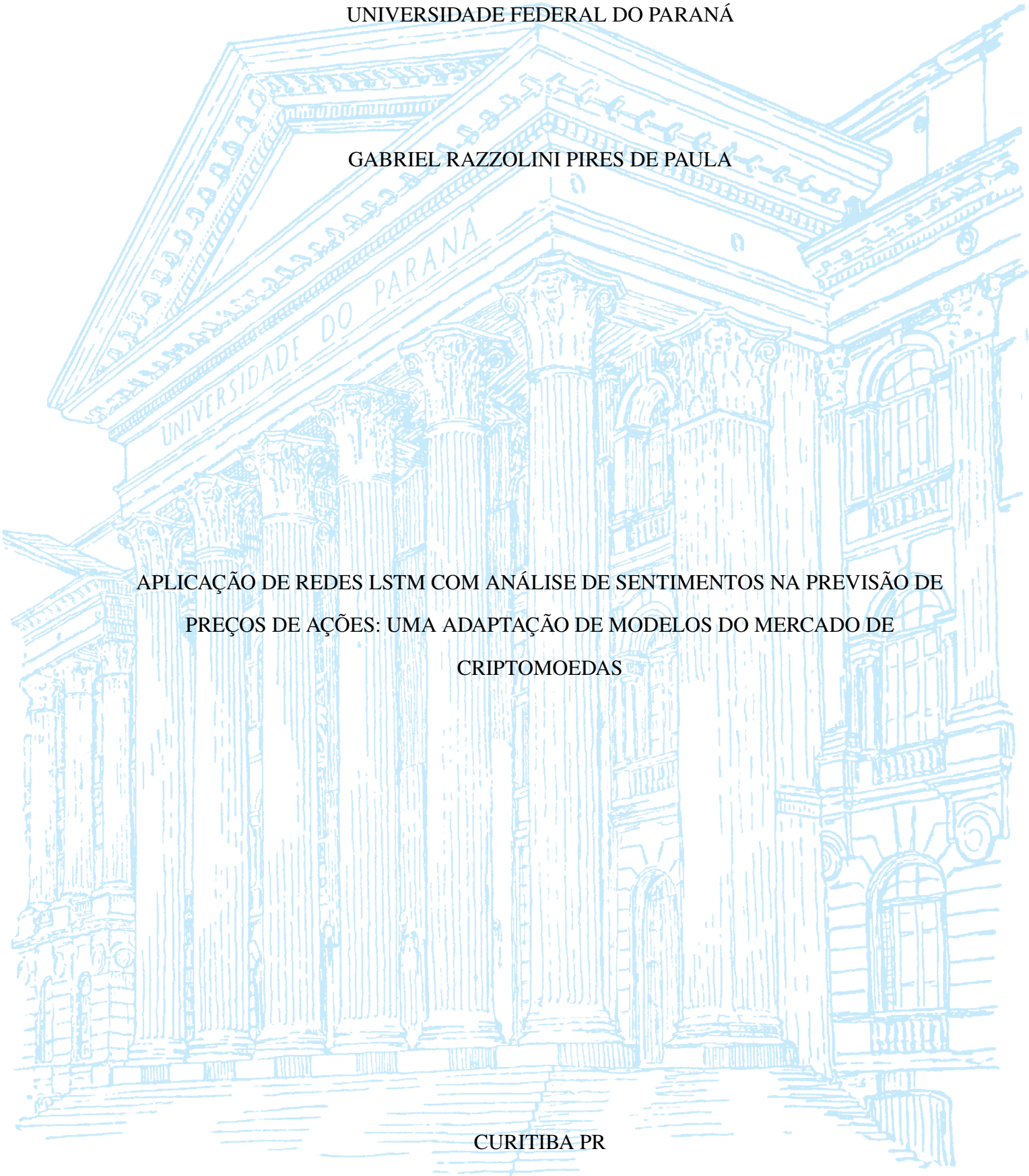
UNIVERSIDADE FEDERAL DO PARANÁ

GABRIEL RAZZOLINI PIRES DE PAULA



APLICAÇÃO DE REDES LSTM COM ANÁLISE DE SENTIMENTOS NA PREVISÃO DE PREÇOS DE AÇÕES: UMA ADAPTAÇÃO DE MODELOS DO MERCADO DE CRIPTOMOEDAS

CURITIBA PR

2025

GABRIEL RAZZOLINI PIRES DE PAULA

APLICAÇÃO DE REDES LSTM COM ANÁLISE DE SENTIMENTOS NA PREVISÃO DE PREÇOS DE AÇÕES: UMA ADAPTAÇÃO DE MODELOS DO MERCADO DE CRIPTOMOEDAS

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Computação*.

Orientador: Fabiano Silva.

CURITIBA PR

2025

Ao Mundo Bem Melhor...

AGRADECIMENTOS

Expresso minha mais profunda gratidão ao meu mestre, Dr. Celso Charuri. Um Grande Homem, seus ensinamentos transcendem o tempo e o espaço, conduzindo não apenas ao conhecimento do Todo, mas à Consciência Cósmica. Foi por meio de suas palavras e de seus ensimantos que encontrei Coragem para seguir com propósito, mesmo diante das incertezas. Sua proposição de um Mundo Bem Melhor permanece como uma luz orientadora em minha jornada.

Aos meus amigos, que são mais que companheiros de afinidade — são irmãos pelo ponto de fidelidade que nos une. Agradeço a cada um pelo exemplo de integridade, pela coragem demonstrada nos momentos desafiadores e pela presença firme ao longo da caminhada. Suas atitudes silenciosas muitas vezes disseram mais do que palavras, e o apoio mútuo que cultivamos foi essencial para que eu me mantivesse em pé, fiel aos meus princípios e ao meu propósito.

À minha mãe e ao meu padrasto, agradeço com o coração cheio de amor e respeito. Foram vocês que, com sensibilidade e firmeza, me ofereceram suporte emocional em todas as ocasiões — nas conquistas e nas dificuldades, nos dias de dúvida e nos momentos de realização. Sem esse apoio, este caminho teria sido muito mais árduo.

Ao meu tio, meu sincero agradecimento por todo o suporte técnico que me foi dado ao longo do curso. Sua disposição em compartilhar conhecimento, esclarecer dúvidas e contribuir ativamente com minha formação foi essencial para que eu pudesse avançar com confiança e profundidade.

A todos vocês, minha eterna Gratidão.

RESUMO

Este trabalho apresenta uma adaptação e expansão da metodologia proposta por Prajapati (2020a), originalmente voltada à previsão de preços de criptomoedas, para o contexto do mercado de ações. Mantendo a mesma métrica de avaliação e a arquitetura baseada em redes neurais LSTM, o estudo redireciona o foco preditivo do Bitcoin para ações tradicionais, por meio de um estudo de caso envolvendo três ativos com características distintas quanto à presença midiática: Apple (AAPL), que possui um volume razoável de notícias, mas não é extremamente midiática; Tesla (TSLA), altamente exposta na mídia e com grande quantidade de notícias; e Electromed (ELMD), com presença praticamente inexistente em fontes noticiosas. Para viabilizar essa mudança de escopo, foram realizadas diversas adaptações técnicas, como a utilização da API do Yahoo Finance para a coleta de dados históricos e o desenvolvimento de um pipeline de análise de sentimentos mais generalista, capaz de processar notícias provenientes de diferentes fontes. Destacam-se, ainda, melhorias na robustez e modularidade dos scripts de coleta e análise em relação às ferramentas originais de Prajapati. A principal contribuição deste trabalho reside em demonstrar a viabilidade da aplicação de modelos preditivos baseados em sentimentos para além do universo das criptomoedas, ampliando sua aplicabilidade no mercado financeiro tradicional.

Palavras-chave: LSTM. Análise de Sentimento. Previsão do Mercado de Ações. Modelagem Preditiva

ABSTRACT

This work presents an adaptation and expansion of the methodology proposed by Prajapati (2020a), originally designed for cryptocurrency price prediction, into the context of the stock market. While maintaining the same evaluation metric and the LSTM-based neural network architecture, the study shifts the predictive focus from Bitcoin to traditional stocks through a case study involving three assets with distinct levels of media exposure: Apple (AAPL), which has a reasonable volume of news but is not highly media-driven; Tesla (TSLA), which receives a high volume of news and enjoys strong media visibility; and Electromed (ELMD), which has virtually no presence in news sources. To enable this change in scope, several technical adaptations were implemented, such as the use of the Yahoo Finance API for historical data collection and the development of a more generalized sentiment analysis pipeline capable of processing news from various sources. Notable improvements were also made in the robustness and modularity of the data collection and analysis scripts when compared to Prajapati's original tools. The main contribution of this work lies in demonstrating the feasibility of applying sentiment-based predictive models beyond the cryptocurrency domain, thereby broadening their applicability within traditional financial markets.

Keywords: LSTM. Sentiment Analysis. Stock Market Prediction. Predictive Modeling.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS

RMSE	Root Mean Square Error
MAE	Mean Absolute Error
MSE	Mean Squared Error
R^2	Coefficient of Determination
MCC	Matthews Correlation Coefficient
VaR	Value at Risk
HAE	Heteroskedasticity Adjusted Error
HSE	Heteroskedasticity Scaled Error
RSI	Relative Strength Index
MACD	Moving Average Convergence Divergence
MOM	Momentum
MR	Mean Reversion
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
1D-CNN	One-Dimensional Convolutional Neural Network
DNN	Deep Neural Network
ANN	Artificial Neural Network
MLP	Multi-Layer Perceptron
RF	Random Forest
SVM	Support Vector Machine
SVR	Support Vector Regression
KNN	K-Nearest Neighbors
MLR	Multiple Linear Regression
RBF	Radial Basis Function
AR	AutoRegressive
ARIMA	AutoRegressive Integrated Moving Average
ARCH	AutoRegressive Conditional Heteroskedasticity
GARCH	Generalized AutoRegressive Conditional Heteroskedasticity
gjrGARCH	Glosten-Jagannathan-Runkle Generalized Autoregressive Conditional Heteroskedasticity

eGARCH	Exponential Generalized Autoregressive Conditional Heteroskedasticity
GBDT	Gradient Boosting Decision Tree
DRL	Deep Reinforcement Learning
VAE	Variational Autoencoder
GNN	Graph Neural Network
GAT	Graph Attention Network
TCN	Temporal Convolutional Network
TFT	Temporal Fusion Transformer
LLM	Large Language Model
GA	Genetic Algorithm
PSO	Particle Swarm Optimization
ACO	Ant Colony Optimization
DE	Differential Evolution
AutoML	Automated Machine Learning
PCA	Principal Component Analysis
NLP	Natural Language Processing
USE	Universal Sentence Encoder
SMI	Social Media Information
CSV	Comma-Separated Values
VADER	Valence Aware Dictionary and sEntiment Reasoner
API	Application Programming Interface
cuDNN	CUDA Deep Neural Network
RMSProp	Root Mean Square Propagation
GPU	Graphics Processing Unit
HTML	HyperText Markup Language
RSS	Really Simple Syndication
JS	JavaScript
UTC	Coordinated Universal Time
PST	Pacific Standard Time
AAPL	Apple Inc. (código de ação)
TSLA	Tesla Inc. (código de ação)

ELMD	Electromed Inc. (código de ação)
BTCUSDT	Bitcoin to Tether (par de negociação na Binance)
ETHUSDT	Ethereum to Tether (par de negociação na Binance)

DINF	Departamento de Informática
UFPR	Universidade Federal do Paraná
NYSE	New York Stock Exchange
NASDAQ	National Association of Securities Dealers Automated Quotations
S&P500	Standard & Poor's 500 - Bolsa de Valores Americana
B3	Brasil, Bolsa, Balcão - Bolsa de Valores Brasileira
ACL	Association for Computational Linguistics
KDD	Knowledge Discovery and Data Mining

SUMÁRIO

1 INTRODUÇÃO

As séries temporais são conjuntos de observações coletadas sequencialmente ao longo do tempo, geralmente com intervalos uniformes entre as medições, sendo amplamente utilizadas para identificar padrões, tendências e comportamentos futuros com base em dados históricos. Sua análise considera quatro componentes principais: a tendência, que revela mudanças de longo prazo; a sazonalidade, que aborda variações periódicas regulares; os ciclos, referentes a oscilações irregulares ligadas a fatores externos, como econômicos ou sociais; e os resíduos, que representam ruídos aleatórios não explicados diretamente pelos demais componentes.

Para prever séries temporais, destacam-se diversos métodos, como o modelo clássico *AutoRegressive Integrated Moving Average (ARIMA)*, amplamente utilizado devido à sua eficácia em séries lineares estacionárias. Segundo (Hyndman and Athanasopoulos, 2018), o *ARIMA* combina três componentes — *AutoRegressive (AR)*, *Moving Average (MA)* e *Integrated (I)* — para modelar a autocorrelação dos dados e torná-los estacionários, permitindo prever valores futuros com base em padrões passados. Além dos métodos estatísticos tradicionais, técnicas mais avançadas baseadas em *deep learning* têm ganhado destaque, especialmente as *Recurrent Neural Networks (RNNs)*. Dentre estas, destacam-se as *Long Short-Term Memory (LSTM)*, um tipo especializado de rede recorrente projetado para lidar com o problema do desvanecimento do gradiente e capturar dependências temporais de longo prazo. De acordo com (Goodfellow et al., 2016), a *LSTM* introduz *input*, *forget* e *output gates* para regular o fluxo de informações ao longo do tempo, tornando-a particularmente eficaz para sequências temporais extensas. Outra variação eficiente dessas redes é a *Gated Recurrent Unit (GRU)*, que simplifica a estrutura das *LSTM* ao fundir as *input* e *forget gates* em uma única *update gate*. Como descrito por (Chollet, 2021), as *GRU* oferecem desempenho comparável às *LSTM* com menor complexidade computacional, sendo especialmente úteis quando há restrições de tempo ou recursos. Ambas as arquiteturas, *LSTM* e *GRU*, são particularmente adequadas para aplicações financeiras como a previsão de preços de ações e criptomoedas, que frequentemente apresentam comportamento não linear e sofrem influência simultânea de múltiplos fatores externos.

O problema central abordado neste trabalho é a predição dos valores futuros das ações negociadas na bolsa de valores, um desafio complexo devido à natureza altamente volátil e não linear do mercado financeiro. A previsão precisa do comportamento dos preços das ações representa um importante diferencial competitivo, possibilitando a investidores e empresas tomadas de decisões mais eficazes e estratégicas. Nesse contexto, torna-se fundamental explorar e desenvolver métodos robustos capazes de modelar adequadamente as séries temporais financeiras, considerando tanto aspectos quantitativos dos dados históricos, quanto qualitativos, como influências externas advindas de notícias, eventos econômicos e sentimentos sociais que impactam diretamente o comportamento do mercado.

Embora a predição de valores de ações envolva técnicas comuns tanto à regressão quanto à previsão temporal *(forecasting)*, é importante salientar a distinção fundamental entre esses dois conceitos no contexto do problema tratado neste trabalho. A *regressão* caracteriza-se pela modelagem da relação entre variáveis independentes e dependentes, assumindo frequentemente que as observações sejam independentes e identicamente distribuídas, sem considerar explicitamente o aspecto temporal. Segundo (James et al., 2013), os métodos de regressão são amplamente utilizados para estimar e inferir relações estatísticas em dados estáticos, sendo particularmente eficazes em contextos onde a independência das observações é válida. Já a previsão temporal *(forecasting)* envolve explicitamente a dimensão temporal, concentrando-se

na identificação de padrões sequenciais e na dependência temporal entre observações, o que é essencial para prever eventos futuros com base no histórico passado. De acordo com (Hyndman and Athanasopoulos, 2018), o *forecasting* é o processo de construir modelos capazes de capturar e extrapolar padrões temporais de dados históricos com o objetivo de prever valores futuros, sendo uma prática fundamental em áreas como economia, finanças e meteorologia. Desta forma, o desafio abordado neste estudo encaixa-se predominantemente no escopo da previsão temporal *(forecasting)*, exigindo modelos que sejam capazes de aprender relações dinâmicas e que levem em consideração a evolução sequencial e temporal dos dados.

Para avaliar a qualidade e o desempenho dos modelos utilizados na previsão de valores de ações, empregam-se métricas estatísticas amplamente difundidas no contexto da previsão temporal *(forecasting)*, como o *Root Mean Square Error (RMSE)* e o *Mean Absolute Error (MAE)*. O *RMSE* mede a magnitude média dos erros ao atribuir um peso maior aos desvios significativos, sendo especialmente sensível a erros elevados. Já o *MAE* avalia o erro médio absoluto, proporcionando uma visão clara e direta da média dos desvios em relação aos valores reais. Essas métricas, ao serem aplicadas aos modelos desenvolvidos, permitem comparar objetivamente seu desempenho, identificar limitações e orientar futuras melhorias, garantindo maior confiabilidade às previsões geradas.

O cenário abordado neste trabalho, voltado à previsão de valores de ações da bolsa, apresenta vantagens conceituais em relação a outros contextos, como o das criptomoedas, por exemplo, principalmente devido à menor volatilidade e maior previsibilidade intrínseca ao mercado acionário tradicional. Isso ocorre porque, diferentemente dos ativos digitais, o valor das ações está diretamente relacionado ao comportamento histórico das empresas, como desempenho financeiro, decisões estratégicas e resultados operacionais, fatores que tendem a ser estáveis e podem ser analisados em profundidade ao longo do tempo. Além disso, a repercussão desses fatores em notícias, relatórios financeiros e eventos públicos permite uma compreensão mais clara e sistemática do comportamento das ações, fornecendo insumos valiosos para a modelagem preditiva, o que tende a resultar em modelos mais robustos e previsões mais confiáveis.

Os dados utilizados neste estudo são compostos pelos preços de fechamento das ações, que representam o último valor negociado do ativo em um determinado período e fornecem um indicador consolidado da percepção do mercado em relação às empresas analisadas. Esses dados, embora essenciais para análises financeiras, apresentam desafios específicos para a modelagem preditiva, como a presença de ruídos de curto prazo, sensibilidade a eventos pontuais e flutuações decorrentes de fatores externos não diretamente mensuráveis. Nesse contexto, o problema atacado aqui consiste precisamente em prever o comportamento futuro do preço das ações, por meio de um estudo de caso envolvendo três empresas especificamente selecionadas: `Apple`, `Tesla` e `Electromed`. Cada uma dessas empresas possui características particulares e diferentes graus de volatilidade, o que permite avaliar e validar a robustez e adaptabilidade dos modelos utilizados, contribuindo para um entendimento mais profundo das dinâmicas envolvidas no mercado acionário.

O problema proposto é resolvido neste trabalho por meio da implementação de um modelo baseado em Redes Neurais Recorrentes *(RNNs)*, do tipo *Long Short-Term Memory (LSTM)*, escolhido devido à sua capacidade comprovada de capturar dependências temporais complexas presentes em séries financeiras. O modelo utiliza como entrada o histórico dos preços de fechamento das ações, adotando uma frequência intradiária, com intervalos regulares, para períodos de análise de até um ano, e frequência diária para períodos maiores que um ano. Adicionalmente, incorpora-se a análise de sentimento extraída de até nove notícias por dia relacionadas às empresas analisadas no mesmo período, com o objetivo de capturar influências qualitativas e externas ao comportamento dos preços. Essa combinação de dados quantitativos e

qualitativos possibilita ao modelo interpretar padrões mais amplos e fornecer previsões mais precisas e fundamentadas sobre os futuros movimentos das ações.

No capítulo de Revisão da Literatura, será dado foco especial aos métodos de aprendizado profundo *(deep learning)* e técnicas de aprendizado de máquina *(machine learning)* aplicadas à previsão de séries temporais financeiras. Serão explorados modelos amplamente utilizados na literatura, como Redes Neurais do tipo *Long Short-Term Memory (LSTM)*, *Gated Recurrent Unit (GRU)*, Redes Neurais Convolucionais *(CNNs)*, bem como abordagens híbridas que combinam algoritmos tradicionais com análise de sentimento ou indicadores técnicos. A revisão buscará contextualizar o avanço dessas técnicas, destacando suas principais características, vantagens e limitações, de forma a fundamentar teoricamente a escolha metodológica adotada neste trabalho.

Diante do contexto apresentado, o objetivo central deste trabalho é desenvolver um modelo preditivo capaz de estimar com precisão os preços futuros de ações negociadas na bolsa de valores, com base em técnicas de aprendizado profundo *(deep learning)*. Para isso, será utilizada uma arquitetura de Rede Neural Recorrente *(RNN)* do tipo *Long Short-Term Memory (LSTM)*, especialmente adequada para lidar com a complexidade e as dependências temporais de longo prazo típicas dos dados financeiros.

Mais especificamente, busca-se avaliar o desempenho da *LSTM* na previsão dos preços de fechamento das ações de três empresas selecionadas — `Apple`, `Tesla` e `Electromed` — por meio da combinação de dados históricos de preços e informações extraídas da análise de sentimento de notícias publicadas ao longo do período de observação. Ao final, pretende-se verificar a eficácia da abordagem proposta e sua viabilidade como ferramenta de suporte à tomada de decisão no mercado financeiro.

A abordagem adotada para alcançar o objetivo proposto baseia-se na construção de uma Rede de Aprendizado Profundo *(Deep Learning Network)* do tipo *LSTM*, projetada para processar e aprender padrões sequenciais a partir de dados financeiros. Essa rede foi alimentada com duas fontes principais de informação: o histórico de preços de fechamento das ações, que oferece uma visão quantitativa e consolidada da variação dos ativos ao longo do tempo, e os dados de sentimentos extraídos de notícias financeiras, que fornecem uma dimensão qualitativa capaz de capturar variações de curto prazo provocadas por eventos externos e repercussões midiáticas.

A escolha por integrar essas duas fontes de dados foi motivada por sua relevância prática no contexto do mercado financeiro. O uso do histórico de preços é uma prática consolidada entre *traders* e analistas técnicos, pois permite identificar tendências, padrões de comportamento e pontos de reversão. Já a incorporação da análise de sentimento das notícias reflete uma visão mais moderna e abrangente do mercado, considerando que a percepção pública e os acontecimentos recentes podem exercer forte influência no valor dos ativos. Dessa forma, os atributos utilizados na modelagem combinam informações objetivas — como os preços anteriores — com sinais subjetivos derivados da linguagem natural, visando melhorar a capacidade preditiva do modelo e torná-lo mais sensível à dinâmica real do mercado.

Os resultados obtidos demonstram que a técnica proposta foi bem-sucedida na previsão de preços de ações, apresentando desempenho variável conforme a empresa analisada, o período considerado e a presença ou não de dados de notícias. No caso da `Apple`, a inclusão de análise de sentimento melhorou a performance no longo prazo, enquanto no curto prazo o modelo baseado apenas em preços obteve melhores resultados. A `Electromed`, mesmo sem notícias, apresentou o menor RMSE geral no período de 2020–2024, evidenciando alta previsibilidade em ativos mais estáveis. Já a `Tesla` apresentou os maiores erros preditivos, com desempenho inferior nos dois cenários, reflexo de sua alta volatilidade. Esses achados confirmam a efetividade do modelo LSTM proposto e indicam que sua performance depende fortemente

das características específicas de cada ativo e da configuração do experimento, reforçando a importância de abordagens personalizadas e adaptativas.

Este trabalho está estruturado em seis capítulos, além desta Introdução. No Capítulo 2, *Revisão da Literatura*, são exploradas as principais abordagens e modelos utilizados na previsão de séries temporais financeiras, com ênfase em técnicas de aprendizado profundo*(deep learning)* e aprendizado de máquina*(machine learning)*. O Capítulo 3, *Trabalho Base*, apresenta o estudo original que serviu como ponto de partida para o desenvolvimento desta pesquisa, destacando sua estrutura e limitações.

Em seguida, o Capítulo 4, *Trabalho Realizado*, descreve detalhadamente as modificações, adaptações e melhorias implementadas em relação ao trabalho base, abrangendo desde a coleta e pré-processamento dos dados até a construção dos modelos. O Capítulo 5, *Resultados*, apresenta e analisa os experimentos realizados, discutindo o desempenho dos modelos sob diferentes condições.

Por fim, o Capítulo 6, *Conclusão*, sintetiza os principais achados, limitações do estudo e possibilidades de trabalhos futuros. As referências utilizadas ao longo do documento estão listadas ao final.

2 REVISÃO DA LITERATURA

Nas últimas décadas, a previsão de preços e tendências em mercados financeiros tem passado por transformações significativas, impulsionadas pelo avanço de técnicas computacionais, especialmente aquelas baseadas em Inteligência Artificial (IA), Aprendizado de Máquina (*Machine Learning (ML)*) e Aprendizado Profundo (*Deep Learning (DL)*). Esse cenário tem estimulado a produção científica e prática de modelos preditivos cada vez mais sofisticados, que buscam lidar com a natureza dinâmica, volátil e não linear dos dados financeiros. A literatura recente aponta para um movimento crescente na direção da adoção de arquiteturas híbridas, que combinam diferentes algoritmos e abordagens estatísticas; do uso de técnicas automatizadas de extração de características, capazes de identificar padrões ocultos nos dados; e da integração de fontes estruturadas e não estruturadas, como séries históricas de preços e dados textuais oriundos de notícias, redes sociais e relatórios financeiros. Tais estratégias têm se mostrado fundamentais para o aumento da precisão preditiva e da robustez dos modelos, contribuindo para o desenvolvimento de sistemas mais adaptáveis e eficazes no apoio à tomada de decisão no contexto do mercado de capitais.

A crescente complexidade do mercado financeiro e a diversidade de abordagens computacionais motivaram a realização de diversos estudos de revisão sistemática, que buscam consolidar o conhecimento acumulado sobre técnicas preditivas aplicadas à previsão de preços e tendências. Esses *surveys* fornecem uma visão abrangente dos principais modelos utilizados, suas evoluções metodológicas, fontes de dados e métricas de avaliação, além de apontarem direções futuras para a pesquisa científica no campo da Inteligência Artificial aplicada às finanças.

Diversos estudos de revisão têm contribuído significativamente para a consolidação do conhecimento sobre técnicas de previsão no mercado financeiro, especialmente no contexto do uso de Inteligência Artificial. Sarker et al. (2024), ao revisarem mais de 50 estudos, destacam as redes *Long Short-Term Memory(LSTM)* como particularmente eficazes na captura de padrões de longo prazo em séries temporais financeiras, ressaltando também o potencial das abordagens híbridas que combinam *CNNs* e análise de sentimentos para aumentar a capacidade de generalização dos modelos. De forma semelhante, Balasubramanian et al. (2024), em uma análise abrangente de 300 estudos, confirmam o protagonismo de modelos como *LSTM*, *SVM*, *ANN* e *CNN*, indicando o uso predominante de bases de dados como *Yahoo Finance*, *NASDAQ* e *NSE*, bem como a ampla aplicação de métricas como *MSE*, *RMSE* e *F1-score*. Esses autores também evidenciam uma tendência crescente de integração de dados multimodais, incluindo sentimentos e notícias.

Complementando esse panorama, Chong et al. (2017) propõem o uso de transformações como *PCA*, *Autoencoders* e *Restricted Boltzmann Machines(RBM)* em redes profundas, as quais demonstraram desempenho superior aos modelos *AR* tradicionais, sobretudo na estimação de covariâncias e na identificação de padrões latentes sob condições de alta volatilidade. No campo das criptomoedas, Otabek and Choi (2024) realizaram uma revisão focada e apontaram que modelos como *LSTM*, *GRU*, *Deep Reinforcement Learning(DRL)* e técnicas híbridas que integram análise técnica e de sentimentos apresentam resultados mais precisos em ambientes extremamente voláteis.

Por fim, Chopra et al. (2024) propuseram uma categorização dos principais métodos em cinco grandes grupos — redes neurais, modelos *fuzzy*, *ARCH/GARCH*, *SVMs* e modelos híbridos — e reforçaram a importância de alinhar essas abordagens com teorias financeiras tradicionais, destacando também a necessidade de futuras pesquisas voltadas à generalização e aplicabilidade prática dos modelos preditivos.

Thakkar and Chaudhari (2024) realizaram uma revisão sistemática das aplicações de algoritmos genéticos*(GAs)* na última década, destacando seu uso na otimização de hiperparâmetros, seleção de atributos e hibridização com modelos como *LSTM*, *ANN*, *SVM* e outras meta-heurísticas, como *PSO*, *ACO* e *DE*. Complementarmente, Lin and Marques (2024) conduziram uma meta-análise de revisões sistemáticas, consolidando *LSTM*, *SVM* e *ANN* como os métodos mais recorrentes em previsões financeiras. O estudo também evidenciou o predomínio do uso de dados de preços históricos, indicadores técnicos e análise de sentimentos, além de apontar como tendências futuras a personalização de modelos e a adoção de métricas de avaliação mais avançadas.

Complementando os avanços metodológicos, Wei et al. (2024) propuseram uma arquitetura baseada na combinação de *Gated Recurrent Units(GRU)* com mecanismos de atenção, focada na extração de características relevantes e na priorização de informações temporais críticas. O modelo é estruturado em três etapas: (1) processamento das sequências temporais via *GRU*, (2) aplicação do mecanismo de atenção para identificar pontos temporais mais influentes e (3) fusão de dados estruturados com informações textuais derivadas de notícias financeiras. Testado em dados do *NYSE* e *NASDAQ* ao longo de uma década, o modelo *GRU-Attention* superou abordagens como *GBDT* e *GRU* puro em métricas como retorno anualizado, razão de *Sharpe* e *drawdown* máximo. Apesar de sua eficácia em tendências de longo prazo, o modelo mostrou limitações para prever pequenas oscilações de curto prazo. A proposta representa uma tendência promissora de combinação entre aprendizado profundo, atenção e multimodalidade, oferecendo maior transparência e aplicabilidade prática para investidores, e os autores defendem que essa integração poderá se tornar central em decisões financeiras baseadas em *DL* no futuro próximo.

A evolução das redes neurais profundas no contexto da previsão de mercados financeiros tem sido amplamente documentada na literatura recente. Dao et al. (2024) realizaram uma análise bibliométrica que evidencia essa trajetória, classificando os modelos em uni-modais — como *LSTM*, *CNN* e *RNN* — e multi-modais, como *GNNs* e *Transformers*, além de destacarem o papel emergente dos *Large Language Models(LLMs)* como ferramentas promissoras para análise textual e de sentimentos em tempo real. Complementarmente, Oyewole et al. (2024) reforçam a eficácia de arquiteturas baseadas em redes neurais, com ênfase em variantes como *LSTM*, *CNN-LSTM* e modelos com mecanismos de atenção. Ambos os estudos convergem na importância da integração de dados estruturados com fontes não estruturadas, bem como no uso de técnicas avançadas de pré-processamento, como normalização e seleção de atributos, para potencializar o desempenho dos modelos preditivos.

Diversos estudos têm destacado a evolução e a importância crescente das técnicas de aprendizado de máquina *(ML)* na previsão de preços de ações, um campo desafiador devido à alta volatilidade e à influência de fatores externos. O artigo de Dubey et al. (2024) apresenta uma análise abrangente das técnicas atuais de *ML* utilizadas para esse fim, evidenciando que modelos como *LSTM*, *GRU*, *Random Forest*, *ARIMA* e *SVM* são os mais recorrentes e eficazes. Os autores enfatizam que modelos híbridos, como *LSTM–GRU* e *ARIMA–GRU*, têm demonstrado desempenho superior ao integrar o poder preditivo de abordagens estatísticas e de aprendizado profundo. A análise mostra que modelos como *LSTM* e *GRU* são adequados para capturar padrões em séries temporais complexas, enquanto *ARIMA* é útil para tendências lineares de curto prazo. *Random Forests* destacam-se por sua robustez frente a dados ruidosos, e *SVMs* por sua precisão em ambientes de alta dimensionalidade. Tais técnicas, quando utilizadas com múltiplas fontes de dados — como históricos de preços, volume de negociação, notícias financeiras e sentimentos extraídos de mídias sociais —, tendem a apresentar previsões mais precisas e contextualizadas. Contudo, desafios persistem, como a presença de dados ruidosos, problemas de não-estacionaridade e risco de *overfitting*, além da dificuldade em integrar informações de

diferentes naturezas em um único modelo eficiente. Para lidar com tais dificuldades, os autores sugerem o uso de técnicas mais robustas de pré-processamento e normalização de dados, além do desenvolvimento contínuo de modelos híbridos mais adaptáveis e precisos. Em suma, o estudo reitera a importância de abordagens integrativas e flexíveis na previsão de preços de ações, apontando para um futuro promissor onde modelos híbridos e multivariados, aliados a fontes de dados diversificadas, poderão gerar decisões de investimento mais assertivas e estratégias mais sólidas (Dubey et al., 2024).

Outra abordagem relevante na previsão de preços de ações é apresentada por Rahman and Akhter (2021), que propõem a utilização de um modelo híbrido de regressão empilhada (*stacked regression*), combinando múltiplas técnicas de aprendizado de máquina. A proposta visa enfrentar os desafios da natureza caótica e não linear dos mercados financeiros, além da influência de fatores externos como economia e comportamento humano. Utilizando dados históricos obtidos da plataforma Quandl, referentes a empresas como Google, Apple, Microsoft, IBM, NIKE, McDonald's, Walt Disney e Intel, os autores realizaram um pré-processamento rigoroso para remoção de atributos irrelevantes e dados ausentes. Em seguida, aplicaram quatro modelos principais — Regressão Linear, *K-Nearest Neighbors(KNN)*, *Support Vector Regression(SVR)* e *Random Forest Regression* — cujas previsões foram combinadas em um modelo final por meio de empilhamento. Esta abordagem permitiu capturar padrões que métodos isolados não conseguem modelar com precisão. Os resultados empíricos demonstraram que o modelo empilhado superou significativamente os modelos individuais, atingindo, por exemplo, 98,7% de precisão para a ação da Google. Resultados similares foram obtidos para outras empresas, como Apple (94,8%) e NIKE (96,8%), validando a robustez da abordagem com métodos como validação cruzada. O artigo conclui que a regressão empilhada representa uma técnica promissora para aplicações reais no mercado financeiro, sugerindo ainda a exploração futura de modelos mais sofisticados como redes neurais profundas para aprimorar ainda mais os resultados (Rahman and Akhter, 2021).

O estudo de Teixeira Zavadzki de Pauli et al. (2020) foca no contexto brasileiro ao comparar diferentes arquiteturas de redes neurais artificiais para a previsão de preços de ações da B3. Foram analisados modelos como regressão linear múltipla *(MLR)*, redes Elman, Jordan, de base radial *(RBF)* e perceptrons multicamadas *(MLP)*, totalizando 620 configurações distintas, com aplicação do método *bootstrap* para gerar intervalos de confiança. A amostra incluiu ações amplamente negociadas, como PETR4 e VALE3, com dados obtidos via o pacote `quantmod` na linguagem R, utilizando séries temporais diárias entre 2013 e 2020. O desempenho dos modelos foi avaliado por meio do *RMSE* no conjunto de validação, e os resultados mostraram que a MLR apresentou desempenho comparável — e, em alguns casos, superior — a modelos mais complexos, sendo a melhor configuração para PETR4, com *RMSE* de 0,79. Em contrapartida, a rede RBF obteve os piores resultados, reforçando a ideia de que maior complexidade arquitetural não garante melhor performance. O artigo descreve em detalhe as arquiteturas implementadas, abordando funções de ativação, camadas e mecanismos de *feedback* nas redes recorrentes, além de discutir limitações de generalização observadas principalmente em períodos de alta volatilidade, como durante a pandemia de COVID-19. Por fim, os autores sugerem que futuras pesquisas considerem o uso de *CNNs*, *LSTMs* e estratégias mais robustas para construção de intervalos de confiança (Teixeira Zavadzki de Pauli et al., 2020).

Al-Ali and Al-Alawi (2024) apresentam uma revisão sistemática sobre a aplicação de técnicas de aprendizado de máquina *(ML)* e aprendizado profundo *(DL)* na previsão do mercado de ações, analisando 11 estudos aplicados a setores diversos, como petróleo e gás, tecnologia da informação e setor bancário, com dados históricos e sentimentais provenientes de fontes como *NASDAQ*, *NYSE*, *Twitter* e *Yahoo Finance*. O estudo evidencia o desempenho superior dos modelos de *DL*, especialmente as redes *LSTM*, eficazes na modelagem de dependências

temporais de longo prazo, e as *CNNs*, com destaque para a detecção de mudanças abruptas em padrões financeiros. Métodos de *ensemble learning*, como o *stacking*, também apresentaram bons resultados ao combinar diferentes algoritmos para aumentar a precisão preditiva. Além disso, modelos como *Random Forest(RF)* e *Naive Bayes* demonstraram desempenho relevante quando integrados com dados de sentimento. Apesar desses avanços, os autores apontam desafios como o *overfitting* observado em modelos como *SVM*, além da sensibilidade dos modelos de *DL* à configuração de hiperparâmetros. Como direções futuras, sugerem ampliar o uso de dados sentimentais, explorar diferentes setores econômicos e avaliar variações híbridas com técnicas de redução de dimensionalidade, como *PCA-DNN*. Em síntese, a revisão reforça que, no cenário atual, *CNN* e *LSTM* são as abordagens mais eficazes para a previsão no mercado financeiro (Al-Ali and Al-Alawi, 2024).

Jain et al. (2024) conduzem uma análise extensiva da evolução dos métodos de previsão de preços no mercado de ações baseados em técnicas de aprendizado de máquina ao longo das últimas duas décadas, organizando a literatura em quatro grandes categorias: redes neurais artificiais *(ANNs)*, máquinas de vetores de suporte *(SVMs)*, algoritmos genéticos *(GAs)* e modelos híbridos. O estudo fornece uma visão sistemática das forças e limitações de cada abordagem, destacando que as *ANNs*, especialmente as variantes recorrentes como *RNN* e *LSTM*, são eficazes na detecção de padrões não lineares e sequenciais, embora enfrentem desafios relacionados ao *overfitting* e à baixa interpretabilidade. As *SVMs*, por sua vez, demonstram bom desempenho em tarefas de classificação de tendências, mas apresentam alto custo computacional em conjuntos de dados extensos. Já os *GAs* são valorizados pela capacidade de otimização de portfólios e ajuste de hiperparâmetros, embora dependam significativamente de uma configuração inicial adequada. O estudo aponta os modelos híbridos como os mais promissores, ao combinar diferentes paradigmas — como *SVM+ANN*, *LSTM* com análise de sentimentos e *GAs* com redes neurais — para capturar múltiplas dimensões do mercado e produzir previsões mais precisas e robustas. Além disso, destaca-se o papel crescente da incorporação de dados qualitativos, como sentimentos de investidores e eventos econômicos, na melhoria da acurácia preditiva. Apesar dos avanços, os autores ressaltam barreiras como a escalabilidade dos modelos, a necessidade de grandes volumes de dados e a complexidade de aplicar essas soluções em ambientes voláteis, concluindo que colaborações entre a academia e o setor financeiro são fundamentais para aprimorar e validar essas abordagens em contextos reais (Jain et al., 2024).

Swamy et al. (2024) realizam uma análise abrangente sobre o uso de técnicas de Aprendizado de Máquina *(ML)* e Aprendizado Profundo *(DL)* na previsão do mercado de ações, enfatizando o papel essencial da engenharia de características, da seleção de variáveis e das estratégias de pré-processamento na construção de modelos preditivos eficazes. O estudo utiliza bases como o *StockNet Dataset*, que integra dados financeiros e sentimentais extraídos do Twitter, o *Historical Stock Market Dataset*, com preços e volumes das bolsas *NYSE* e *NASDAQ*, e o *News and Stock Data Dataset*, que combina notícias do *Reddit* com o desempenho do índice Dow Jones. Entre os principais modelos avaliados estão *Support Vector Machines(SVM)*, *Random Forest(RF)* e *K-Nearest Neighbors(KNN)*. O *SVM* apresentou acurácia de até 93,7% e mostrou-se eficaz para dados de alta dimensionalidade, embora sensível a ruídos; o *Random Forest* destacou-se por sua robustez e desempenho em ambientes voláteis, superando 90% de precisão; enquanto o *KNN* atingiu 97,2%, apesar de depender fortemente da escolha do parâmetro k e da escala dos dados. Além disso, o estudo destaca como características preditivas mais relevantes os preços históricos (abertura, fechamento, volume), indicadores técnicos como *RSI*, *MACD* e Bandas de Bollinger, além de variáveis macroeconômicas como PIB, inflação e taxas de juros. Ferramentas como *PCA*, *autoencoders* e análise de sentimentos com *NLP* também são citadas como estratégias promissoras para o enriquecimento dos modelos. Por fim, os autores concluem que, embora

abordagens como *RF* e *SVM* superem métodos estatísticos tradicionais como o *ARIMA*, desafios como *overfitting*, qualidade dos dados e ajuste de hiperparâmetros ainda limitam sua aplicação prática, recomendando que futuras pesquisas invistam na integração de múltiplas fontes de dados e no aprimoramento da capacidade de generalização e interpretabilidade dos modelos (Swamy et al., 2024).

Feng et al. (2018) introduzem uma arquitetura inovadora para previsão de movimentos do mercado de ações por meio do modelo *Adv-ALSTM*, que combina redes *LSTM* com mecanismos de atenção e técnicas de treinamento adversarial. O objetivo é superar a limitação de generalização comum em modelos de aprendizado profundo, especialmente diante da alta volatilidade e estocasticidade dos dados financeiros. A estrutura do modelo envolve uma camada de mapeamento de características, uma camada *LSTM* para capturar dependências temporais, uma camada de atenção que atribui pesos diferenciados a momentos históricos relevantes e, por fim, uma camada de predição. A principal inovação reside na inserção de perturbações adversariais na última camada latente, tornando o modelo mais robusto a ruídos e flutuações inesperadas. Os experimentos foram conduzidos sobre dois conjuntos de dados amplamente utilizados: o *ACL18*, com ações da *NASDAQ* e *NYSE* entre 2014 e 2016, e o *KDD17*, que abrange dados de ações dos EUA entre 2007 e 2016. Os resultados revelaram ganhos significativos em termos de acurácia e coeficiente de correlação de Matthews*(MCC)* em relação a métodos tradicionais como *Momentum(MOM)* , *Mean Reversion(MR)* e Autoencoders Variacionais*(VAE)* . No conjunto ACL18, o *Adv-ALSTM* apresentou acréscimos de 4,02% em acurácia e 42,19% em *MCC*, enquanto no KDD17 os ganhos foram de 2,14% e 56,12%, respectivamente. O estudo conclui que a incorporação de treinamento adversarial em redes com atenção temporal oferece uma estratégia promissora para mitigar o *overfitting*, ampliar a estabilidade das previsões e aumentar a confiabilidade das decisões de investimento em cenários marcados por alta incerteza (Feng et al., 2018).

Sawhney et al. (2020) propõem o modelo *Multipronged Attention Network for Stock Forecasting(MAN-SF)*, que integra informações de preços históricos, textos de redes sociais e correlações estruturais entre empresas para aprimorar a previsão de movimentos no mercado de ações. A arquitetura do modelo é composta por três módulos principais: o *Price Encoder*, baseado em *GRU* com mecanismo de atenção temporal, responsável por capturar padrões históricos de preços; o *Social Media Information Encoder(SMI)*, que utiliza o *Universal Sentence Encoder(USE)* combinado com atenção hierárquica para avaliar a influência e relevância temporal de tweets; e a *Graph Attention Network(GAT)*, que modela as relações entre empresas a partir de dados da *Wikidata*, atribuindo pesos diferenciados conforme a importância das conexões. Esses três componentes são integrados por uma camada bilinear, capaz de aprender interações não-lineares entre os sinais multimodais, resultando em uma representação conjunta mais robusta para fins preditivos. O modelo foi avaliado com o dataset *StockNet*, composto por dados de ações do S&P 500 e tweets associados a cada empresa, e a tarefa de previsão foi formulada como um problema de classificação binária. Os resultados demonstraram que o *MAN-SF* superou modelos de referência como *StockNet*, *HATS* e *Adversarial LSTM*, alcançando *F1-score* de 0,605, acurácia de 0,608, *MCC* de 0,195 e um *Sharpe Ratio* anualizado de 1,05. Os estudos de ablação confirmaram que a fusão bilinear de dados multimodais é mais eficaz do que abordagens tradicionais como concatenação simples ou atenção isolada. O artigo conclui que a integração de informações contextuais derivadas de redes sociais e das relações estruturais entre empresas oferece ganhos substanciais de desempenho e contribui para decisões de investimento mais informadas e robustas (Sawhney et al., 2020).

Ghosh et al. (2022) investigam a previsão direcional de preços de ações para operações intradiárias, comparando dois modelos: *Random Forest(RF)* e *Long Short-Term Memory(LSTM)*

com aceleração via *cuDNN*. Utilizando dados do S&P 500 entre 1990 e 2018, estruturados em janelas móveis de quatro anos (três anos para treino e um para teste), os autores analisam três tipos de retorno — intradiário *(ir)*, de fechamento *(cr)* e de abertura *(or)* —, o que resulta em 93 variáveis para a *RF* e em sequências de 240 *timesteps* com 3 *features* para a *LSTM*. A arquitetura da rede *LSTM* inclui 25 células `CuDNNLSTM`, seguidas por *dropout* de 0,1 e uma camada densa com ativação *softmax*, treinada com o otimizador *RMSProp* e função de perda *cross-entropy* categórica. A *Random Forest*, por sua vez, é composta por 1000 árvores com profundidade máxima de 10 e utiliza a raiz quadrada do número total de atributos ($\sqrt{p}$) como critério de divisão. As previsões são formuladas como uma tarefa de classificação binária, classificando cada ação com base na mediana transversal do retorno do mercado no mesmo dia. Os resultados demonstram que o modelo *LSTM* multivariado proporciona retorno médio diário de 0,64% antes de custos (0,44% líquido), superando tanto o modelo *LSTM* univariado (0,41%) quanto a *RF* (0,54%). Além do retorno superior, o modelo *LSTM* apresentou menor risco — medido por *VaR*, *drawdown* e desvio padrão — e maior frequência de dias positivos, especialmente no subperíodo de 1993 a 2010. Apesar do maior custo computacional, o uso de GPU via *cuDNN* tornou o treinamento do modelo *LSTM* eficiente e viável. O estudo reforça a eficácia da previsão intradiária por meio de técnicas de aprendizado profundo e evidencia a superioridade de abordagens multivariadas que integram múltiplos tipos de retorno para estratégias de *trading* de curto prazo (Ghosh et al., 2022).

García-Medina and Aguayo-Moreno (2024) propõem um modelo híbrido baseado em *LSTM* com *GARCH* para previsão da volatilidade em portfólios de criptomoedas, destacando-se pela incorporação de variáveis financeiras de alta frequência. Utilizando dados de preços e volume das dez principais criptomoedas entre janeiro e junho de 2020, o estudo modelou séries temporais com frequência de 5 minutos, convertidas em retornos logarítmicos horários e variância realizada, totalizando 4.368 observações. O modelo híbrido mais eficaz foi o *LSTM–gjrGARCH*, que combinou a arquitetura clássica da *LSTM* com coeficientes estimados do *gjrGARCH* — como tendência, resíduo, volatilidade condicional e assimetria — e apresentou os melhores resultados em termos de erro heterocedástico (*HAE* e *HSE*), embora sem diferenças estatisticamente significativas em relação a outros modelos híbridos. Treinado com janelas móveis de 504 horas e deslocamento de 24 horas, o modelo foi avaliado por métricas como *Sharpe Ratio*, *Value at Risk (VaR)* e o teste de Diebold-Mariano. Apesar da sofisticação do *LSTM–GARCH*, o modelo com melhor desempenho global foi o *MLP*, que superou os demais em simplicidade, precisão e custo computacional. Para previsões de longo prazo, o modelo *DCC-eGARCH-Vol* apresentou o maior *Sharpe Ratio*, sendo mais indicado para estratégias com horizonte temporal ampliado. A análise de alocação de ativos revelou uma predominância do Bitcoin antes do anúncio da pandemia, seguida de maior diversificação posterior. O estudo reforça que, embora modelos híbridos como o *LSTM–GARCH* ofereçam recursos avançados ao considerar a volatilidade condicional, arquiteturas mais simples como o *MLP*, quando bem calibradas, podem alcançar desempenho superior. Além disso, a inclusão de variáveis exógenas como o volume mostrou-se eficaz para reduzir riscos e aumentar a precisão nas previsões de volatilidade (García-Medina and Aguayo-Moreno, 2024).

Murray et al. (2023) conduziram uma análise comparativa entre diferentes abordagens para previsão de preços de criptomoedas, destacando o desempenho superior dos modelos *LSTM* em relação a outras técnicas de aprendizado de máquina, aprendizado profundo e modelos híbridos. O modelo proposto utilizou uma arquitetura composta por uma camada convolucional com 64 filtros e ativação ReLU, seguida de uma camada *LSTM* com 75 unidades e uma camada densa com 16 neurônios, sendo treinado com taxa de aprendizado de 1×10^{-4} e regularização via *early stopping*. O estudo utilizou dados históricos coletados entre junho de 2017 e maio de

2022 para cinco criptomoedas — Bitcoin, Ethereum, Litecoin, Monero e Ripple — considerando preços de abertura, fechamento, máxima e mínima diárias, com a tarefa formulada como uma previsão univariada do fechamento do dia seguinte. As séries temporais foram transformadas para estacionariedade por *differencing* e normalizadas via *Min-Max Scaling*. A avaliação foi realizada com uma estratégia *rolling window* mensal, em que os dados de cada mês de teste eram incorporados ao conjunto de treinamento, o modelo re-treinado do zero e os erros registrados. O modelo *LSTM* obteve os melhores resultados entre todos os testados, com *RMSE* médio de 0,0222, *MAE* de 0,0173, *MAPE* de 3,86% e R^2 de 0,735, superando inclusive modelos híbridos, *GRU*, *CNNs temporais (TCN)* e modelos baseados em *transformadores (TFT)*. A tentativa de aplicar *ensembles* também não foi eficaz, pois a inclusão de modelos menos precisos comprometeu o desempenho geral. Os autores concluíram que o *LSTM* representa o melhor equilíbrio entre precisão preditiva e custo computacional, sendo indicado como principal abordagem para previsão de preços diários em cenários com recursos computacionais moderados e dados históricos confiáveis, além de reforçarem a importância da reprodutibilidade científica ao disponibilizar publicamente o código e os dados utilizados nos experimentos (Murray et al., 2023).

O artigo de Prajapati (2020a), intitulado *"Predictive Analysis of Bitcoin Price Considering Social Sentiments"*, investiga o impacto de sentimentos sociais na predição do preço do Bitcoin. O estudo propõe um modelo preditivo robusto baseado em aprendizado profundo, combinando dados históricos de preços com análises de sentimento extraídas de fontes como *Google News* e *Reddit*. A etapa de coleta de dados foi inteiramente automatizada por meio de scripts desenvolvidos em Python, abrangendo três principais fontes: (1) notícias diárias extraídas do *Google News*, (2) postagens horárias provenientes do subreddit r/Bitcoin, obtidas via *Pushshift API*, e (3) dados históricos de preço e volume das criptomoedas Bitcoin, Litecoin e Ethereum. Para garantir a consistência temporal entre as fontes, todos os dados foram padronizados com granularidade horária, replicando as notícias ao longo das 24 horas do dia.

O funcionamento do modelo preditivo desenvolvido por (Prajapati, 2020a) pode ser descrito em quatro etapas principais. Primeiramente, ocorre a coleta automatizada das três fontes de dados mencionadas. Em seguida, realiza-se a análise de sentimentos com o uso de ferramentas como *Flair*, *VADER* e *TextBlob*, que quantificam o sentimento presente nos textos. Posteriormente, os dados estruturados (preço e volume) e não estruturados (sentimentos) são fundidos e padronizados em um único dataset temporalmente alinhado. Por fim, modelos de aprendizado profundo, como *LSTM* e *GRU*, são treinados e avaliados com o objetivo de prever os movimentos de preço do Bitcoin a partir da combinação dessas informações heterogêneas. As análises de sentimento resultaram em múltiplas variáveis por hora, que foram integradas ao conjunto de dados final.

Esse dataset unificado passou a conter 24 colunas, englobando tanto indicadores quantitativos (como preço e volume) quanto qualitativos (relacionados aos sentimentos extraídos das fontes textuais). Essa estrutura permitiu o treinamento de diferentes arquiteturas de redes neurais, como *LSTM*, *GRU*, *1D-CNN*, além de combinações híbridas como *LSTM→GRU* e *CNN→LSTM*, totalizando 17 experimentos. Dentre os testes realizados, o Experimento #4 destacou-se como o mais eficaz, pois utilizou todos os dados históricos e variáveis de sentimento, alcançando um *RMSE* de 434,87. Em comparação, o Experimento #5, embora tenha obtido o menor *RMSE* (173,72), não utilizou dados de sentimento, o que, segundo os autores, compromete a confiabilidade do modelo diante da natureza especulativa do mercado de criptomoedas.

A conclusão do estudo reforça a relevância da incorporação de sentimentos sociais em modelos de previsão de preços, especialmente no caso do Bitcoin, cujas variações são fortemente influenciadas por fatores especulativos e pela opinião pública. O *Reddit* demonstrou maior poder

preditivo que o *Google News*, indicando que plataformas com conteúdo gerado por usuários podem capturar melhor as percepções do mercado.

A seleção dos estudos relevantes foi realizada com base em uma revisão prévia da literatura, fundamentada nos *surveys* e artigos de revisão mencionados anteriormente. A partir desses trabalhos, foram identificadas diversas pesquisas empíricas sobre previsão de preços com técnicas de aprendizado de máquina e análise de sentimentos. Durante essa etapa, buscou-se priorizar estudos que, além de apresentarem abordagens metodológicas robustas, disponibilizassem também seus códigos-fonte, de modo a viabilizar a reprodução e posterior adaptação dos experimentos. Foi justamente nesse processo de levantamento e triagem que se encontrou o trabalho-base de (Prajapati, 2020a) utilizado neste estudo, o qual atendeu aos critérios de acessibilidade, relevância metodológica e compatibilidade com os objetivos propostos.

2.1 JUSTIFICATIVA DA ESCOLHA DO TRABALHO BASE E ADAPTAÇÃO METODOLÓGICA

Diante da clareza metodológica, da efetiva integração entre dados estruturados e não estruturados, e dos resultados promissores obtidos, este artigo de (Prajapati, 2020a) será adotado como base conceitual e metodológica para o desenvolvimento deste Trabalho de Conclusão de Curso (TCC). Sua abordagem híbrida, aliada à viabilidade de reprodução e à disponibilidade do código-fonte, oferece um ponto de partida sólido para a adaptação do modelo ao contexto do mercado de ações, ampliando assim seu escopo de aplicação e relevância acadêmica.

Nos próximos capítulos — Trabalho Base e Trabalho Realizado — serão apresentados em detalhes tanto o funcionamento do modelo original de Prajapati (2020a) quanto as modificações propostas neste TCC. A escolha desse estudo como base não é meramente teórica, mas resultado de uma análise crítica fundamentada em critérios de viabilidade, reprodutibilidade e aplicabilidade. Dentre todos os trabalhos analisados ao longo da revisão de literatura, este foi o único que pôde ser reproduzido fielmente conforme descrito, o que o torna não apenas relevante, mas também acessível para fins acadêmicos e experimentais. Além disso, o código-fonte está integralmente disponível em repositórios públicos no GitHub, fator essencial para garantir transparência e aprofundamento técnico. Outro ponto determinante para essa escolha foi a possibilidade concreta de transpor o modelo original — focado em criptomoedas — para o contexto de previsão de preços de ações, que constitui o escopo deste TCC. Essa capacidade de adaptação e reaproveitamento metodológico amplia significativamente o valor do estudo de Prajapati, ao contrário de outros trabalhos revisados, que, embora teoricamente robustos, não oferecem código aberto, não são reprodutíveis ou apresentam forte dependência de dados inacessíveis. Portanto, a decisão de utilizar este artigo como base está diretamente associada ao seu potencial de aplicação prática, adaptabilidade metodológica e ao seu alinhamento com os objetivos específicos deste trabalho.

Diante desse cenário, o presente trabalho propõe uma adaptação metodológica do estudo original de Prajapati (2020a) para o contexto de previsão de preços de ações, redirecionando o foco das criptomoedas para ativos do mercado financeiro tradicional. Essa transposição exige não apenas uma mudança no escopo dos dados, mas também reformulações substanciais na estrutura dos scripts utilizados. Entre as principais modificações implementadas, destaca-se o script `scraper.py`, que passou a realizar coletas diárias por meio do *RSS* do *Google News*, garantindo maior estabilidade e previsibilidade no processo. Para lidar com os redirecionamentos presentes nos links dessas notícias, foi incorporado o uso do Selenium em modo *headless*, assegurando a obtenção do conteúdo final da página acessada. No que diz respeito à análise de sentimentos, o script `sentiment_analysis.py` foi reestruturado para adotar boas práticas de programação, com supressão de mensagens de advertência, automação no download de léxicos e maior robustez

na interpretação dos resultados do modelo Flair. A coleta dos dados de preços históricos das ações foi viabilizada por meio do script `stock.py`, que utiliza a biblioteca `yfinance` e permite parametrização via linha de comando, ajustando automaticamente a granularidade conforme o período analisado. Para consolidar os dados em uma estrutura unificada, o script `merge.py` foi desenvolvido com uma abordagem modular e flexível, superando as limitações da versão original ao lidar com diferentes formatos e frequências temporais. Por fim, o script `LSTM.py` representa uma reinterpretação abrangente do `Expr4-LSTM.py`, promovendo avanços tanto em organização quanto em aplicabilidade, ao permitir o uso de dados de ações como Apple (AAPL) e Tesla (TSLA), sem restringir-se a criptomoedas. Essas modificações não apenas ampliam a aplicabilidade do modelo original, mas também adaptam sua arquitetura às especificidades do mercado acionário, tornando-o mais alinhado com os objetivos deste TCC.

3 TRABALHO BASE

Este capítulo tem como objetivo detalhar a estrutura e o funcionamento dos repositórios que compõem a base do estudo de previsão de preços do Bitcoin com *Long Short-Term Memory (LSTM)*. Cada um dos repositórios analisados desempenha um papel específico no processo de coleta, processamento e integração de dados, permitindo a construção de um *pipeline* completo para modelagem preditiva. O trabalho base do repositório (Prajapati, 2020d) é estruturado a partir da integração de três repositórios complementares. O repositório (Prajapati, 2020e) coleta e analisa sentimentos de postagens do *Reddit*, enquanto (Prajapati, 2020c) realiza a mesma tarefa com notícias do *Google News*. Já o (Prajapati, 2020b) fornece os dados históricos de preços de criptomoedas. Esses dados são unificados e utilizados para treinar um modelo *LSTM* de previsão do preço do Bitcoin. As seções a seguir descrevem individualmente os quatro repositórios utilizados: `reddit_scraper_and_sentiment_analyzer`, responsável pela extração e análise de sentimentos de postagens no *Reddit*; `google_news_scraper_and_sentiment_analyzer`, voltado para a coleta e análise de sentimentos de notícias do *Google*; *cryptocurrency_data_downloader*, que fornece os dados históricos de preços das criptomoedas; e o repositório principal *predicting_bitcoin_market*, onde todos os dados são integrados e utilizados no treinamento do modelo *LSTM*.

As redes Neurais do tipo *Long Short-Term Memory (LSTM)* são um tipo de Rede Neural Recorrente*(RNN)* especialmente projetado para lidar com sequências temporais e dependências de longo prazo. Elas se destacam por sua capacidade de armazenar e recuperar informações ao longo do tempo, superando limitações comuns de redes recorrentes tradicionais, como o desaparecimento ou explosão do gradiente. Isso torna as *LSTM* particularmente adequadas para tarefas como previsão de séries temporais, onde a ordem e a persistência de padrões passados são essenciais para estimar eventos futuros.

3.1 REPOSITÓRIO *REDDIT_SCRAPER_AND_SENTIMENT_ANALYZER*

O repositório (Prajapati, 2020e) disponibiliza dois scripts principais para coleta e análise de sentimentos em postagens do *Reddit*.

O primeiro script, `download_data_from_reddit.py`, é responsável por coletar dados de postagens usando a *Application Programming Interface (API) Pushshift*. Ele permite buscar publicações em *subreddits* específicos a partir de palavras-chave, respeitando janelas de tempo definidas e um número máximo de registros. Os dados extraídos incluem informações como título, conteúdo, data de publicação, autor e *subreddit*, sendo todos salvos em um arquivo *Comma-Separated Values (CSV)*.

O segundo script, *reddit_post_sentiment_analysis.py*, realiza a *análise de sentimento (sentiment analysis)* das postagens coletadas. Ele carrega o *CSV* gerado anteriormente e combina o título com o corpo da publicação para formar o texto a ser analisado. Em seguida, aplica três modelos distintos de análise de sentimento — *Flair*, *TextBlob* e *Valence Aware Dictionary and sEntiment Reasoner (VADER)* — para gerar classificações e *scores* de sentimento. Os resultados são organizados por hora, com médias calculadas para cada janela temporal. O script gera um novo *CSV* contendo os *scores* de sentimento por modelo e a média por hora.

Ao final, os dois arquivos produzidos — um com os dados brutos e outro com os sentimentos agregados — oferecem uma base sólida para análises temporais de sentimento em comunidades do *Reddit*.

3.2 REPOSITÓRIO *GOOGLE_NEWS_SCRAPER_AND_SENTIMENT_ANALYZER*

O repositório (Prajapati, 2020c) disponibiliza dois scripts principais para coleta de notícias e análise de sentimentos a partir do *Google News*.

O primeiro script, `google_news_scraper.py`, é responsável por realizar buscas no *Google News* com base em palavras-chave definidas pelo usuário, dentro de intervalos de datas específicas. Ele extrai informações como título, conteúdo da notícia, data de publicação, fonte e link direto. Os dados coletados são organizados e armazenados em um arquivo *Comma-Separated Values (CSV)*, permitindo a análise posterior.

O segundo script, `google_news_sentiment_analysis.py`, realiza a *análise de sentimento (sentiment analysis)* das notícias previamente coletadas. Ele carrega o *CSV* gerado pelo `scraper` e aplica três métodos distintos de análise de sentimento: *Flair*, *TextBlob* e *Valence Aware Dictionary and sEntiment Reasoner (VADER)*. Para cada notícia, são calculadas métricas como polaridade, subjetividade e intensidade do sentimento (positivo, negativo ou neutro). Os resultados são agrupados por data e exportados em um novo *CSV* que apresenta, para cada dia, a média dos sentimentos detectados por cada modelo.

O resultado final consiste em dois arquivos principais: um contendo os artigos coletados e outro com os sentimentos agregados por data. Essa estrutura oferece uma base sólida para analisar a evolução do sentimento presente nas notícias ao longo do tempo, podendo ser usada em estudos de opinião pública, impacto de eventos na mídia ou como variável auxiliar em modelos preditivos.

3.3 REPOSITÓRIO *CRYPTOCURRENCY_DATA_DOWNLOADER*

O script `download_data_from_binance.py`, presente no repositório (Prajapati, 2020b), tem como objetivo coletar dados históricos de criptomoedas diretamente da *Application Programming Interface (API)* da *Binance*, utilizando a biblioteca `python-binance`.

O funcionamento do script envolve a definição de uma lista de símbolos de criptomoedas, como `BTCUSDT` e `ETHUSDT`, além do intervalo de tempo desejado para a coleta dos dados. Para cada símbolo, o script realiza chamadas à *API* da *Binance* e obtém informações como preços de abertura, máxima, mínima, fechamento e volume. Os dados são inicialmente salvos em arquivos *Comma-Separated Values (CSV)* individuais para cada criptomoeda.

Após a coleta, os arquivos gerados são concatenados em um único *dataset* consolidado. O script também realiza uma etapa de limpeza nos dados finais, especialmente útil para evitar duplicações em execuções múltiplas ou lidar com eventuais limitações impostas pela *API* da *Binance*.

Como resultado, é gerado um arquivo *CSV* com dados históricos limpos e organizados de várias criptomoedas, que pode ser utilizado em análises financeiras, estudos de mercado ou modelos de previsão.

3.4 REPOSITÓRIO MAIN *PREDICTING_BITCOIN_MARKET*

O repositório (Prajapati, 2020d) tem como objetivo prever o preço do Bitcoin a partir da combinação de dados de *sentimento (sentiment)* extraídos de notícias e postagens no *Reddit*, juntamente com dados históricos de preços de criptomoedas. Ele consolida os resultados gerados pelos repositórios auxiliares em um fluxo integrado de análise e previsão.

O script `merge_data_files.py` é responsável por unificar os arquivos *Comma-Separated Values (CSV)* gerados anteriormente, contendo informações de *sentimento* provenientes

do *Reddit* e do *Google News*, além dos dados de mercado obtidos da *Binance*. O resultado desse processo é um único arquivo chamado `crypto_data_news_reddit_final.csv`, que organiza variáveis como preços de abertura e fechamento de criptomoedas, além de *scores* de *sentimento* gerados por diferentes modelos.

A segunda etapa do projeto é conduzida pelo *notebook* `Expr4-LSTM.ipynb`, que treina um modelo de *Rede Neural do tipo Long Short-Term Memory (LSTM)* utilizando os dados integrados. O modelo considera janelas temporais com múltiplas variáveis para aprender padrões históricos e prever os valores futuros do Bitcoin. Ao final do treinamento, os resultados são avaliados e comparados com os valores reais, sendo exibidos visualmente na imagem `expr4_results.png`, disponível no diretório `images/`. Essa visualização mostra que o modelo é capaz de capturar tendências relevantes, indicando o potencial do uso combinado de dados de *sentimento* e históricos de preço para previsões no mercado de criptomoedas.

4 TRABALHO REALIZADO

Este capítulo apresenta, de forma detalhada, o conjunto de adaptações, melhorias e desenvolvimentos implementados ao longo deste estudo com o objetivo de viabilizar a aplicação de técnicas de previsão baseadas em *Redes Neurais do tipo Long Short-Term Memory (LSTM)* ao mercado de ações. Cada seção do capítulo descreve individualmente os scripts que compõem o *pipeline* experimental, evidenciando suas funcionalidades, inovações técnicas e relação com os códigos originais utilizados como base. A seção `scraper.py` aborda o processo de coleta automatizada de notícias a partir do *Google News*, com ênfase na resolução de redirecionamentos e extração de conteúdo completo. Em seguida, `sentiment_analysis.py` trata da *análise de sentimento (sentiment analysis)* aplicada ao conteúdo textual, destacando a modularidade do código e a integração com diferentes ferramentas semânticas. A seção `stock.py` descreve o processo de obtenção de dados históricos das ações por meio da *Application Programming Interface (API)* do *Yahoo Finance*, enquanto `merge.py` apresenta a lógica de fusão entre os dados de preços e os dados de sentimento, adaptável a múltiplas frequências temporais. Já em `LSTM.py`, são exploradas as melhorias estruturais no script de treinamento do modelo preditivo, com foco na generalização do uso para diferentes ativos financeiros. Todo esse processo está ilustrado na figura 4.1. Por fim, a seção *Diferença de Escopo* discute a principal contribuição do trabalho, que consiste na adaptação metodológica da proposta original de previsão de preços de criptomoedas para o domínio das ações, ampliando a aplicabilidade e relevância do estudo no contexto financeiro tradicional.

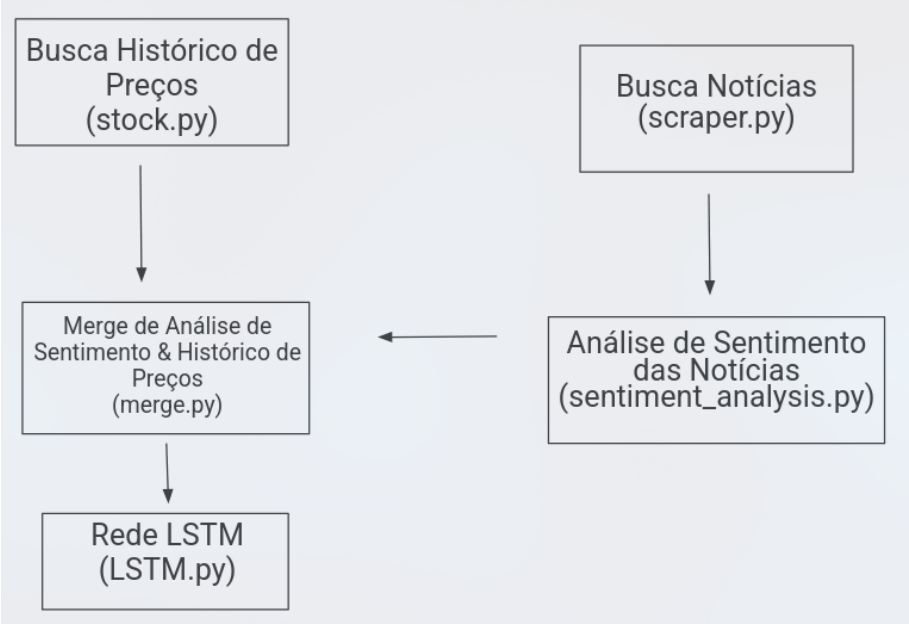


Figure 4.1: Processo de Previsão do Preço de Ações

4.1 SCRAPER.PY

O script `scraper.py`, desenvolvido com o suporte do módulo auxiliar `extract.js`, representa um avanço substancial em relação aos scripts anteriores

google_news_scraper.py e download_data_from_reddit.py, destacando-se por sua arquitetura modular, maior resiliência a falhas e capacidade de capturar textos mais completos e consistentes.

O google_news_scraper.py, apresentado por (Prajapati, 2020a) no repositório (Prajapati, 2020c), tem como finalidade a coleta de notícias diárias a partir de pesquisas no *Google News*. Para isso, emprega as bibliotecas requests, BeautifulSoup e newspaper3k na localização e extração de conteúdo textual dos artigos. No entanto, essa abordagem não contempla o tratamento de redirecionamentos de URLs — característica comum no *Google News* — além de depender integralmente da newspaper3k, cuja eficácia pode ser comprometida em páginas com carregamento dinâmico ou uso intenso de *JavaScript* (Prajapati, 2020a).

Já o download_data_from_reddit.py, disponível no repositório (Prajapati, 2020e), também de autoria de (Prajapati, 2020a), é voltado à coleta de postagens publicadas no *Reddit* que contenham termos específicos, como *"bitcoin"*, utilizando a *Application Programming Interface (API)* Pushshift. Embora eficiente na captura de metadados e conteúdo textual dos posts, o escopo deste script é restrito à plataforma *Reddit*, não incorporando fontes jornalísticas externas, o que limita sua aplicabilidade em análises que demandam maior diversidade de fontes e profundidade textual.

O scraper.py, por sua vez, propõe uma abordagem mais robusta e versátil ao realizar buscas diárias via *Really Simple Syndication (RSS)* do *Google News*, o que assegura maior estabilidade na coleta de dados. Para lidar com os redirecionamentos que comprometem o acesso ao conteúdo real das notícias, o script incorpora o uso do Selenium em modo *headless*, que automatiza a navegação até a URL final efetiva. Dessa forma, garante-se que a informação extraída corresponda fielmente ao conteúdo da página original.

A etapa de extração textual é delegada ao módulo externo extract.js, desenvolvido em Node.js. Esse módulo emprega o *Mercury Parser* para acessar o corpo do texto da notícia e realizar sua limpeza, removendo integralmente as marcações *HTML*. A separação clara entre as etapas de coleta, resolução de redirecionamentos e *parsing* de texto confere ao scraper.py um elevado grau de organização e manutenibilidade. Além disso, o script implementa paralelismo com ThreadPoolExecutor, otimizando o tempo de execução, e conta com mecanismos para evitar reprocessamentos, como logs de datas processadas e *flags* opcionais de controle (-resume, -force).

Dessa forma, o scraper.py supera as limitações identificadas nos scripts google_news_scraper.py (Prajapati, 2020c) e download_data_from_reddit.py (Prajapati, 2020e), apresentando-se como uma ferramenta mais confiável, flexível e eficiente para a coleta de textos jornalísticos. Sua arquitetura é especialmente adequada para aplicações em *análise de sentimentos (sentiment analysis)*, previsão de séries temporais e mineração de notícias financeiras.

4.2 SENTIMENT_ANALYSIS.PY

O script sentiment_analysis.py representa um aprimoramento substancial em relação aos códigos anteriores google_news_sentiment_analysis.py e reddit_post_sentiment_analysis.py, ambos disponibilizados por (Prajapati, 2020a) com o propósito de realizar *análise de sentimentos (sentiment analysis)* em dados oriundos de notícias e redes sociais. O primeiro, presente no repositório (Prajapati, 2020c), aplica-se especificamente a conteúdos do *Google News*, enquanto o segundo, disponível no repositório (Prajapati, 2020e), foca em postagens da plataforma *Reddit*.

Esses scripts foram concebidos com estruturações fixas, exigindo formatos rígidos nos arquivos de entrada e apresentando baixa flexibilidade para adaptação a diferentes tipos de dados. Em contraste, o `sentiment_analysis.py` adota uma arquitetura modular e genérica, o que possibilita sua aplicação em diversos contextos temáticos — como finanças, política ou saúde — sem necessidade de modificações estruturais. Essa generalização torna o novo script uma ferramenta mais versátil e reutilizável.

Do ponto de vista técnico, destaca-se a melhoria na organização interna do código e na adoção de boas práticas de desenvolvimento. O `sentiment_analysis.py` suprime mensagens desnecessárias e automatiza o download do léxico *Valence Aware Dictionary and sEntiment Reasoner (VADER)*, além de reformular a função `get_sentiment_val_for_flair()`, tornando-a mais segura e clara ao tratar casos de ausência de rótulos nos dados analisados.

Embora os três scripts empreguem os mesmos métodos — *Flair*, *TextBlob* e *VADER* — para a *análise de sentimentos*, eles diferem na forma de processamento temporal. O script `reddit_post_sentiment_analysis.py` (Prajapati, 2020a) utiliza uma granularidade horária, o que demanda uma etapa adicional de agregação por meio da função `bucketize_sentiment_report()`. Já os scripts `google_news_sentiment_analysis.py` (Prajapati, 2020a) e `sentiment_analysis.py` realizam as análises com base em datas diárias, tornando o fluxo mais direto.

Além disso, o novo script introduz funcionalidades inéditas, como os parâmetros `start_date`, que delimita o início da análise a partir de uma data específica, e `simulate`, que permite a simulação sem carregamento dos modelos reais — ambos ausentes nas versões anteriores. Esses recursos ampliam o controle sobre o processo de execução e são especialmente úteis em cenários de teste e depuração.

A padronização na nomenclatura das colunas de saída também foi preservada, com prefixos como `gnews_` ou `reddit_`, conforme a origem dos dados. Ademais, a função `clean_sentiment_report()`, presente nas três versões, permanece responsável pela remoção de duplicidades e organização cronológica dos registros gerados.

Conclui-se, portanto, que o script `sentiment_analysis.py` consolida e aprimora as soluções propostas nos repositórios anteriores de (Prajapati, 2020a), ao oferecer uma ferramenta mais robusta, flexível e adaptável à *análise de sentimentos* em séries temporais. Sua estrutura generalista o torna especialmente adequado para projetos acadêmicos e aplicações em larga escala.

4.3 STOCK.PY

O script `stock.py` apresenta uma solução simplificada para a coleta de dados históricos de ações por meio da biblioteca `yfinance`, bastante utilizada em contextos analíticos e educacionais. Sua estrutura prioriza a praticidade, com entrada de parâmetros via linha de comando (`-ticker`, `-start`, `-end`) e definição automática do intervalo temporal com base na extensão do período informado. Caso esse intervalo ultrapasse 730 dias, utiliza-se granularidade diária; caso contrário, aplica-se intervalo *intradiário* de uma hora, o que torna o script adequado a diferentes cenários de análise de séries temporais.

Em contraste, o script `download_data_from_binance.py`, disponibilizado por Prajapati (2020a), foi desenvolvido com foco na coleta de dados de criptomoedas por meio da *Application Programming Interface (API)* da *Binance*. Sua configuração exige autenticação com chaves de *API* e permite maior controle sobre o processo de extração, com funcionalidades

como divisão da coleta em lotes diários (`step`), ajustes na frequência de requisições (`pause`) e simulação de chamadas para testes. Essas características o tornam mais apropriado para uso em ambientes operacionais e projetos de larga escala.

No que se refere ao processamento dos dados, o `stock.py` realiza a limpeza e organização do conjunto extraído, descartando colunas irrelevantes como `Adj_Close`, renomeando os campos com base no *ticker* selecionado e mantendo apenas as informações essenciais para análise: `timestamp`, `open`, `high`, `low`, `close` e `volume`. Já o script de Prajapati opera com uma estrutura mais detalhada, típica de plataformas de criptoativos, contendo colunas como `quote_av`, `tb_base_av` e `trades`, além de lidar com `timestamps` em milissegundos e ajustar automaticamente os fusos horários entre *UTC* e *PST*.

Complementarmente, o `download_data_from_binance.py` incorpora rotinas auxiliares que expandem seu potencial de uso, como a concatenação de dados de diferentes ativos (`concat_binance_data`), eliminação de registros duplicados por índice temporal (`remove_dup_by_index`) e atualização incremental de bases consolidadas (`append_binance_data`). Esses recursos conferem ao script uma maior escalabilidade e automação, não contempladas na proposta mais enxuta do `stock.py`.

Conclui-se, portanto, que ambos os scripts são eficazes dentro de seus respectivos propósitos. O `stock.py` destaca-se por sua simplicidade e adequação a aplicações acadêmicas voltadas à análise de ações, enquanto o `download_data_from_binance.py`, conforme descrito por Prajapati (2020a), oferece uma infraestrutura mais robusta, voltada ao monitoramento e à análise avançada de mercados de criptomoedas.

4.4 MERGE.PY

O script `merge.py` apresenta uma versão mais aprimorada e generalizada do processo de junção de dados em comparação ao original `merge_data_files.py`, disponível no repositório de (Prajapati, 2020d). Enquanto o código de origem foi concebido para tratar exclusivamente da concatenação entre dados de preços de criptomoedas e sentimentos extraídos de fontes específicas, como *Google News* e *Reddit*, sua estrutura rígida limita a adaptabilidade a diferentes formatos e cenários experimentais. Em contrapartida, `merge.py` adota uma abordagem mais flexível e automatizada, capaz de lidar com múltiplas configurações de dados temporais.

Dentre as melhorias introduzidas, destaca-se a abstração da lógica de mesclagem por meio da função genérica `merge_price_news_sentiment()`, que se adapta dinamicamente à frequência dos dados — seja diária ou horária — por meio da função `pandas.infer_freq()`. Essa capacidade elimina a necessidade de replicações manuais dos dados de sentimento, como ocorre em `merge_data_files.py` (Prajapati, 2020a), onde os registros diários são artificialmente expandidos para cada hora do dia.

Outro aspecto relevante diz respeito ao tratamento das colunas temporais. O script `merge.py` detecta e ajusta automaticamente os nomes de colunas como `'timestamp'` ou `'date'`, além de padronizar a estrutura dos dados de sentimento para garantir compatibilidade no momento da junção. Essa lógica contrasta com a do script original, que pressupõe nomenclaturas fixas e exige transformações explícitas sobre os índices temporais.

A inclusão de uma interface de linha de comando com a biblioteca `argparse` também representa um avanço importante, permitindo que o usuário especifique os arquivos de entrada e saída diretamente pelo terminal. Isso confere maior versatilidade ao uso do script, superando a limitação presente em `merge_data_files.py` (Prajapati, 2020a), que opera com caminhos definidos diretamente no código-fonte.

Além dessas melhorias técnicas, observa-se também uma maior preocupação com a legibilidade e organização do código. Comentários explicativos, tratamento mais estruturado de exceções e eliminação de redundâncias tornam `merge.py` mais adequado a aplicações em ambientes de pesquisa e desenvolvimento que exigem manutenibilidade e integração eficiente com *pipelines* de processamento de dados.

Conclui-se, portanto, que o script `merge.py` representa uma evolução significativa em relação ao modelo original proposto por (Prajapati, 2020a), oferecendo uma solução mais robusta, reutilizável e alinhada às boas práticas de *ciência de dados (data science)* no contexto da previsão de preços com base em *análise de sentimentos (sentiment analysis)*.

4.5 LSTM.PY

O script `LSTM.py` representa uma evolução significativa do código original `Expr4_LSTM.py`, disponibilizado por Prajapati (2020a), promovendo não apenas melhorias estruturais e técnicas, mas também uma ampliação clara no escopo da aplicação. Enquanto o `Expr4_LSTM.py` foi concebido com foco exclusivo na previsão do preço do Bitcoin, utilizando dados de *sentimento (sentiment)* provenientes de plataformas como *Reddit* e *Google News*, combinados com indicadores de criptomoedas como *Ethereum* e *Litecoin*, o novo script expande essa abordagem para contemplar também ativos do mercado acionário tradicional, como ações da Apple (AAPL) e Tesla (TSLA), entre outros.

Essa ampliação de escopo é viabilizada pela generalização da entrada de dados. O `LSTM.py` permite que o usuário especifique dinamicamente o arquivo *Comma-Separated Values (CSV)* por meio de um argumento de linha de comando (`-csv`), adaptando-se automaticamente ao ativo analisado. A partir do nome do arquivo, o script identifica o *ticker* correspondente e monta as colunas de preços e volume de forma programática. Essa abordagem contrasta com a versão original, na qual essas colunas são pré-definidas e rigidamente atreladas ao mercado de criptomoedas.

Além disso, o novo script realiza a detecção automática da granularidade temporal dos dados (horária ou diária), ajustando o parâmetro `look_back` de maneira proporcional à resolução da série temporal. Tal funcionalidade torna o modelo mais adequado à variabilidade de frequência encontrada em dados de ações, ao passo que o `Expr4_LSTM.py` utiliza um valor fixo, assumindo implicitamente dados horários.

No que tange à configuração experimental e estrutura da rede, o `LSTM.py` oferece maior flexibilidade e controle. Parâmetros como número de épocas (`epochs`), tamanho do lote (`batch_size`) e proporção da base de treino são facilmente ajustáveis. Complementarmente, o script incorpora práticas modernas de controle do ambiente de execução, como verificação da presença de *Graphics Processing Unit (GPU)*, configuração do uso de memória e limitação do número de *threads*, contribuindo para maior estabilidade e eficiência em ambientes computacionais otimizados — elementos ausentes na versão original.

Por fim, destaca-se a automatização da nomenclatura dos experimentos e dos arquivos gerados, que passam a incorporar o nome do ativo e a frequência dos dados analisados. Essa organização favorece a replicação e análise dos resultados em múltiplos cenários. Em suma, ao substituir uma estrutura voltada exclusivamente à previsão do Bitcoin por uma ferramenta genérica e reutilizável para diferentes ativos financeiros, o `LSTM.py` amplia substancialmente o escopo metodológico proposto por Prajapati (2020a), tornando-se aplicável tanto ao mercado de criptomoedas quanto ao mercado de ações.

4.6 DIFERENÇA DE ESCOPO

Embora ambos os estudos — tanto o original de (Prajapati, 2020a) quanto o trabalho aqui desenvolvido — compartilhem a mesma métrica de avaliação e estrutura metodológica baseada em *Redes Neurais do tipo Long Short-Term Memory (LSTM)*, observa-se uma diferença fundamental no escopo de aplicação de cada um. O artigo *Predictive Analysis of Bitcoin Price Considering Social Sentiments* (Prajapati, 2020a), assim como os scripts a ele associados — tais como `download_data_from_binance.py`, `reddit_post_sentiment_analysis.py` e `Expr4_LSTM.py` — são voltados exclusivamente à previsão do preço de criptomoedas, com ênfase no Bitcoin. Essa previsão é construída a partir da integração de dados históricos de preço e *sentimento (sentiment)* extraídos de fontes como *Google News* e *Reddit*, visando compreender o impacto de manifestações sociais e midiáticas sobre os movimentos do mercado de criptoativos.

Em contraste, o presente trabalho amplia esse arcabouço metodológico ao aplicá-lo ao mercado de ações, concentrando-se na previsão de preços de ativos como Apple (AAPL) e Tesla (TSLA). Essa mudança de escopo é refletida em diversas adaptações técnicas, como a utilização do script `stock.py` para obtenção de dados históricos via *Application Programming Interface (API)* da *Yahoo Finance*, substituindo a *API* da *Binance* originalmente utilizada por Prajapati, bem como na modificação do script `LSTM.py`, que foi reestruturado para lidar com diferentes granularidades temporais e diferentes ativos financeiros.

Além disso, o novo conjunto de scripts introduz melhorias substanciais na coleta e análise de sentimentos. O `scraper.py`, ao incorporar o uso do `Selenium` e do *Mercury Parser*, supera limitações técnicas dos scripts `google_news_scraper.py` e `download_data_from_reddit.py` (Prajapati, 2020a), garantindo maior robustez e fidelidade na extração textual de notícias. A *análise de sentimentos (sentiment analysis)*, por sua vez, é generalizada com o script `sentiment_analysis.py`, que adota uma arquitetura modular e é capaz de processar diferentes tipos de fontes com flexibilidade, diferentemente das versões rigidamente acopladas a *Google News* ou *Reddit* propostas por Prajapati.

Portanto, a principal contribuição deste trabalho, em relação ao estudo original de (Prajapati, 2020a), reside na adaptação e expansão da metodologia de previsão com base em sentimentos para o contexto do mercado de ações. Essa mudança de domínio amplia significativamente a aplicabilidade do modelo, mantendo sua fundamentação técnica, mas adaptando sua arquitetura a um novo conjunto de dados, dinâmicas de mercado e exigências operacionais.

A fim de validar empiricamente essa nova proposta metodológica, serão apresentados a seguir os resultados obtidos a partir da aplicação experimental do modelo no contexto do mercado acionário. Essa etapa busca avaliar o desempenho preditivo da abordagem adaptada, utilizando dados reais de preços e sentimentos associados a ações específicas, permitindo assim uma análise comparativa e crítica da eficácia do modelo frente ao novo escopo adotado.

Com base na reformulação metodológica e nos aprimoramentos técnicos implementados, este trabalho contribui significativamente para o avanço das pesquisas em previsão de preços no mercado financeiro, ao demonstrar a viabilidade de aplicar modelos originalmente concebidos para criptomoedas em um novo domínio: o mercado acionário. Ao oferecer uma solução mais robusta, modular e generalista para a coleta, análise de sentimentos e integração de dados, a proposta aqui desenvolvida amplia o escopo de aplicação da arquitetura *LSTM*, permitindo sua adaptação a diferentes ativos e contextos de mercado. Dessa forma, este estudo não apenas reproduz, mas também estende e qualifica a abordagem de (Prajapati, 2020a), consolidando-se

como uma base sólida para futuras investigações e aplicações práticas na área de finanças computacionais.

5 RESULTADOS

Este capítulo apresenta os resultados obtidos com os experimentos realizados, descrevendo de forma detalhada as etapas que compõem o *pipeline* de previsão de preços de ações com e sem a utilização de dados de *sentimento (sentiment)* extraídos de notícias. O conteúdo está estruturado em seções específicas que abordam, respectivamente: a *organização dos dados*, a *coleta* e a *preparação* das informações utilizadas, a avaliação da *qualidade* desses dados, a *parametrização* adotada para medir o desempenho dos modelos, e os *resultados obtidos*. As seções finais se concentram na apresentação comparativa dos cenários *sem* e com *uso* de notícias, incluindo tabelas com os valores de *Root Mean Square Error (RMSE)* e gráficos que destacam os menores erros obtidos para as empresas Apple, Tesla e Electromed. Também são apresentados gráficos de previsão que ilustram a performance dos modelos sobre os dados reais, destacando as séries previstas nos conjuntos de treino e teste em relação à série histórica original. Esses gráficos auxiliam na análise visual da capacidade preditiva do modelo ao longo do tempo.

5.1 ORGANIZAÇÃO DOS DADOS

Os experimentos realizados neste trabalho foram aplicados a três empresas distintas: Apple, Tesla e Electromed. A escolha dessas empresas visou contemplar diferentes níveis de exposição midiática e disponibilidade de dados noticiosos, de forma a avaliar o impacto da presença ou ausência de informações de *sentimento (sentiment)* na performance dos modelos preditivos. A Apple representa uma empresa com um volume considerável de notícias, embora nem sempre com forte apelo midiático. A Tesla, por outro lado, é uma empresa altamente midiática, frequentemente presente em manchetes e com grande volume de cobertura jornalística. Já a Electromed foi escolhida por ser uma empresa com presença midiática extremamente limitada, contando com pouquíssimas ou nenhuma notícia relevante ao longo do período analisado. Essa diversidade de perfis permitiu explorar como a variabilidade na densidade informacional influencia a eficácia da *análise de sentimentos (sentiment analysis)* na previsão de preços de ações.

A organização do processo experimental foi estruturada em etapas independentes, porém integradas, com o objetivo de garantir modularidade, reprodutibilidade e clareza metodológica. As fases envolveram, inicialmente, a coleta de notícias e preços históricos para cada empresa, seguidas pela *análise de sentimentos* dos textos obtidos, a fusão dos dados temporais em um único conjunto e, por fim, a aplicação do modelo preditivo. Cada etapa foi concebida de forma autônoma, permitindo que as transformações e ajustes fossem feitas de maneira isolada, sem comprometer a consistência do fluxo completo. Essa separação entre as fases do experimento também contribuiu para uma execução mais organizada, facilitando tanto a replicação dos testes quanto a manutenção do código.

As notícias foram coletadas automaticamente por meio de consultas diárias a agregadores jornalísticos, e o conteúdo completo foi extraído com o auxílio de técnicas de *parsing* específicas para cada fonte. Já os dados de preços foram obtidos de serviços financeiros amplamente reconhecidos, com precisão temporal adequada ao intervalo de análise. Após a coleta, os textos foram submetidos a algoritmos de *análise de sentimentos* que atribuíram uma polaridade a cada notícia com base no seu conteúdo. Em seguida, esses valores foram agrupados e integrados às séries de preços por meio de procedimentos de fusão temporal. Ao final desse processo, os dados estavam prontos para alimentar os modelos baseados em Redes Neurais Recorrentes do tipo *Long Short-Term Memory (LSTM)*.

Por fim, os resultados foram organizados de forma a permitir a comparação entre os cenários com e sem a utilização de dados de *sentimento*. Para cada empresa e período de análise, foram realizadas múltiplas execuções dos modelos a fim de garantir estabilidade e confiabilidade estatística nas métricas de erro obtidas. Essa abordagem permitiu uma análise robusta dos impactos reais da inclusão de variáveis exógenas de natureza textual na previsão do comportamento do mercado acionário.

5.2 COLETA DOS DADOS

A coleta dos dados foi realizada por meio dos scripts `scraper.py` e `stock.py`. O script `scraper.py` foi responsável por buscar, diariamente, até nove notícias relacionadas a uma determinada empresa, utilizando como base o *Really Simple Syndication (RSS)* do *Google News* e técnicas de *web scraping* para extrair o conteúdo textual completo de cada matéria. Já o script `stock.py` realizou a obtenção do histórico de preços de ações a partir da biblioteca `yfinance`, com base no *ticker* informado pelo usuário. Essa separação clara entre os módulos de coleta permitiu automatizar e padronizar a extração de dados, assegurando consistência entre os períodos de análise de preços e de notícias.

5.3 PREPARAÇÃO DOS DADOS

Antes de alimentar os modelos, todos os dados foram devidamente preparados. Isso incluiu a padronização dos formatos de data, a sincronização temporal entre os históricos de preços e as notícias coletadas, a normalização das colunas numéricas e o tratamento de valores ausentes. Além disso, os dados de *sentimento (sentiment)* foram agregados de forma a manter apenas uma entrada por data, compatível com a granularidade dos dados de preços. Essas etapas de preparação foram fundamentais para garantir a consistência do *conjunto de dados (dataset)* final e possibilitar o correto funcionamento dos modelos baseados em Redes Neurais do tipo *Long Short-Term Memory (LSTM)* utilizados nas previsões.

5.4 QUALIDADE DOS DADOS

A fim de garantir a confiabilidade dos resultados obtidos, foi realizada também uma avaliação da qualidade dos dados utilizados ao longo do processo. Essa avaliação foi conduzida por meio de métricas descritivas e inspeções exploratórias, com o objetivo de verificar a completude, consistência e variabilidade dos dados coletados. No caso das séries de preços, analisou-se a ausência de lacunas temporais e a regularidade das amostragens. Para os dados de notícias e *sentimentos (sentiment)*, a qualidade foi verificada com base na densidade informacional por período (quantidade de notícias por dia), na coerência dos valores de *sentimento* atribuídos e na diversidade lexical dos textos extraídos. Além disso, foram observadas possíveis correlações entre os dados de *sentimento* e os preços das ações, com o intuito de identificar relações causais ou comportamentais. Essa abordagem permitiu não apenas garantir a integridade dos dados empregados nos experimentos, mas também compreender em que medida a qualidade desses dados influenciou o desempenho preditivo dos modelos.

5.5 PARAMETRIZAÇÃO

A parametrização utilizada para avaliar o desempenho dos modelos de previsão foi o *Root Mean Squared Error (RMSE)*. Essa métrica permite quantificar o erro médio entre os valores previstos e os valores reais, penalizando maiores discrepâncias.

5.6 RESULTADOS OBTIDOS

Os resultados obtidos dos experimentos estão nas tabelas abaixo. Para cada conjunto de dados, são apresentados os valores de *Root Mean Squared Error (RMSE)* mínimo, mediano e máximo, refletindo respectivamente o melhor, o desempenho intermediário e o pior resultado entre os testes realizados. Os dados utilizados abrangeram dois tipos de granularidade temporal, dependendo do período analisado: para o intervalo de 2020 a 2024, foram utilizados dados com frequência diária, enquanto para o período de 2024 a 2024, os dados foram coletados em frequência *intradiária*, com registros espaçados a cada 6 horas e janelas deslizantes de 30 minutos de intervalo. Essa estratégia permitiu avaliar o comportamento dos modelos em diferentes resoluções temporais, conciliando maior densidade de informação no curto prazo com maior amplitude histórica no longo prazo.

5.6.1 Sem Notícias

A Tabela 5.1 apresenta os resultados obtidos pelos modelos de previsão de preços de ações sem a inclusão de variáveis exógenas, como dados de notícias, considerando três empresas distintas — Apple, Tesla e Electromed — ao longo de dois intervalos temporais: um período mais extenso (2020–2024) e um período mais recente e concentrado (2024–2024). A avaliação dos modelos foi realizada por meio da métrica *Root Mean Squared Error (RMSE)*, sendo apresentados os valores mínimo, mediano, máximo, médio e o desvio padrão para cada combinação de empresa e período.

No caso da Apple, observou-se que o modelo apresentou desempenho razoável no período mais longo, com *RMSE* médio de 9,53 e desvio padrão de 1,34. O *RMSE* mínimo foi de 7,98, indicando que algumas execuções alcançaram bons resultados, embora o *RMSE* máximo de 12,16 revele a existência de limitações em outras execuções. Por outro lado, no período mais curto (2024–2024), o desempenho foi significativamente superior, com *RMSE* médio de 3,46 e desvio padrão de apenas 1,08. O *RMSE* mínimo de 2,22 demonstra que o modelo foi bastante eficaz nesse cenário. Esses dados sugerem que, para a Apple, modelos baseados apenas em preços históricos são mais precisos quando aplicados a janelas temporais mais recentes e condensadas, possivelmente devido à menor influência acumulada de variáveis externas ao longo do tempo.

A Tesla, por sua vez, apresentou os piores resultados gerais. No intervalo de 2020 a 2024, o modelo obteve *RMSE* médio de 19,11, com valores oscilando entre 17,14 (mínimo) e 21,01 (máximo), ainda que o desvio padrão tenha sido moderado (1,28). No período mais recente (2024–2024), os resultados pioraram em termos de instabilidade: o desvio padrão aumentou para 6,44, com um *RMSE* máximo de 31,38 — o maior de toda a tabela — e um *RMSE* médio de 20,03. Apesar da leve melhora no *RMSE* mínimo (11,88), os resultados revelam que a ação da Tesla é altamente volátil e sensível a fatores externos. Isso indica que modelos que se baseiam exclusivamente em séries históricas de preços não são suficientes para capturar a complexidade do comportamento dessa ação, evidenciando a necessidade de incorporar variáveis exógenas, como dados de sentimento e cobertura midiática.

Já a Electromed apresentou os melhores resultados entre todas as empresas analisadas. No período de 2020 a 2024, o modelo alcançou um *RMSE* médio de apenas 1,17, com valores variando de 0,90 (mínimo) a 1,41 (máximo) e um desvio padrão extremamente baixo (0,14), indicando alta estabilidade e previsibilidade. Mesmo no período mais curto (2024–2024), o desempenho manteve-se satisfatório, com *RMSE* médio de 3,31 e desvio padrão de 0,50. Esses resultados sugerem que o comportamento da ação da Electromed é mais regular e menos sujeito a flutuações imprevisíveis, o que torna os modelos baseados exclusivamente em dados históricos suficientemente eficazes para fins preditivos, mesmo na ausência de informações provindas de notícias.

Table 5.1: Desempenho RMSE para ações sem notícias

Empresa	Período	RMSE Mínimo	RMSE Mediano	RMSE Máximo	RMSE Médio	Desvio Padrão
Apple	2020–2024	7,9800	9,2900	12,1600	9,53	1,34
Apple	2024–2024	2,2200	3,5700	5,5600	3,46	1,08
Tesla	2020–2024	17,1400	19,0400	21,0100	19,11	1,28
Tesla	2024–2024	11,8800	18,3400	31,3800	20,03	6,44
Electromed	2020–2024	0,9000	1,1600	1,4100	1,17	0,14
Electromed	2024–2024	2,0400	3,3400	4,1500	3,31	0,50

Essa tendência é reforçada pela análise gráfica do *Root Mean Squared Error (RMSE)* mínimo por empresa e período. O Gráfico de Barras 5.1 evidencia visualmente o contraste entre os ativos: Tesla, em ambos os períodos, apresenta os maiores valores de *RMSE* mínimo, com destaque para o intervalo de 2020–2024, que atinge o pico da distribuição. A Apple aparece em posição intermediária, com *RMSEs* mínimos consideravelmente mais baixos, sobretudo no período curto. Por fim, a Electromed se destaca com os menores valores mínimos de erro em ambos os cenários, confirmando sua elevada previsibilidade. A disposição visual dos dados contribui para consolidar a interpretação de que ativos com menor exposição midiática e menor volatilidade — como a Electromed — são mais bem modelados por abordagens baseadas exclusivamente em séries temporais de preços.

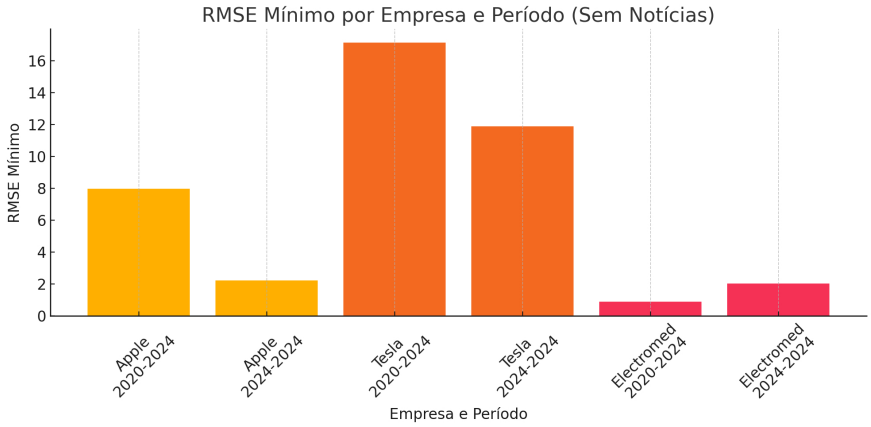


Figure 5.1: RMSE Mínimo por Empresa e Período — Sem Notícias

A Figura 5.2 ilustra o desempenho do modelo baseado em redes neurais *Long Short-Term Memory (LSTM)* no processo de previsão dos preços da ação da *Apple* (AAPL), considerando o cenário sem uso de dados de notícias e restrito ao ano de 2024. O gráfico apresenta a série temporal dos valores reais (linha azul), os valores previstos durante o treinamento (linha laranja) e os valores previstos na fase de teste (linha verde). Observa-se que o modelo é capaz de acompanhar de forma coerente a dinâmica dos preços reais, especialmente no período de teste, onde as previsões mostram boa aderência à trajetória observada. Esse comportamento sugere que o modelo conseguiu capturar adequadamente os padrões históricos da série de preços, mesmo na ausência de variáveis exógenas como o sentimento de notícias.

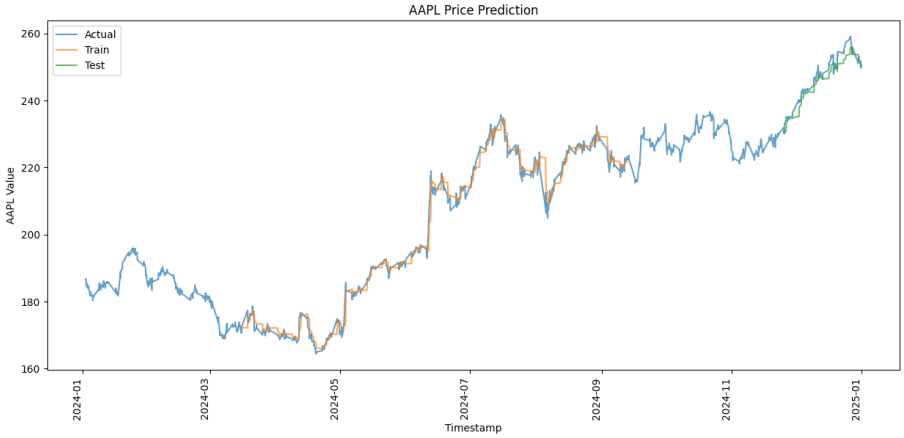


Figure 5.2: Treinamento da Rede com RMSE Mínimo Apple, Período 2024 - 2024 — Sem Notícias

A Figura 5.3 apresenta o desempenho do modelo baseado em redes neurais *Long Short-Term Memory (LSTM)* na previsão dos preços da ação da *Tesla* (TSLA), no ano de 2024, sem a utilização de variáveis de sentimento provenientes de notícias. O gráfico mostra as séries temporais dos valores reais (linha azul), das previsões durante o treinamento (linha laranja) e das previsões na fase de teste (linha verde). Nota-se que, apesar da alta volatilidade dos preços da Tesla, especialmente no último trimestre do ano, o modelo conseguiu acompanhar de forma satisfatória a tendência geral da série, com previsões bastante próximas dos valores reais, sobretudo no trecho de teste. Isso indica que, mesmo sem o apoio de variáveis externas, o modelo foi capaz de capturar padrões relevantes do comportamento do ativo.

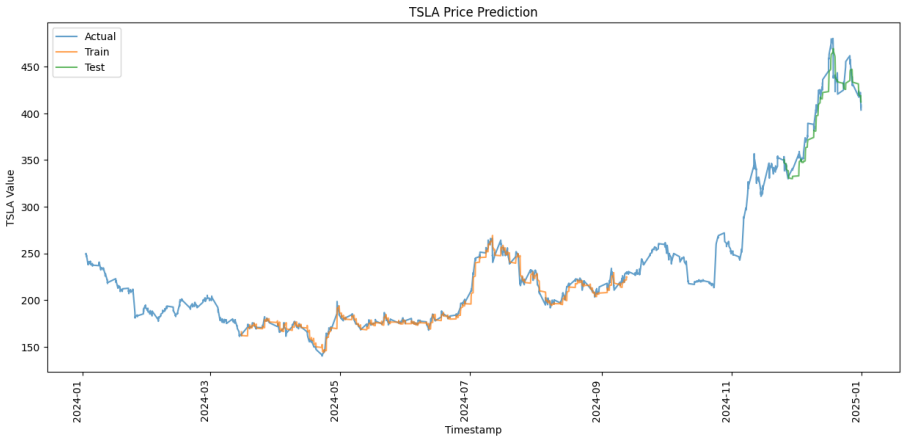


Figure 5.3: Treinamento da Rede com RMSE Mínimo Tesla, Período 2024 - 2024 — Sem Notícias

A Figura 5.4 exibe o desempenho do modelo baseado em redes neurais *Long Short-Term Memory (LSTM)* na previsão dos preços da ação da *Electromed* (ELMD) no intervalo de 2020 a 2024, desconsiderando o uso de variáveis de sentimento. O gráfico apresenta os valores reais (linha azul), os valores previstos durante o treinamento (linha laranja) e os valores de teste (linha verde). Nota-se uma forte aderência entre as séries prevista e observada, tanto no treinamento quanto na fase de teste, com destaque para a acurácia nas previsões mesmo durante a acentuada valorização da ação em 2024. A baixa volatilidade durante grande parte do período e a estabilidade estrutural da série podem ter contribuído para o bom desempenho do modelo, evidenciado pelo *Root Mean Squared Error (RMSE)* mínimo entre todos os experimentos analisados.

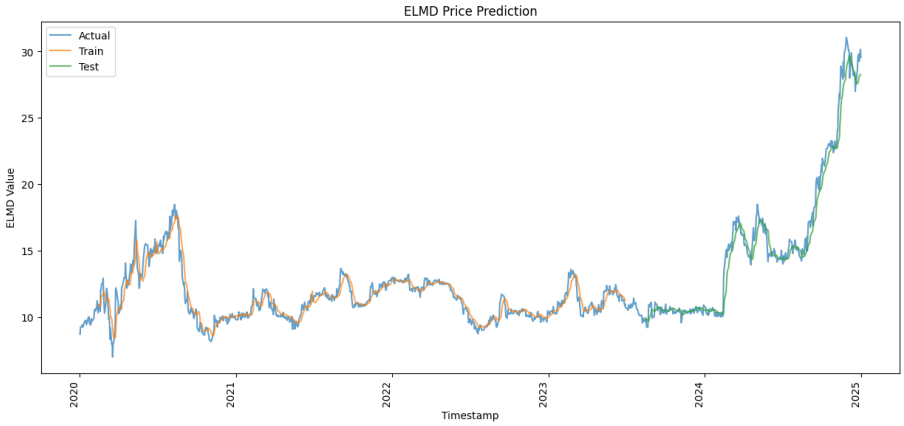


Figure 5.4: Treinamento da Rede com RMSE Mínimo Electromed, Período 2020 - 2024 — Sem Notícias

De modo geral, os dados demonstram que o desempenho dos modelos varia significativamente em função da ação analisada, do período de previsão e da ausência de variáveis exógenas. Enquanto ações mais estáveis, como a da *Electromed*, mantêm alta previsibilidade mesmo sem a utilização de dados de notícias, outras mais voláteis e midiáticas, como a *Tesla*, evidenciam limitações claras sob essas mesmas condições. Esses achados reforçam a importância de considerar as características específicas de cada ativo ao projetar modelos preditivos no contexto do mercado financeiro.

5.6.2 Com Notícias

A Tabela 5.2 apresenta os resultados dos modelos preditivos que consideram variáveis de sentimento oriundas de notícias, aplicados às ações das empresas *Apple* e *Tesla* em dois intervalos temporais distintos: longo prazo (2020–2024) e curto prazo (2024–2024). Para a avaliação do desempenho dos modelos, foi utilizada a métrica *Root Mean Squared Error (RMSE)*, expressa por seus valores mínimo, mediano, máximo, médio e pelo desvio padrão dos resultados experimentais.

No caso da *Apple*, os resultados mostram que a inclusão de notícias no período mais longo (2020–2024) contribuiu para uma redução significativa do RMSE mínimo, que passou de 7,98 (sem notícias) para 5,25 (com notícias), evidenciando que, em alguns experimentos, o modelo obteve melhor performance ao incorporar dados de sentimento. Entretanto, o RMSE mediano aumentou de 9,29 para 10,41 e o desvio padrão passou de 1,34 para 2,06, indicando que a variabilidade entre as execuções também se intensificou. Isso sugere que, embora as notícias tenham potencial para melhorar a previsão em alguns casos, também podem introduzir incertezas adicionais.

No período mais curto (2024–2024), os resultados para a *Apple* indicam que a inclusão de notícias não trouxe ganhos de desempenho. O RMSE médio aumentou de 3,46 (sem notícias) para 4,62 (com notícias), enquanto o RMSE máximo passou de 5,56 para 6,39. O RMSE mediano também subiu de 3,57 para 4,88, mantendo o desvio padrão praticamente inalterado. Esses resultados apontam que, para janelas temporais recentes e concentradas, o histórico de preços por si só oferece sinais suficientes para a previsão, e a introdução de dados de sentimento pode inclusive comprometer a eficácia do modelo ao introduzir ruído.

Em relação à *Tesla*, observa-se um comportamento distinto. No período de 2020–2024, houve uma leve melhora nos resultados com a inclusão de notícias: o RMSE médio caiu de 19,11 (sem notícias) para 18,66 (com notícias), e tanto o RMSE mínimo quanto o máximo apresentaram ligeiras reduções. O desvio padrão também permaneceu em níveis semelhantes (1,28 sem notícias e 1,80 com notícias). Essa melhora modesta sugere que a presença de variáveis exógenas contribuiu de forma pontual para aprimorar a capacidade preditiva do modelo em janelas mais amplas, ainda que os ganhos não tenham sido substanciais.

No entanto, o cenário para o período curto (2024–2024) revela o pior desempenho entre todos os experimentos realizados. A introdução de notícias nesse contexto resultou em um RMSE médio de 26,76, substancialmente superior ao valor observado sem notícias (20,03). O RMSE máximo atingiu 38,89 — o maior da tabela — e o desvio padrão aumentou consideravelmente, passando de 6,44 para 8,73. Além disso, o RMSE mediano saltou de 18,34 para 26,23. Esses dados indicam que, para a *Tesla*, a inclusão de variáveis de sentimento em janelas curtas compromete severamente a performance do modelo, tornando-o menos estável e mais suscetível a erros significativos.

Table 5.2: Desempenho RMSE para ações com notícias

Empresa	Período	RMSE Mínimo	RMSE Mediano	RMSE Máximo	RMSE Médio	Desvio Padrão
Apple	2020–2024	5,2500	10,4100	12,4600	9,69	2,06
Apple	2024–2024	3,0300	4,8800	6,3900	4,62	1,10
Tesla	2020–2024	16,4400	18,9400	21,8000	18,66	1,80
Tesla	2024–2024	13,1600	26,2300	38,8900	26,76	8,73

Essas observações são reforçadas pela análise gráfica dos valores mínimos de *Root Mean Squared Error (RMSE)* com inclusão de notícias. O Gráfico 5.5 evidencia que, mesmo nos melhores cenários de desempenho, os modelos aplicados à *Tesla* apresentam erros significativamente maiores do que aqueles aplicados à *Apple*. A ação da *Tesla* alcança um RMSE mínimo de 16,44 no período longo e 13,16 no curto, superando amplamente os valores da *Apple*, que obteve 5,25 e 3,03, respectivamente. Essa visualização torna explícito que, mesmo com a inclusão de dados de sentimento, o comportamento da ação da *Tesla* permanece de difícil modelagem, enquanto a *Apple* apresenta um padrão mais estável e previsível.

A Figura 5.6 apresenta os resultados do modelo *Long Short-Term Memory (LSTM)* treinado com dados da ação da *Apple* (AAPL) no ano de 2024, considerando a inclusão de variáveis de sentimento provenientes de notícias. O gráfico exibe os valores reais (linha azul), os valores previstos durante o treinamento (linha laranja) e as previsões na fase de teste (linha verde). Observa-se uma forte aderência entre os valores previstos e os valores reais, com o modelo capturando com precisão as oscilações do preço da ação ao longo do tempo. A presença dos dados de sentimento parece contribuir para um refinamento adicional nas previsões, especialmente no período de teste, onde a aproximação com a série real é bastante consistente.

A Figura 5.7 mostra o desempenho do modelo *Long Short-Term Memory (LSTM)* na previsão dos preços da ação da *Tesla* (TSLA) ao longo do ano de 2024, incorporando variáveis de sentimento extraídas de notícias. O gráfico exibe os valores reais (linha azul), as previsões obtidas

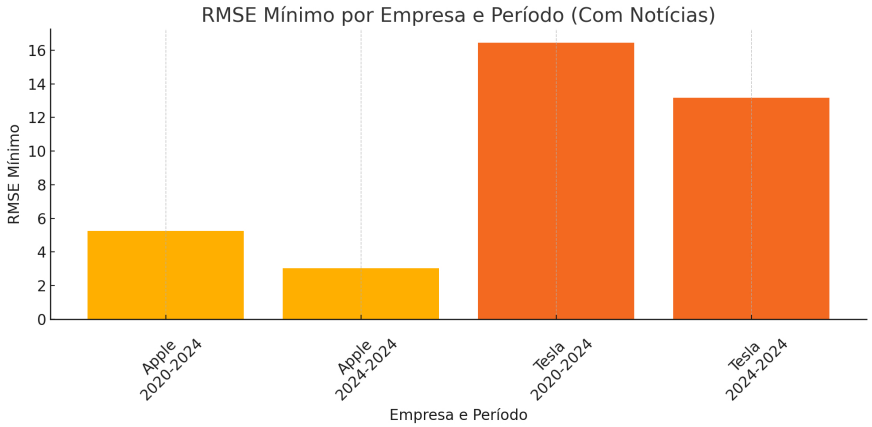


Figure 5.5: RMSE Mínimo por Empresa e Período — Com Notícias

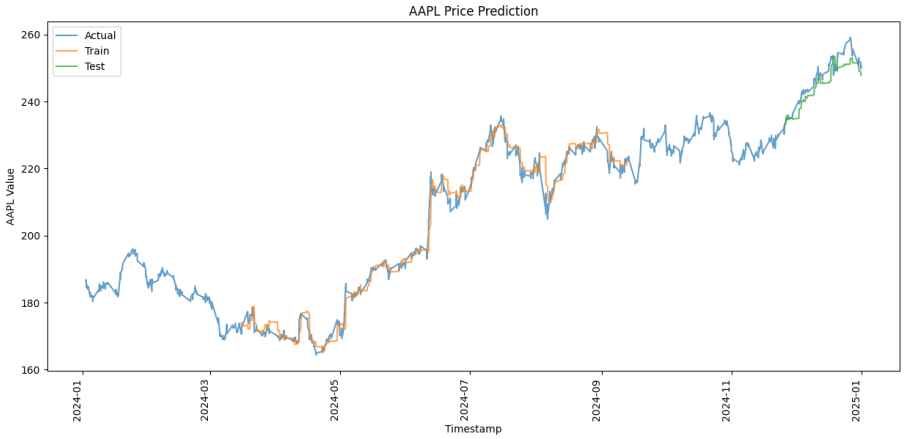


Figure 5.6: Treinamento da Rede com RMSE Mínimo Apple, Período 2024 - 2024 — Com Notícias

durante o treinamento (linha laranja) e os valores previstos na fase de teste (linha verde). Nota-se que o modelo consegue acompanhar de forma satisfatória a tendência dos preços, inclusive durante o forte movimento de alta no final do período analisado. A proximidade entre as curvas, especialmente na fase de teste, sugere que a inclusão de dados de sentimento pode ter contribuído positivamente para capturar mudanças abruptas no comportamento do ativo.

Dessa forma, os resultados demonstram que o impacto da análise de sentimentos na previsão de preços de ações depende diretamente do ativo analisado e do horizonte temporal. Enquanto, para a *Apple*, as notícias se mostram úteis em janelas longas e contraproducentes em períodos curtos, para a *Tesla* os efeitos são ainda mais pronunciados: as notícias pouco contribuem no longo prazo e prejudicam fortemente o desempenho no curto prazo. Essas evidências reforçam que a inclusão de variáveis exógenas deve ser cuidadosamente avaliada conforme o contexto da aplicação, o perfil da ação e os objetivos da previsão.

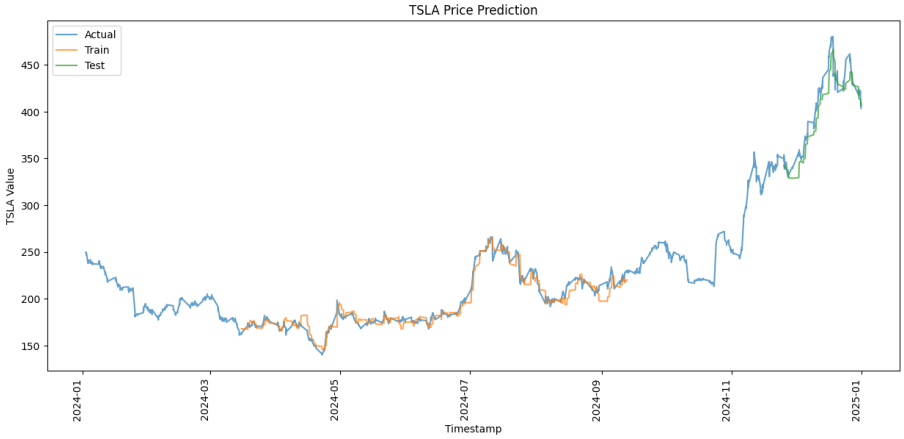


Figure 5.7: Treinamento da Rede com RMSE Mínimo Tesla, Período 2024 - 2024 — Com Notícias

6 CONCLUSÃO

Este trabalho apresentou uma adaptação metodológica de um modelo originalmente desenvolvido para previsão de preços de criptomoedas com base em sentimentos extraídos de notícias, aplicando-o ao contexto do mercado de ações. Inicialmente, a proposta foi contextualizada na introdução, onde se destacou que a eficácia do modelo *Long Short-Term Memory* (LSTM) depende fortemente das características do ativo, do horizonte temporal e da inclusão ou não de variáveis exógenas, como sentimentos oriundos de notícias. Os resultados empíricos demonstraram que, enquanto a inclusão de notícias melhora a performance em períodos longos — como no caso da *Apple* entre 2020 e 2024 —, esse benefício nem sempre se mantém em janelas temporais curtas, podendo inclusive gerar ruído, como observado em 2024–2024. A ação da *Electromed*, por sua vez, mostrou grande previsibilidade mesmo sem dados de sentimento, enquanto a *Tesla*, conhecida por sua volatilidade, apresentou maiores desafios, reforçando a importância de parametrizações customizadas para cada ativo e cenário.

Na *Revisão da Literatura*, argumentou-se que, embora diversos trabalhos apliquem redes neurais recorrentes, como *Long Short-Term Memory* (LSTM), à previsão de preços financeiros, poucos exploram a integração de sentimentos de maneira sistemática no contexto do mercado acionário. Por isso, o presente trabalho redirecionou o foco do estudo original de (Prajapati, 2020a), que atuava sobre criptomoedas, para ações listadas em bolsa, implementando uma série de aprimoramentos técnicos nos scripts envolvidos. Tais aprimoramentos incluíram a substituição das coletas baseadas em *scraping* por métodos mais robustos via RSS, melhorias no tratamento de redirecionamentos com *Selenium*, otimizações na análise de sentimentos com o modelo *Flair*, além de maior modularidade e parametrização na coleta, fusão e modelagem dos dados.

No capítulo *Trabalho Base*, foram descritos os repositórios que compõem a estrutura do projeto original, os quais forneceram inspiração e base metodológica para o presente estudo. Esses repositórios realizam a coleta e análise de sentimentos de fontes como *Reddit* e *Google News*, e o fornecimento de séries temporais de preços de criptomoedas, unificando os dados para alimentar um modelo *Long Short-Term Memory* (LSTM) de previsão. Essa estrutura foi mantida conceitualmente, mas reconfigurada para se adequar aos dados e à dinâmica do mercado de ações.

No capítulo *Trabalho Realizado*, detalhou-se como cada etapa foi reformulada para acomodar as novas exigências do domínio acionário. A principal contribuição foi justamente essa transposição de escopo, mantendo a essência técnica do modelo, mas adaptando-a para lidar com dados diferentes, fontes distintas e características operacionais próprias do novo mercado-alvo. Com isso, a aplicabilidade do modelo foi significativamente ampliada, abrindo novas possibilidades para análises futuras.

Por fim, o capítulo de *Resultados* evidenciou a importância de considerar diferentes granularidades temporais na avaliação do desempenho preditivo. Foram utilizados dados diários para o período de 2020 a 2024 e dados intradiários (com registros a cada 6 horas e janelas de 30 minutos) para o ano de 2024. Essa abordagem permitiu analisar os modelos em múltiplas escalas, fornecendo *insights* relevantes sobre a sensibilidade do desempenho em relação à frequência dos dados.

Como sugestão para trabalhos futuros, propõe-se a ampliação do modelo para considerar mais fontes de dados textuais, como *Twitter* e relatórios financeiros, além da exploração de arquiteturas mais complexas de redes neurais, como *Transformers*. Outra possível linha de pesquisa é a automatização do processo de seleção de variáveis e parametrização de modelos,

utilizando técnicas de *AutoML* (Automated Machine Learning) ou algoritmos genéticos. Também seria interessante expandir o número de ativos analisados e estudar a generalização dos modelos em contextos distintos, como mercados emergentes, setores específicos ou ativos de baixa liquidez. Essas perspectivas futuras podem contribuir para o desenvolvimento de sistemas de previsão mais robustos, adaptativos e sensíveis ao contexto do mercado financeiro.

REFERÊNCIAS

Al-Ali, A. M. and Al-Alawi, A. I. (2024). Stock market forecasting using machine learning techniques: A literature review. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, pages 466–471. IEEE.

Balasubramanian, P., Chinthan, P., Badarudeen, S., and Sriraman, H. (2024). A systematic literature survey on recent trends in stock market prediction. *PeerJ Computer Science*, 10:e1700.

Chollet, F. (2021). *Deep learning with Python*. simon and schuster.

Chong, E., Han, C., and Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, 83:187–205.

Chopra, R., Sharma, G. D., and Pereira, V. (2024). Identifying bulls and bears? a bibliometric review of applying artificial intelligence innovations for stock market prediction. *Technovation*, 135:103067.

Dao, H. N., ChuanYuan, W., Suzuki, A., Sudo, H., Ye, L., and Roy, D. (2024). Ai in stock market forecasting: A bibliometric analysis. In *SHS Web of Conferences*, volume 194, page 01003. EDP Sciences.

Dubey, A., Singh, S., Mishra, A. K., et al. (2024). A survey on machine learning techniques for stock market price prediction. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, pages 682–691. IEEE.

Feng, F., Chen, H., He, X., Ding, J., Sun, M., and Chua, T.-S. (2018). Enhancing stock movement prediction with adversarial training. *arXiv preprint arXiv:1810.09936*.

García-Medina, A. and Aguayo-Moreno, E. (2024). Lstm–garch hybrid model for the prediction of volatility in cryptocurrency portfolios. *Computational Economics*, 63(4):1511–1542.

Ghosh, P., Neufeld, A., and Sahoo, J. K. (2022). Forecasting directional movements of stock prices for intraday trading using lstm and random forests. *Finance Research Letters*, 46:102280.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Jain, S., Saluja, D. N., Pimplapure, D. A., and Sahu, D. R. (2024). Exploring the future of stock market prediction through machine learning: An extensive review and outlook. *International Journal of Innovative Science and Modern Engineering*, 12(4):1–10.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

Lin, C. Y. and Marques, J. A. L. (2024). Stock market prediction using artificial intelligence: A systematic review of systematic reviews. *Social Sciences & Humanities Open*, 9:100864.

Murray, K., Rossi, A., Carraro, D., and Visentin, A. (2023). On forecasting cryptocurrency prices: A comparison of machine learning, deep learning, and ensembles. *Forecasting*, 5(1):196–209.

Otabek, S. and Choi, J. (2024). From prediction to profit: A comprehensive review of cryptocurrency trading strategies and price forecasting techniques. *Ieee Access*, 12:87039–87064.

Oyewole, A. T., Adeoye, O. B., Addy, W. A., Okoye, C. C., Ofodile, O. C., and Ugochukwu, C. E. (2024). Predicting stock market movements using neural networks: A review and application study. *Computer Science & IT Research Journal*, 5(3):651–670.

Prajapati, P. (2020a). Predictive analysis of bitcoin price considering social sentiments. *arXiv preprint arXiv:2001.10343*.

Prajapati, P. (2020b). `cryptocurrency_data_downloader`. `https://github.com/pratikpv/cryptocurrency_data_downloader`.

Prajapati, P. (2020c). `google_news_scraper_and_sentiment_analyzer`. `https://github.com/pratikpv/google_news_scraper_and_sentiment_analyzer`.

Prajapati, P. (2020d). `predicting_bitcoin_market`. `https://github.com/pratikpv/predicting_bitcoin_market`.

Prajapati, P. (2020e). `reddit_scraper_and_sentiment_analyzer`. `https://github.com/pratikpv/reddit_scraper_and_sentiment_analyzer`.

Rahman, M. T. and Akhter, R. (2021). Forecasting stock market price using multiple machine learning technique. *Preprint*.

Sarker, P., Sayed, A., Apu, A. S., Tasnim, S. A., Mahmud, R., et al. (2024). A comparative review on stock market prediction using artificial intelligence. *Malaysian Journal of Science and Advanced Technology*, pages 383–404.

Sawhney, R., Agarwal, S., Wadhwa, A., and Shah, R. (2020). Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 8415–8426.

Swamy, S., Rajgoli, S., and Hegde, T. (2024). Stock market prediction with machine learning: A comprehensive review. *Indiana J. Multidiscip. Res*, 4:265–271.

Teixeira Zavadzki de Pauli, S., Kleina, M., and Bonat, W. H. (2020). Comparing artificial neural network architectures for brazilian stock market prediction. *Annals of Data Science*, 7(4):613–628.

Thakkar, A. and Chaudhari, K. (2024). Applicability of genetic algorithms for stock market prediction: A systematic survey of the last decade. *Computer Science Review*, 53:100652.

Wei, Y., Gu, X., Feng, Z., Li, Z., and Sun, M. (2024). Feature extraction and model optimization of deep learning in stock market prediction. *Journal of Computer Technology and Software*, 3(4).

APÊNDICE A – CÓDIGO TRABALHO REALIZADO

O código fonte para o trabalho realizado está disponível em: Application of LSTM Networks with Sentiment Analysis in Stock Price Prediction: An Adaptation of Cryptocurrency Market Models