

**APRENDIZADO PROFUNDO PARA PREVISÃO EM SÉRIES TEMPORAIS:
UM ESTUDO COMPARATIVO ENTRE MODELOS DE REDES NEURAI**

Relatório Científico Final do projeto na modalidade - Iniciação Científica e Tecnológica Sem Remuneração(ICTSR), fomentado pela Coordenadoria dos Programas de Iniciação Científica e Tecnológica.

Projeto CoPICT #2024/4720

Nome do Aluno: João Eduardo Batelochi Altarugio, Graduando
Nome do Orientador: Alexandre Luis Magalhaes Levada, Dr.

São Carlos, 1 de agosto de 2025

Sumário

| | | |
|----------|---|-----------|
| 1 | Resumo | 1 |
| 2 | Introdução | 1 |
| 2.1 | Objetivos | 2 |
| 3 | Metodologia | 3 |
| 3.1 | Dataset | 3 |
| 3.2 | Large Language Models (LLMs) | 5 |
| 3.3 | Convolutional Neural Networks(CNNs) | 7 |
| 3.4 | Long short-term memory(LSTM) | 8 |
| 4 | Resultados e Discussão | 10 |
| 4.1 | Cenário (i): Modelos Univariados | 10 |
| 4.2 | Cenário (ii): Modelos Multivariados | 11 |
| 5 | Conclusões | 12 |
| | Referências | 13 |
| | Autoavaliação do(a) Aluno(a) | 14 |
| | Apreciação do(a) Orientador(a) | 14 |

1 RESUMO

A análise de séries temporais financeiras é fundamental para a compreensão e previsão de comportamentos de mercado, oferecendo suporte à tomada de decisões estratégicas. Nesse contexto, técnicas baseadas em aprendizado profundo têm-se mostrado promissoras, devido à sua capacidade de capturar padrões complexos e não lineares presentes nos dados. Assim, este estudo explorou dois cenários de aprendizado profundo para previsão em séries temporais: (i) modelos univariados baseados em Redes Neurais Convolucionais (CNNs) e Redes de Memória de Longo Prazo (LSTMs), com janela de 5 dias para a previsão dos 5 dias seguintes; e (ii) um modelo multivariado LSTM enriquecido com variáveis qualitativas derivadas da análise de sentimentos de notícias macroeconômicas, processadas pelo modelo de linguagem *FinBERT-pt-Br* e coletadas da base pública GDELT, utilizando uma janela de 8 dias para prever o dia seguinte. Os experimentos foram conduzidos com uma base de dados composta por aproximadamente 3 mil registros diários do ETF BOVA11, coletados desde 2015 via Yahoo Finance, e mais de 300 mil notícias macroeconômicas associadas. Os resultados demonstraram que os modelos baseados exclusivamente em CNN e LSTM superaram o desempenho da linha de base, que apenas repetia a janela de dados de entrada. Além disso, verificou-se que a adição da análise de sentimentos contribuiu para a melhoria da capacidade preditiva do modelo multivariado.

2 INTRODUÇÃO

Séries temporais, compreendidas como sequências ordenadas de observações coletadas ao longo do tempo, apresentam grande relevância em domínios como demografia, climatologia e mercados financeiros, nos quais a antecipação de tendências futuras configura uma vantagem estratégica substancial [1]. No contexto financeiro, a previsão de preços de ativos revela-se uma tarefa complexa, caracterizada por alta volatilidade, não linearidade e presença de ruído nos dados. Diante desses desafios, optou-se por investigar modelos de aprendizado profundo (*deep learning*), dada sua eficácia na modelagem de padrões complexos em diferentes domínios aplicados [2].

Neste estudo, buscou-se avaliar o desempenho de diferentes arquiteturas de redes neurais profundas na previsão do preço de fechamento do Exchange Traded Fund (ETF) BOVA11, representativo do índice Ibovespa. Os dados históricos desse ativo, compostos por aproximadamente 3 mil registros diários desde 2015, foram extraídos da plataforma Yahoo Finance, que oferece uma API acessível para a coleta automatizada de dados financeiros.

Este trabalho aborda dois cenários distintos de modelagem.

No cenário (i), denominado cenário univariado, foram investigadas duas arquiteturas principais de redes neurais: Redes Neurais Convolucionais (CNNs) e Redes de Memória de Longo Prazo (LSTMs), utilizando exclusivamente a série temporal de preços de fechamento (*Close*) como entrada para os modelos. A configuração adotada consistiu em uma janela de entrada de 5 dias para prever os 5 dias seguintes. Como linha de base (*baseline*) para aferir o desempenho desses modelos, utilizou-se uma abordagem trivial que simplesmente repete os valores da janela de entrada como predição para os dias futuros. Essa comparação permite verificar se os modelos são capazes de aprender padrões relevantes, superando a simples persistência histórica.

No cenário (ii), com o objetivo de enriquecer os modelos com informações qualitativas, incorporou-se a análise de sentimentos extraída de notícias macroeconômicas. Essas informações foram utilizadas como variáveis adicionais em uma arquitetura baseada em LSTM, agora configurada como multivariada. Nesse modelo, adotou-se uma janela de entrada de 8 dias para

a previsão do dia subsequente, configuração distinta daquela empregada nas abordagens univariadas.

Para investigar se a inclusão da análise qualitativa das notícias de fato contribui para o aprimoramento do desempenho preditivo, explorou-se, como modelo comparativo dentro do mesmo cenário (ii), uma LSTM multivariada alimentada exclusivamente pela série de preços de fechamento. Essa estratégia permite isolar o efeito das variáveis semânticas derivadas das notícias, avaliando se sua incorporação resulta em ganhos efetivos de acurácia na previsão.

As notícias relacionadas à macroeconomia foram obtidas por meio do serviço Google Cloud BigQuery, utilizando a base de dados pública do projeto GDELT (*Global Database of Events, Language, and Tone*). O GDELT monitora continuamente fontes de notícias de todo o mundo, incluindo mídias impressas, televisivas e online, em mais de 100 idiomas, estruturando dados sobre eventos, atores envolvidos, temas abordados, emoções expressas e intensidade dos sentimentos. As informações são atualizadas em intervalos de aproximadamente 15 minutos, permitindo uma análise quase em tempo real. Essa abrangência e granularidade tornam o GDELT uma fonte especialmente valiosa para aplicações em análise de mercado, pois possibilita a incorporação de variáveis qualitativas atualizadas com elevada frequência e ampla cobertura geográfica e temática [3].

Para converter o conteúdo textual das notícias em variáveis quantitativas relevantes aos modelos preditivos, foi empregado o modelo de linguagem *FinBERT-pt-Br*, uma adaptação para a língua portuguesa do modelo FinBERT, originalmente desenvolvido para o domínio financeiro [4]. Modelos de Linguagem Massivos (LLMs — *Large Language Models*), como o FinBERT, são projetados para interpretar textos e extrair sentimentos associados [5]. O *FinBERT-pt-Br*, por sua vez, apresenta-se como especialmente adequado para essa tarefa por ter sido ajustado tanto ao contexto financeiro quanto ao idioma português. Com isso, tornou-se possível classificar cada notícia segundo seu conteúdo emocional (positivo, negativo ou neutro) e, a partir dessas classificações, gerar variáveis semânticas adicionais que enriqueceram a representação informacional dos dados, contribuindo para uma modelagem preditiva mais robusta [6].

As implementações foram desenvolvidas em linguagem Python, com o uso das bibliotecas Keras e PyTorch. Para a avaliação dos modelos, foi utilizada a métrica *Mean Absolute Error* (MAE), escolhida por sua simplicidade interpretativa e sensibilidade a desvios absolutos médios.

Conclui-se que o presente estudo possui relevância científica ao propor uma comparação sistemática entre diferentes abordagens de aprendizado profundo na tarefa de previsão de séries temporais financeiras, além de integrar informações qualitativas oriundas de fontes textuais. Os resultados obtidos permitiram avaliar a efetividade de cada abordagem em condições realistas de mercado, contribuindo para o avanço das aplicações de inteligência artificial em finanças.

2.1 Objetivos

Para viabilizar o desenvolvimento desta pesquisa, foram definidos os seguintes objetivos específicos:

- Coletar e processar dados financeiros do ETF Ibovespa por meio da plataforma Yahoo Finance;
- Coletar e processar notícias macroeconômicas utilizando o Google Cloud BigQuery do projeto GDELT;

- Aplicar e avaliar diferentes arquiteturas univariadas e multivariadas de redes neurais, incluindo LSTM (Long Short-Term Memory) e CNN (Redes Neurais Convolucionais), na previsão de séries temporais financeiras;
- Empregar Modelos de Linguagem Massivos (LLMs) na análise de sentimentos das notícias com o objetivo de aprimorar a capacidade preditiva das redes neurais profundas aplicadas às séries temporais;
- Disponibilizar um framework de código aberto no GitHub para reprodutibilidade e uso pela comunidade científica.

3 METODOLOGIA

Esta seção apresenta o método de pesquisa adotado neste estudo. As próximas subseções descrevem o conjunto de dados utilizado na avaliação experimental, tanto os dados de valor de fechamento do índice Bovespa quanto o conjunto de notícias, além de apresentar as arquiteturas abordadas para aprendizado profundo.

3.1 Dataset

Nesta subseção, são descritos os procedimentos de extração e pré-processamento dos dois conjuntos de dados utilizados nesta pesquisa: (i) a série temporal do índice Ibovespa e (ii) um conjunto de notícias macroeconômicas.

Em relação ao primeiro conjunto de dados, foi utilizada a API do Yahoo Finance, por meio da biblioteca `yfinance` (versão 0.2.63), para extrair os dados históricos do ticker `BOVA11.SA`, cobrindo o período de 2015 a 2025. Essa extração resultou em um total de 2.542 amostras diárias. Após a coleta, realizou-se a análise e remoção de valores nulos. Além disso, manteve-se apenas a coluna correspondente ao valor de fechamento (*Close*), por ser a variável de interesse para a análise. As demais colunas, como *Adj Close*, *High*, *Low*, *Open* e *Volume*, foram descartadas, resultando em uma série temporal univariada.

O resultado dessa extração pode ser visualizado na Figura 1.

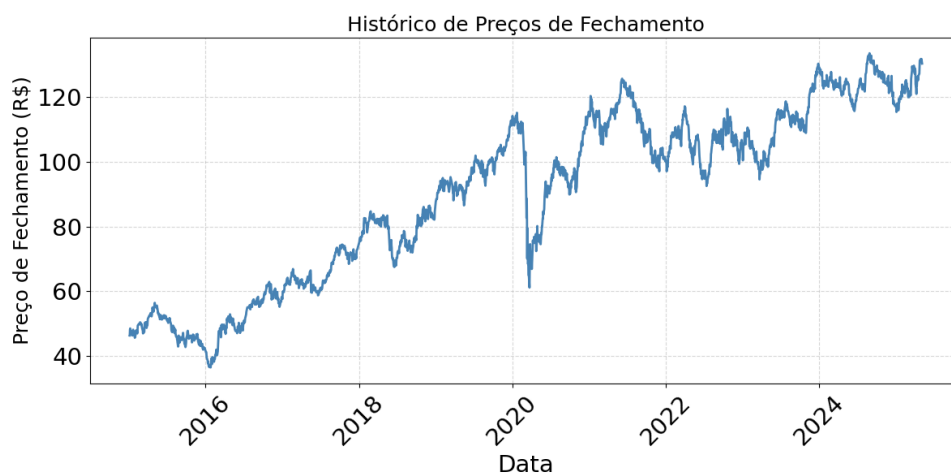


Figura 1: Série temporal do valor de fechamento diário do ETF BOVA11 (2015–2025).

Antes de serem utilizados como entrada para os modelos de aprendizado profundo, os dados de *Close* foram devidamente normalizados, de modo a garantir uma escala adequada e facilitar a convergência durante o treinamento. Optou-se pelo uso do *RobustScaler*, uma vez que séries temporais financeiras frequentemente apresentam distribuições assimétricas e não estacionaridade. Esse método de normalização oferece maior robustez frente a essas características, preservando a estrutura dos dados sem ser excessivamente influenciado por valores atípicos [7].

O segundo conjunto de dados utilizado neste estudo é composto por manchetes de notícias macroeconômicas. Para a obtenção desse material textual, foi utilizado o GDEL T (*Global Database of Events, Language, and Tone*), uma base de dados global que fornece cobertura abrangente de eventos e notícias de todo o mundo. O GDEL T é particularmente relevante por permitir a coleta em tempo quase real de eventos associados a variáveis políticas, econômicas e sociais, com uma ampla variedade de atributos e metadados.

Devido ao seu tamanho e complexidade, a utilização local do GDEL T é limitada. Apenas o conjunto de dados de 2015 ultrapassa 2,5 TB. Para viabilizar a manipulação em larga escala, o GDEL T é disponibilizado por meio do Google BigQuery, que permite consultas *ad hoc* com resultados em tempo quase real.

Dessa forma, a análise desenvolvida neste estudo baseou-se na extração de registros do período de 2015 a 2025, utilizando a plataforma Google BigQuery. Cada registro inclui informações como data, atores envolvidos, código do evento, escala de Goldstein (uma medida do impacto do evento) e a URL da fonte da notícia. Para focar em desenvolvimentos econômicos e políticos, os eventos foram filtrados utilizando códigos de evento específicos do GDEL T (faixa 100–199), que abrangem consultas, declarações e engajamentos diplomáticos ou econômicos. Além disso, a filtragem foi restrita à região geográfica do Brasil, com foco em notícias macroeconômicas relevantes para o índice Bovespa.

Para cada dia do período analisado, os eventos filtrados foram ordenados pelo número de artigos que os mencionavam, sendo retidas as 100 manchetes mais amplamente divulgadas. Essa heurística visa priorizar notícias com maior visibilidade e, portanto, com maior potencial de impacto sobre os preços dos ativos. As manchetes foram extraídas diretamente das URLs associadas aos eventos; casos em que a extração falhou ou resultou em texto vazio foram descartados. Esse processo resultou em um conjunto de até 100 manchetes macroeconômicas por dia, com alta cobertura midiática.

O resultado dessa consulta por notícias gerou um conjunto de dados com 317.732 linhas. A partir disso, foi realizado um pré-processamento textual. Inicialmente, selecionaram-se apenas as manchetes escritas em português ou em inglês, utilizando a biblioteca `langdetect`. Para os casos em que as manchetes estavam em inglês, aplicou-se a tradução automática para o português por meio da biblioteca `googletrans`. Essa etapa foi necessária, uma vez que o modelo de agregação de sentimento utilizado posteriormente foi treinado exclusivamente para o idioma português.

Em seguida, todo o texto foi convertido para letras minúsculas (*lowercase*), com o objetivo de eliminar distinções entre caracteres maiúsculos e minúsculos que poderiam influenciar negativamente a análise lexical. Também foram removidos os caracteres não alfanuméricos, com exceção dos espaços em branco e dos caracteres acentuados comumente encontrados na língua portuguesa, de modo a preservar a integridade semântica do texto.

Com o objetivo de refinar ainda mais a análise textual, concentrando-se nas notícias mais relevantes para o comportamento do mercado financeiro brasileiro — especialmente no que se refere ao índice Ibovespa —, adotou-se uma abordagem baseada em padrões de *regex* para a detecção de palavras-chave. Esse tipo de técnica é amplamente empregado em pesquisa sobre

mineração de texto em finanças, pois permite filtrar com precisão termos específicos (como nomes de índices, indicadores econômicos ou palavras-chave setoriais) e extrair informações qualitativas relevantes para modelagem preditiva [8].

Três conjuntos temáticos de padrões foram definidos: (i) termos amplos relacionados ao Ibovespa e ao mercado financeiro nacional e internacional; (ii) vocábulos associados à macroeconomia brasileira; e (iii) entidades políticas com impacto direto na economia. Cada conjunto abrange entidades institucionais, empresas listadas na bolsa, eventos econômicos, índices macroeconômicos, autoridades políticas e econômicas, além de conceitos-chave como 'inflação', 'taxa de juros', 'dólar', entre outros.

Após o pré-processamento descrito, obteve-se um conjunto final com 29.138 amostras, abrangendo o período de 2015 a 2025. Esse subconjunto representa uma seleção altamente refinada de manchetes, com foco específico em temas de macroeconomia brasileira e no índice Bovespa, livre de conteúdos ruidosos ou irrelevantes para os objetivos da pesquisa.

3.2 Large Language Models (LLMs)

Nesta etapa da metodologia, exploramos as particularidades da aplicação de modelos de linguagem de grande escala (LLMs) ao domínio financeiro brasileiro. O pipeline proposto tem início com a tokenização das manchetes em português, utilizando o tokenizador do modelo *FinBERT-PT-BR*. Esse processo inclui o truncamento e o preenchimento (*padding*) das sequências textuais, respeitando o limite de 512 tokens imposto pela arquitetura BERT.

Uma vez tokenizadas, as sequências são processadas pelo modelo *FinBERT-PT-BR*, que gera representações vetoriais contextuais com base na arquitetura BERT. Esse modelo é uma adaptação do *FinBERT* original, ajustado por meio de *fine-tuning* com textos financeiros em português do Brasil, com o objetivo de capturar nuances linguísticas e contextuais específicas do mercado nacional [4]. Tanto o modelo quanto seu tokenizador foram obtidos a partir do repositório *lucas-leme/FinBERT-PT-BR*, disponibilizado por meio da biblioteca *Transformers*, da Hugging Face.

O *FinBERT* é um modelo de linguagem baseado na arquitetura BERT (*Bidirectional Encoder Representations from Transformers*), uma das inovações mais significativas no campo do Processamento de Linguagem Natural (PLN). O BERT adota uma abordagem de pré-treinamento bidirecional, permitindo que o modelo considere simultaneamente os contextos à esquerda e à direita de cada palavra. Essa característica o torna especialmente eficaz na interpretação de relações semânticas complexas e na resolução de ambiguidades textuais, em contraste com modelos unidirecionais [9].

A arquitetura do BERT é composta por múltiplas camadas do tipo *Transformer encoder*, cujo componente central é o mecanismo de *atenção multi-cabeças* (*multi-head self-attention*). Esse mecanismo permite que o modelo identifique, para cada palavra de entrada, quais outras palavras no texto são mais relevantes para sua interpretação. A atenção é calculada por meio de projeções lineares que geram vetores de *query*, *key* e *value*, seguidas de operações de similaridade e combinações ponderadas, capturando interdependências contextuais de forma eficiente e paralelizável [10].

O BERT é pré-treinado em duas tarefas principais: *Masked Language Modeling (MLM)*, na qual algumas palavras da entrada são ocultadas aleatoriamente e o modelo deve prever essas lacunas; e *Next Sentence Prediction (NSP)*, onde o objetivo é aprender relações semânticas entre pares de sentenças. Após o pré-treinamento, o modelo pode ser ajustado (*fine-tuned*) para

tarefas específicas, como classificação de sentimentos, reconhecimento de entidades ou análise de similaridade textual.

Nesta pesquisa, o modelo *FinBERT-PT-BR* foi utilizado juntamente com seu sistema de segmentação lexical, sendo adaptado para processar sequências de texto e gerar representações numéricas associadas a diferentes classes de sentimento. Essa versão em português é particularmente relevante, pois o domínio econômico-financeiro brasileiro apresenta expressões idiomáticas, construções sintáticas e vocabulário próprios, frequentemente distintos dos corpora financeiros internacionais. O *fine-tuning* realizado com textos nacionais proporciona maior aderência semântica e melhora o desempenho do modelo em tarefas de análise de sentimento voltadas ao mercado brasileiro.

Como saída, o modelo retorna *logits* (pontuações não normalizadas), os quais são transformados em probabilidades por meio da função *softmax*, aplicada às classes. As três categorias previstas pelo modelo são: *Negativo* (P_{Neg}), *Neutro* (P_{Neu}) e *Positivo* (P_{Pos}). Além das probabilidades individuais, calcula-se um escore composto de sentimento, definido como:

$$\text{Sentimento Composto} = P_{Pos} - P_{Neg} \quad (1)$$

A partir dessas probabilidades e do escore composto, foram extraídas características adicionais para capturar a dinâmica temporal do sentimento nas notícias diárias. A composição de novas variáveis temporais foi baseada no artigo de Zhang [11].

A principal métrica é o Sentimento Médio diário, S_t , que representa a média dos escores de sentimento de todas as manchetes válidas (N_t) no dia t , calculado por:

$$S_t = \frac{1}{N_t} \sum_{i=1}^{N_t} s_{i,t} \quad (2)$$

onde $s_{i,t}$ é o escore individual da manchete i no dia t . Valores elevados de S_t indicam um sentimento positivo predominante, frequentemente associado a previsões de alta no mercado, enquanto valores baixos sugerem o contrário.

Além disso, foi computado o desvio padrão diário do sentimento, σ_t , que mensura a dispersão dos escores em um mesmo dia:

$$\sigma_t = \sqrt{\frac{1}{N_t} \sum_{i=1}^{N_t} (s_{i,t} - S_t)^2} \quad (3)$$

Altos valores de σ_t indicam diversidade ou conflito no sentimento das notícias, sinalizando menor consenso, o que tende a estar associado a movimentos de baixa no mercado. Já baixos valores refletem consenso claro, geralmente ligado a movimentos de alta.

Para capturar a persistência e a evolução temporal do sentimento, foram incorporadas defasagens de 1 a 3 dias dos valores de S_t e σ_t , representadas por S_{t-1} , S_{t-2} , S_{t-3} e σ_{t-1} , σ_{t-2} , σ_{t-3} . Essas defasagens permitem ao modelo considerar tendências recentes e possíveis reversões no sentimento agregado.

Também foram calculadas médias móveis de 5 e 20 dias do Sentimento Médio, denotadas por $MA5_t$ e $MA20_t$, definidas respectivamente como:

$$MA5_t = \frac{1}{5} \sum_{k=0}^4 S_{t-k}, \quad MA20_t = \frac{1}{20} \sum_{k=0}^{19} S_{t-k} \quad (4)$$

Essas médias suavizam as flutuações de curto prazo e destacam tendências mais robustas do sentimento ao longo de diferentes horizontes temporais.

A aceleração do sentimento foi então definida pela diferença entre essas médias móveis:

$$\Delta S_t = MA5_t - MA20_t \quad (5)$$

Um valor positivo de ΔS_t indica que o sentimento de curto prazo está superando o de longo prazo, sugerindo uma aceleração otimista. Valores negativos apontam para desaceleração, possivelmente indicando reversões ou fraquezas na tendência.

Por fim, a estabilidade do sentimento foi mensurada por meio dos desvios padrão móveis do Sentimento Médio em janelas de 5 e 10 dias, $rolling_std_5_t$ e $rolling_std_10_t$, que indicam a volatilidade do sentimento ao longo do tempo, com altos valores sugerindo instabilidade e baixos valores sinalizando consistência.

Para garantir a correta sincronização entre os valores diários de fechamento do índice e as variáveis de sentimento extraídas de mídias sociais e notícias, foi adotada a seguinte estratégia de alinhamento temporal.

Como as variáveis de sentimento podem estar associadas a datas não úteis (como finais de semana ou feriados), nas quais não há pregão por conta do fechamento da B3, cada data do conjunto de sentimentos foi mapeada para o próximo dia útil disponível de fechamento. Essa realocação assegura que informações qualitativas capturadas em dias sem mercado aberto sejam associadas ao próximo dia financeiro relevante. Posteriormente, os valores de sentimento realocados para o mesmo dia útil foram agregados pela média. Dessa forma, o conjunto final de dados apresenta alinhamento temporal preciso entre as variáveis quantitativas e qualitativas, permitindo que os modelos de previsão considerem ambas as fontes de informação em cada dia útil de mercado.

3.3 Convolutional Neural Networks(CNNs)

Nesta seção será tratado o uso de redes neurais profundas do tipo CNN (Convolutional Neural Networks). A CNN será utilizada apenas no cenário (i), ou seja, em um contexto univariado, considerando somente os valores de fechamento (*Close*). Essas redes utilizam operações de convolução para extrair características locais das sequências, o que pode ser útil na identificação de padrões temporais [12].

Um modelo de CNN padrão pode ser dividido em duas partes principais: a primeira compreende as camadas de convolução e pooling; a segunda envolve a aplicação de uma rede neural totalmente conectada (Multi-Layer Perceptron – MLP). Na primeira parte, a camada convolucional aplica filtros sobre a entrada para gerar mapas de características (*feature maps*) que descrevem as principais características do dado inicial. Esses filtros deslizam pela entrada, ajudando a destacar padrões específicos [13]. Geralmente, sobre o mapa de características, aplica-se uma função de ativação como a ReLU (Rectified Linear Unit), que zera os valores negativos, tornando o mapa mais esparsa e auxiliando a manter apenas as ativações mais relevantes.

Após a convolução, uma camada de pooling é aplicada, podendo ser do tipo *max pooling* ou *average pooling*. O *max pooling* mantém apenas o valor máximo dentro de uma região específica, enquanto o *average pooling* calcula a média dos valores nessa região. Essa etapa reduz a dimensionalidade e preserva as informações mais importantes, permitindo que o modelo reconheça padrões locais na entrada, independentemente de pequenas variações de posição.

Na segunda parte, a versão compactada dos dados é processada por uma rede neural totalmente conectada, que combina as características identificadas nas etapas anteriores para gerar a previsão final.

No contexto de séries temporais, a convolução é realizada sobre um vetor unidimensional (1D), em vez de uma matriz bidimensional (2D).

Em relação à arquitetura de CNN abordada neste trabalho, ela foi projetada para receber como entrada os valores de abertura dos últimos cinco dias úteis (uma semana), organizados em um vetor univariado com formato (5, 1), onde 5 corresponde ao número de dias e 1 representa a única variável utilizada (preço de fechamento). A arquitetura da CNN é composta por uma camada convolucional 1D, que aplica filtros sobre a sequência temporal para extrair padrões locais relevantes, seguida por uma camada de *max pooling*, que reduz a dimensionalidade, mantendo as características mais importantes e promovendo invariância a pequenas variações temporais. A saída da camada de pooling é então achatada (*flatten*) em um vetor unidimensional.

Esse vetor é alimentado em uma camada totalmente conectada, que combina as informações extraídas para gerar a previsão. Por fim, a camada de saída produz as previsões dos valores de abertura para os próximos cinco dias, em uma abordagem *multi-step walk-forward*. O modelo é treinado por 20 épocas, utilizando a função de ativação ReLU e o otimizador Adam para a atualização dos pesos.

Tal modelo foi baseado na arquitetura apresentada por Mehtab [14], utilizando a CNN que apresentou os melhores resultados para a tarefa de previsão multi-passo. Essa arquitetura também pode ser visualizada na Figura 2.

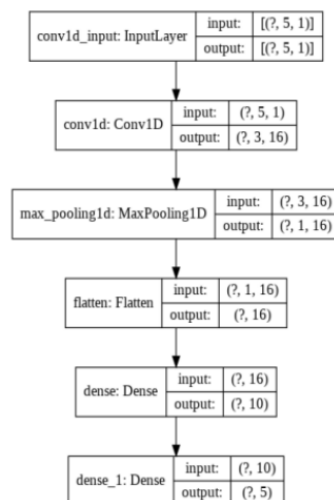


Figura 2: Arquitetura da rede neural convolucional (CNN) usada para previsão multipasso,

3.4 Long short-term memory(LSTM)

Nesta subseção, são abordadas duas arquiteturas distintas baseadas em LSTM. A primeira delas, associada ao Cenário (i), corresponde a uma abordagem univariada, na qual apenas o valor de fechamento da ação (*Close*) é utilizado como entrada do modelo. A segunda arquitetura, vinculada ao Cenário (ii), consiste em uma abordagem capaz de utilizar múltiplas variáveis como entrada, incluindo variáveis relacionadas à agregação de sentimento. Essa extensão visa

avaliar o impacto de informações adicionais de natureza qualitativa na previsão dos preços. Na abordagem univariada, utiliza-se uma janela de entrada de 5 dias para prever os valores de fechamento dos 5 dias subsequentes. Já na abordagem multivariada, considera-se uma janela de 8 dias como entrada para realizar a previsão de apenas 1 dia no futuro.

A arquitetura LSTM é composta por unidades especiais chamadas de células de memória, as quais facilitam a propagação dos gradientes ao longo do tempo, mitigando problemas como o desaparecimento do gradiente [15]. O funcionamento dessas células é governado por três portões multiplicativos: o portão de entrada, o portão de esquecimento e o portão de saída. Esses portões controlam, respectivamente, o fluxo de novas informações para a célula, o quanto da informação antiga será descartada e o quanto da informação armazenada será transmitida para os próximos passos. Esses mecanismos permitem atualizar o estado interno da célula com base em informações novas e passadas, bem como gerar a saída apropriada para o próximo passo temporal [16].

Essa estrutura possibilita que a LSTM armazene e manipule informações de forma eficaz ao longo do tempo, com maior resiliência a perdas de informação causadas por intervalos temporais longos. Entretanto, essa maior capacidade de retenção de informação tem como custo uma complexidade computacional mais elevada, já que a cada passo de tempo a rede realiza diversas operações envolvendo multiplicações de matrizes e ativações não lineares.

A arquitetura univariada adotada neste trabalho segue, mais uma vez, a estrutura proposta por Mehtab [14], utilizando a LSTM que apresentou os melhores resultados para a tarefa de previsão multi-passo. Nessa configuração, os dados de entrada têm formato (5, 1), correspondendo a cinco dias de preços de fechamento. Os dados passam por uma camada LSTM com 200 unidades, cuja saída é encaminhada para uma camada densa com 200 neurônios na entrada e 100 na saída. Por fim, a previsão para os cinco dias futuros é produzida por uma camada de saída com 100 neurônios de entrada e 5 de saída. Tal estrutura está ilustrada na Figura 3. O modelo é treinado por 20 épocas, utilizando a função de ativação *tanh*.

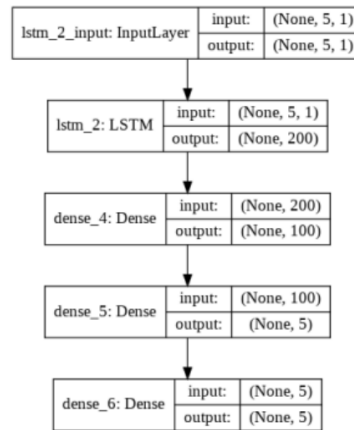


Figura 3: Arquitetura da rede neural recorrente (LSTM) univariada.

A arquitetura multivariada, por sua vez, foi baseada na utilizada no artigo [17], e incorpora um número maior de camadas LSTM. Especificamente, são utilizadas três camadas LSTM sequenciais: as duas primeiras com 50 neurônios cada, retornando sequências completas, e a terceira também com 50 neurônios, retornando apenas a saída do último passo temporal. Essa saída é conectada a uma camada totalmente conectada com um único neurônio, responsável por

gerar a previsão final para o próximo dia. O modelo é treinado por 100 épocas, utilizando o otimizador Adam. A Figura 4 ilustra essa estrutura.

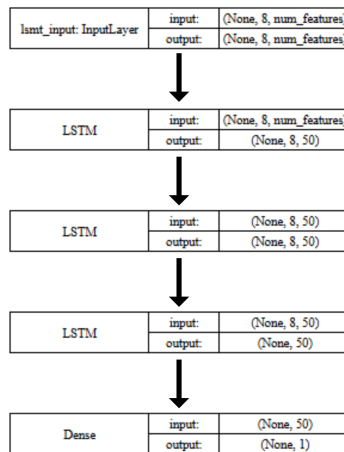


Figura 4: Arquitetura da rede neural recorrente (LSTM) multivariada.

4 RESULTADOS E DISCUSSÃO

Nesta seção, são apresentados e discutidos os resultados obtidos a partir da aplicação dos modelos em dois cenários distintos: (i) um cenário univariado, em que apenas o valor de fechamento do índice Bovespa (*Close*) é utilizado como entrada; e (ii) um cenário multivariado, no qual são incorporadas variáveis adicionais relacionadas ao sentimento extraído de mídias sociais e notícias, visando enriquecer a previsão com informações qualitativas.

Para avaliar o desempenho dos modelos, utilizou-se como métrica principal o erro absoluto médio (MAE – *Mean Absolute Error*), que representa a média das diferenças absolutas entre os valores previstos e os valores reais. Essa métrica foi escolhida por sua interpretabilidade direta no contexto financeiro, pois expressa o desvio médio entre a previsão e o valor observado [6].

Os dados foram divididos em conjuntos de treinamento, validação e teste, seguindo a mesma proporção para ambos os cenários. A divisão considerou 70% dos dados para treinamento, 20% para validação e 10% para teste. Essa abordagem garante que o modelo seja treinado com uma ampla base de dados, ajustado e calibrado com dados de validação, e finalmente avaliado em dados não vistos para medir seu desempenho real.

Com o objetivo de promover a reprodutibilidade dos experimentos e incentivar o uso e aprimoramento da metodologia por parte da comunidade científica, todo o código-fonte desenvolvido neste trabalho foi disponibilizado como software de código aberto em repositório público. A implementação completa, incluindo os scripts de pré-processamento, modelagem e avaliação, pode ser acessada no GitHub¹.

4.1 Cenário (i): Modelos Univariados

No cenário univariado, foram aplicadas redes LSTM e CNN utilizando como entrada exclusivamente os valores de fechamento do índice Bovespa. A janela de entrada utilizada foi de 5 dias, com o objetivo de prever os valores dos 5 dias subsequentes.

¹Repositório disponível em: <https://github.com/JoaoAltarugio/Iniciacao-Cientifica---CoPICT---4720>

Como baseline para este cenário, foi adotada uma abordagem simples: a repetição da própria janela de entrada como previsão para os próximos 5 dias. Ou seja, o modelo base simplesmente repete os últimos valores observados, assumindo que o comportamento futuro será idêntico ao mais recente. Os resultados obtidos por cada modelo podem ser visualizados na Figura 5.

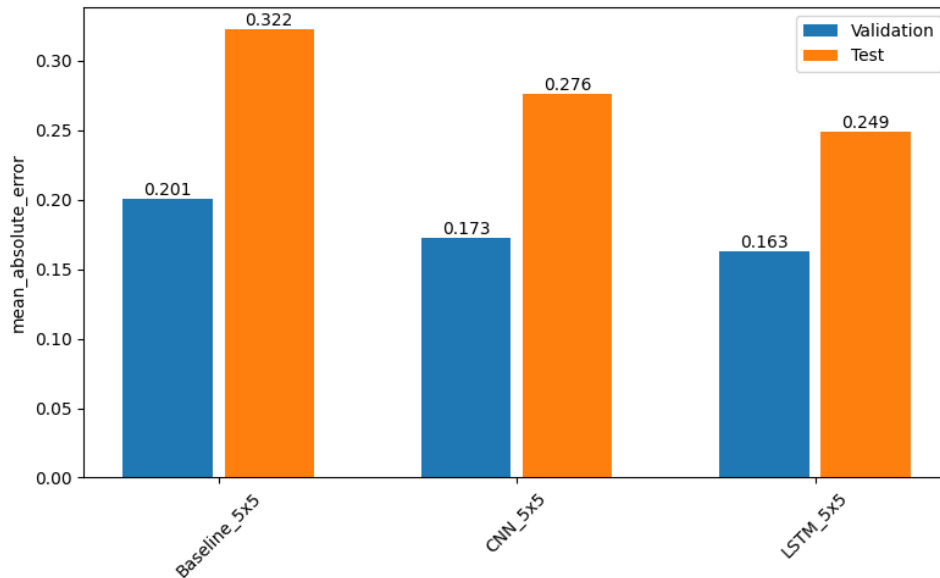


Figura 5: Comparação do desempenho dos modelos LSTM, CNN e da baseline no cenário univariado, com base na métrica MAE.

Os resultados evidenciam que tanto a rede LSTM quanto a CNN superaram o desempenho da baseline, indicando que os modelos foram capazes de aprender padrões relevantes dos dados históricos e realizar previsões mais precisas do que uma abordagem trivial.

Entre os modelos analisados, a LSTM apresentou o menor valor de MAE, superando inclusive a CNN. Esse resultado era esperado, uma vez que a LSTM é projetada especificamente para lidar com dependências temporais de longo prazo, característica fundamental em séries temporais financeiras. Sua estrutura com células de memória permite capturar tendências e correlações que se estendem por períodos mais longos. Por outro lado, embora a CNN seja eficaz na detecção de padrões locais e variações rápidas, apresenta limitações para modelar relações temporais de maior abrangência.

4.2 Cenário (ii): Modelos Multivariados

Neste segundo cenário, optou-se por utilizar apenas a rede LSTM, dada sua capacidade superior em modelar dependências temporais de longo prazo em séries financeiras. Diferentemente do cenário anterior, que utilizava apenas o valor de fechamento do índice Bovespa como entrada, aqui foram utilizadas 14 variáveis como entrada para o modelo, incluindo a série temporal dos valores de fechamento e as variáveis de sentimento definidas na Seção 3.2.

O objetivo deste experimento foi avaliar se a inclusão de variáveis de sentimento realmente contribui para melhorar o desempenho preditivo da LSTM. Aproveitando o fato de que a arquitetura multivariada da LSTM permite a inserção de uma ou mais variáveis de entrada, foi

realizada uma comparação com uma versão da mesma arquitetura recebendo apenas a variável *Close*. Dessa forma, foi possível isolar o efeito das variáveis de sentimento e verificar se elas agregam valor ao modelo. Os resultados podem ser visualizados na Figura 6.

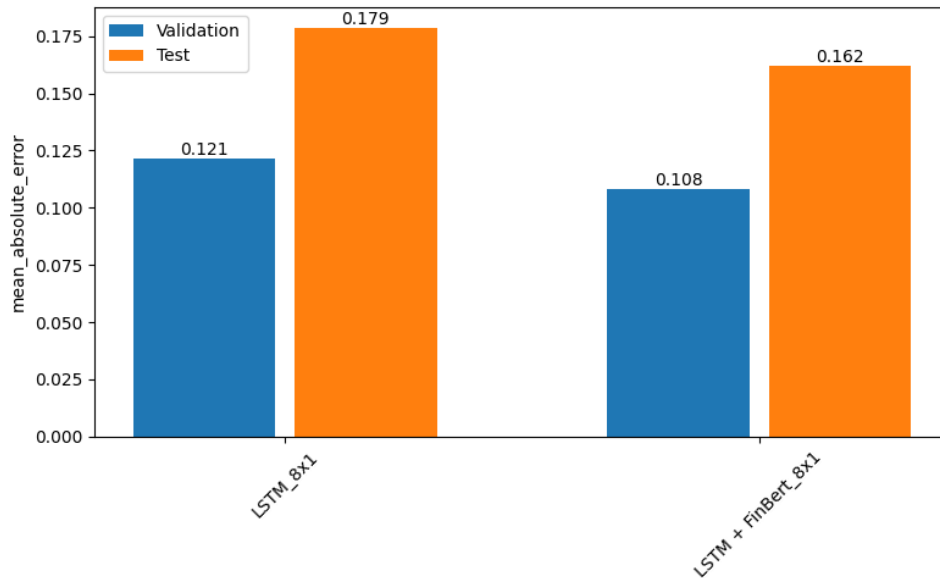


Figura 6: Desempenho da LSTM multivariada com e sem variáveis de sentimento,

É possível observar que a LSTM com variáveis de sentimento obteve resultados significativamente melhores do que sua versão sem essas variáveis. Isso indica que o uso de informações qualitativas extraídas de mídias sociais e notícias auxilia o modelo a capturar aspectos do comportamento do mercado que não estão presentes apenas nos dados históricos numéricos.

Além disso, a versão da LSTM multivariada que utilizou exclusivamente a variável *Close* como entrada ainda assim superou os modelos univariados aplicados no cenário (i). Esse resultado reforça que a arquitetura multivariada da LSTM possui maior capacidade de modelar séries temporais financeiras, mesmo quando utilizada com uma única variável de entrada. Tal desempenho superior pode estar relacionado à sua estrutura composta por múltiplas camadas LSTM conectadas em série, em contraste com a arquitetura univariada, que conta com apenas uma camada LSTM. Essa configuração empilhada favorece a extração de padrões temporais mais complexos e profundos, o que pode ter contribuído significativamente para os melhores resultados observados.

5 CONCLUSÕES

Como esperado, o modelo LSTM agregado às variáveis de sentimento extraídas por meio da análise semântica apresentou o melhor desempenho na métrica MAE. Além disso, observou-se que a arquitetura multivariada superou a univariada, mesmo sem a inclusão das variáveis de sentimento, evidenciando sua maior capacidade de modelagem. Esses resultados demonstram que a escolha adequada da arquitetura, em função do cenário e da natureza dos dados disponíveis, desempenha um papel fundamental no desempenho dos modelos preditivos, sendo crucial

considerar o nível de complexidade necessário para capturar os padrões relevantes em séries temporais financeiras.

Como trabalhos futuros, seria interessante investigar o impacto de outras arquiteturas de redes neurais recorrentes, como GRU, LSTM bidirecional, entre outros modelos. Além disso, também seria relevante analisar como a inclusão de outras variáveis disponíveis pela API do Yahoo Finance, como *High*, *Low*, *Volume* e *Open*, pode influenciar a capacidade preditiva dos modelos.

REFERÊNCIAS

- [1] MALIK, P. et al. An analysis of time series analysis and forecasting techniques. *International Journal of Advance Research and Innovative Ideas in Education*, v. 9, n. 5, p. 1364–1374, 2023.
- [2] BENGIO, Y. Deep learning of representations for unsupervised and transfer learning. In: JMLR WORKSHOP AND CONFERENCE PROCEEDINGS. *Proceedings of ICML workshop on unsupervised and transfer learning*. [S.l.], 2012. p. 17–36.
- [3] THE GDELT Project: Global Database of Events, Language, and Tone. 2025. <https://www.gdeltproject.org/>. Acesso em: jul. 2025.
- [4] SANTOS, L. L.; BIANCHI, R. A.; COSTA, A. H. Finbert-pt-br: Análise de sentimentos de textos em português do mercado financeiro. In: SBC. *Brazilian Workshop on Artificial Intelligence in Finance (BWAIF)*. [S.l.], 2023. p. 144–155.
- [5] ARACI, D. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.
- [6] HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. *International journal of forecasting*, Elsevier, v. 22, n. 4, p. 679–688, 2006.
- [7] Scikit-learn developers. *RobustScaler - scikit-learn documentation*. 2025. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html>. Acesso em: 31 jul. 2025.
- [8] NIKFARJAM, A.; AL. et. Comprehensive review of text-mining applications in finance. *Financial Innovation*, SpringerOpen, v. 6, n. 1, 2020.
- [9] DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. [S.l.: s.n.], 2019. p. 4171–4186.
- [10] VASWANI, A. et al. Attention is all you need [j]. *Advances in neural information processing systems*, v. 30, n. 1, p. 261–272, 2017.
- [11] ZHANG, Y. Interpretable machine learning for macro alpha: A news sentiment case study. *arXiv preprint arXiv:2505.16136*, 2025.
- [12] LI, Z. et al. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, IEEE, v. 33, n. 12, p. 6999–7019, 2021.

- [13] LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Ieee, v. 86, n. 11, p. 2278–2324, 2002.
- [14] MEHTAB, S.; SEN, J.; DASGUPTA, S. Robust analysis of stock price time series using cnn and lstm-based deep learning models. In: IEEE. *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. [S.l.], 2020. p. 1481–1486.
- [15] NICHOLSON, C. *A beginner's guide to LSTMs and recurrent neural networks (2019)*. 2019.
- [16] KILIAN, J.; SIEGELMANN, H. T. On the power of sigmoid neural networks. In: *Proceedings of the sixth annual conference on Computational learning theory*. [S.l.: s.n.], 1993. p. 137–143.
- [17] GU, W. jun et al. Predicting stock prices with finbert-lstm: Integrating news sentiment analysis. In: *Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing*. [S.l.: s.n.], 2024. p. 67–72.

AUTOAVALIAÇÃO DO(A) ALUNO(A)

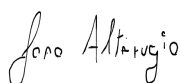
A participação no programa de Iniciação Científica foi extremamente enriquecedora, tanto no aspecto técnico quanto no desenvolvimento pessoal. Um dos principais desafios enfrentados foi encontrar uma maneira eficiente de coletar um grande volume de notícias relevantes utilizando apenas plataformas gratuitas, o que exigiu diversas tentativas e adaptações até chegarmos a uma solução viável com o uso do GDELT e do Google BigQuery. Além disso, também foi desafiador selecionar arquiteturas de redes neurais que atendessem aos requisitos específicos da pesquisa, especialmente em relação à previsão em séries temporais com dados financeiros e de sentimento. Apesar dessas dificuldades, o projeto proporcionou uma oportunidade valiosa de aprendizado prático em ciência de dados, processamento de linguagem natural e aprendizado profundo, além de ter ampliado meu entendimento sobre a importância da experimentação e da resiliência no processo científico.

APRECIÇÃO DO(A) ORIENTADOR(A)

O aluno demonstrou grande interesse e dedicação ao projeto de iniciação científica, tendo conseguido desenvolver com sucesso os métodos propostos. Os resultados obtidos indicam que a abordagem investigada é promissora. Na minha opinião, o desempenho do aluno foi muito bom.

João Eduardo Batelochi Altarugio
Assinatura do(a) Aluno(a)

Prof. Dr. Alexandre Luis Magalhaes Levada
Assinatura do(a) Orientador(a)



Digitally signed via ZapSign by
Joao Eduardo Batelochi Altarugio
Date 01/08/2025 12:11:06.290 (UTC-0300) *Alexandre Luis Magalhaes Levada*

Digitally signed via ZapSign by
Alexandre Luis Magalhaes Levada
Date 01/08/2025 12:11:56.339 (UTC-0300)

Signatures Log

Dates and times in UTC-0300 (America/Sao_Paulo)

Last updated on August 1 2025, 12:11:57

Status: Signed

Document: Relatório_Final_IC - João Eduardo Batelochi Altarugio_VERSAO_02.Pdf

Number: 414b71db-5c5b-4dda-8e01-7d9a46055322

Date of creation: August 1, 2025, 12:08:12

Original document hash (SHA256): 9921b02dab2cf0a33959f462072ae683ae28a1324b9f3fc7af59a75e49a618d2



Signatures

2 de 2 Signatures

| | | |
|--|--|--|
| <div>Signed  via ZapSign by Truora</div> <div>JOAO EDUARDO BATELOCHI ALTARUGIO Date and time of signature: 08/01/2025 12:11:06 Token: 06a2a5a3-e7e4-4032-a268-00ab73df8b45</div> | | Signature  Joao Eduardo Batelochi Altarugio |
| <div>Authentication points:</div> <div>Phone: + 5519998160910 Email: joao.altarugio@estudante.ufscar.br Security Level: Validado por código único enviado por e-mail</div> <div>IP: 138.122.166.253 Device: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/138.0.0.0 Safari/537.36</div> | | |
| <div>Signed  via ZapSign by Truora</div> <div>ALEXANDRE LUIS MAGALHAES LEVADA Date and time of signature: 08/01/2025 12:11:56 Token: 9cfd9315-51d4-4aae-be60-1747002aba1d</div> | | Signature  Alexandre Luis Magalhaes Levada |
| <div>Authentication points:</div> <div>Phone: + 5519988104993 Email: alexandre.levada@ufscar.br Security Level: Validado por código único enviado por e-mail</div> <div>IP: 187.183.59.200 Device: Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/138.0.0.0 Safari/537.36</div> | | |

CERTIFIED INTEGRITY - ICP-BRASIL

Electronic and physical signatures have equal legal validity, according to MP 2.200-2/2001 e Lei 14.063/2020.

[Confirm the integrity of the document here.](#)



This Log is exclusive and an integral part of document number 414b71db-5c5b-4dda-8e01-7d9a46055322, according to the [ZapSign Terms of Use](#), available at zapsign.com.br

ZapSign 414b71db-5c5b-4dda-8e01-7d9a46055322