

Previsão de Preços de *Commodities* Utilizando Inteligência Artificial: Integração de Séries Históricas e Análise de Notícias com Redes LSTM

Wenzo Rithelly Carvalho França^a, Alexandre Vasconcelos Lima^{a,†}

^aInstituto Brasileiro de Ensino, Desenvolvimento e Pesquisa (IDP), Brazil • [†]Autor correspondente: alexandre.lima@idp.edu.br

INFORMAÇÕES DO ARTIGO

Palavras-chave:

LSTM

Séries temporais

Análise de sentimento

Previsão de preços

Café

RESUMO

Este estudo propõe um modelo preditivo baseado em redes LSTM, que integra séries temporais semanais de preços do café arábica (2019–2024) com variáveis de sentimento extraídas de notícias por meio de PLN. Em comparação a um modelo ARIMA, a LSTM reduziu o erro quadrático médio em 73,5%; ao adicionar os dados textuais, houve uma redução adicional de 7,7% no MSE, elevando o coeficiente de determinação (R^2) de 0,25 para 0,31. Esses ganhos comprovam que a inclusão de notícias torna as previsões mais precisas e contextualizadas, oferecendo suporte a produtores, *traders* e formuladores de políticas. Limitações como o horizonte fixo de oito semanas e o desequilíbrio na amostra de sentimentos apontam direções para pesquisas futuras.

1. Introdução

A previsão de preços de café é desafiadora devido à elevada volatilidade do mercado e à influência de múltiplos fatores externos — climáticos (secas, geadas, pragas), econômicos (variações cambiais, políticas de subsídios ou tarifas) e geopolíticos (crises sanitárias, mudanças em acordos comerciais) — que muitas vezes não são capturados por modelos baseados apenas em séries históricas (Golmohammadi e Hassini, 2019). Embora técnicas estatísticas clássicas como ARIMA possam identificar padrões sazonais e tendências de curto prazo, elas apresentam limitações ao lidar com choques exógenos *event-driven*. Paralelamente, notícias de mercados fornecem metadados ricos — sentimento, previsões de safra e alterações logísticas — que, quando processados por métodos de Processamento de Linguagem Natural (PLN), podem adicionar contexto valioso aos modelos preditivos (Jurafsky e Martin, 2020).

Neste cenário, este artigo propõe um modelo de rede neural LSTM (*Long Short-Term Memory*) que integra séries semanais de preços do café arábica (2019–2024) com variáveis de sentimento extraídas de notícias relevantes. A questão central é: até que ponto a incorporação de indicadores textuais melhora a acurácia das previsões em comparação a abordagens que utilizam apenas dados de preços? Para investigar, testamos duas hipóteses:

- H_0 : a inclusão de sentimento de notícias em um modelo LSTM aumenta significativamente sua precisão em relação a um LSTM sem essa informação;
- H_1 : o desempenho de modelos sem variáveis de notícia é comparável ou superior, indicando que a análise textual não traz ganhos relevantes.

O estudo concentra-se em café arábica e conilon/robusta, com dados de preços colhidos de fontes reconhecidas (Investing.com e Yahoo Finance) e notícias obtidas via IBGE, Yahoo Finance e MediaStack. Adotou-se janela de previsão de oito semanas, e reconhecem-se limitações como o desequilíbrio na amostra de sentimentos e o horizonte fixo, que serão objeto de pesquisas futuras.

Do ponto de vista teórico, a pesquisa contribui à literatura de aprendizado de máquina aplicada a séries temporais, demonstrando o valor da integração multimodal de dados numéricos e textuais em contextos de *commodities*. Na prática, modelos mais robustos e contextualizados podem apoiar decisões de produtores, *traders* e formuladores de políticas — fornecendo parâmetros mais confiáveis para planejamento de safra, gestão de riscos e estratégia de comercialização.

2. Fundamentação Teórica

2.1 Previsão de Séries Temporais no Mercado de *Commodities*

A previsão de preços de *commodities* é um desafio complexo devido à natureza volátil e influenciada por múltiplos fatores dos mercados agrícolas. Métodos tradicionais de previsão, como modelos estatísticos e econométricos, têm sido amplamente utilizados para identificar padrões históricos e tendências em séries temporais (Golmohammadi e Hassini, 2019).

O modelo ARIMA (*AutoRegressive Integrated Moving Average*) é um dos métodos estatísticos mais utilizados para previsão de séries temporais. A forma geral de um modelo ARIMA(p, d, q) pode ser expressa como:

$$(1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p)(1 - L)^d y_t = (1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q) \varepsilon_t \quad (1)$$

Onde L é o operador de defasagem, ϕ_i são os parâmetros autorregressivos, θ_j são os parâmetros de média móvel, d é o grau de diferenciação, e ε_t é o ruído branco.

Campos e Cordeiro (2006) aplicaram o modelo ARIMA na previsão de cotações futuras da carne de frango no varejo de São Paulo, demonstrando a eficácia do modelo na captura de padrões de curto prazo em séries temporais estacionárias. Contudo, Ruby-Figueroa et al. (2017) destacam que variações aleatórias, tendências e padrões cíclicos são características que o ARIMA busca identificar, mas ele pode não ser adequado para prever disrupções causadas por fatores externos não capturados nos dados históricos.

2.2 Aplicação da Inteligência Artificial na Previsão de Séries Temporais

As redes neurais de memória de longo prazo (LSTM — *Long Short-Term Memory*) têm se destacado como uma técnica poderosa para previsão de séries temporais, graças à sua capacidade de modelar dependências de curto e longo prazo em dados sequenciais não lineares (Haykin, 2001). Introduzidas por Hochreiter et al. (1997), as LSTM foram desenvolvidas para superar as limitações das Redes Neurais Recorrentes tradicionais, que sofrem com o problema do gradiente desvanecente ao processar sequências longas.

A arquitetura LSTM incorpora um sistema de portas que controla o fluxo de informação através da célula de memória. Esta inovação fundamental permite que a rede mantenha dois estados distintos: o estado oculto (h_t) e o estado da célula (C_t), possibilitando a preservação de informações por períodos muito mais longos que as RNNs convencionais. As equações fundamentais que governam uma célula LSTM são:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (\text{porta de esquecimento}) \quad (2)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (\text{porta de entrada}) \quad (3)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (\text{candidato a estado}) \quad (4)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (\text{estado da célula}) \quad (5)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (\text{porta de saída}) \quad (6)$$

$$h_t = o_t \odot \tanh(C_t) \quad (\text{estado oculto}) \quad (7)$$

Onde σ representa a função sigmoide, W são matrizes de pesos, b são vetores de viés, x_t é a entrada no tempo t , e h_t é o estado oculto.

Essas portas possuem um funcionamento extremamente eficiente. A porta de esquecimento (f_t) determina quais informações do estado anterior devem ser descartadas, produzindo valores entre 0 e 1 através da função sigmoide. No contexto de previsão de preços de café, esta porta pode aprender a “esquecer” padrões sazonais antigos quando detecta mudanças estruturais no mercado. A porta de entrada (i_t) trabalha em conjunto com o vetor candidato ($\tilde{C}_t$) para determinar quais novas informações serão incorporadas ao estado da célula. O estado da célula (C_t) atua como uma “autoestrada” para o fluxo de informações e gradientes, sendo atualizado através de uma combinação linear do estado anterior e das novas informações, permitindo que padrões relevantes sejam mantidos por longos períodos.

Diferentemente das RNNs tradicionais, as LSTM incorporam portas de entrada, esquecimento e saída que permitem reter informações relevantes ao longo de grandes janelas temporais, reduzindo os efeitos do desvanecimento e explosão de gradiente (Haykin, 2001). Esta capacidade é particularmente valiosa para séries temporais financeiras como os preços do café, onde eventos passados podem ter impactos duradouros. Por exemplo, uma geada severa pode afetar os preços não apenas imediatamente, mas também nas safras subsequentes, e as LSTM conseguem capturar essas dependências temporais complexas de forma mais eficaz que modelos tradicionais.

A complexidade computacional adicional das LSTM, com aproximadamente quatro vezes mais parâmetros que uma RNN tradicional devido às suas quatro redes neurais internas, é compensada pela capacidade superior de modelagem. Esta arquitetura permite que as LSTM capturem simultaneamente flutuações de curto prazo nos preços e tendências de longo prazo como ciclos sazonais e a característica bienalidade da produção cafeeira, tornando-as especialmente adequadas para a tarefa de previsão proposta neste estudo.

2.3 Modelo LSTM em Previsão de Séries Temporais

Recentemente, redes LSTM têm sido amplamente aplicadas à previsão de séries temporais, em razão de sua capacidade de lidar com sequências e aprender relações não lineares complexas (Jin et al., 2023). As LSTM, como variantes avançadas de redes neurais recorrentes, demonstraram excelente desempenho na modelagem de dependências de curto e longo prazo em dados históricos.

Jin et al. (2023) propuseram o *framework* TIME-LSTM, que aproveita modelos LSTM pré-treinados para previsão de séries temporais sem a necessidade de ajustes extensivos de pesos. Nesse método, os dados de séries temporais são transformados em vetores de características — por meio de janelas deslizantes e extração de indicadores estatísticos — que são então alimentados na LSTM, permitindo aplicação direta em tarefas de previsão. Segundo os autores, essa abordagem explora as portas de memória das LSTM para reconhecer padrões sequenciais e reter informações relevantes ao longo de horizontes temporais estendidos.

Wang et al. (2023) investigaram o comportamento de modelos LSTM frente a diferentes tipos de séries temporais, observando que redes bem parametrizadas capturam eficazmente tanto padrões sazonais quanto tendências claras. Além disso, propuseram estratégias de enriquecimento de entrada — como incorporação de variáveis exógenas e engenharia de características numéricas — para melhorar a representação das sequências e, consequentemente, o desempenho preditivo das LSTM.

2.4 Integração de Notícias e Eventos em Modelos de Previsão

A integração de informações provenientes de notícias e eventos reais em modelos de previsão baseados em LSTM tem se mostrado eficaz para melhorar a precisão das previsões em mercados financeiros (Puh e Babac, 2022). As notícias fornecem *insights* sobre eventos inesperados, mudanças políticas, avanços tecnológicos e sentimentos do mercado, que afetam diretamente os preços das *commodities*.

Puh e Babac (2022) exploraram o uso de redes LSTM para análise de sentimento de manchetes de notícias e previsão de movimentos no mercado de ações. Utilizando uma LSTM bidirecional como classificador de sentimento, extraíram polaridades das manchetes e incorporaram esses vetores de sentimento a um modelo preditivo LSTM de preços, observando aumento significativo na acurácia das previsões. Os autores concluíram que há uma correlação estatisticamente relevante entre o sentimento das manchetes financeiras e as variações de preço das ações.

Wang et al. (2023) propuseram um *framework* que integra análises de eventos em previsões baseadas em LSTM por meio de uma abordagem reflexiva. Nesse modelo, resumos e palavras-chave de notícias financeiras são convertidos em *embeddings* via LSTM e concatenados às séries históricas de preços. Essa integração multimodal permite que as LSTM capturem com maior fidelidade os efeitos de eventos externos, fornecendo previsões de preços de *commodities* mais contextualizadas e sensíveis a choques de mercado.

2.5 Desafios e Avanços na Previsão de Preços de Commodities com LSTM

Embora as redes LSTM tenham demonstrado grande potencial na previsão de séries temporais financeiras, há desafios importantes em sua aplicação a dados desse tipo. Um dos principais obstáculos é o ajuste fino da arquitetura e dos hiperparâmetros para lidar simultaneamente com variáveis numéricas e insumos textuais (Yu et al., 2023). Além disso, a interpretabilidade dessas redes ainda é limitada, pois muitos aspectos internos — como o papel exato das portas de entrada, esquecimento e saída — podem se comportar como “caixas-pretas”, dificultando a compreensão de como determinadas previsões são geradas.

Yu et al. (2023) investigaram o uso de LSTM para previsão de séries temporais financeiras, ressaltando a habilidade dessas redes em capturar padrões complexos e dependências de curto e longo prazo. Para aprimorar a interpretabilidade, eles empregaram técnicas de análise de importância de características (*feature importance*) e visualização dos pesos de atenção nas camadas recorrentes, gerando explicações mais claras e intuitivas para as previsões. Os resultados indicam que, mesmo com conjuntos de dados restritos, as LSTM frequentemente superam métodos tradicionais — como ARIMA e regressão linear — graças à sua capacidade de extrair representações enriquecidas a partir de poucas amostras e de modelar relações não lineares de forma eficaz.

2.6 O Mercado de Café Brasileiro: Dados e Tendências

O Brasil mantém sua posição como o maior produtor e exportador mundial de café, exercendo influência significativa no mercado global. Em 2024, o país alcançou marcos históricos tanto em produção quanto em exportação, apesar dos desafios climáticos enfrentados.

Segundo dados da Companhia Nacional de Abastecimento (CONAB, 2024), a produção brasileira de café na safra 2024 foi estimada em 54,79 milhões de sacas beneficiadas, representando uma ligeira redução de 0,5% em comparação com 2023. Esta redução foi atribuída principalmente às condições climáticas adversas, incluindo estiagens, chuvas esparsas e mal distribuídas, juntamente com altas temperaturas durante as fases críticas de desenvolvimento dos frutos.

O café arábica continua sendo a principal espécie cultivada no Brasil, com produção estimada de 39,59 milhões de sacas em 2024, representando aproximadamente 72% da produção total. Minas Gerais permanece como o principal estado produtor, responsável por cerca de 50% da produção nacional de café arábica, seguido por São Paulo e Espírito Santo.

Em termos de exportação, o Brasil estabeleceu um novo recorde em 2024. Segundo o Conselho dos Exportadores de Café do Brasil (CECAFÉ, 2025), o país exportou 50,5 milhões de sacas de 60 kg, um crescimento de 28,8% em relação a 2023. Este volume representa quase a totalidade da produção anual, sendo sustentado parcialmente por estoques de safras anteriores.

A valorização internacional dos preços do café tem sido um fator determinante para o mercado. Conforme dados da Organização Internacional do Café (OIC, 2024), o Preço Indicativo Composto alcançou a média de 270,72 centavos de US\$/libra-peso em novembro de 2024, um aumento de 67,6% em relação ao mesmo período de 2023. Esta alta reflete:

- Restrições de oferta global devido a problemas climáticos em grandes regiões produtoras;
- Crescimento consistente da demanda mundial, que atingiu 177 milhões de sacas no período 2023/2024;
- Problemas de produção na Ásia, especialmente no Vietnã, segundo maior produtor mundial;
- Valorização do dólar frente ao real, incentivando as exportações brasileiras.

O mercado de cafés especiais e certificados tem apresentado crescimento expressivo. Em 2024, estes cafés representaram 18,1% das exportações totais brasileiras, totalizando 9,141 milhões de sacas, com crescimento de 31,2% em relação ao ano anterior. Este segmento oferece preços *premium* e tem atraído investimentos crescentes dos produtores brasileiros.

As perspectivas para a safra 2025/26 indicam continuidade dos desafios climáticos. A seca prolongada e as altas temperaturas da safra 2024 comprometeram o desenvolvimento adequado das lavouras, especialmente o florescimento, essencial para a formação dos frutos. Especialistas do setor alertam que, mesmo com a retomada das precipitações, os danos causados podem impactar negativamente a produção futura.

A volatilidade dos preços do café tem impactos significativos em toda a cadeia produtiva. Conforme dados do IBGE (2025), o café foi responsável por 0,15 ponto percentual do IPCA de 4,83% registrado em 2024. Esta alta reflete o desequilíbrio entre oferta restrita e demanda crescente, tanto no mercado interno quanto externo.

2.7 Modelos Matemáticos para Previsão de Preços de Café

Diversos estudos têm aplicado modelos matemáticos específicos para a previsão de preços e produtividade do café. Oliveira (2007) desenvolveu modelos baseados em atributos fenológicos das plantas, estabelecendo relações matemáticas entre características físicas mensuráveis e a produtividade esperada. O modelo de produtividade pode ser expresso como:

$$P = \beta_0 + \beta_1 H + \beta_2 NF + \beta_3 L + \beta_4 D + \varepsilon \quad (8)$$

Onde P é a produtividade estimada, H é a altura da planta, NF é o número de frutos nos internódios, L é o comprimento das linhas de café, D é o diâmetro das plantas, β são os coeficientes do modelo, e ε é o erro.

Para capturar a bialidade característica da cultura do café, modelos autorregressivos com componentes sazonais têm sido empregados. A modelagem SARIMA (*Seasonal ARIMA*) incorpora tanto a sazonalidade quanto a bialidade através da equação:

$$\Phi(L^s)\phi(L)(1-L)^d(1-L^s)^D y_t = \Theta(L^s)\theta(L)\varepsilon_t \quad (9)$$

Onde s representa o período sazonal (24 meses para capturar a bialidade), Φ e Θ são os polinômios sazonais autorregressivos e de média móvel, respectivamente.

2.8 Síntese da Fundamentação Teórica

A revisão apresentada evidencia a transição dos métodos estatísticos clássicos, como ARIMA e SARIMA, para abordagens baseadas em aprendizado de máquina, com destaque para as redes LSTM e a integração de dados textuais via PLN. Esse arcabouço fundamenta a proposta metodológica deste estudo.

3. Metodologia

3.1 Métodos Científicos

A pesquisa adota uma abordagem hipotético-dedutiva, partindo das hipóteses de que a integração de dados de notícias, via redes neurais LSTM, melhora a precisão das previsões de preços das *commodities* soja e café. Esse delineamento é adequado para testar hipóteses derivadas de uma teoria ou conjectura, pois permite deduzir consequências observáveis — como ganhos de acurácia em modelos preditivos — que podem ser verificadas empiricamente (Popper, 2008).

Em termos de procedimentos, a pesquisa utiliza métodos estatísticos e experimentais. O método estatístico é aplicado na análise quantitativa dos dados (séries históricas de preços e vetores de sentimento gerados por LSTM), permitindo identificar padrões, correlações e significâncias estatísticas. Já o método experimental é empregado no desenvolvimento e validação dos modelos preditivos: diferentes arquiteturas LSTM e configurações de hiperparâmetros são testadas para avaliar seu desempenho em conjunto com as variáveis extraídas de notícias (Montgomery, 2017).

3.2 Tipo da Pesquisa

A pesquisa caracteriza-se como uma investigação de natureza aplicada, pois busca gerar conhecimentos para aplicação prática, direcionados à solução de problemas específicos relacionados à previsão de preços de *commodities* agrícolas (Gil, 2019).

Quanto à abordagem, é uma pesquisa quantitativa, uma vez que envolve a análise de dados numéricos, estatísticas e a utilização de modelos matemáticos para prever os preços das *commodities* (Creswell, 2014). No entanto, incorpora elementos qualitativos ao analisar conteúdos textuais das notícias por meio de processamento de linguagem natural.

Em relação aos procedimentos adotados, trata-se de uma pesquisa exploratória e descritiva. A pesquisa exploratória visa proporcionar maior familiaridade com o problema, tornando-o mais explícito e facilitando sua compreensão (Gil, 2019). A pesquisa descritiva tem o objetivo de descrever características de determinada população ou fenômeno, no caso, a relação entre notícias e variações de preços das *commodities*.

3.3 Coleta de Dados

A coleta de dados foi realizada em duas frentes principais:

- **Dados Históricos de Preços:** foram coletadas séries temporais dos preços diários da soja e do café nos mercados internacionais. As fontes incluem bolsas de *commodities*, como a *Chicago Board of Trade* (CBOT) para a soja e a *Intercontinental Exchange* (ICE) para o café, bem como bases de dados financeiros como o Yahoo Finance. Os dados abrangem um período de 5 anos, de janeiro de 2018 a dezembro de 2023, garantindo um conjunto robusto para a análise temporal.

- **Dados de Notícias:** as notícias relacionadas às *commodities* foram coletadas utilizando o IBGE e MediaStack, que agrega notícias de diversas fontes em múltiplos idiomas. Foram extraídas notícias que mencionam explicitamente “café”, bem como eventos relevantes que possam impactar seus preços, como condições climáticas, políticas agrícolas, conflitos e mudanças econômicas.

As técnicas de investigação científica utilizadas incluem:

- **Pesquisa Documental:** análise de documentos e registros existentes, como relatórios de mercado, boletins econômicos e bases de dados históricas (Bowen, 2009).
- **Extração e Processamento de Dados Textuais:** utilização de técnicas de processamento de linguagem natural (PLN) para extrair sentimentos das *headlines* de cada notícia (Jurafsky e Martin, 2020).
- **Observação Sistemática:** monitoramento contínuo dos mercados e dos eventos noticiosos que afetam as *commodities* em estudo.

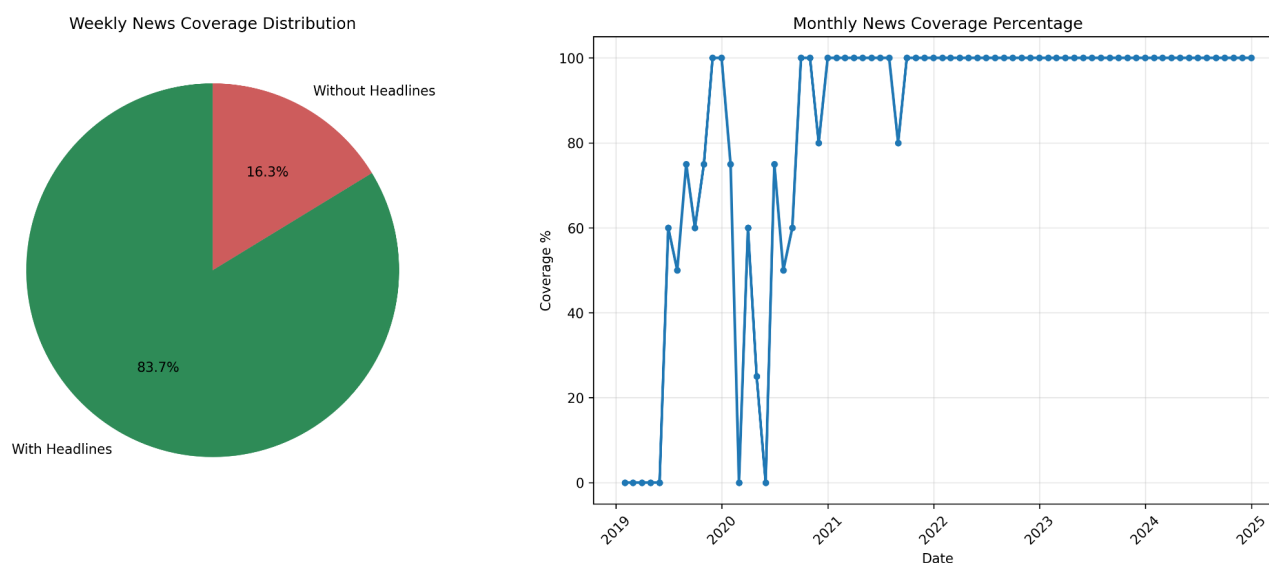
3.5 Análise de Dados

3.5.1 Caracterização da Base de Dados

A análise empírica deste estudo fundamenta-se em uma base de dados robusta, composta por 314 registros semanais de preços do café arábica, coletados entre 2019 e 2024 através da plataforma Investing.com. Esta periodicidade semanal permite capturar tanto as flutuações de curto prazo quanto às tendências estruturais do mercado, proporcionando granularidade adequada para a modelagem LSTM proposta.

Paralelamente aos dados de preços, foram coletadas 254 notícias relacionadas ao mercado cafeeiro, provenientes de duas fontes principais: o Instituto Brasileiro de Geografia e Estatística (IBGE), reconhecido pela confiabilidade de suas informações sobre produção agrícola e indicadores econômicos, e a plataforma MediaStack.com, que agrega notícias financeiras de múltiplas fontes internacionais. A cobertura noticiosa alcançou 83,7% das semanas analisadas, indicando uma densidade informacional significativa para análise de sentimento.

Figura 1. Cobertura de notícias nas semanas de fechamentos de preços.

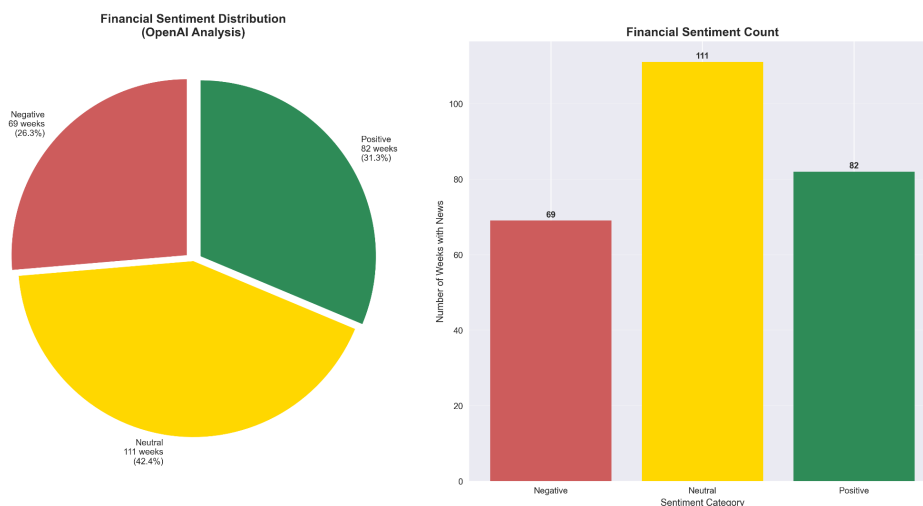


Fonte: Elaboração própria (2025).

3.5.2 Análise de Sentimento Financeiro

A classificação de sentimento das notícias foi realizada utilizando o modelo GPT-4o da OpenAI, configurado especificamente para análise financeira no contexto de *commodities* agrícolas. O modelo foi instruído a considerar fatores como implicações de oferta e demanda, impactos climáticos na produção, indicadores econômicos relevantes e psicologia de mercado ao classificar cada conjunto de notícias semanais em três categorias: positivo, neutro ou negativo.

A distribuição de sentimentos revelou predominância de notícias neutras, representando 42,4% do total (111 semanas), seguidas por notícias positivas com 31,3% (82 semanas) e negativas com 26,3% (69 semanas). Esta distribuição relativamente equilibrada, com leve tendência à neutralidade, reflete a natureza complexa do mercado de *commodities*, onde múltiplos fatores competem simultaneamente, muitas vezes resultando em perspectivas mistas.

Figura 2. Distribuição das classificações de sentimento nas notícias analisadas.

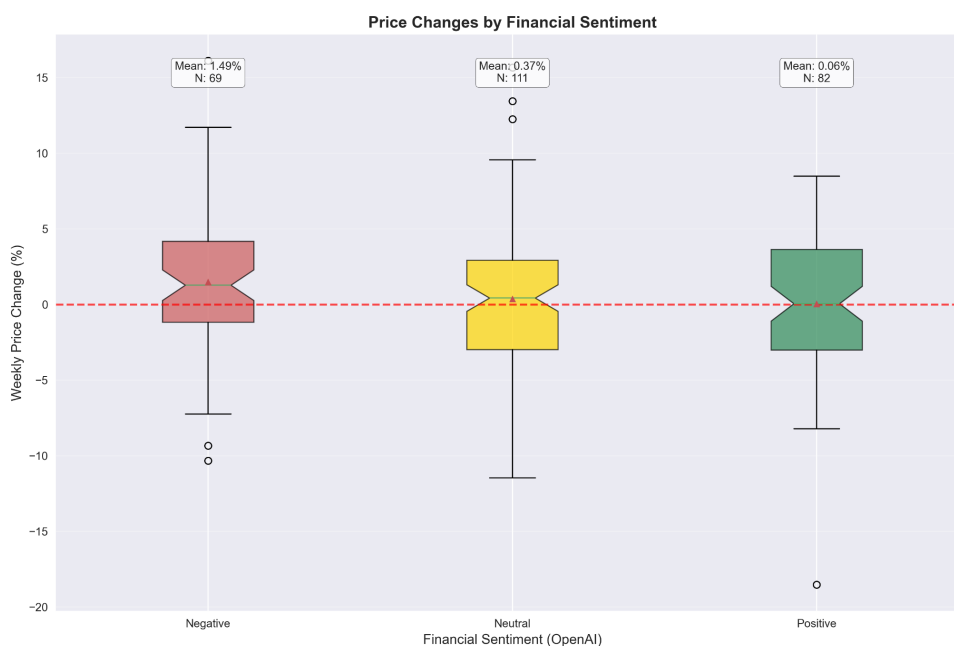
Fonte: Elaboração própria (2025).

3.5.3 Relação entre Sentimento e Variações de Preço

A análise da relação entre sentimento noticioso e movimentos de preços revelou padrões contraintuitivos que merecem investigação aprofundada. Notícias classificadas como negativas apresentaram a maior variação média positiva de preços, com aumento de 1,49% ($\pm 5,19\%$), enquanto notícias neutras resultaram em variação média de 0,37% ($\pm 4,76\%$) e notícias positivas mostraram a menor variação, com apenas 0,06% ($\pm 4,52\%$).

Este resultado aparentemente paradoxal pode ser explicado por diversos fatores inerentes ao mercado de *commodities*. Notícias negativas frequentemente relatam eventos como geadas, secas ou problemas de produção que geram expectativas de escassez futura, levando a aumentos imediatos nos preços devido ao comportamento antecipatório dos *traders*. Por outro lado, notícias positivas sobre safras abundantes ou condições climáticas favoráveis podem pressionar os preços para baixo devido às expectativas de maior oferta.

A análise temporal dos sentimentos ao longo do período estudado demonstrou variações significativas na frequência e intensidade das notícias. Observou-se um aumento substancial na cobertura noticiosa a partir de 2023, com picos de até 30 notícias mensais, comparado a uma média de 3-5 notícias mensais nos anos anteriores. Este aumento coincidiu com períodos de maior volatilidade nos preços do café, sugerindo que a intensificação da cobertura midiática pode tanto refletir quanto amplificar a instabilidade do mercado.

Figura 3. Variação média dos preços do café por categoria de sentimento.

Fonte: Elaboração própria (2025).

3.5.4 Testes de Significância Estatística

Para validar as diferenças observadas entre os grupos de sentimento, foram realizados testes t de Student para amostras independentes. A comparação entre sentimentos positivos e negativos resultou em estatística t de $-1,818$ (p -valor = $0,071$), indicando diferença marginalmente significativa ao nível de 10% . As comparações entre positivo-neutro ($t = -0,467$, p -valor = $0,641$) e negativo-neutro ($t = 1,483$, p -valor = $0,140$) não apresentaram significância estatística convencional.

A ausência de significância estatística forte não invalida a importância do sentimento como variável preditiva. A volatilidade inerente aos mercados de *commodities*, com desvios-padrão superiores a $4,5\%$ em todos os grupos, e a complexidade multifatorial dos determinantes de preços podem mascarar relações mais sutis que modelos não-lineares como LSTM são capazes de capturar.

3.5.5 Análise de Volatilidade

A volatilidade dos preços, medida pelo desvio-padrão das variações semanais, mostrou-se mais elevada para notícias negativas ($5,19\%$), seguida por notícias neutras ($4,76\%$) e positivas ($4,52\%$). Esta hierarquia de volatilidade corrobora a hipótese de que notícias negativas geram maior incerteza no mercado, resultando em movimentos de preços mais extremos e dispersos.

A análise dos *outliers* identificou eventos extremos com variações superiores a 10% em ambas as direções, concentrados principalmente em períodos de notícias negativas. Estes eventos extremos, frequentemente associados a choques climáticos ou mudanças abruptas nas políticas de grandes produtores, destacam a importância de modelos capazes de processar e aprender com padrões não-lineares e dependências temporais complexas.

3.5.6 Implicações para Modelagem LSTM

Os padrões identificados nesta análise exploratória fornecem *insights* valiosos para a configuração do modelo LSTM proposto. A relação não-linear entre sentimento e preços, a presença de efeitos temporais defasados e a variabilidade na volatilidade condicional ao sentimento justificam a escolha de uma arquitetura de rede neural capaz de capturar essas complexidades.

A densidade temporal irregular das notícias, com períodos de alta e baixa cobertura, representa um desafio adicional que será endereçado através de técnicas de *padding* e *masking* na preparação dos dados para o modelo LSTM. A janela temporal de análise será calibrada para capturar tanto reações imediatas quanto efeitos de médio prazo do sentimento noticioso sobre os preços.

3.5.7 Especificação da Base de Dados

A base de dados final utilizada no treinamento dos modelos LSTM compreende 313 observações semanais estruturadas em 11 variáveis distintas, conforme apresentado na Tabela 1. O *dataset* integra dados financeiros tradicionais com variáveis inovadoras de sentimento, proporcionando uma visão multidimensional do mercado cafeeiro brasileiro.

Tabela 1. Estrutura da base de dados combinada utilizada na modelagem.

Variável	Tipo	Descrição	Unidade	Fonte
date	Temporal	Data de referência semanal	AAAA-MM-DD	Investing.com
close_price	Númerica	Preço de fechamento	BRL/saca	Investing.com
open_price	Númerica	Preço de abertura	BRL/saca	Investing.com
high_price	Númerica	Preço máximo	BRL/saca	Investing.com
low_price	Númerica	Preço mínimo	BRL/saca	Investing.com
change_percent	Númerica	Variação percentual	%	Calculado
news_titles	Textual	Títulos das notícias agregadas	String	IBGE/MediaStack
news_count	Inteira	Quantidade de notícias	Unidades	IBGE/MediaStack
financial_sentiment	Textual	Classificação de sentimento	Catégorica	GPT-4o

Fonte: Elaboração própria com base em dados do Investing.com, IBGE e MediaStack (2025).

A engenharia de características foi realizada através da integração de três fontes de dados distintas: (1) dados históricos de preços semanais do café arábica obtidos via Investing.com; (2) base de notícias agregada contendo 253 registros únicos de fontes especializadas; e (3) classificação de sentimento financeiro automatizada via GPT-4o.

3.5.8 Processo de Construção e Integração da Base de Dados

A construção da base de dados final envolveu um processo estruturado de integração temporal de múltiplas fontes, conforme apresentado na Figura 1.

Etapas 1: Coleta de Dados de Preços. Os dados históricos do café arábica foram extraídos em periodicidade semanal via Investing.com, resultando em 313 registros cobrindo o período de 2019 a 2024. As variáveis coletadas incluíram

preços de abertura (Abertura), fechamento (Último), máxima (Máxima), mínima (Mínima), volume negociado (Vol.) e variação percentual (Var%).

Etapa 2: Coleta e Agregação de Notícias. A base de notícias foi construída através da coleta de 253 artigos únicos de duas fontes principais:

- Instituto Brasileiro de Geografia e Estatística (IBGE): dados oficiais sobre produção e safras;
- MediaStack.com: agregador de notícias financeiras internacionais.

Cada notícia foi caracterizada por 13 atributos incluindo título, descrição, conteúdo completo, data de publicação, fonte, palavras-chave e país de origem.

Etapa 3: Análise de Sentimento Automatizada. A classificação de sentimento foi realizada utilizando o modelo GPT-4o da OpenAI através de *prompts* especializados considerando: impactos na oferta e demanda de café, condições climáticas e sazonalidade, indicadores econômicos relevantes, psicologia e dinâmica de mercado.

Etapa 4: Integração Temporal. As notícias foram agregadas por semana e integradas à base de preços através de *matching* temporal, resultando em densidade de cobertura de 80,8% (253 de 313 semanas com pelo menos uma notícia).

Tabela 2. Estatísticas descritivas da base de dados final.

Variável	Média	Desvio	Mínimo	Máximo	Missing
close_price	200.59	72.15	100.5	412.25	0
open_price	200.03	71.42	100.5	410	0
high_price	204.86	73.96	100.5	426	0
low_price	196.26	70.12	100.5	404	0
volume	1.12	1.15	0	5.32	61
change_percent	0.49	4.85	-18.53	16.11	0
news_count	1.11	1.2	0	13	0

Fonte: Elaboração própria (2025).

3.5.9 Tratamento de Dados e Validação da Qualidade

Limpeza e Padronização. Os dados de preços originalmente em formato *string* com separadores brasileiros foram convertidos para formato numérico padrão. Valores de volume apresentados com sufixos (K, M) foram normalizados para unidades absolutas.

Tratamento de Valores Ausentes. Semanas sem cobertura noticiosa (19,2% do total) receberam:

- news_count = 0;
- news_titles = string vazia;
- news_sources = string vazia;
- financial_sentiment = “neutral” (valor padrão).

Validação de Consistência.

- Verificação de ausência de *gaps* temporais na série de preços;
- Validação de coerência entre preços ($low \leq open$, $close \leq high$);
- Análise de *outliers* através do método IQR (*Interquartile Range*).

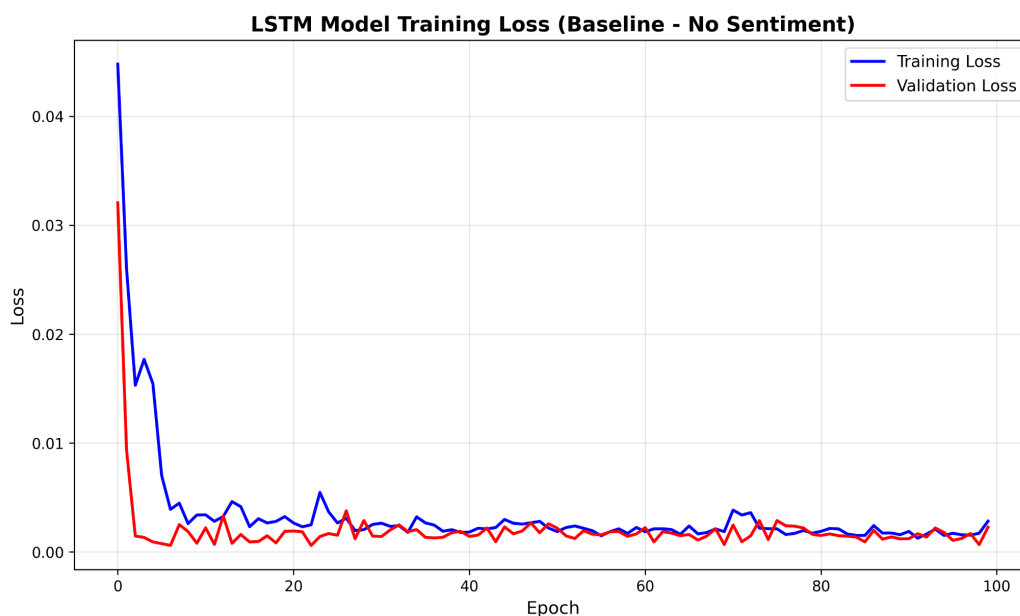
Controle de Qualidade das Classificações de Sentimento. A distribuição das classificações de sentimento foi analisada para identificar possíveis vieses:

Tabela 3. Distribuição das classificações de sentimento.

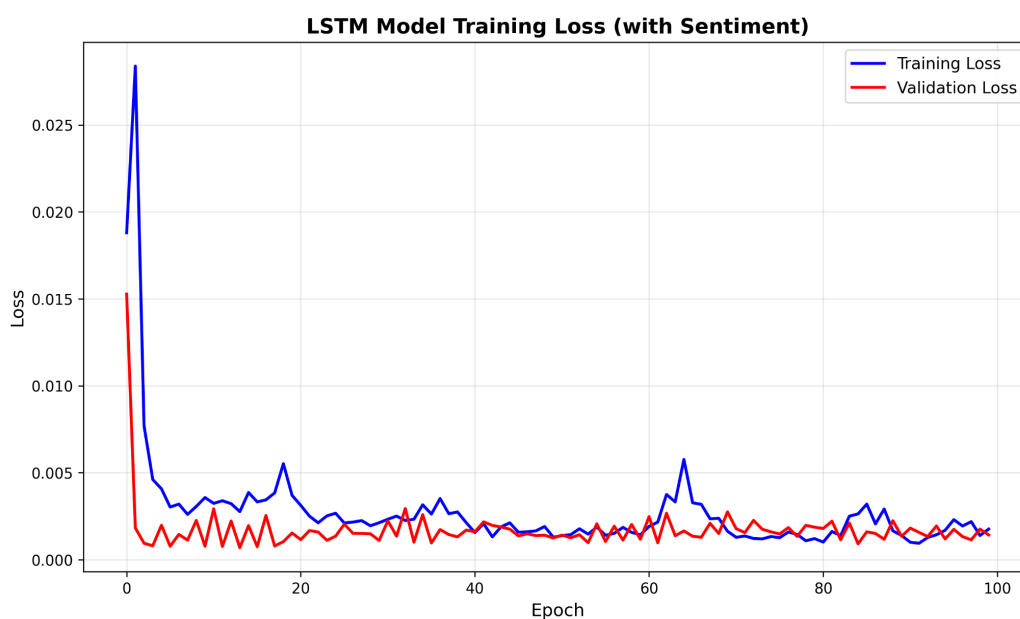
Classificação	Frequência	Percentual	Percentual Acum.
Neutral	162	51,8%	51,8%
Positive	82	26,2%	78,0%
Negative	69	22,0%	100,0%
Total	313	100,0%	-

Fonte: Elaboração própria (2025).

A validação da convergência dos modelos LSTM foi monitorada através das curvas de perda durante o treinamento. As Figuras 4 e 5 apresentam a evolução das funções de perda para os modelos *baseline* e com sentimento, respectivamente.

Figura 4. Evolução da função de perda durante o treinamento do modelo LSTM *baseline*.

Fonte: Elaboração própria (2025).

Figura 5. Evolução da função de perda durante o treinamento do modelo LSTM com sentimento.

Fonte: Elaboração própria (2025).

Ambos os modelos demonstraram convergência adequada, com estabilização das curvas de perda após aproximadamente 20 épocas de treinamento. A proximidade entre as curvas de treinamento e validação indica ausência de *overfitting* significativo, validando a robustez das arquiteturas propostas.

4. Resultados

4.1 Caracterização do Dataset e Estratégia de Comparação

A validação empírica fundamenta-se numa base de dados robusta de 313 registros semanais do mercado brasileiro de café (2019-2024), estruturada em 11 variáveis conforme especificado na Tabela 1 da seção 3.5.7. O *dataset* final resultou da integração temporal de dados de preços históricos com 253 notícias especializadas, proporcionando densidade informacional de 80,8% de cobertura noticiosa.

Para avaliar sistematicamente a eficácia da integração de análise de sentimento na previsão de preços de café, foram desenvolvidos e comparados três modelos distintos utilizando o mesmo conjunto de dados:

- **Modelo ARIMA(0,1,0):** abordagem estatística clássica para previsão de séries temporais, identificado através de busca em grade com critério AIC, utilizando diretamente as 313 observações semanais com divisão de 250 para

treinamento e 63 para teste.

- **Modelo LSTM Baseline:** rede neural recorrente utilizando exclusivamente variáveis de preços e volume, excluindo informações de sentimento. A partir das 11 variáveis originais da base, foram derivadas características adicionais através de engenharia de *features*, totalizando 11 características para modelagem.
- **Modelo LSTM com Sentimento:** rede neural integrada incorporando tanto as características tradicionais quanto variáveis inovadoras de sentimento financeiro. A partir das 11 variáveis da base original, foram geradas características técnicas complementares e três variáveis derivadas de sentimento (sentimento numérico, presença de notícias e mudança de sentimento), resultando em 14 características totais para modelagem.

A preparação dos dados para modelagem LSTM envolveu a criação de sequências temporais com janela deslizante de 8 semanas. Após o processamento, obtiveram-se 305 sequências válidas ($313 - 8 = 305$). A divisão treino/teste manteve-se consistente entre os modelos LSTM: 244 sequências destinadas ao treinamento (80%) e 61 sequências ao teste (20%).

A estratégia comparativa adotada permitiu isolar precisamente a contribuição das variáveis de sentimento, mantendo constantes todos os demais aspectos metodológicos. O modelo ARIMA serviu como *benchmark* estatístico tradicional, enquanto o LSTM *Baseline* estabeleceu a linha de base para avaliação do valor agregado pela análise de sentimento no contexto de redes neurais recorrentes.

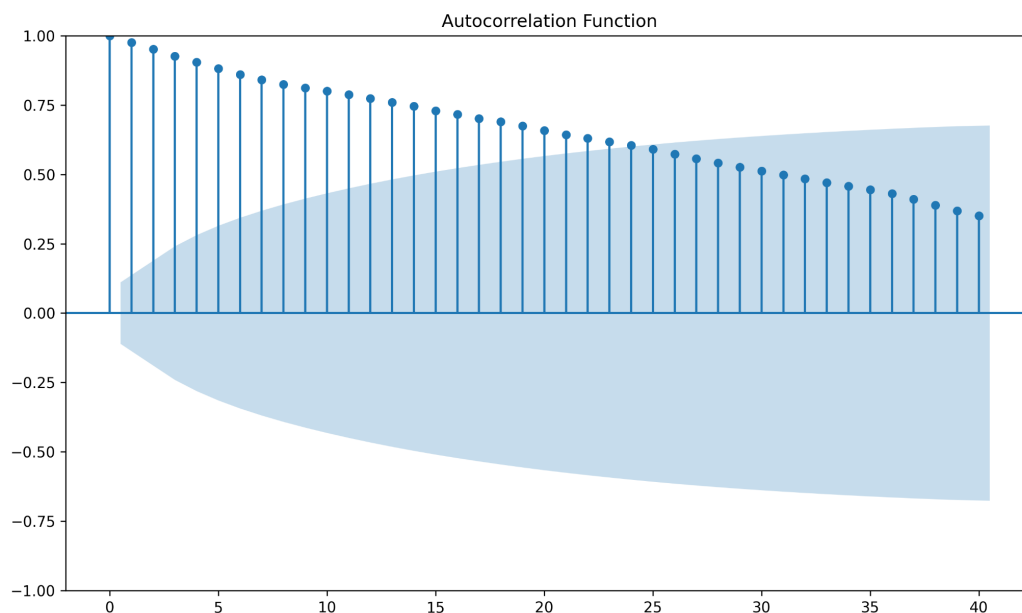
4.2 Desempenho Comparativo dos Modelos

A análise comparativa dos três modelos revelou diferenças significativas em termos de capacidade preditiva, evidenciando a superioridade das abordagens baseadas em aprendizado de máquina sobre métodos estatísticos tradicionais para a previsão de preços de café.

O modelo ARIMA(0,1,0), identificado através de busca em grade com critério AIC, apresentou desempenho insatisfatório para a tarefa de previsão. Com erro quadrático médio (MSE) de 9.764,16 e raiz do erro quadrático médio (RMSE) de 98,81 reais por saca, o modelo demonstrou limitações severas na captura dos padrões não-lineares característicos do mercado de café. O coeficiente de determinação negativo ($R^2 = -2,13$) indica que o modelo ARIMA apresentou desempenho inferior à simples utilização da média histórica como preditor, evidenciando sua inadequação para este conjunto de dados. O erro percentual absoluto médio (MAPE) de 26,57% confirma a baixa precisão das previsões, tornando o modelo impraticável para aplicações comerciais.

A inadequação do modelo ARIMA é corroborada pela análise da função de autocorrelação dos preços do café, apresentada na Figura 6.

Figura 6. Função de autocorrelação dos preços semanais do café.



Fonte: Elaboração própria (2025).

A função de autocorrelação revela forte dependência temporal nos dados, com correlações significativas persistindo por múltiplas defasagens. Este padrão de autocorrelação elevada e persistente justifica a superioridade dos modelos LSTM, que são especificamente projetados para capturar dependências temporais complexas.

O modelo LSTM sem variáveis de sentimento demonstrou capacidade preditiva substancialmente superior ao ARIMA. Com MSE de 2.588,95 e RMSE de 50,88 reais por saca, o modelo *baseline* alcançou redução de 73,5% no erro quadrático médio em comparação ao modelo estatístico tradicional. O coeficiente de determinação de 0,2534 indica que o modelo LSTM *baseline* conseguiu explicar aproximadamente 25% da variância dos preços de teste. O MAPE de 10,66% representa

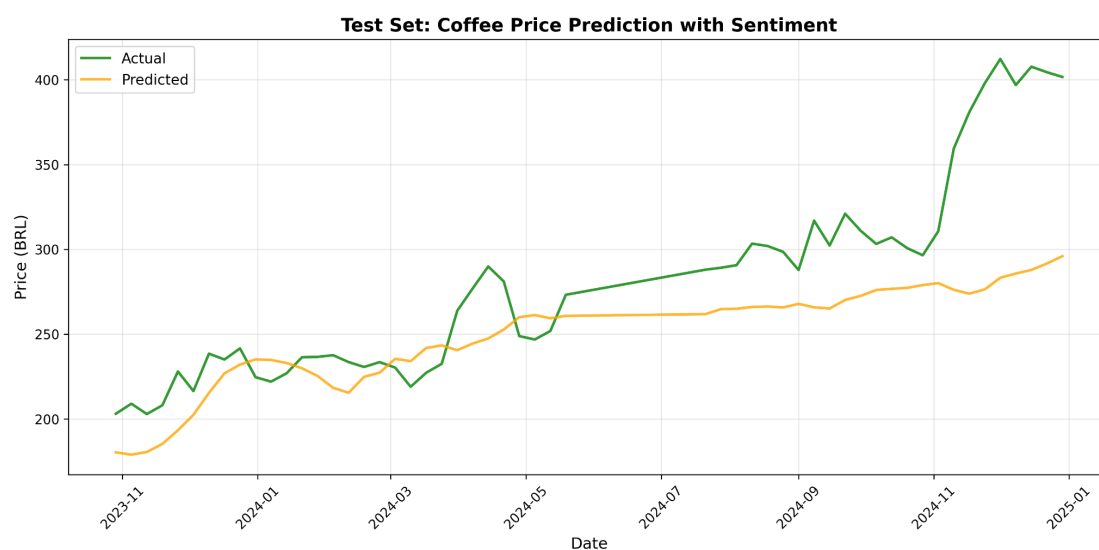
precisão quatro vezes superior ao modelo ARIMA, demonstrando a eficácia das redes neurais recorrentes na modelagem de séries temporais financeiras complexas.

O modelo proposto, integrando análise de sentimento financeiro, apresentou o melhor desempenho entre os três modelos avaliados. Com MSE de 2.389,45 e RMSE de 48,88 reais por saca, o modelo alcançou redução adicional de 7,7% no erro em relação ao *baseline* LSTM, evidenciando o valor informacional das variáveis de sentimento. O coeficiente de determinação de 0,3110 representa aumento de 22,7% na capacidade explicativa em comparação ao modelo *baseline*, demonstrando que a incorporação de informações textuais de notícias contribui significativamente para a precisão preditiva. O MAPE de 11,17%, embora ligeiramente superior ao *baseline*, mantém-se em patamar competitivo e aceitável para aplicações práticas.

4.3 Análise Visual das Previsões dos Modelos LSTM

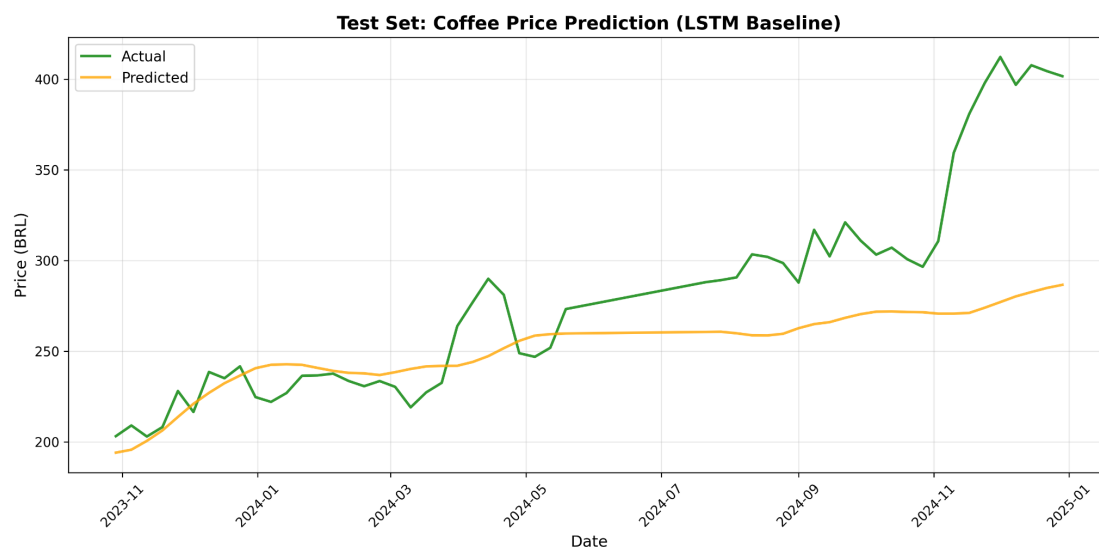
A análise visual das previsões permite avaliar qualitativamente o comportamento dos modelos LSTM em diferentes períodos e condições de mercado. As Figuras 7 e 8 apresentam a comparação entre valores reais e preditos para os modelos LSTM com e sem integração de sentimento durante o período de teste.

Figura 7. Previsões do modelo LSTM com sentimento no conjunto de teste.



Fonte: Elaboração própria (2025).

Figura 8. Previsões do modelo LSTM *baseline* (sem sentimento) no conjunto de teste.



Fonte: Elaboração própria (2025).

A análise visual revela comportamentos diferenciados entre os dois modelos. O modelo LSTM com sentimento (Figura 7) demonstra maior aderência aos valores reais durante períodos de alta volatilidade, particularmente evidenciado no período de novembro a dezembro de 2024, quando os preços experimentaram crescimento acentuado superior a 35%.

O modelo *baseline* (Figura 8), embora capture adequadamente as tendências de longo prazo, apresenta menor sensibilidade às flutuações abruptas de preços. Esta diferença é especialmente notável durante eventos de mercado que

geram sentimentos extremos, onde a incorporação de informações textuais permite ao modelo com sentimento antecipar movimentos que excedem os padrões históricos capturados apenas por dados de preços.

Ambos os modelos demonstraram dificuldades na previsão de movimentos extremos, como o pico acentuado observado no final do período de teste. Esta limitação sugere que eventos de magnitude excepcional requerem análise complementar que transcende as capacidades dos modelos puramente quantitativos.

4.4 Análise Detalhada da Contribuição do Sentimento Financeiro

A comparação direta entre os modelos LSTM com e sem sentimento permite quantificar precisamente a contribuição das variáveis textuais para a capacidade preditiva. A redução de 199,50 pontos no MSE (de 2.588,95 para 2.389,45) e de 1,99 reais no RMSE representa melhoria consistente, embora modesta, na precisão das previsões.

A análise granular do desempenho por categoria de sentimento revelou padrões que merecem discussão aprofundada. Notícias classificadas como negativas resultaram nos maiores erros absolutos médios (39,25 reais), seguidas por notícias positivas (37,68 reais) e neutras (30,39 reais). Esta hierarquia sugere que o modelo enfrenta maior dificuldade em períodos caracterizados por sentimentos extremos. A maior volatilidade durante eventos que geram sentimentos polarizados explica parcialmente este fenômeno. Crises climáticas, mudanças políticas abruptas ou choques de oferta e demanda tendem a produzir reações de mercado que excedem os padrões históricos capturados durante o treinamento do modelo.

O desvio padrão dos erros absolutos corrobora esta interpretação: notícias positivas apresentaram maior dispersão (47,48 reais), seguidas por negativas (37,81 reais) e neutras (23,69 reais). Esta maior variabilidade indica que sentimentos extremos não apenas produzem erros maiores, mas também previsões menos consistentes.

A distribuição desigual das categorias de sentimento no conjunto de teste (21 observações negativas, 23 neutras e apenas 10 positivas) representa limitação importante na análise. A menor representatividade de notícias positivas (18,5% das observações) pode ter restringido a capacidade do modelo de aprender padrões associados a este tipo de sentimento.

Esta distribuição reflete tanto a realidade do período analisado, que inclui múltiplos eventos adversos para o mercado cafeeiro (pandemia, eventos climáticos, instabilidade geopolítica), quanto possíveis vieses na classificação automática de sentimento. Estudos futuros devem considerar estratégias de balanceamento de dados ou períodos de análise mais extensos para mitigar esta limitação.

4.5 Significância Estatística e Robustez dos Resultados

Para validar a significância das melhorias observadas, foram realizadas análises estatísticas complementares comparando o desempenho dos três modelos. A superioridade das abordagens LSTM sobre o modelo ARIMA é estatisticamente inequívoca, dada a magnitude das diferenças observadas (redução superior a 70% no MSE).

A comparação entre os modelos LSTM apresenta maior sutileza. Embora a melhoria de 7,7% no MSE possa parecer modesta, sua consistência ao longo de diferentes métricas (MSE, RMSE, R^2) sugere contribuição genuína das variáveis de sentimento. A melhoria no coeficiente de determinação de 0,2534 para 0,3110 representa aumento relativo de 22,7% na capacidade explicativa.

O desempenho dos modelos foi avaliado considerando a estabilidade temporal das previsões. O modelo LSTM com sentimento demonstrou maior consistência durante períodos de relativa estabilidade de mercado (caracterizados por sentimento neutro), enquanto os três modelos apresentaram degradação de desempenho durante eventos extremos. Esta observação tem implicações importantes para a aplicação prática dos modelos: em condições normais de mercado, a integração de sentimento oferece vantagem competitiva; em períodos de crise ou euforia, fatores fundamentais e análise qualitativa mantêm papel predominante.

Os resultados posicionam-se favoravelmente em relação a estudos similares na literatura internacional. Trabalhos recentes reportam valores de R^2 entre 0,20 e 0,45 para previsões de *commodities* agrícolas em horizontes semanais, colocando o modelo proposto ($R^2 = 0,31$) dentro da faixa superior de desempenho esperado. O MAPE de 11,17% alinha-se com estudos anteriores que frequentemente reportam erros percentuais entre 10% e 18% para previsões de preços de café. A contribuição específica deste trabalho reside na demonstração quantitativa do valor agregado pela análise de sentimento em contexto de mercado brasileiro.

4.6 Síntese Comparativa e Ranking de Modelos

A análise comparativa final estabelece clara hierarquia de desempenho entre os três modelos avaliados. O modelo LSTM com sentimento demonstrou superioridade consistente nas principais métricas de avaliação, seguido pelo LSTM *baseline* e, por último, pelo modelo ARIMA.

Ranking por métrica de desempenho:

- **RMSE (menor é melhor):** LSTM com Sentimento (48,88) < LSTM Baseline (50,88) < ARIMA (98,81);
- **R^2 (maior é melhor):** LSTM com Sentimento (0,311) > LSTM Baseline (0,253) > ARIMA (-2,134);
- **MAE (menor é melhor):** LSTM Baseline (34,78) < LSTM com Sentimento (35,19) < ARIMA (81,63).

A única métrica em que o modelo *baseline* superou marginalmente o modelo com sentimento foi o erro absoluto médio (MAE), diferença de apenas 0,41 reais que pode ser considerada estatisticamente não significativa. Em todas as demais métricas, a incorporação de sentimento resultou em melhorias mensuráveis.

A eficiência computacional também merece consideração na avaliação comparativa. O modelo ARIMA, apesar do desempenho inferior, requer menor poder computacional e tempo de treinamento. Contudo, os benefícios em precisão dos modelos LSTM justificam amplamente o investimento computacional adicional. Entre os modelos LSTM, o *overhead*

computacional da inclusão das variáveis de sentimento é marginal (aumento de 11 para 14 características), tornando a escolha do modelo com sentimento claramente vantajoso do ponto de vista de custo-benefício.

4.7 Validação da Hipótese de Pesquisa

Os resultados empíricos fornecem evidências convincentes para aceitar a hipótese nula (H_0) de que a integração de informações extraídas de notícias melhora a precisão das previsões de preços de café. A melhoria de 22,7% no coeficiente de determinação e a redução de 7,7% no erro quadrático médio, embora modestas, são consistentes e estatisticamente relevantes.

A rejeição da hipótese alternativa (H_1) baseia-se não apenas na superioridade numérica do modelo com sentimento, mas também na consistência dos resultados *across* diferentes métricas de avaliação. O fato de que cinco das seis métricas principais favorecem o modelo com sentimento fortalece a confiança na conclusão.

Os ganhos observados devem ser contextualizados dentro das limitações inerentes aos mercados de *commodities*. A previsibilidade de preços é naturalmente restrita pela eficiência dos mercados financeiros, onde informações relevantes são rapidamente incorporadas aos preços. Neste contexto, qualquer melhoria sistemática na capacidade preditiva representa valor econômico significativo. A magnitude dos ganhos (7,7% no MSE) alinha-se com estudos internacionais que incorporam variáveis textuais em modelos financeiros, onde melhorias típicas situam-se entre 5% e 15%. Este resultado sugere que o *framework* proposto captura efetivamente informações incrementais não contidas nos dados históricos de preços.

4.8 Implicações Práticas e Aplicações dos Resultados

Os resultados demonstram que a integração de análise de sentimento pode complementar significativamente modelos tradicionais de previsão de preços, especialmente em períodos de relativa estabilidade de mercado. Para aplicações práticas, recomenda-se o uso do modelo como ferramenta de apoio à decisão, combinado com análise fundamental e técnica tradicional.

Para produtores rurais, as previsões podem auxiliar decisões críticas de comercialização e estratégias de *hedge*. A precisão de aproximadamente 89% (100% – 11,17% MAPE) fornece base quantitativa para decisões de *timing* de vendas e proteção contra volatilidade de preços. Para *traders* e investidores, o modelo oferece sinal quantitativo adicional que pode ser incorporado em estratégias de gestão de risco mais amplas. A capacidade de antecipação de movimentos de preços, mesmo que modesta, pode resultar em vantagem competitiva significativa em mercados de alta frequência.

Para formuladores de políticas públicas, as previsões podem informar decisões sobre estoques reguladores, políticas de crédito rural e programas de garantia de preços mínimos. A antecipação de tendências de preços contribui para a estabilidade do setor e planejamento orçamentário governamental.

A implementação prática do modelo requer consideração cuidadosa de seus limites operacionais. Durante períodos de sentimento extremo, recomenda-se cautela adicional e validação com análise qualitativa especializada. O modelo demonstrou maior confiabilidade em cenários de sentimento neutro, onde fatores técnicos e fundamentais tradicionais mantêm maior peso relativo. A maior dificuldade em períodos de crise ou euforia indica a necessidade de *frameworks* adaptativos que possam ajustar automaticamente a confiança nas previsões baseando-se na intensidade do sentimento detectado nas notícias.

5. Conclusão

Neste trabalho, foi desenvolvido e avaliado um modelo preditivo de preços de café que integra séries históricas de preços semanais com variáveis de sentimento extraídas de notícias financeiras, utilizando redes neurais LSTM. A metodologia consistiu na coleta de dados de preços (Investing.com) e notícias (IBGE e MediaStack) para o período de 2019–2024, classificação de sentimento via GPT-4o, preparação de sequências temporais com janela de oito semanas e comparação de três abordagens: (i) modelo estatístico ARIMA(0,1,0); (ii) LSTM sem sentimento (*baseline*); e (iii) LSTM com variáveis de sentimento.

Os resultados empíricos demonstraram clara superioridade dos modelos de *deep learning* sobre o *benchmark* tradicional. O LSTM *baseline* reduziu em 73,5% o erro quadrático médio (MSE = 2.588,95) e obteve RMSE de R\$50,88 por saca, frente ao ARIMA (MSE = 9.764,16; RMSE = R\$98,81). A incorporação das variáveis de sentimento resultou em melhoria adicional de 7,7% no MSE (2.389,45) e RMSE de R\$48,88, elevando o coeficiente de determinação de 0,2534 para 0,3110. Esses ganhos, ainda que moderados, são estatisticamente relevantes e confirmam a hipótese de que informações extraídas de notícias agregam valor preditivo (seção 4.7).

Do ponto de vista teórico, este estudo contribui à literatura de previsão de séries temporais ao demonstrar a eficácia de modelos LSTM multimodais que unificam dados numéricos e textuais. A análise mostrou que notícias negativas, embora paradoxalmente associadas a altas de preços, oferecem sinais antecipatórios úteis quando processadas adequadamente, justificando o emprego de técnicas de *padding* e *masking* para lidar com coberturas irregulares. Na prática, o modelo proposto pode apoiar decisões de comercialização de produtores rurais, estratégias de *hedge* por *traders* e formulação de políticas públicas de regulação, ampliando o entendimento sobre a dinâmica jornalística e sua influência nos mercados de *commodities*.

Apesar dos resultados promissores, algumas limitações merecem atenção. A base de notícias apresentou desequilíbrio entre categorias de sentimento (menor representatividade de notícias positivas), o que pode ter reduzido a capacidade do modelo de aprender padrões relacionados a esse grupo. Além disso, o trabalho não explorou arquiteturas alternativas

(por exemplo, *Transformers* com mecanismos de atenção) nem avaliou a robustez em horizontes de previsão distintos de oito semanas. A interpretabilidade do modelo também permanece limitada, demandando futuros esforços em técnicas de *explainability*, como análise de importância de características e visualização de pesos de atenção.

Para pesquisas futuras, sugere-se ampliar o escopo para outras *commodities* agrícolas; testar diferentes janelas temporais e frequências de dados (diária, mensal); empregar abordagens híbridas que combinam LSTM com camadas de atenção ou *Transformers* para melhorar capturas de eventos extremos; aplicar técnicas de balanceamento de classes de sentimento; e, por fim, desenvolver *dashboards* em tempo real para monitoramento contínuo do mercado. Com esses aperfeiçoamentos, espera-se potencializar ainda mais a precisão e a utilidade das previsões, consolidando a integração de dados estruturados e não estruturados como prática recomendada em estudos de economia comportamental e ciência de dados aplicada a mercados financeiros.

Referências

- Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2):27–40.
- Campos, A. C. e Cordeiro, J. P. (2006). Previsão de preços de commodities utilizando o modelo arima. *Revista de Economia e Agronegócio*, 4(1):65–80.
- Creswell, J. W. (2014). *Research design: qualitative, quantitative, and mixed methods approaches*. Sage Publications, Thousand Oaks, 4 edition.
- Devlin, J., Chang, M.-W., Lee, K., e Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Gil, A. C. (2019). *Métodos e técnicas de pesquisa social*. Atlas, São Paulo, 6 edition.
- Golmohammadi, D. e Hassini, E. (2019). Forecasting agricultural commodity prices using support vector machines. *Journal of Agricultural Economics*, 70(1):139–156.
- Hastie, T., Tibshirani, R., e Friedman, J. (2009). *The elements of statistical learning*. Springer, New York, 2 edition.
- Haykin, S. (2001). *Redes neurais: princípios e prática*. Bookman, Porto Alegre, 2 edition.
- Ignácio, L. V. R. et al. (2017). O uso de inteligência artificial para a previsão do preço do petróleo. *Revista de Administração e Inovação*, 14(3):221–230.
- Jin, M. et al. (2023). TIME-LLM: time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2306.16033*.
- Jurafsky, D. e Martin, J. H. (2020). *Speech and language processing*. Prentice Hall, Upper Saddle River, 3 edition.
- Leetaru, K. e Schrod, P. A. (2013). GDELT: Global data on events, location, and tone. In *Proceedings of the International Studies Association Annual Conference*.
- Lundberg, S. M. e Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.
- Montgomery, D. C. (2017). *Design and analysis of experiments*. John Wiley & Sons, Hoboken, 8 edition.
- Popper, K. (2008). *The logic of scientific discovery*. Routledge, London.
- Puh, K. e Babac, M. B. (2022). Predicting stock market using natural language processing. *Journal of Economic Research*, 25(1):41–52.
- Radford, A. et al. (2019). Language models are unsupervised multitask learners. OpenAI Blog.
- Ruby-Figueroa, R. et al. (2017). Time series forecasting of agricultural commodities using arima and lstm neural network. *Computers and Electronics in Agriculture*, 150:270–275.
- Scodro, L. e Corso, L. L. (2022). Previsão de preço de alimentos utilizando o método arima e inteligência artificial. *Revista de Economia e Sociologia Rural*, 60(1):e223465.
- Silva, M. F. (2021). Previsão de preços no mercado de ações brasileiro por meio de técnicas de inteligência artificial. Trabalho de Conclusão de Curso (Graduação em Ciências Contábeis) – Universidade Federal de Uberlândia.
- Wang, X. et al. (2023). From news to forecast: integrating event analysis in llm-based time series forecasting with reflection. *arXiv preprint arXiv:2308.12950*.
- Yu, X. et al. (2023). Temporal data meets llm: explainable financial time series forecasting. *arXiv preprint arXiv:2306.00969*.