UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

MARCO ANTONIO SQUINE MARTINS

COMBINAÇÃO DE TÉCNICAS DE ANÁLISE DE SENTIMENTOS DE TWEETS E DADOS HISTÓRICOS PARA PREVISÃO DE COTAÇÃO DE CRIPTOMOEDAS

DOIS VIZINHOS

2022

MARCO ANTONIO SQUINE MARTINS

COMBINAÇÃO DE TÉCNICAS DE ANÁLISE DE SENTIMENTOS DE TWEETS E DADOS HISTÓRICOS PARA PREVISÃO DE COTAÇÃO DE CRIPTOMOEDAS

COMBINATION OF TECHNIQUES FOR ANALYSIS OF FEELINGS TWEETS AND HISTORICAL DATA FOR PRICE FORECAST OF CRYPTOCURRENCIES

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Software do Curso de Bacharelado em Engenharia de Software da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Marlon Marcon

DOIS VIZINHOS
2022

MARCO ANTONIO SQUINE MARTINS

COMBINAÇÃO DE TÉCNICAS DE ANÁLISE DE SENTIMENTOS DE TWEETS E DADOS HISTÓRICOS PARA PREVISÃO DE COTAÇÃO DE CRIPTOMOEDAS

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Software do Curso de Bacharelado em Engenharia de Software da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 22/junho/2022

Marlon Marcon
doutorado
Universidade Tecnológica Federal do Paraná

André Roberto Ortoncelli
doutorado
Universidade Tecnológica Federal do Paraná

Franciele Beal
doutorado
Universidade Tecnológica Federal do Paraná

DOIS VIZINHOS
2022

AGRADECIMENTOS

Gostaria de agradecer a minha família, pois acredito que sem o apoio deles seria muito difícil vencer esse desafio.

Agradeço ao meu orientador Prof. Dr. Marlon Marcon, pela paciência e ensinamentos, suas contribuições foram de suma importância para que o desenvolvimento deste trabalho fosse realizado.

Aos meus amigos e colegas que fiz durante esta etapa acadêmica. Gostaria de deixar registrado também, o meu reconhecimento à minha família, pois acredito que sem o apoio deles seria muito difícil vencer esse desafio.

Enfim, a todos os que por algum motivo contribuíram para a realização desta pesquisa

RESUMO

O crescimento da popularidade das criptomoedas se deve muito as redes sociais e por isso a mesma vem impactando em muitos aspectos durante a “vida” de uma criptomoeda, com isso este trabalho propôs realizar uma análise para identificar se é possível predizer a cotação de uma criptomoeda com o uso de análise de sentimento das postagens do twitter em conjunto de técnicas de aprendizagem de máquinas posteriormente os comparando entre eles e com métodos clássicos para a predição de valores futuros.

Palavras-chave: bitcoin; mineração de dados; análise de sentimentos; aprendizagem de máquina; criptomoedas.

ABSTRACT

The growth in the popularity of cryptocurrencies is largely due to social networks and therefore it has been impacting in many ways during the “life” of a cryptocurrency, so this work proposed to carry out an analysis to identify whether it is possible to predict the price of a cryptocurrency with the use of sentiment analysis of twitter posts in conjunction with machine learning techniques, later comparing them with each other and with classical methods for predicting future values.

Keywords: bitcoin; data mining; sentiment analysis; machine learning; cryptocurrencies.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

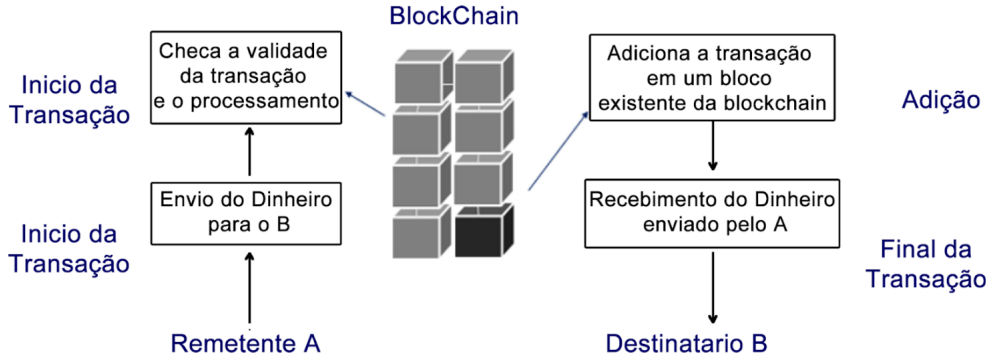
1 INTRODUÇÃO

Com o lançamento da primeira criptomoeda, o Bitcoin em 2009 junto com a blockchain pelo pseudônimo de Satoshi Nakamoto, o mercado de criptomoedas cresceu rapidamente, atraindo atenção tanto na mídia, redes sociais, meios empresariais, meios políticos e representantes de governos em todo o mundo. Também despertou forte interesse em investidores e instituições financeiras, especialmente por algumas de suas características, como o uso de criptografia para o processo de validação de transações, sendo anônimas, mais rápidas e mais simples que o uso de cartões de crédito (NAKAMOTO, 2008; LAZO *et al.*, 2021). Muito deste interesse se deve ao grande mercado de criptomoedas e ao volume de dinheiro envolvendo tais transações, que movimentam diariamente cerca de US$ 140 bilhões (R$ 700 bilhões), e totalizam cerca de US$ 1,25 trilhões (ou cerca de R$ 6,2 trilhões) (COINMARKETCAP, 2021).

A tecnologia blockchain foi disseminada como solução para o problema de gastos duplos da criptomoeda Bitcoin. Funciona a partir de uma chave privada criptografada (secreta como uma senha) e uma chave pública compartilhada com todas as outras partes da rede, cada bloco criado por uma transação possui uma espécie de impressão digital chamada hash, que se trata de um algoritmo matemático criptografado e extremamente difícil de ser revertido (NAKAMOTO, 2008). Se o bloqueio se referir a uma transação financeira, então cada transação no blockchain, por definição, inclui informações sobre transações anteriores e, assim, verifica a propriedade do ativo financeiro que está sendo transferido.

A Figura 1 apresenta um exemplo de uma transação na blockchain, na qual existem dois clientes (do inglês *peer-to-peer, ou P2P*) A e B, sendo que o cliente A deseja enviar dinheiro para o cliente B. A autenticação de cada transação do cliente A para o cliente B pode ser verificada por meio de um livro razão distribuído ou blockchain que é mantido por todos os participantes. Ele verifica a validade da transação no remetente e adiciona as informações da transação ao armazenamento global na extremidade do destinatário e, em seguida, a transação é fechada (KHEDR, 2021).

Figura 1 – Workflow de uma criptomoeda utilizando o mecanismo da blockchain



Fonte: Adaptada de Khedr (2021).

Como não apresentam uma unidade centralizadora e controladora, criptomoedas em geral possuem uma alta volatilidade nas cotações, e esta alta variação pode ser influenciada por movimentações em redes sociais. Eventualmente postagens de pessoas com alta influência sobre a opinião das outras (do inglês influencers) podem alterar positiva ou negativamente as cotações destas.

A predição das tendências de valorização ou desvalorização de criptomoedas se torna muito difícil devido a esta alta volatilidade de cotação e alta influência externa (FERREIRA *et al.*, 2019). Moedas tradicionais como dólar americano e euro, por exemplo, podem ser estimadas por meio de técnicas de predição de séries temporais, que avaliam dados históricos para estimar tendências de valorização ou desvalorização. Entretanto, a estimativa de tendência quando considera-se criptomoedas é muito mais complicada, muito pelos fatores externos que influenciam seus valores.

Diversos autores, podendo-se citar Mohanty *et al.* (2018) e Mittal *et al.* (2019), apontam tais problemas e sugerem a utilização das postagens de redes sociais, como o Twitter, para adicionar complexidade e robustez aos modelos de predição. Estes autores utilizam técnicas de análise de sentimentos para avaliar se uma postagem tem caráter positivo, negativo ou neutro e correlacionam estas observações com dados históricos para estimar melhor as alterações nos valores das criptomoedas.

Um ponto importante a ser citado é que ao analisar postagens, os autores consideram somente o quantitativo das postagens e respectivos sentimentos, porém como citado anteriormente, os influencers podem ter um papel muito importante perante a essas variações. Neste caso, o engajamento gerado a partir de um comentário ou postagem pode influenciar na valorização ou desvalorização de uma criptomoeda. A avaliação do teor de uma postagem pode ser avaliada por meio da análise de sentimentos e o engajamento pelo número de postagens, de repostagens e da quantidade de seguidores que determinado usuário pode ter. Com base no exposto anteriormente, este trabalho propõe verificar a influência da combinação de fatores internos (dados históricos de cotação) e externos (engajamento em redes sociais e teor das postagens) para para a predição da cotação futura de criptomoedas.

O presente trabalho foca na criptomoeda Bitcoin e na rede social Twitter, porém a metodologia aqui apresentada pode ser aplicada em outras moedas e/ou redes sociais.

1.0.1 Ojetivo Geral

Desenvolver um modelo de predição do valor de criptomoedas, que combine dados históricos das cotações de criptomoedas, análise de sentimentos e engajamento de postagens em redes sociais.

Para atingir o objetivo geral, os seguintes objetivos específicos são vislumbrados:

1. Realizar a extração e modelagem dos dados a serem utilizados no treinamento/teste dos modelos.
2. Explorar ferramentas para analisar postagens na rede social Twitter.
3. Testar o modelos de Aprendizagem de Máquina para predição de tendência no valor de criptomoedas.
4. Analisar e comparar os resultados de modelos de Aprendizagem de Máquina testados anteriormente.

1.0.2 Organização do Trabalho

Este trabalho está organizado em cinco capítulos. No Capítulo 2 são apresentados aspectos conceituais para fundamentar a monografia: conceitos acerca do BTC; sobre a rede social Twitter, foco neste trabalho; sobre técnicas e ferramentas para a análise de sentimentos em textos; e por fim estratégias para predição de séries temporais.

No Capítulo **??**, é descrita a proposta deste trabalho, referente a criação de modelos de predição de tendência de valorização de criptomoedas. No Capítulo 4, são apresentados os resultados obtidos no decorrer deste trabalho. Por fim, o Capítulo 5 apresenta as principais conclusões do trabalho .

2 ESTADO DA ARTE

Neste capítulo são apresentados conceitos fundamentais para melhor compreensão da presente Trabalho de Conclusão de Curso, no qual são apresentados, nesta ordem, aspectos relacionados às criptomoedas, à rede social Twitter, à análise de sentimentos em textos, técnicas e modelos de predição e à predição do valor de criptomoedas por meio de técnicas de aprendizagem de máquina.

2.0.1 Criptomoedas

Criptomoedas são moedas digitais usadas para a transição de ativos[1] pelas *exchanges*[2] que agem como corretoras que propõem trocas de diferentes tipos de moedas, podendo ser as nacionais e virtuais. Podem ser usadas para uso comum, assim como as moedas físicas para a compra ou venda de itens ou até mesmo para registrar itens dependendo de sua característica. Atualmente existem milhares de criptomoedas sendo comercializadas, cada uma com suas especificidades. A Figura 2 apresenta alguns exemplos.

Figura 2 – Imagem das Logos das Criptomoedas mais populares do momento

Fonte: Autoria própria.

A Tabela 1 apresenta um comparativo entre as principais criptomoedas, apresentando a participação de mercado destas, e do valor de capitalização em bilhões de dólares. Estas moedas em conjunto representam valor de mercado de mais de um trilhão de dólares.

O presente trabalho focou na predição da cotação da criptomoeda Bitcoin, que é mais difundida no mercado e também a com maior valor de mercado. O Bitcoin, devido a sua importância, também tende a influenciar as outras criptomoedas, o que também justifica sua adoção.

[1] Ativos: Algo que gera um valor direto, como ações, serviços, criptomoedas, no geral algo que vira dinheiro

[2] Exchange: São como corretoras de ativos que atuam em sua maioria com criptomoedas;

Tabela 1 – Criptomoedas mais utilizadas atualmente

Criptomoeda	% mercado	Capitalização total (US$ bi)
Bitcoin	44,84	614,57
Ethereum	17,60	222,67
Tether	4,20	62,67
Binance Coin	3,51	44,03
Cardano	3,07	39,26
Dogecoin	2,45	29,50
XRP	2,41	28,98
Total	**78,08**	**1.041,68**

Fonte: (COINMARKETCAP, 2021).

2.0.1.1 Bitcoin

É uma moeda online ponto a ponto, o que significa que todas as transações acontecem diretamente entre participantes iguais e independentes da rede, sem a necessidade de nenhum intermediário para permitir ou facilitar.

O Bitcoin foi criado, de acordo com as próprias palavras de Nakamoto, para permitir "que os pagamentos online sejam enviados diretamente de uma parte para outra, sem passar por uma instituição financeira".

Por serem descentralizadas cada país ou grandes empresas tendem a estipular regras sobre o uso das criptomoedas podendo aceitar ou negar possíveis comercialização das quais bem desejar com isso os valores destas criptomoedas podem se alterar visto que sua comercialização foi expandida ou reduzida.

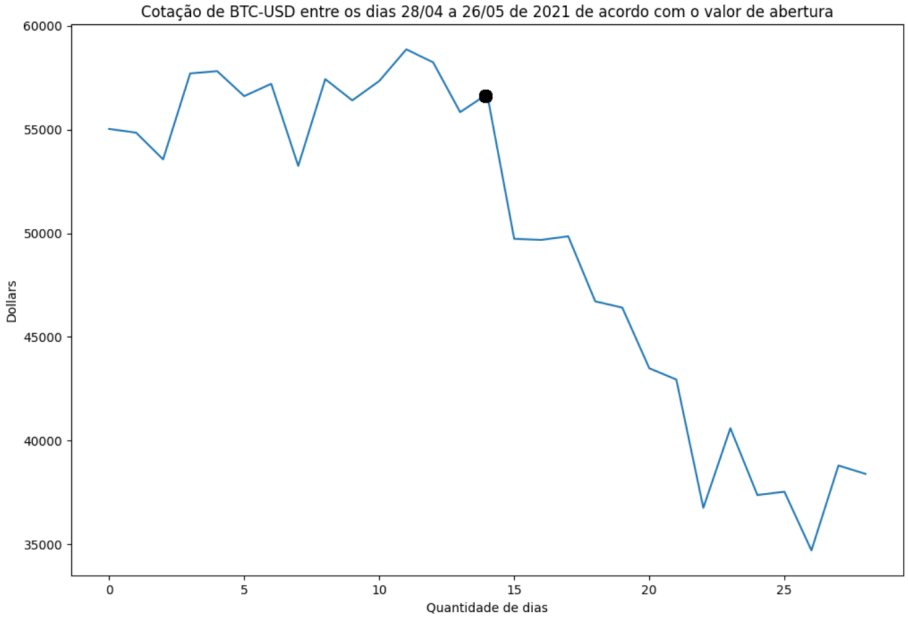
Além disso, por serem naturalmente digitais e desenvolvidas na era das redes sociais, o valor de tais moedas é facilmente influenciado por menções em redes sociais. Por exemplo do empresário Elon Musk, no dia 12/05/2021 (MUSK, 2021), fez uma publicação na rede social Twitter anunciando que sua empresa Tesla não aceitaria mais o Bitcoin em suas transações, devido a isso a moeda teve uma queda de 12,06% no seu valor de mercado em um período de 24 horas. A Figura 3 apresenta o histórico de cotações do Bitcoin 14 dias antes e depois do tweet do empresário, demonstrando a queda substancial na cotação da moeda.

2.0.2 Twitter

O Twitter foi lançado em julho de 2006 como uma mistura de redes sociais com blogs permitindo publicações atualmente de até 280 caracteres. Além disso, os usuários podem adicionar "hashtags"a um tweet, que é o símbolo ""seguido de uma sequência consecutiva de caracteres. Isso é usado para identificar o tópico ou tema de um tweet e torná-los pesquisáveis.a popularidade do Twitter cresceu rapidamente.

Um dos primeiros exemplos de seu alcance e poder foi em 15 de janeiro de 2009, quando um vôo da US Airways caiu no rio Hudson. Uma imagem postada no Twitter deu a notícia antes

Figura 3 – Gráfico de linha gerado a partir de dados coletados pela API "yfinance" no período de 28 de abril a 26/05 do mês de maio de 2021. Dia 12 demarcado no gráfico com (•)



Fonte: Autoria própria.

que os meios de comunicação tradicionais o fizessem. O Twitter tem 330 milhões de usuários ativos mensais, 1,3 bilhão de contas foram criadas, 83% dos líderes mundiais têm uma conta no Twitter, aproximadamente 23 milhões de usuários ativos do Twitter (ABRAHAM, 2018).

Portanto o Twitter pode ser uma fonte muito rica de dados sobre como as pessoas se sentem em relação a praticamente qualquer tópico. Com a capacidade de ver quando um tweet foi postado, também é possível dizer como esses sentimentos mudam com o tempo. Isso torna o Twitter um excelente recurso para coletar dados de texto sobre um tópico como criptomoedas para explorar as possíveis relações entre isso e os preços (ABRAHAM, 2018).

2.0.3 Análise de Sentimentos

A análise de sentimentos é uma área que utiliza processamento da linguagem natural (PLN), análise de textos e linguística computacional para identificar, extrair e quantificar os estados afetivos vinculados à informação subjetiva presentes em textos. Técnicas de análise de sentimentos são amplamente aplicadas na avaliação de opiniões em sistemas de mercado eletrônico e redes sociais, por exemplo.

Estima-se que 90% dos dados do mundo foram gerados nos últimos dois anos. Muitos desses dados estão na forma de dados de texto não estruturados, seja na forma de tweets, artigos postados na Internet, mensagens de texto, e-mails ou outras formas. Essa vasta quantidade de dados não estruturados permitiu o crescimento do PLN como uma área de estudo ou desenvolvimento. PLN é uma coleção de métodos para computadores analisarem e entenderem textos (ABRAHAM, 2018).

Para realização da análise de um sentimento associado a um texto não estruturado, existem algumas ferramentas computacionais já consolidadas, tais como o VADER e o TextBlob. O VADER é uma ferramenta computacional para análise de sentimentos, que utiliza uma lista de palavras, rotuladas como negativas e positivas, de acordo com seu significado semântico, para calcular o grau de sentimento do texto (HUTTO; GILBERT, 2014) . A ferramenta retorna a probabilidade de uma determinada sentença ser positiva, negativa ou neutra.

O Textblob utiliza três parâmetros para realizar a análise dos sentimentos: polaridade, subjetividade e intensidade (LORIA, 2020). A polaridade, definida entre os valores [-1,1] está relacionada ao grau do sentimento (-1 como negativo e +1 como positivo). A subjetividade quantifica a informação pessoal contida no texto, sendo que, um texto com alto nível de subjetividade contém informação pessoal em vez de um fato. A intensidade determina, a partir de palavras chave como: “very good”, “too boring” etc. A partir destes parâmetros, a ferramenta calcula uma escala de satisfação ou insatisfação no texto.

2.0.4 Predição de Séries Temporais

Uma série temporal é qualquer conjunto de dados extraídos de forma ordenada no tempo. Exemplos de séries temporais podem ser: valores diários de temperatura em uma determinada cidade, índices de cotação de ações no mercado financeiro, índices de poluição entre outros. Conhecer o comportamento da distribuição ao longo do tempo se torna importante para que a predição destas amostras em tempo futuro possa ser realizada (MORETTIN; TOLOI, 2018). A seguir são apresentados os modelos para predição de séries temporais explorados neste trabalho.

2.0.4.1 Modelos de Predição

Abaixo serão descritos algumas técnicas de avaliação usadas para o trabalho:

- ***Linear Regression:*** Essa forma de análise estima os coeficientes da equação linear, envolvendo uma ou mais variáveis independentes que melhor preveem o valor da variável dependente. A regressão linear se ajusta a uma linha reta ou superficial que minimiza as discrepâncias entre os valores de saída previstos e reais (IBM, 2021).
- **ARIMA:** é baseada no modelo ARMA que combina os modelos de auto regressivo(AR), integrados(I) e média móvel(MA) que converte dados não estacionários em dados estacionários em seus modelos (MONDAL; SAPTARSI, 2014).
- **KNN (IBk):** O k-Nearest-Neighbours (KNN) é um método de classificação não paramétrica, é simples porém eficaz. Para que um registro de dado seja classificado é asso-

ciado aos valores(k) mais próximos fazendo assim uma vizinhança (GUO GONGDE, 2003).

- **MLP:** As redes neurais MLP consistem em unidades dispostas em camadas. Cada camada é composta por nós e nas redes totalmente conectadas consideradas neste artigo, cada nó se conecta a cada nó nas camadas subsequentes. A camada de entrada distribui as entradas para as camadas subsequentes (OLIVEIRA, 2007).
- **HoltWinters:** O procedimento de previsão de Holt-Winters é um método de projeção simples e amplamente utilizado que pode lidar com tendências e variações sazonais. No entanto, estudos empíricos tendem a mostrar que o método não é muito preciso, em média, quanto os métodos mais complicados (CHATFIELD, 1978).

2.0.4.2 Modelos de Tendência

Modelos de tendência é um estilo de padrão oculto na variação de um dado, existem vários tipos de modelos que podem ser aplicados em diversos campos os usados neste trabalho são:

- **Sazonal:** Uma série temporal é sazonal (periódica) quando os fenômenos se repetem a cada período idêntico de tempo – por exemplo, fenômenos que ocorrem diariamente em uma certa hora, todos os dias, ou em um certo mês em todos os anos.
- **Estacionário:** são flutuações nos valores da variável com duração inferior a um ano, e que se repetem todos os anos, geralmente em função das estações do ano (ou em função de feriados ou festas populares, ou por exigências legais, como o período para entrega da declaração de Imposto de Renda); se os dados forem registrados anualmente NÃO haverá influência da sazonalidade na série.

2.0.5 Predição de Valor de Criptomoedas por meio de Aprendizagem de Máquina

A predição de cotações de mercado é considerada uma tarefa de alta complexidade, muito devido ao caráter especulativo que as mesmas possuem. Por conta disso, muitos trabalhos têm proposto soluções aplicadas ao mercado de ações, cotações de moedas nacionais e estrangeiras e, mais recentemente, de criptomoedas (BEKIROS; GUPTA; KYEI, 2016; LAHMIRI; BEKIROS, 2018).

Graças a um movimento crescente na indústria em conduzir análises e fazer previsões com base em dados de mídia social. Com o acúmulo de dados e desenvolvimento de novas ferramentas para analisar e conectar grandes conjuntos de dados, técnicas de Big Data, mineração de dados, análise preditiva e aprendizado de máquina, estão sendo utilizadas na tentativa

de entender a relação entre o comportamento humano e as tendências do mercado financeiro (KARPPI; CRAWFORD, 2016).

Nas últimas décadas, estratégias baseadas em aprendizagem de máquina têm se mostrado muito eficientes para tarefas de predição de séries temporais, que caracterizam cotações de mercado. Tal tendência se aplica também ao mercado de criptomoedas, das quais podem-se destacar técnicas de regressão linear e logística, redes neurais artificiais, máquinas de vetores de suporte, entre outras, para predição de Bitcoin e Ethereum (GREAVES; AU, 2015; CHEN; NARWAL, 2017; URAS *et al.*, 2020). A Tabela 2 apresenta um comparativo entre propostas, apresentando um comparativo entre alguns trabalhos, apresentando as técnicas de Aprendizagem de Máquina bem como as criptomoedas utilizadas.

Tabela 2 – Comparativo de trabalhos que propuseram a predição de valor de criptomoedas

Técnica de Aprendizagem de Máquina	**Criptomoeda**	**Referência**
Árvore binária autoregressiva	Bitcoin, ripple, and ethereum	(DERBENTSEV *et al.*, 2001)
Regressão linear, regressão logística, ANN, SVM	Bitcoin	(GREAVES; AU, 2015)
Modelo de árvore de regressão de aumento de gradiente extremo (XGBoost)	ZClassic, ZCash, and bitcoin private	(LI *et al.*, 2019)
Regressão linear, regressão polinomial, RNN e análise baseada em LSTM	Bitcoin	(MITTAL *et al.*, 2019)
Abordagem de séries temporais estruturais bayesianas	Bitcoin	(POYSER, 2019)
Análise de séries temporais usando LSTM bidirecional	Bitcoin and others	(MOHANTY *et al.*, 2018)
Regressão logística	Bitcoin, ripple, ethereum, litecoin, nem, dash, and stellar	(BOURI; SHAHZAD; ROUBAUD, 2019)
Regressão linear, regressão linear múltipla, perceptron multicamada NN e memória de curto prazo longo NN	Bitcoin	(URAS *et al.*, 2020)

Fonte: Khedr (2021).

3 METODOLOGIA

O presente trabalho apresenta duas contribuições básicas: a primeira consiste na construção de uma metodologia para a análise de sentimentos de uma base e associação temporal com o valor da criptomoeda; a segunda consiste no desenvolvimento de modelos de predição baseados em aprendizagem de máquina para estimar tendências de valorização e desvalorização de criptomoedas.

3.0.1 Criação da base de dados

Para criação da base de dados utilizada neste trabalho, foram primeiramente extraídos dados provenientes de três fontes: o histórico de tweets; o processamento destes tweets de modo a obter a análise de sentimentos atrelada a cada um deles; o histórico de cotações do Bitcoin em período compatível aos tweets coletados. Estes processos são descritos a seguir.

3.0.1.1 Base de Dados de Tweets

Para a base dos Tweets foi usado um repositório de nome "Bitcoin tweets - 16M tweets" disponível no Kaggle que coletou os tweets com o auxílio da api Tweepy e Twint que possibilitou a obtenção de dados importantes para o desenvolvimento do trabalho como: usuário que realizou a postagem; quantidade de likes[1], quantidade de compartilhamentos realizado a partir de uma postagem (retweets).

Para composição desta base, os autores filtraram a coleta para salvar os dados referentes à tweets que apresentam os seguintes termos "Bitcoin" ou "BTC" da língua inglesa contendo tweets de 2009/01/11 até 2019/11/23 às 15 h com dados regulares após 2016/01/01 totalizando 18,8 milhões de tweets e totalizando 3,90 GB de dados.

A Tabela 3 apresenta os atributos retornados pela base de dados coletados. Nela, são detalhados os campos relacionados a uma publicação no Twitter, dentre eles se destacam a data de criação (timestamp), o conteúdo do "tweet" (text), qual o usuário que realizou a postagem (user), a quantidade de compartilhamento que o tweet teve (retweets) e quantidade de curtidas (likes). Tais atributos são fundamentais para, além da análise do sentimento do texto postado, analisar o impacto que uma determinada postagem tem na rede social.

3.0.1.2 Análise de sentimentos Tweets

Para a análise de sentimento foram usadas as bibliotecas Textblob(LORIA, 2020) e VADER (HUTTO; GILBERT, 2014), escritas na linguagem Python. Com auxílio destas, após reali-

[1] Expressão de que o comentário agradou algum usuário.

Tabela 3 – Tabela de descrição do retorno da base de dados do kaggle

Campos da base de dados	Descrição
tweet-id	Identificador único do usuário
fullname	Identificador único em string do usuário
text	Tweet Postado
user	Nome do usuário que realizou a postagem
url	Link direto para a postagem no twitter.com/
retweets	Total de retweets que o tweet teve
likes	Total de favotivo que o tweet teve

Fonte: Autoria própria.

zada a coleta dos tweets, foram extraídos dados de classificação quanto ao sentimento associado a cada postagem. Como citado anteriormente, tais bibliotecas definem uma pontuação para cada texto postado. Para o presente trabalho, serão utilizados textos postados somente em inglês. Na pontuação do sentimento foram usados a variação de 1 a -1 para a intensidade do sentimento do tweet, sendo o -1 o mais negativo e 1 o mais positivo, tanto para o Textblob quanto para o VADER, conforme apresentado na Tabela 4.

Apesar de apresentarem funcionalidades semelhantes para análise de sentimentos, é possível notar que em todos os exemplos apresentados na Tabela 4 os dados tendem a ser diferentes sendo a análise realizada com o VADER diferente consideravelmente ao do TextBlob. A Figura 4 apresenta de forma agrupada a classificação média dos tweets em um mesmo dia da base de dados. O valor médio é calculado com base na quantidade de tweets realizados no período de cada amostra, neste caso de hora em hora.

A partir da figura é possível analisar que, segundo os métodos testados para análise de sentimento, em média os tweets tendem a ser mais positivos que negativos, visto que a maior parte das classificações está acima de zero no eixo y. Em algumas situações específicas é possível notar que os comentários podem ser altamente favoráveis.

3.0.1.3 Extração dos dados

Para a extração de dados de cotação foi usada a api yfinance[2] para a linguagem Python, que coleta dados históricos de vários tipos de ativos mercadológicos, tais como, ações da bolsa de valores, moedas nacionais e estrangeiras e criptomoedas. A partir do seu uso, é possível selecionar o período de coleta, o intervalo de tempo e o ativo desejado. O retorno se dá em uma lista de cotações, que representa uma série temporal, na qual é possível extrair tendências de valorização, desvalorização ou estabilidade a partir dos dados.

A Figura 5 apresenta a série de cotações entre os dias 01 de janeiro de 2017 e 23 de novembro de 2019, período compatível com os dados da coleta de tweets. Estes dados dizem respeito à cotação de abertura (open) do Bitcoin

[2] yfinance - API de coleta de dados de cotações de ativos. https://pypi.org/project/yfinance/.

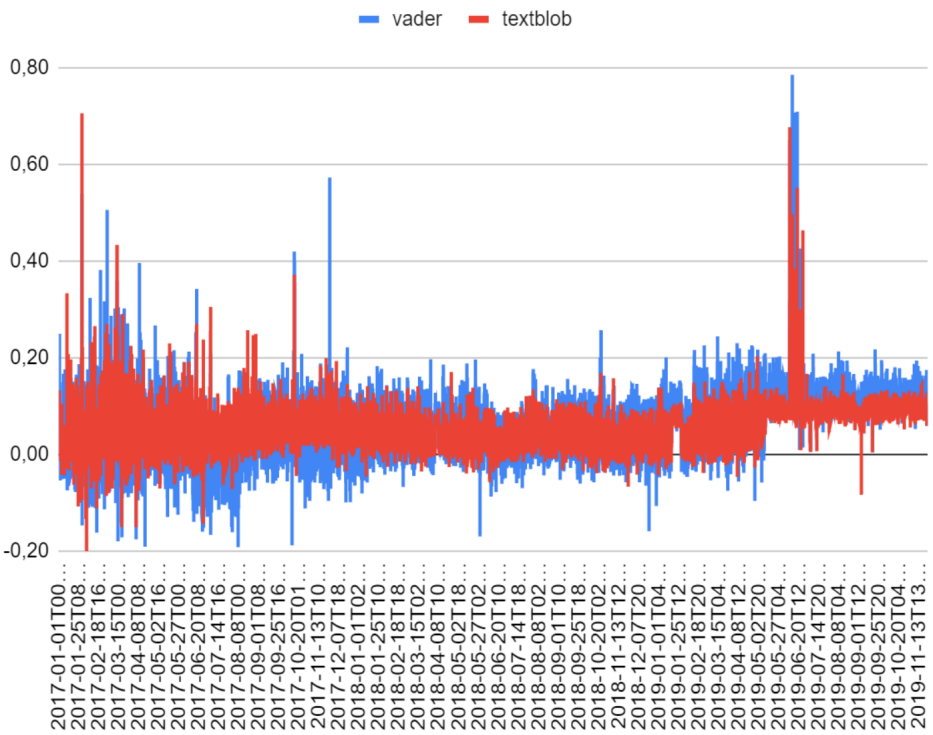
Tabela 4 – Tabela de tweets exemplo extraídos da base do kaggle classificado por meio da biblioteca vade e Textblob

Data	Tweet	Vader	Textblob	Likes	Retweets
2019-05-15 20:33:05	-=[576.214]=- Txs: 3.048 Size: 1.24 MB Stripped: 0.86 MB Time: 1557959525 Reward: 12.5 BTC Fees: 1.65802485 BTC Miner: AntPool Mempool: 55.367 txs	0,785	0	0	0
2019-05-15 20:33:06	Visual FoxPro 9.0 License for 25 PCs https://t.co/401EWUc1DB #BloggersBlast #bloggerstribe #digitalgoods #rocketrnet #bitcoin #windows #software	0	0	0	0
2019-05-15 20:33:06	Daily profit for HODLING BTC since 2013 Data taken since 29/10/2018 @ 10:45am (UTC) Updated every hour #bitcoin #btc #cryptocurrency #blockchain #cryptotrade #profit #digitalmine https://t.co/DWJZWigJMM	0,440	0	0	0
2019-05-15 20:33:08	RT Ultrascan419 "Exclusive: Scammed Porn Watchers Have Paid Nearly $1 Million in Bitcoin Blackmail: Unfortunately, thousands of others have fallen prey to the same email scam, which instructs the victims to send Bitcoin or else see intimate photos from	-0,872	-0,05	0	0
2019-11-01 11:56:55	it’s official. this meme is DEAD! nice try, @Snowden #Bitcoin #StoreOfDrama	-0,546	0,175	0	0
2019-11-01 12:00:17	#bitcoin Earn FREE bitcoin with this easy strategy https://t.co/K41USsQ7H8 https://t.co/fNc3pgYYzT	0,869	0,417	0	0

Fonte: Autoria própria.

A partir da análise da Figura 5 é possível verificar qual volátil é o mercado do Bitcoin, visto que no início da série apresentada, a cotação média era em torno de US$ 1.000,00 e um ano após chegava a um pico de quase US$ 20.000,00, o que corresponde a valorização de cerca de 2.000%. Da mesma forma, cerca de 1 ano e meio após o pico da série a moeda apresentou desvalorização para patamares abaixo de US$ 5.000,00. Tal volatilidade torna a predição deste

Figura 4 – Gráfico de média móvel da análise de sentimento



Fonte: Autoria própria.

tipo de moeda extremamente difícil e melhorias, mesmo que pequenas, em modelos de predição, podem impactar em ganhos maiores ou perdas menores nestas transações.

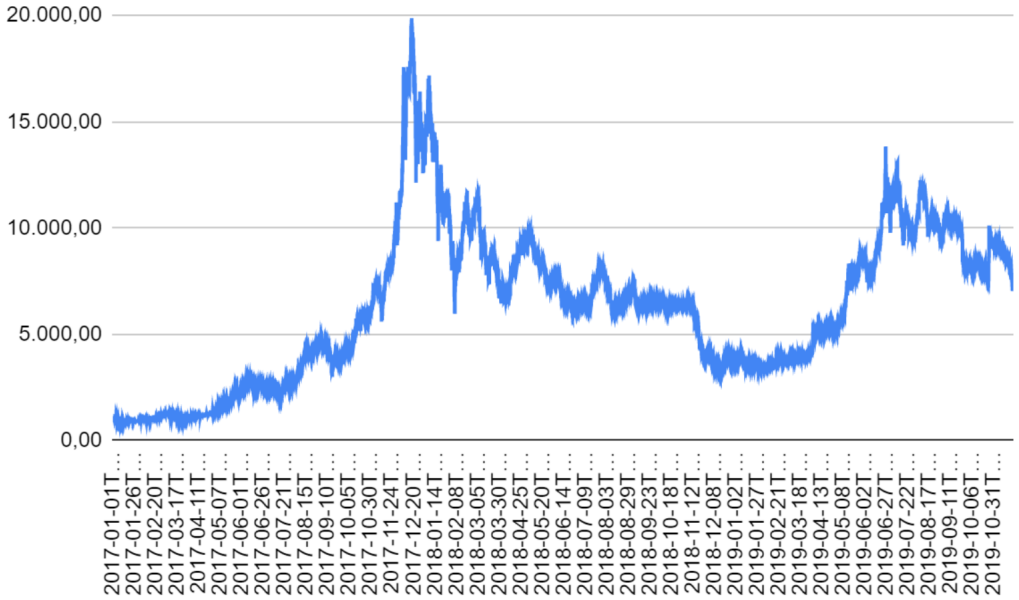
3.0.1.4 Construção da Base de Dados

Para a construção da base de dados usada foi executada em 3 partes sendo a primeira a parte da coleta e processamento dos tweets, a segunda a coleta da cotação da criptomoeda e a terceira a junção destas duas bases através da hora, conforme apresentado na Figura 6.

A partir da coleta da base de dados foram removidos os campos que não foram usados durante o trabalho, sendo deixados apenas os campos: timestamp, text, likes e retweets. Com o campo “text” foram aplicadas técnicas de análise de sentimentos das postagens e salvas em dois modelos sendo o “vader” e “textblob”, após a análise do sentimento o campo “text” foi removido. A seguir é feita uma ordenação para evitar que algum dado fique agrupado ou associado indevidamente com outros, Após a ordenação é realizado a concatenação destes dados na variação de uma hora e assim adicionando uma variável de contagem de quantos itens teve naquele período de 1 hora e assim finalizado a primeira etapa.

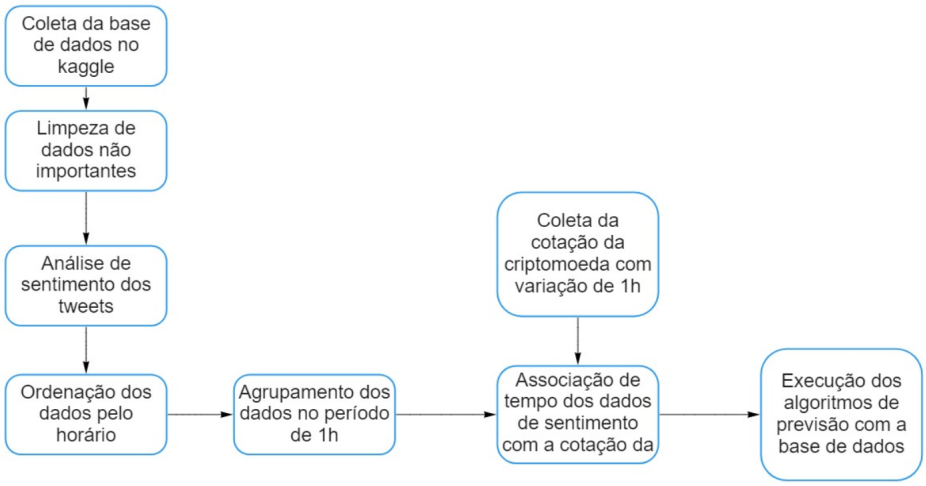
Para a segunda etapa foram coletados os dados da criptomoeda com os filtros de tempo e variação. A terceira etapa ela associa a coleta da cotação da criptomoeda e das médias dos tweets e salva os dados em seu modelo final que será utilizado para treinamento e testes dos modelos de predição.

Figura 5 – Gráfico do valor de entrada "open" da cotação em US$ de 2017/01/01 até 2019/11/23 com variação de 1h



Fonte: Autoria própria.

Figura 6 – Ilustração de modelo de criação da base de dados e a predição do valor da criptomoeda



Fonte: Autoria própria.

3.0.2 Análise de Predição de Cotação

Para a análise foi usado o software de aprendizagem de máquina WEKA(WEKA, 2021), com o módulo Forecast, para os modelos regressão linear, KNN, MLP, HoltWinters. E para o ARIMA o software IBM SPSS STATISTICS(IBM, 2021), onde ambos oferecem diversas opções de técnicas para a predição de séries temporais.

Visando verificar o real benefício de se utilizar dados provenientes da análise de sentimentos em detrimento de somente a cotação, para predizer a cotação do Bitcoin, foram utilizados diversos cenários de treinamento. A partir da análise dos dados destes cenários é possível

comparar a real influência de postagens na cotação de criptomoedas. Os seguintes cenários foram utilizados:

- **Open:** Cotação de abertura;
- **Open + VADER:** Cotação de abertura combinada com a análise de sentimentos realizada pelo VADER;
- **Open + Textblob:** Cotação de abertura combinada com a análise de sentimentos realizada pelo TextBlob;
- **Open + VADER + Textblob:** Cotação de abertura combinada com ambas técnicas de análise de sentimentos;
- **Open + VADER + Textblob + likes + tweets + quantidade:** Cotação de abertura, combinada com as análises de sentimentos e com o engajamento das postagens, sejam pela quantidade postagens, de likes e retweets.

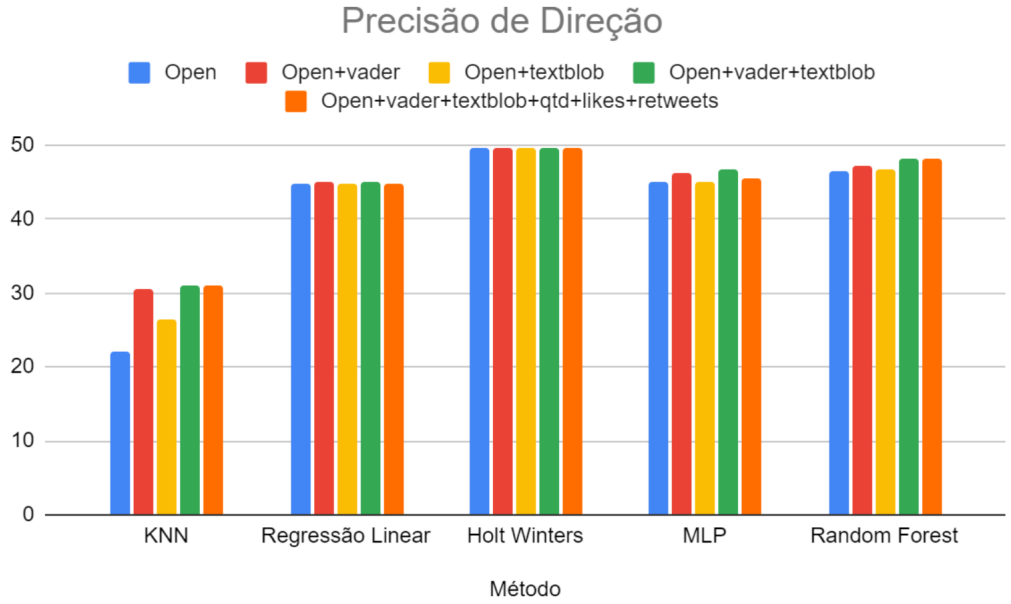
4 RESULTADOS

Neste capítulo são apresentados os principais resultados obtidos durante o desenvolvimento deste trabalho no qual são apresentados, nesta ordem, a precisão obtida dos modelos de predição usados, e o erro médio absoluto (*Mean Absolute Error*).

4.0.1 Predição dos modelos de predição

Ao testar os modelos de previsão no WEKA obtivemos resultados promissores que além de definir a diferença de deficiências dos modelos também pode se obter os resultados das diferentes associações (Figura 7.

Figura 7 – Gráfico de precisão da direção predita do valor



Fonte: Autoria própria.

Tabela 5 – Porcentagem de acertos dos modelos testados

Métodos	Open	Open+vader	Open+textblob	Open+vader+textblob	Open+vader+textblob+qtd+likes+retweets
KNN	22,04	30,48	26,34	30,91	31,03
Regressão Linear	44,87	44,91	44,83	44,91	44,87
HoltWinters	**49,64**	**49,64**	**49,64**	**49,64**	**49,64**
MLP	45,03	46,09	45,03	46,56	45,54
Random Forest	46,33	47,12	46,72	48,18	48,22
Média Geral	**41,58**	**43,65**	**42,51**	**44,04**	**43,86**

Fonte: Autoria própria.

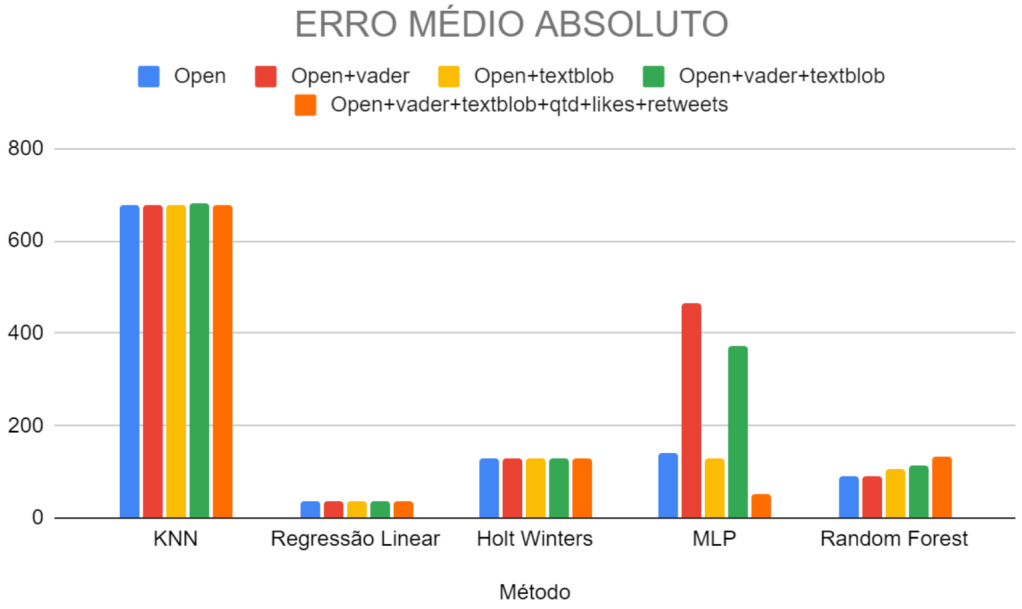
Com os resultados apresentados na Tabela 5 podemos identificar que a porcentagem média de acerto da direção da previsão inicialmente é de 41,5%, mas com a adição dos avaliadores de sentimentos do twitter obtivemos uma ligeira melhoria de 2,46% para a previsão dos valores quando utilizada cotação mais o uso do VADER.

O algoritmo KNN apresentou um crescimento de quase 8,99% com todos os valores em comparação com o primeiro método que continha apenas os dados de cotação, mas ainda sim com o crescimento ainda ficou abaixo dos outros métodos. O Random Forest apresentou um crescimento de 1,89% com todos os valores em comparação ao primeiro modelo. No que diz respeito a MLP, teve um crescimento de 1,53% no modelo onde é usado os valores da análise de sentimento do vader e textblob juntas. Para finalizar, os métodos de Regressão linear e HoltWinter não apresentaram melhorias significativas em suas precisões, embora o método HotWinter apresentou o melhor resultado independente dos dados adicionados.

4.0.2 *Mean Absolute Error*

O erro absoluto médio é uma medida de erros entre observações observadas, em geral todos algoritmos se mantiveram em uma variação aceitável de erros com exceção do MLP que apresentou uma variação notável em todos os resultados. Já o Regressão linear apresentou a menor taxa de erro entre todos os casos e assim como a precisão à medida que adicionamos as demais variáveis os erros tendem a diminuir (Figura 8.

Figura 8 – Gráfico da média de erros absolutos dos modelos testados



Fonte: Autoria própria.

Com os resultados apresentados na Tabela 6 podemos identificar que a média de erro médio inicialmente é de 263,32, mas com a adição dos avaliadores de sentimentos do twitter obtivemos uma ligeira melhoria de -9,11 para a previsão dos valores.

Tabela 6 – Média de erros absolutos dos modelos testados

Métodos	Open	Open+vader	Open+textblob	Open+vader+textblob	Open+vader+textblob+qtd+likes+retweets
KNN	765,66	767,81	767,96	768,42	768,38
Regressão Linear	**63,34**	**63,05**	**63,00**	**63,05**	**63,05**
Holt Winters	197,35	197,35	197,35	197,35	197,35
MLP	163,77	557,72	151,04	465,31	80,29
Random Forest	126,50	127,46	141,56	148,66	162,00
Média Geral	**263,32**	**342,68**	**264,18**	**328,56**	**254,21**

Fonte: Autoria própria.

O método MLP apresentou tanta melhoria no resultado quanto piora, o melhor resultado foi com todos os dados da base de dados e o pior resultado foi com apenas a cotação + vader, tendo -83,47 de melhoria em relação ao primeiro modelo e o +393,94 com pior caso.

Os demais métodos não apresentaram variações significativas e o método de Regressão Linear apresentou o melhor resultado independente dos dados adicionados.

4.0.3 Modelo ARIMA

Como o modelo ARIMA é amplamente usado para predição de valores, foi realizados alguns testes no software estatístico IBM SPSS ESTATISTIC pois o software WEKA não apresenta este modelo em sua coletânea, com isso não podemos comparar diretamente os resultados dos testes no WEKA pois o processamento ocorre de forma diferente (Figura 9.

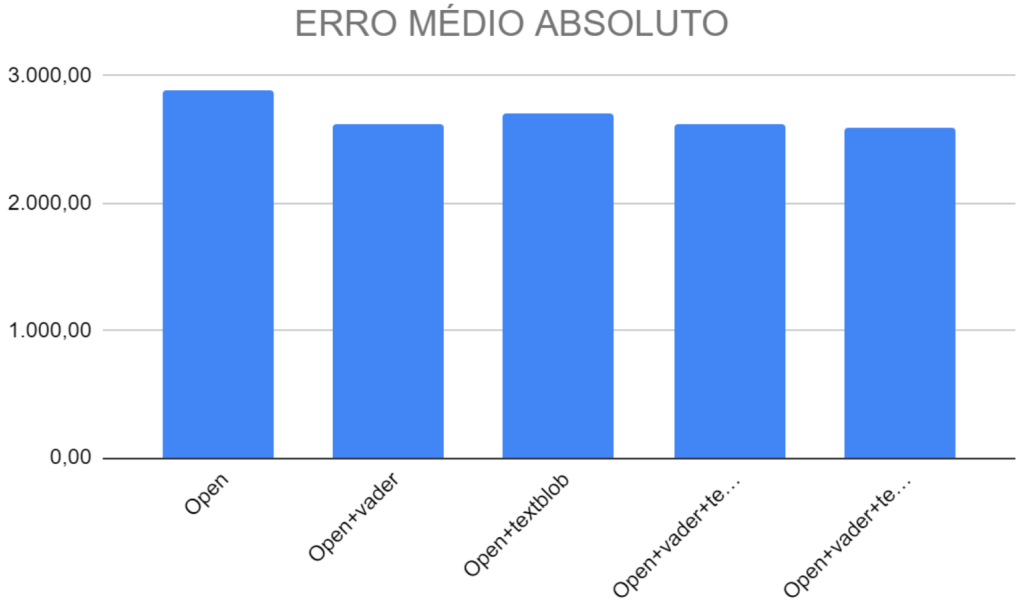
Com os resultados apresentados na Tabela 7 encontrou-se uma queda nos resultados de erro médio de 287,94 do modelo com o todos os dados em comparação ao primeiro modelo que continha apenas os dados de cotação. Entretanto, o IBM SPSS STATISTICS não apresenta um dado específico de precisão da previsão, por esse motivo iremos focar apenas nos casos de erro médio absoluto.

Tabela 7 – Resultados médios de erros absolutos dos modelos testados com ARIMA

Métodos	Open	Open+vader	Open+textblob	Open+vader+textblob	Open+vader+textblob+qtd+likes+retweets
ARIMA	**2880,31**	**2624,08**	**2711,62**	**2617,53**	**2592,37**

Fonte: Autoria própria.

Figura 9 – Gráfico de média de erros absolutos dos modelos testados com ARIMA



Fonte: Autoria própria.

5 CONCLUSÃO

Analisando os resultados obtidos até aqui, o desenvolvimento completo de uma aplicação, com todas as funcionalidades e testes em produção ainda não foram realizados, mas já é o suficiente para saber que é possível desenvolver e utilizar as tecnologias alavancadas para o projeto. O início da prototipagem do design, definindo o fluxo principal da aplicação e definindo alguns requisitos iniciais foram essenciais para o desenvolvimento do aplicativo.

Também foram realizadas pesquisas quanto ao Flutter para entender as suas capacidades quanto a utilização do GPS para a localização em tempo real, também foi preciso analisar a viabilidade do uso da API de Geolocalização, pois essa versão inicial não terá verbas financeiras para o desenvolvimento.

A definição da Framework Flutter para o desenvolvimento da aplicação multiplataforma e a arquitetura, foram de extrema importância, pois permite a escalabilidade da aplicação caso ela venha a ser disponibilizada para mais usuários, visando a evolução do projeto para outros municípios e regiões.

A escolha para armazenamento de dados foi o Firebase, pois sua velocidade é muito importante na aplicação e pôde ser utilizado para servir tanto como um sistema de *login* quanto para o armazenamento das requisições dos processos de entrega.

O projeto também visou atingir o mercado local das cidades onde forem escolhidas para funcionarem, pois, necessita de uma pesquisa para conhecer o mercado local, saber se eles utilizam o transporte para as cidades vizinhas com uma certa frequência, e também por isso foi estimado uma distância máxima de 80 (oitenta) quilômetros de raio para realizar a entrega.

Por fim, conclui-se que o projeto pode ser totalmente desenvolvido em Flutter e não existem limitações atualmente para que uma aplicação não possa ser desenvolvida para produção com essas ferramentas escolhidas.

REFERÊNCIAS

ABRAHAM, J. e. a. Cryptocurrency price prediction using tweet volumes and sentiment analysis. **SMU Data Science Review**, v. 1, n. 3, p. 1–22, 2018.

BEKIROS, S.; GUPTA, R.; KYEI, C. A non-linear approach for predicting stock returns and volatility with the use of investor sentiment indices. **Applied Economics**, v. 48, n. 31, p. 2895–2898, 2016.

BOURI, E.; SHAHZAD, S. J. H.; ROUBAUD, D. Co-explosivity in the cryptocurrency market. **Finance Research Letters**, v. 29, p. 178–183, 2019.

CHATFIELD, C. The holt-winters forecasting procedure. **Journal of the Royal Statistical Society: Series C**, v. 27, n. 3, p. 264–279, 1978.

CHEN, M.; NARWAL, N. Predicting price changes in ethereum. *In*: . [*s.n.*], 2017. Disponível em: https://api.semanticscholar.org/CorpusID:40823912.

COINMARKETCAP. **Today's Cryptocurrency Prices by Market Cap**. 2021. Disponível em: https://coinmarketcap.com. Acesso em: 25 jun. 2021.

DERBENTSEV, V. *et al.* Forecasting cryptocurrency prices time series using machine learning approach. *In*: **SHS Web of Conferences**. [*S.l.*: *s.n.*], 2001. p. 1–7.

FERREIRA, M. *et al.* Blockchain: A tale of two applications. **Applied Sciences**, v. 8, n. 9, p. 1–24, 2019.

GREAVES, A.; AU, B. Using the bitcoin transaction graph to predict the price of bitcoin. *In*: . [*s.n.*], 2015. Disponível em: https://api.semanticscholar.org/CorpusID:18038866.

GUO GONGDE, e. a. Knn model-based approach in classification. *In*: **OTM Confederated International Conferences "On the Move to Meaningful Internet Systems**. [*S.l.*: *s.n.*], 2003. p. 986–996.

HUTTO, C.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *In*: **8th International AAAI conference on weblogs and social media (IDWSM)**. [*S.l.*: *s.n.*], 2014. p. 216–225.

IBM. **IBM SPSS Statistics(2022)**. 2021. Disponível em: https://www.ibm.com/products/spss-statistics. Acesso em: 22 set. 2021.

KARPPI, T.; CRAWFORD, K. Social media, financial algorithms and the hack crash. **Theory, Culture & Society**, v. 33, n. 1, p. 73–92, 2016.

KHEDR, A. M. e. a. Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. **Intelligent Systems in Accounting, Finance and Management**, v. 28, n. 1, p. 3–34, 2021.

LAHMIRI, S.; BEKIROS, S. Chaos, randomness and multi-fractality in bitcoin market. **Chaos, Solitons Fractals**, v. 106, p. 28–34, 2018. ISSN 0960-0779.

LAZO, J. G. L. *et al.* Sistema híbrido para tomada de decisão em investimentos no mercado de criptomoedas. **Brazilian Journal of Development**, v. 7, n. 2, p. 19577–19593, 2021.

LI, L. *et al.* Bitcoin options pricing using lstm-based prediction model and blockchain statistics. *In*: **IEEE international conference on blockchain**. [*S.l.*: *s.n.*], 2019. p. 67–74.

LORIA, S. **textblob Documentation Release 0.16.0**. 2020. Disponível em: hhttps://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf. Acesso em: 22 set. 2020.

MITTAL, A. *et al.* Short-term bitcoin price fluctuation prediction using social media and web search data. *In*: **Twelfth International Conference on Contemporary Computing (IC3)**. [*S.l.*: *s.n.*], 2019. p. 1–6.

MOHANTY, P. *et al.* Predicting fluctuations in cryptocurrencies' price using users' comments and real-time prices. *In*: **7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)**. [*S.l.*: *s.n.*], 2018. p. 77–482.

MONDAL, P.; SAPTARSI, G. Study of effectiveness of time series modeling (arima) in forecasting stock prices. **International Journal of Computer Science, Engineering and Applications**, v. 4, n. 2, p. 13–29, 2014.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de séries temporais: modelos lineares univariados**. Rio de janeiro: Editora Blucher, 2018. 474 p. ISBN 978-8521213512.

MUSK, E. **Tesla Bitcoin**. 2021. Disponível em: https://twitter.com/elonmusk/status/1392602041025843203. Acesso em: 25 jun. 2021.

NAKAMOTO, S. **Bitcoin: A peer-to-peer electronic cash system**. 2008. Disponível em: https://bitcoin.org/bitcoin.pdf. Acesso em: 21 set. 2020.

OLIVEIRA, P. C. de. **Séries Temporais: Analisar o Passado, Predizer o Futuro**. 2007. Disponível em: https://student.dei.uc.pt/~pcoliv/reports/ct_timeseries.pdf. Acesso em: 21 set. 2020.

POYSER, O. Fxploring the dynamics of bitcoin's price: A bayesian structural time series approach. **Eurasian Economic Review**, v. 9, n. 1, p. 29–60, 2019.

URAS, N. *et al.* Forecasting bitcoin closing price series using linear regression and neural networks models. **Peer J Computer Science**, v. 6, n. e279, p. 1–25, 2020.

WEKA. **Weka 3: Machine Learning Software in Java**. 2021. Disponível em: https://www.cs.waikato.ac.nz/ml/weka/. Acesso em: 22 set. 2021.