (pp. 1-7). IEEE.

Pratik Nabriya, “O que são médias móveis?”, https://towardsdatascience.com/making-a-trade-call-using-simple-moving-average-sma-crossover-strategy-python-implementation-29963326da7a, acessado em 20 de Março de 2022.

Ran Aroussi, “yfinance”, https://github.com/ranaroussi/yfinance , acessado em 20 de Março de 2022.

Royal, James, “Cons of investing in stocks”, <https://www.bankrate.com/investing/crypto-vs-stocks/> , acessado em 20 de Março de 2022.

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In 9th Python in Science Conference..

Treleaven, P., Galas, M., & Lalchand, V. (2013). Algorithmic trading review. Communications of the ACM, 56(11), 76–85. doi:10.1145/2500117

Valle-Cruz, D., Fernandez-Cortez, V., López-Chau, A., & Sandoval-Almazán, R. (2021). Does twitter affect stock market decisions? financial sentiment analysis during pandemics: A comparative study of the h1n1 and the covid-19 periods. *Cognitive computation*, 1-16.

**Anexo**

Lista de pacotes instalados no WSL 2 com Ubuntu 20.04.3 LTS, pela ordem de instalação:

|  |  |  |  |
| --- | --- | --- | --- |
| Ordem | Pacote | Release | Descrição |
| 1 | python3.8 | 3.8.10-0ubuntu1~20.04.2 | Linguagem de programação Python 3 |
| 2 | python3-pip | 20.0.2-5ubuntu1.6 | Instalador de pacotes do Python 3 |
| 3 | pandas | 1.3.5 | Instalado via pip. Biblioteca para suporte a dataframes. |
| 4 | yfinance | 0.1.69 | Instalado via pip. Biblioteca para comunicação com a API Yahoo Finance. |
| 5 | snscrape | 0.4.3.20220107. dev22+ga192dc6 | Instalado via pip. Biblioteca para suporte a web scrapping em diversos sites e redes sociais. |
| 6 | afinn | 0.1 | Instalado via pip. Biblioteca para Sentiment Analysis. |
| 7 | textblob | 0.17.1 | Instalado via pip. Biblioteca para Sentiment Analysis. |
| 8 | vaderSentiment | 3.3.2 | Instalado via pip. Biblioteca para Sentiment Analysis. |
| 9 | statsmodels | 0.13.1 | Instalado via pip. Biblioteca com diversos métodos estatísticos prontos como o MAD, por exemplo. |