FUNDAÇÃO GETULIO VARGAS

ESCOLA DE ECONOMIA DE SÃO PAULO

JOÃO EDUARDO APARECIDO LUISI VIEIRA

MODELO PREDITIVO PARA PRECIFICAÇÃO DE ATIVOS COM INTEGRAÇÃO DE NOTÍCIAS DO MERCADO FINANCEIRO E TÉCNICAS DE MACHINE LEARNING

SÃO PAULO

2025

JOÃO EDUARDO APARECIDO LUISI VIEIRA

MODELO PREDITIVO PARA PRECIFICAÇÃO DE ATIVOS COM INTEGRAÇÃO DE NOTÍCIAS DO MERCADO FINANCEIRO E TÉCNICAS DE MACHINE LEARNING

Dissertação apresentada ao Programa de Mestrado Profissional da Escola de Economia de São Paulo da Fundação Getulio Vargas, como requisito para a obtenção do título de Mestre em Economia.

Área de concentração:
Engenharia Financeira.

Orientador:
Afonso de Campos Pinto
Coorientador:
Marcos Antonio Botelho

SÃO PAULO

2025

Dados Internacionais de Catalogação na Publicação (CIP)
Ficha catalográfica elaborada pelo Sistema de Bibliotecas/FGV

Luisi, João.
Modelo preditivo para precificação de ativos com integração de notícias do mercado financeiro e técnicas de machine learning/ João Eduardo Aparecido Luisi Vieira. – 2025.
49 f.

Orientador: Afonso de Campos Pinto.
Coorientador: Marcos Antonio Botelho.
Dissertação (mestrado profissional MPFE) – Fundação Getulio Vargas, Escola de Economia de São Paulo.

1. Modelo de precificação de ativos. 2. Processo decisório. 3. Análise de séries temporais. 4. Processamento da linguagem natural (Computação). 5. Aprendizado de computador. I. Afonso de Campos Pinto. II. Marcos Antonio Botelho. III. Dissertação (mestrado profissional MPFE) – Escola de Economia de São Paulo. IV. Fundação Getúlio Vargas. V. Título.

CDU 336.767

Ficha Catalográfica elaborada por: Isabele Oliveira dos Santos Garcia CRB SP-010191/O

JOÃO EDUARDO APARECIDO LUISI VIEIRA

MODELO PREDITIVO PARA PRECIFICAÇÃO DE ATIVOS COM INTEGRAÇÃO DE NOTÍCIAS DO MERCADO FINANCEIRO E TÉCNICAS DE MACHINE LEARNING

Dissertação apresentada ao Programa de Mestrado Profissional da Escola de Economia de São Paulo da Fundação Getulio Vargas, como requisito para a obtenção do título de Mestre em Economia.

Área de concentração:
Engenharia Financeira.

Data da Aprovação: 27 / 06 / 2025

Banca Examinadora:

Afonso de Campos Pinto
(Orientador)
FGV - EESP

Marcos Antonio Botelho
(Coorientador)
FGV - EESP

Prof. Dr. Douglas Coimbra de Andrade
ISI-SE

Dedico este trabalho, primeiramente, à minha família, pelo apoio incondicional em todas as minhas decisões. A mim mesmo, pela coragem de começar e pela curiosidade incansável de me aprofundar. Aos professores da FGV, que foram guias fundamentais nesta jornada, e aos colegas do Insper, professores como eu, que me inspiraram e motivaram a seguir na carreira acadêmica. E, por fim, a Deus, por ter me concedido todas essas oportunidades e forças ao longo do caminho.

Agradecimentos

Gostaria de expressar minha sincera gratidão à minha esposa, Isabella, pelo suporte incondicional ao longo de todo o processo de concepção e desenvolvimento deste trabalho. Sua paciência, compreensão e incentivo foram fundamentais para que eu pudesse me dedicar a esta pesquisa.

Agradeço também ao Prof. Dr. Marcos Antonio Botelho e ao Prof. Dr. Afonso de Campos Pinto pelas orientações valiosas e pela inspiração que proporcionaram, tanto no âmbito acadêmico quanto pessoal. Um agradecimento especial ao Prof. Dr. Douglas Coimbra de Andrade, cujo profundo conhecimento da área foi essencial para me ajudar a superar os diversos obstáculos e desafios encontrados durante o percurso.

Minha gratidão também ao Nícolas Behar Cambraia, que, desde os tempos da graduação, tem sido uma fonte constante de incentivo, motivação e amizade, impulsionando meu crescimento profissional e acadêmico.

Por fim, não poderia deixar de mencionar meus gatos, que, de maneira única, ofereceram apoio emocional em todas as etapas e entregas deste trabalho acadêmico.

"Sapere aude!"

Horácio (Quintus Horatius Flaccus)

RESUMO

Este estudo propõe um modelo preditivo para a precificação de ativos financeiros, integrando análise de sentimento de notícias econômicas e dados financeiros históricos por meio de técnicas de aprendizado de máquina. O processo inclui a automação da coleta de dados, aplicação de métodos de processamento de linguagem natural (NLP) para classificação de notícias e uso de algoritmos como LSTM e LightGBM para previsão de preços. Os resultados indicam que a inclusão de variáveis textuais melhora o desempenho preditivo em relação aos modelos baseados apenas em variáveis técnicas, com ganhos observados em métricas como R^2, MAE, MAPE e RMSE. A pesquisa reforça o potencial da análise de dados não estruturados como complemento à modelagem tradicional de séries temporais em finanças.

Palavras-chave: Precificação de Ativos. Análise de Sentimento. Aprendizado de Máquina. Séries Temporais. Processamento de Linguagem Natural.

ABSTRACT

This study proposes a predictive model for the pricing of financial assets by integrating sentiment analysis of economic news and historical financial data through machine learning techniques. The process includes automated data collection, the application of natural language processing (NLP) methods for news classification, and the use of algorithms such as LSTM and LightGBM for price forecasting. The results indicate that the inclusion of textual variables improves predictive performance compared to models based solely on technical variables, with gains observed in metrics such as R^2, MAE, MAPE, and RMSE. The research reinforces the potential of unstructured data analysis as a complement to traditional time series modeling in finance.

Keywords: Asset Pricing. Sentiment Analysis. Machine Learning. Time Series. Natural Language Processing.

Lista de ilustrações

Figura 1 – Arquitetura LSTM. Fonte: adaptado de Medium DataDrivenInvestor. 19
Figura 2 – Estratégia de crescimento *leaf-wise* no *LightGBM*. Fonte: documentação oficial do LightGBM. 21
Figura 3 – Distribuição das classificações de sentimento das notícias brasileiras. 28
Figura 4 – Distribuição contínua dos scores de sentimento das notícias brasileiras. 28
Figura 5 – Distribuição das classificações de sentimento para notícias relacionadas à Tesla. . . . 31
Figura 6 – Distribuição das classificações de sentimento para notícias relacionadas à inflação. . . 31
Figura 7 – Distribuição das classificações de sentimento para notícias relacionadas à taxa de juros. 32
Figura 8 – Comparação das classificações de sentimento entre os diferentes conjuntos de notícias do New York Times. 33
Figura 9 – VALE3 — LSTM com variáveis técnicas 37
Figura 10 – VALE3 — LSTM + Score de Sentimento 38
Figura 11 – VALE3 — Comparação geral 38
Figura 12 – PETR4 — LSTM com variáveis técnicas 39
Figura 13 – PETR4 — LSTM + Score de Sentimento 39
Figura 14 – PETR4 — Comparação geral 39
Figura 15 – ITUB4 — LSTM com variáveis técnicas 40
Figura 16 – ITUB4 — LSTM + Score de Sentimento 40
Figura 17 – ITUB4 — Comparação geral 40
Figura 18 – BOVA11 — LSTM com variáveis técnicas 41
Figura 19 – BOVA11 — LSTM + Score de Sentimento 41
Figura 20 – BOVA11 — Comparação geral 41
Figura 21 – Importância das features no modelo LightGBM para Tesla 43
Figura 22 – Tesla — Previsão do LSTM vs. Valores Reais 44
Figura 23 – Tesla — Previsão do LSTM com sentimentos vs. Valores Reais 44
Figura 24 – Tesla — Previsão do LightGBM vs. Valores Reais 44
Figura 25 – Tesla — Previsão do LightGBM com sentimentos vs. Valores Reais 45
Figura 26 – Tesla — Comparação final: LSTM puro, LSTM com sentimentos e LightGBM 45

Lista de tabelas

Tabela 1 – Exemplo de notícias com scores de sentimento (FinBERT-PT-BR). 27
Tabela 2 – Exemplo de notícias (Tesla) com scores gerados pelo FinBERT. 29
Tabela 3 – Exemplo de notícias (Inflação) com scores gerados pelo FinBERT. 30
Tabela 4 – Exemplo de notícias (Taxa de Juros) com scores gerados pelo FinBERT. 30
Tabela 5 – Métricas do LSTM puro e LSTM + Score de Sentimento para ativos brasileiros . . . 42
Tabela 6 – Métricas comparativas dos modelos para Tesla 45

Lista de abreviaturas e siglas

API	Interface de Programação de Aplicações (*Application Programming Interface*)
ARCH	Heterocedasticidade Condicional Autorregressiva (*Autoregressive Conditional Heteroskedasticity*)
ARIMA	Modelo Autorregressivo Integrado de Médias Móveis (*AutoRegressive Integrated Moving Average*)
BERT	Representações Bidirecionais Codificadas com Transformadores (*Bidirectional Encoder Representations from Transformers*)
BOVA11	Fundo de Índice do Ibovespa (*iShares Ibovespa Index ETF*)
CNN	Canal de Notícias CNN Brasil (*Cable News Network*)
CSV	Valores Separados por Vírgula (*Comma-Separated Values*)
EMA	Média Móvel Exponencial (*Exponential Moving Average*)
FinBERT	Variante do BERT adaptada ao domínio financeiro (*Financial Bidirectional Encoder Representations from Transformers*)
GARCH	Modelo Generalizado de Heterocedasticidade Condicional Autorregressiva (*Generalized Autoregressive Conditional Heteroskedasticity*)
HTTP	Protocolo de Transferência de Hipertexto (*HyperText Transfer Protocol*)
ITUB4	Ação do Itaú Unibanco (*Itaú Preferred Stock*)
JSON	Notação de Objetos JavaScript (*JavaScript Object Notation*)
LSTM	Memória de Longo e Curto Prazo (*Long Short-Term Memory*)
MAE	Erro Absoluto Médio (*Mean Absolute Error*)
MACD	Convergência/Divergência de Médias Móveis (*Moving Average Convergence Divergence*)
MAPE	Erro Percentual Absoluto Médio (*Mean Absolute Percentage Error*)
ML	Aprendizado de Máquina (*Machine Learning*)
MSE	Erro Quadrático Médio (*Mean Squared Error*)
NLP	Processamento de Linguagem Natural (*Natural Language Processing*)
PETR4	Ação Preferencial da Petrobras (*Petrobras Preferred Stock*)
R^2	Coeficiente de Determinação (*Coefficient of Determination*)
RMSE	Raiz do Erro Quadrático Médio (*Root Mean Squared Error*)
RSI	Índice de Força Relativa (*Relative Strength Index*)
SMA	Média Móvel Simples (*Simple Moving Average*)
TSLA	Ação da Tesla Inc. (*Tesla Stock*)
URL	Localizador Uniforme de Recursos (*Uniform Resource Locator*)
VALE3	Ação da Vale (*Vale Common Stock*)

Sumário

1 Introdução . . . 14

2 Aspectos Gerais da Configuração do Modelo . . . 16

2.1 Modelo Preditivo de Precificação de Ativos . . . 16

2.1.1 Estrutura Temporal do Modelo . . . 16

2.1.2 Incorporação de Variáveis de Sentimento . . . 17

2.1.3 Objetivo de Otimização . . . 17

2.1.4 Generalização para Diferentes Modelos . . . 17

2.1.5 Resumo da Abordagem . . . 17

2.2 Modelos Tradicionais de Séries Temporais . . . 18

2.3 Processamento de Linguagem Natural (NLP) em Finanças . . . 18

2.4 Modelos de Linguagem: BERT e FinBERT . . . 18

2.5 Modelos de Machine Learning . . . 19

2.5.1 Redes Neurais: LSTM . . . 19

2.5.2 Árvore de Decisão: LightGBM . . . 20

2.6 Desafios e Limitações na Literatura . . . 21

3 Métricas de Avaliação . . . 22

3.1 Mean Squared Error (MSE) . . . 22

3.2 Root Mean Squared Error (RMSE) . . . 22

3.3 Mean Absolute Error (MAE) . . . 22

3.4 Mean Absolute Percentage Error (MAPE) . . . 23

3.5 R^2 (Coeficiente de Determinação) . . . 23

4 Metodologia . . . 24

4.1 Coleta de Dados . . . 24

4.1.1 Coleta de Notícias Brasileiras . . . 24

4.1.2 Coleta de Notícias do New York Times . . . 25

4.1.3 Coleta de Preços dos Ativos . . . 26

4.2 Processamento de Dados Brasileiros . . . 26

4.2.1 Notícias Brasileiras . . . 26

4.2.2 Pipeline de Processamento . . . 27

4.2.3 Exemplo de Dados Gerados . . . 27

4.2.4 Resumo das Classificações . . . 27

4.3 Processamento de Dados Internacionais . . . 28

4.3.1 Notícias do New York Times . . . 28

4.3.2 Pipeline de Processamento . . . 29

4.3.3 Exemplo de Dados Gerados . . . 29

4.3.4 Resumo das Classificações . . . 30

4.4 Modelagem . . . 33

4.4.1 Modelagem com Notícias Brasileiras . . . 33

4.4.2 Modelagem com Notícias do New York Times . . . 35

5 Resultados . . . 37

5.1 Ativos Brasileiros . . . 37

5.1.1 Gráficos Comparativos - Vale . . . 37
5.1.2 Gráficos Comparativos - Petrobrás . . . 38
5.1.3 Gráficos Comparativos - Itaú . . . 40
5.1.4 Gráficos Comparativos - Ibovespa . . . 41
5.1.5 Desempenho dos Modelos . . . 41
5.1.6 Análise Geral . . . 42
5.2 Ativo Americano — Tesla (TSLA) . . . 42
5.2.1 Importância das Variáveis . . . 43
5.2.2 Comparação Visual dos Modelos . . . 43
5.2.3 Desempenho Quantitativo . . . 45
5.2.4 Interpretação dos Resultados . . . 46

6 Conclusão . . . 47
6.1 Análise Crítica dos Resultados . . . 47
6.2 Considerações Finais e Trabalhos Futuros . . . 47

Referências . . . 49

1 Introdução

O mercado financeiro é profundamente influenciado por informações, sejam elas provenientes de eventos macroeconômicos, políticas governamentais, mudanças regulatórias ou notícias corporativas. A teoria dos mercados eficientes, apresentada por Fama (FAMA, 1970), argumenta que os preços dos ativos refletem todas as informações disponíveis no mercado, tornando extremamente relevante a análise de novos fluxos informacionais para identificar potenciais desvios, anomalias ou oportunidades exploráveis.

Nas últimas décadas, o crescimento exponencial do volume de dados, impulsionado pela digitalização, globalização e a constante atualização de fluxos informacionais, tornou-se um desafio crescente para analistas e investidores. Processar, interpretar e incorporar tais dados em modelos quantitativos passou a exigir metodologias mais robustas e adaptáveis. Neste cenário, os avanços em *machine learning* (ML) e *natural language processing* (NLP) transformaram profundamente a maneira como dados não estruturados, como textos de notícias e relatórios financeiros, são tratados e analisados. Ferramentas como o *FinBERT* (ARACI, 2019) e técnicas descritas por López de Prado (PRADO, 2018) oferecem novas possibilidades para identificar padrões ocultos e relações complexas em grandes volumes de dados financeiros, indo além das limitações da análise tradicional.

A literatura recente reforça a relevância da análise de sentimento no contexto financeiro. Tetlock (TETLOCK, 2007) demonstrou que notícias com tom negativo tendem a provocar reações mais intensas nos preços dos ativos, evidenciando uma assimetria no comportamento dos investidores. Loughran e McDonald (LOUGHRAN; MCDONALD, 2011) destacaram a importância da personalização de dicionários financeiros para melhorar a acurácia da análise de sentimentos. Modelos baseados em aprendizado profundo, como o *BERT* (DEVLIN et al., 2019) e o *FinBERT* (ARACI, 2019), ampliaram ainda mais a capacidade de interpretar o conteúdo semântico de textos financeiros. Estudos mais recentes, como os de Chen et al. (CHEN; PELGER; ZHU, 2022) e Alberg et al. (ALBERG; LIPTON et al., 2021), avançaram ao integrar essas variáveis textuais com modelos quantitativos preditivos, evidenciando ganhos significativos de *performance*.

Apesar dos avanços mencionados, ainda há uma lacuna relevante na literatura no que se refere à integração sistemática entre dados de sentimento extraídos de notícias e modelos de previsão de preços de ativos financeiros. Grande parte dos estudos concentra-se na previsão de preços com base exclusivamente em variáveis quantitativas ou em dados históricos, sem incorporar de forma robusta o impacto qualitativo das informações veiculadas pela mídia. Mesmo entre os trabalhos que utilizam análise de sentimentos, muitos se restringem a eventos específicos ou ativos isolados, deixando de lado uma visão mais ampla e comparativa. Assim, esta dissertação se propõe a avançar na compreensão de como o fluxo noticioso influencia a precificação de ativos, avaliando a eficácia da combinação entre modelos de linguagem e algoritmos de previsão em diferentes cenários de mercado.

O objetivo central desta dissertação é investigar como o sentimento extraído de notícias financeiras pode ser incorporado, de forma eficiente, em modelos de previsão de preços de ativos financeiros. Para isso, propõe-se uma abordagem interdisciplinar que combina fundamentos de econometria (WOOLDRIDGE, 2006), técnicas de aprendizado de máquina supervisionado e métodos avançados de *processamento de linguagem natural*. A proposta visa explorar tanto relações lineares quanto não lineares, considerando a dinâmica temporal das séries financeiras e o impacto de informações qualitativas.

Do ponto de vista aplicado, o estudo implementa e testa o modelo em diferentes contextos de mercado, abrangendo ativos de mercados emergentes e desenvolvidos. Embora os experimentos incluam

exemplos práticos, como ativos do mercado brasileiro e internacional, a ênfase principal está na avaliação metodológica da integração de dados de sentimento com séries temporais de preços.

A estrutura deste trabalho está organizada da seguinte forma: no próximo capítulo, é apresentado o modelo preditivo de forma geral e abstrata, incluindo sua formulação matemática e arquitetura. O Capítulo 3 discute as métricas utilizadas para avaliar o desempenho preditivo. No Capítulo 4, são detalhados os procedimentos de implementação, incluindo a coleta e o processamento dos dados. O Capítulo 5 apresenta os resultados obtidos e suas análises críticas. Por fim, o Capítulo 6 expõe as conclusões, limitações do estudo e sugestões para pesquisas futuras.

2 Aspectos Gerais da Configuração do Modelo

2.1 Modelo Preditivo de Precificação de Ativos

O problema de previsão de preços de ativos financeiros pode ser formalizado como uma tarefa de modelagem supervisionada, em que o objetivo é encontrar uma função preditiva $f(\cdot)$ capaz de mapear um conjunto de variáveis explicativas para o valor futuro de um ativo.

Sejam T_t e S_t, respectivamente, as séries temporais de variáveis técnicas e qualitativas observadas até o instante $t > 0$. O vetor T_t é composto por variáveis derivadas exclusivamente da própria série de preços do ativo, como médias móveis, volatilidade, RSI e MACD — indicadores amplamente utilizados na análise técnica por refletirem padrões estatísticos ao longo do tempo. Por sua vez, S_t representa variáveis qualitativas extraídas de fontes não estruturadas, como scores de sentimento gerados a partir de notícias financeiras, comunicados oficiais e análises de mercado. Essas variáveis capturam dimensões subjetivas que influenciam a percepção dos agentes econômicos.

A separação entre T_t e S_t permite ao modelo integrar tanto aspectos quantitativos quanto qualitativos do mercado, favorecendo uma abordagem mais abrangente para a previsão de preços. Ambos os vetores são considerados como colunas, ou seja, $T_t \in \mathbb{R}^{n\times 1}$ e $S_t \in \mathbb{R}^{m\times 1}$, e juntos formam o vetor de entrada do modelo:

$$X_t = \begin{bmatrix} T_t \\ S_t \end{bmatrix} \quad \text{com } T_t \in \mathbb{R}^{n\times 1},\ S_t \in \mathbb{R}^{m\times 1},\ X_t \in \mathbb{R}^{(n+m)\times 1} \tag{2.1}$$

É importante destacar que não há necessidade de que o número de variáveis técnicas seja igual ao número de variáveis qualitativas, ou seja, $n \neq m$ não compromete a estrutura do modelo. Como X_t é construído por meio da concatenação vertical de T_t e S_t, o vetor resultante $X_t \in \mathbb{R}^{(n+m)\times 1}$ permanece bem definido independentemente da quantidade de variáveis em cada subconjunto. Essa flexibilidade é especialmente útil quando diferentes conjuntos de dados possuem granularidades ou relevâncias distintas para a tarefa preditiva.

Assim, o objetivo central da modelagem é encontrar uma função $f(\cdot)$ tal que:

$$\hat{y}_{t+h} = f(X_t) \tag{2.2}$$

Onde $\hat{y}_{t+h}$ representa a previsão do preço (ou retorno) no horizonte h.

2.1.1 Estrutura Temporal do Modelo

Dada a natureza sequencial dos dados financeiros, a modelagem incorpora um mecanismo de janelas temporais (*sliding windows*) para capturar dependências ao longo do tempo. Define-se uma janela de tamanho k, de forma que, para cada instante t, o vetor de entrada X_t seja construído a partir das k observações anteriores $\{X_{t-k+1}, X_{t-k+2}, \ldots, X_t\}$. Com isso, o modelo aprende padrões de evolução temporal das variáveis técnicas e de sentimento.

2.1.2 Incorporação de Variáveis de Sentimento

A inclusão de informações textuais no modelo ocorre por meio da transformação de textos em variáveis numéricas. Para cada t, considera-se um *score* de sentimento $s_t \in \mathbb{R}$, atribuído ao conjunto de notícias divulgadas até aquele momento. Em contextos de alta frequência, o modelo pode utilizar o último *score* disponível (*fill forward*) para preencher lacunas de tempo entre novas publicações.

Opcionalmente, podem ser incluídas defasagens desses *scores*, de modo a capturar efeitos retardados das notícias, de forma que S_t seja construída em termos das defasagens dos *scores* de sentimentos $\{s_t, s_{t-1}, s_{t-2}, \ldots, s_{t-p}\}$, onde p representa o número de defasagens consideradas.

2.1.3 Objetivo de Otimização

O treinamento do modelo consiste em minimizar uma função de perda $\mathcal{L}(\theta)$, onde θ representa o vetor de parâmetros ajustáveis do modelo (e.g., pesos da rede LSTM ou folhas das árvores do LightGBM). Um exemplo clássico de função de perda para problemas de regressão é o erro quadrático médio (MSE):

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2.3}$$

Onde N é o número de amostras utilizadas no treinamento.

2.1.4 Generalização para Diferentes Modelos

A função $f(\cdot)$ pode ser implementada por diferentes arquiteturas de aprendizado de máquina, de acordo com as características dos dados disponíveis. Nesta dissertação, dois tipos principais de modelos foram considerados:

- **Redes Neurais Recorrentes do tipo LSTM:** adequadas para capturar relações temporais de curto e longo prazo, modelando sequências de dados com múltiplas variáveis de entrada.
- **Modelos de Árvore de Decisão com Boosting (LightGBM):** eficientes para lidar com grandes conjuntos de dados tabulares e para capturar interações não lineares entre as variáveis técnicas e os scores de sentimento, especialmente quando enriquecidos com *lags* explícitos.

A escolha entre as arquiteturas depende da granularidade, volume e estrutura dos dados disponíveis.

2.1.5 Resumo da Abordagem

Em síntese, o modelo proposto combina informações estruturadas (preços históricos) e não estruturadas (sentimentos extraídos de textos), integrando-as em um arcabouço de aprendizado de máquina com o objetivo de melhorar a acurácia das previsões de preços de ativos financeiros. A etapa seguinte detalha, de maneira empírica, a implementação específica do modelo para os ativos selecionados neste estudo.

2.2 Modelos Tradicionais de Séries Temporais

Os modelos clássicos de séries temporais, como ARIMA, ARCH e GARCH, foram amplamente utilizados nas últimas décadas para previsão de preços de ativos financeiros. O modelo *ARIMA* (*AutoRegressive Integrated Moving Average*), formalizado por Box e Jenkins (BOX; JENKINS; REINSEL, 1976), combina componentes autorregressivos (dependência entre valores passados), médias móveis (ajuste de choques aleatórios) e diferenciação (remoção de tendências) para lidar com séries estacionárias ou transformadas. Já os modelos *ARCH* e *GARCH*, propostos por Engle (ENGLE, 1982) e Bollerslev (BOLLERSLEV, 1986), foram desenvolvidos para modelar volatilidade condicional e heterocedasticidade — aspectos cruciais em séries financeiras com alta variabilidade.

Apesar de serem robustos para padrões lineares, essas abordagens não capturam bem relações não lineares e não conseguem incorporar fontes não estruturadas, como textos de notícias. Como destacou Diebold (DIEBOLD, 2012), há um potencial promissor em expandir modelos tradicionais incorporando variáveis qualitativas, como sentimentos de mercado extraídos de texto.

2.3 Processamento de Linguagem Natural (NLP) em Finanças

O Processamento de Linguagem Natural (*Natural Language Processing*, *NLP*) refere-se ao campo de estudo que busca capacitar máquinas a entender, interpretar e gerar linguagem humana. Em finanças, o *NLP* é aplicado para extrair informações relevantes de documentos não estruturados, como notícias, relatórios financeiros, discussões em mídias sociais e comunicados de empresas. Técnicas de NLP incluem tokenização, lematização, análise sintática, extração de entidades nomeadas (*Named Entity Recognition*, *NER*), análise de sentimento e construção de embeddings de palavras ou frases.

Estudos como Tetlock (TETLOCK, 2007) demonstraram que o tom emocional das notícias afeta diretamente os retornos de mercado, enquanto Loughran e McDonald (LOUGHRAN; MCDONALD, 2011) mostraram que dicionários financeiros especializados são essenciais para evitar falsas classificações em análises de sentimento. Com o avanço das arquiteturas de deep learning, especialmente os Transformers, surgiram modelos como BERT e FinBERT, capazes de capturar nuances semânticas complexas, como sarcasmo, ambiguidade e jargões específicos do domínio financeiro.

O uso de dados textuais revolucionou a modelagem financeira. Tetlock (TETLOCK, 2007) demonstrou que o tom emocional das notícias afeta diretamente retornos, com notícias negativas gerando reações mais acentuadas. Loughran e McDonald (LOUGHRAN; MCDONALD, 2011) mostraram que dicionários financeiros especializados são essenciais para evitar falsas classificações em análises de sentimento.

Estudos recentes, como Chen et al. (CHEN; PELGER; ZHU, 2022) e Alberg et al. (ALBERG; LIPTON et al., 2021), exploraram a integração entre dados textuais e quantitativos, mostrando ganhos de desempenho preditivo, principalmente em contextos de alta volatilidade e incerteza.

2.4 Modelos de Linguagem: BERT e FinBERT

Com o avanço dos Transformers, surgiram arquiteturas como o BERT (Bidirectional Encoder Representations from Transformers) (DEVLIN et al., 2019), capazes de gerar embeddings contextuais profundos, considerando interdependências bidirecionais em texto. Porém, como ressaltado por Araci (ARACI, 2019), o BERT genérico apresenta limitações no vocabulário financeiro, o que motivou o

surgimento do FinBERT: uma versão treinada especificamente em corpora econômicos, capaz de capturar nuances e jargões financeiros em tarefas como classificação de sentimento.

2.5 Modelos de Machine Learning

Além dos modelos econométricos tradicionais, algoritmos de machine learning têm sido amplamente explorados na previsão de preços e retornos de ativos. Entre os grupos mais relevantes estão as redes neurais profundas e os modelos baseados em árvores de decisão.

2.5.1 Redes Neurais: LSTM

O *Long Short-Term Memory (LSTM)*, introduzido por Hochreiter e Schmidhuber (HOCHREITER; SCHMIDHUBER, 1997), supera limitações das redes neurais recorrentes tradicionais ao resolver o problema do desaparecimento do gradiente. Ele faz isso usando uma arquitetura composta por portas que controlam o fluxo de informações.

As principais componentes são:

- Porta de esquecimento (f_t): decide o que deve ser descartado da memória antiga,
- Porta de entrada (i_t): seleciona quais novos dados entram na memória,
- Vetor candidato ($\tilde{C}_t$): gera novos conteúdos candidatos a serem armazenados,
- Memória da célula (C_t): estado interno que armazena informações ao longo do tempo,
- Porta de saída (o_t): decide o que será passado para o próximo estado oculto,
- Estado oculto (h_t): saída visível da célula naquele instante.

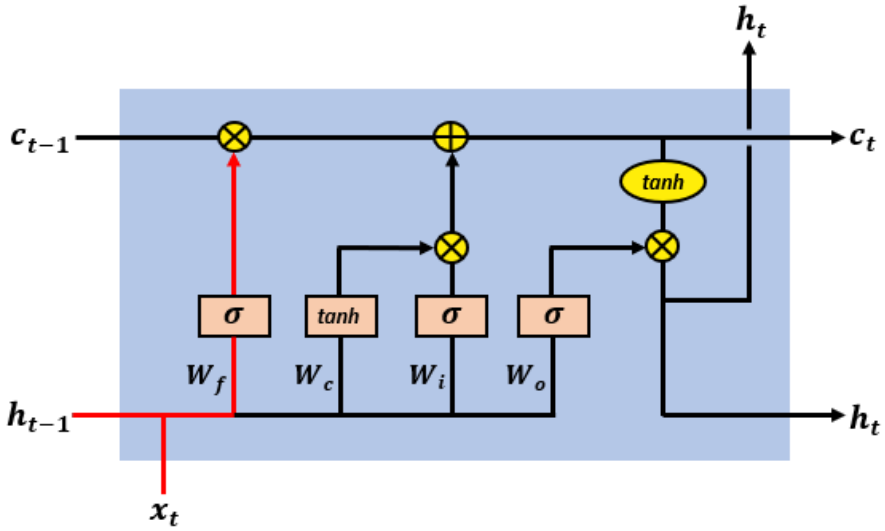

Figura 1 – Arquitetura LSTM. Fonte: adaptado de Medium DataDrivenInvestor[1]

Formalmente:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.4}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.5}$$

$$\tilde{C}t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{2.6}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}t \tag{2.7}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.8}$$

$$h_t = o_t * \tanh(C_t) \tag{2.9}$$

As equações acima descrevem o funcionamento interno de uma célula LSTM (Long Short-Term Memory), um tipo de rede neural recorrente projetada para capturar dependências de longo prazo em séries temporais. Cada célula processa a entrada atual x_t, combinada com o estado oculto anterior h_{t-1}, por meio de diferentes portas que controlam o fluxo de informações.

A concatenação $[h_{t-1}, x_t]$ representa a junção dos vetores de estado oculto anterior e da entrada atual, resultando em um vetor único que é multiplicado por diferentes matrizes de pesos W_f, W_i, W_C, W_o, dependendo da operação. O símbolo $\cdot$ denota a multiplicação matricial padrão, enquanto o símbolo $*$ representa a multiplicação elemento a elemento (Hadamard product), típica em mecanismos de portas.

A porta de esquecimento f_t (Eq. 2.4) determina qual fração da memória anterior C_{t-1} deve ser mantida. Em seguida, a porta de entrada i_t (Eq. 2.5) decide quais informações novas serão adicionadas, e $\tilde{C}_t$ (Eq. 2.6) representa a candidata a novo conteúdo da célula. A atualização da célula de memória C_t (Eq. 2.7) combina essas operações, balanceando memória antiga e nova. Finalmente, a porta de saída o_t (Eq. 2.8) regula quais partes da memória atual serão expostas como estado oculto h_t, o qual é usado tanto como saída da célula quanto como entrada para o próximo instante de tempo (Eq. 2.9).

Essas operações tornam o LSTM capaz de preservar informações relevantes ao longo do tempo, reduzindo o impacto do problema do desvanecimento do gradiente frequentemente observado em redes recorrentes tradicionais.

2.5.2 Árvore de Decisão: LightGBM

O *LightGBM*, introduzido por Ke et al. (KE et al., 2017), é um algoritmo de *boosting* eficiente que constrói árvores de decisão de forma iterativa. Ele adota um crescimento *leaf-wise* (crescimento por folha), onde a divisão ocorre sempre no nó que maximiza a redução de perda, ao invés de seguir uma estrutura nivelada (*level-wise*).

Além disso, trabalhos como Fischer e Krauss (FISCHER; KRAUSS, 2018) demonstraram que modelos de *ensemble*, como *Random Forest* e *Gradient Boosting*, conseguem superar *benchmarks* tradicionais, principalmente quando combinados com features derivadas de texto, como *scores* de sentimento e volume de notícias.

A função objetivo pode ser escrita como:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.10}$$

em que:

- $l(y_i, \hat{y}_i)$: função de perda (como o erro quadrático médio MSE),
- $\Omega(f_k)$: termo de regularização para penalizar complexidade,
- f_k: cada árvore do ensemble.

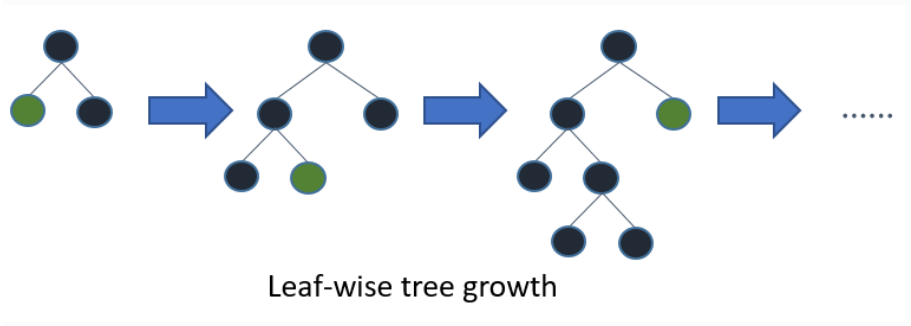

Figura 2 – Estratégia de crescimento *leaf-wise* no *LightGBM*. Fonte: documentação oficial do LightGBM.

2.6 Desafios e Limitações na Literatura

Apesar dos avanços significativos, a literatura aponta desafios importantes:

- **Ruído e Redundância nos Dados Textuais:** Nem toda notícia tem impacto financeiro relevante, e incluir textos irrelevantes pode confundir os modelos.
- **Risco de Overfitting:** Modelos complexos, como *LSTMs* e *Transformers*, são suscetíveis a overfitting, exigindo regularização cuidadosa.
- **Horizonte Temporal:** O impacto do texto pode ser distinto dependendo se a previsão é intradiária, diária ou semanal, exigindo modelos calibrados para diferentes janelas.

3 Métricas de Avaliação

A avaliação dos modelos preditivos é feita usando métricas clássicas de regressão, que quantificam o quão próximos estão os valores preditos dos valores reais. As métricas selecionadas foram escolhidas por capturarem diferentes aspectos do erro.

3.1 Mean Squared Error (MSE)

O Erro Quadrático Médio (MSE) mede a média dos quadrados das diferenças entre os valores preditos $\hat{y}_i$ e os valores reais y_i:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.1}$$

em que:

- y_i: valor real no ponto i.
- $\hat{y}_i$: valor predito no ponto i.
- n: número total de observações.

O MSE penaliza mais fortemente erros grandes devido à elevação ao quadrado, sendo particularmente útil quando queremos evitar grandes desvios pontuais. Ele é sensível a outliers, o que pode ser vantajoso em mercados financeiros voláteis.

3.2 Root Mean Squared Error (RMSE)

A Raiz do Erro Quadrático Médio (RMSE) é a raiz quadrada do MSE:

$$RMSE = \sqrt{MSE} \tag{3.2}$$

Isso devolve o erro médio à mesma escala dos dados originais, facilitando a interpretação. Um RMSE baixo indica que, em média, as previsões estão próximas dos valores reais.

3.3 Mean Absolute Error (MAE)

O Erro Absoluto Médio (MAE) calcula a média das diferenças absolutas:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.3}$$

Diferente do MSE, o MAE não penaliza fortemente grandes erros, oferecendo uma visão mais estável quando outliers não são de interesse. Ele é útil para entender o erro médio "puro" entre as previsões.

3.4 Mean Absolute Percentage Error (MAPE)

O Erro Percentual Absoluto Médio (MAPE) mede o erro médio em termos percentuais:

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{3.4}$$

Ele facilita comparações entre séries de diferentes magnitudes, permitindo dizer, por exemplo, que o modelo errou em média "X%" em relação ao valor real.

3.5 R^2 (Coeficiente de Determinação)

O coeficiente de determinação R^2 mede a proporção da variância nos dados que é explicada pelo modelo:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{3.5}$$

Onde $\bar{y}$ é a média dos valores reais. Valores próximos de 1 indicam que o modelo explica bem a variabilidade dos dados, enquanto valores próximos de 0 ou negativos indicam baixo poder explicativo.

4 Metodologia

A metodologia adotada neste estudo foi projetada para integrar a análise de dados não estruturados, como notícias econômicas, com dados estruturados de preços de ativos financeiros, com o objetivo de desenvolver um modelo preditivo eficiente para a precificação de ativos. Essa abordagem combina técnicas de processamento de linguagem natural (NLP), aprendizado de máquina e coleta automatizada de dados para criar uma solução robusta e automatizada.

O processo metodológico é estruturado em etapas, abrangendo desde a coleta de dados até a validação do modelo. Primeiramente, um pipeline foi desenvolvido para automatizar a coleta de notícias econômicas de portais relevantes, junto com os preços dos ativos mais mencionados nesses contextos. Essas informações são armazenadas em um banco de dados centralizado para garantir consistência e acessibilidade durante as análises.

Em seguida, os dados textuais das notícias são processados por meio de técnicas de NLP, incluindo análise de sentimento e classificação, com o objetivo de avaliar a polaridade (positiva, neutra ou negativa) e a relevância do conteúdo. Paralelamente, os dados históricos de preços são analisados e integrados ao modelo. Algoritmos de aprendizado de máquina são então utilizados para correlacionar os sentimentos das notícias com as variações nos preços dos ativos.

Por fim, a eficácia do modelo é avaliada com base em métricas quantitativas, validando sua capacidade de prever movimentos de mercado de alta ou baixa. Este capítulo detalha cada uma dessas etapas, apresentando as ferramentas, técnicas e abordagens utilizadas para atingir os objetivos propostos.

A metodologia adotada neste trabalho visa conectar dados qualitativos, provenientes de notícias e informações de mercado, com dados quantitativos, como preços de ativos financeiros, para construir modelos preditivos que avaliem o impacto de notícias no comportamento de mercado. Este processo é dividido em múltiplas etapas interligadas, descritas a seguir.

4.1 Coleta de Dados

A coleta de dados deste estudo foi dividida em duas etapas principais: (i) obtenção de notícias financeiras brasileiras e (ii) coleta de notícias internacionais a partir do New York Times. Para uma descrição detalhada dos códigos utilizados na etapa de coleta de dados, consulte o Apêndice A.

4.1.1 Coleta de Notícias Brasileiras

As notícias brasileiras foram coletadas entre dezembro de 2024 e março de 2025, totalizando aproximadamente 3000 notícias. As fontes selecionadas foram:

G1: especificamente da seção Economia/Negócios (<https://g1.globo.com/economia/negocios/>).

CNN Brasil: seção Economia (<https://www.cnnbrasil.com.br/economia/>).

Bloomberg Línea: seções Mercado e Negócios (<https://www.bloomberglinea.com.br/mercados/> e <https://www.bloomberglinea.com.br/negocios/>).

O processo foi realizado por meio de um web crawler desenvolvido em *Python*, utilizando as seguintes bibliotecas:

- **requests**: para fazer requisições HTTP e obter o conteúdo HTML das páginas.
- **BeautifulSoup** (do pacote `bs4`): para fazer o parsing do HTML e extrair links e textos relevantes.
- **sqlite3**: para armazenar localmente as notícias coletadas em um banco de dados relacional leve.
- **datetime**: para registrar e manipular datas associadas às notícias.

O *crawler* foi programado para percorrer as páginas principais das seções-alvo e extrair todos os links que correspondessem a notícias relevantes, removendo duplicados. Em seguida, cada link era visitado individualmente para capturar o conteúdo completo: título, corpo da matéria, data e fonte. Esses dados eram então inseridos no banco SQLite para posterior análise.

O script principal incluía:

- Um módulo para configurar o banco de dados e garantir a criação da tabela `news`, com campos para título, link, conteúdo, data e fonte.
- Uma função `coletar_links_bloomberg()` que fazia requisições às páginas da Bloomberg Línea e filtrava os links de interesse.
- Uma função `coletar_conteudo_bloomberg(link)` que abria cada link, extraía o título (geralmente do elemento `<h1>`) e concatenava todos os parágrafos do artigo (`<p>` dentro de `<article>`) para formar o conteúdo.

Esse processo foi replicado, com adaptações, para os portais G1 e CNN Brasil, respeitando as diferenças na estrutura HTML de cada site.

4.1.2 Coleta de Notícias do New York Times

A segunda etapa consistiu na coleta de notícias do New York Times, usando a **New York Times API**. Diferente do método aplicado aos portais brasileiros (via scraping direto), a API permitiu filtrar as notícias por:

- **Palavras-chave**: como “Tesla”, “inflação” e “taxa de juros”.
- **Seções específicas**, incluindo:
  - Business (Economia em geral)
  - Business Day (Notícias econômicas do dia)
  - Entrepreneurs (Startups e inovação)
  - Financial (Finanças e mercado financeiro)
  - Technology (Setor tecnológico)
  - U.S. (Política e governo dos EUA)
  - Politics (Campanhas, eleições, decisões políticas)
  - Washington (Governo federal dos EUA)

– World (Política e economia internacionais)

As notícias foram coletadas entre 2018 e 2022, resultando em mais de 10.000 entradas. Cada requisição à API retornava dados estruturados em formato JSON, contendo informações como título, resumo, URL, data de publicação e seção, que foram armazenados para posterior alinhamento com os dados de preços.

4.1.3 Coleta de Preços dos Ativos

Para alinhar os dados de sentimento com os preços de mercado, foi realizada a coleta de preços históricos dos ativos utilizados no estudo. Foram considerados tanto os preços das ações da Tesla (TSLA), negociadas no mercado norte-americano, quanto os preços dos ativos brasileiros (VALE3, PETR4, ITUB4 e BOVA11).

No caso da Tesla, os dados foram obtidos através da biblioteca **yfinance**, que acessa a API do Yahoo Finance, para o período de 1º de janeiro de 2018 a 31 de dezembro de 2022. O script em Python realizou o download das séries históricas de preços diários, incluindo informações como preço de abertura, fechamento, máxima, mínima e volume, e os armazenou em formato CSV para posterior integração com os dados de sentimento.

Já para os ativos brasileiros, foi utilizada também a biblioteca **yfinance**, porém com uma configuração específica para coletar dados em intervalos de 5 minutos, permitindo uma análise mais granular durante o período entre 18 de janeiro de 2025 e 4 de março de 2025. Além das métricas tradicionais (abertura, fechamento, máxima, mínima e volume), foram calculadas variáveis adicionais, como:

- **Retornos logarítmicos de 5 minutos**: computados com base na variação percentual entre preços consecutivos.
- **Volatilidade intradiária**: obtida como o desvio padrão móvel das séries de retornos em janelas de 12 períodos (aproximadamente uma hora).
- **Volatilidade diária**: anualizada a partir de uma janela móvel de 5 dias, para capturar oscilações de curto prazo no contexto diário.

Os dados coletados passaram por processos de tratamento, como remoção de valores ausentes e sincronização temporal com as séries de notícias, garantindo consistência na etapa de modelagem. Dessa forma, foi possível construir uma base robusta para treinar os modelos preditivos, tanto para os ativos nacionais quanto para o ativo internacional Tesla.

4.2 Processamento de Dados Brasileiros

4.2.1 Notícias Brasileiras

Para as notícias brasileiras, foi utilizado o modelo pré-treinado **FinBERT-PT-BR**, desenvolvido especificamente para análise de sentimentos em textos financeiros em português. Esse modelo foi construído em duas etapas: (i) treinamento de linguagem com mais de 1,4 milhão de textos financeiros, (ii) treinamento de classificação de sentimentos com cerca de 500 textos rotulados.

O modelo mostrou desempenho superior a outros modelos de estado da arte em português, sendo adequado para aplicações como construção de índices de sentimento, estratégias de investimento e análise macroeconômica.

4.2.2 Pipeline de Processamento

O código Python desenvolveu um pipeline que seguiu as etapas:

- **Leitura do dataset**: as notícias foram carregadas de um arquivo CSV contendo título e conteúdo.
- **Combinação de título e conteúdo**: os textos foram unificados em uma única string para maximizar o contexto.
- **Tokenização e classificação**: usando a biblioteca Transformers (`AutoTokenizer` e `BertForSequenceClassifica` o modelo produziu uma predição de sentimento (POSITIVO, NEGATIVO ou NEUTRO).
- **Mapeamento para score contínuo**: as classes foram convertidas em um score numérico no intervalo [-1, 1], onde positivo variava de 0 a +1, negativo de 0 a -1, e neutro era zero.
- **Armazenamento**: os resultados foram salvos em um novo arquivo CSV chamado `sentiment.csv`.

4.2.3 Exemplo de Dados Gerados

A Tabela 1 mostra exemplos do dataset final com as análises de sentimento:

Tabela 1 – Exemplo de notícias com scores de sentimento (FinBERT-PT-BR).

Data	**Título**	**Conteúdo (resumo)**	**Score**
2025-01-18	Sephora acumula reclamações	...relacionadas à Black Friday 2024...	-0.82
2025-01-18	Trabalhadores chineses são resgatados	...de situação análoga à escravidão...	-0.80
2025-01-18	Controladora do iFood compra Decola	...expansão para novas áreas de atuação...	+0.69
2025-01-18	Viagens, cesta de Natal, salários extras	...como empresas premiam funcionários...	0.00
2025-01-18	MacKenzie Scott anuncia doação	...doação de US$ 2 bilhões a ONGs...	0.00

4.2.4 Resumo das Classificações

Das notícias brasileiras analisadas, o modelo FinBERT-PT-BR gerou as seguintes classificações:

- 468 notícias foram classificadas como neutras.
- 1170 notícias foram classificadas como negativas.
- 446 notícias foram classificadas como positivas.

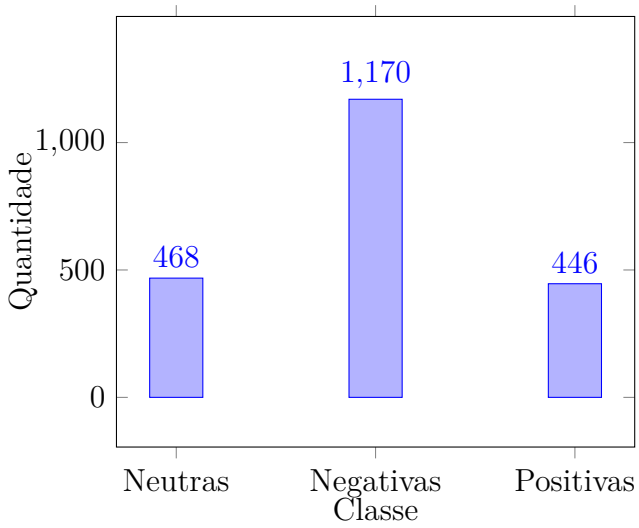

Figura 3 – Distribuição das classificações de sentimento das notícias brasileiras.

Além das classificações discretas (neutras, negativas, positivas), foi gerada a distribuição contínua dos scores de sentimento atribuídos pelo modelo FinBERT-PT-BR.

A Figura 4 mostra o histograma das frequências de *score*, que variam de -1 a +1, acompanhado de uma linha de densidade estimada. Observa-se uma concentração clara de notícias em torno de valores fortemente negativos, refletindo o viés típico de notícias financeiras, onde eventos negativos tendem a receber maior cobertura. Em contrapartida, há uma menor quantidade de notícias altamente positivas, enquanto a região neutra se destaca como um segundo pico, confirmando o número significativo de textos classificados sem polarização emocional.

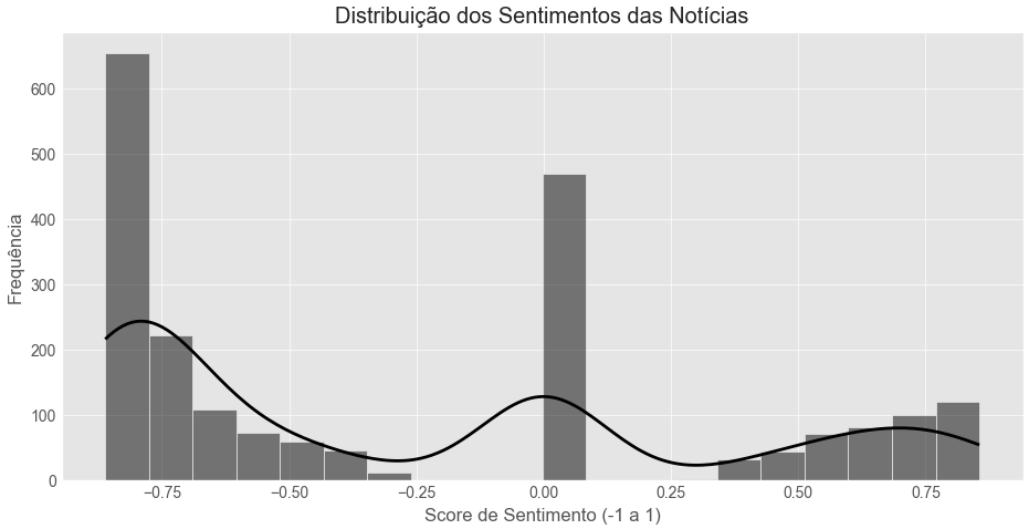

Figura 4 – Distribuição contínua dos scores de sentimento das notícias brasileiras.

4.3 Processamento de Dados Internacionais

4.3.1 Notícias do New York Times

Para as notícias extraídas do New York Times, foi utilizado o modelo **FinBERT** original (ProsusAI/finbert), treinado para textos financeiros em inglês. As notícias foram processadas em três

grupos distintos, com base em palavras-chave:

- **Tesla**: notícias contendo a palavra-chave "tesla".
- **Inflação**: notícias contendo "inflation".
- **Taxa de Juros**: notícias contendo "interest rate".

4.3.2 Pipeline de Processamento

O código Python desenvolveu um pipeline que seguiu as etapas:

- Tokenização e truncamento de textos longos para até 512 tokens.
- Predição dos scores (positive, neutral, negative) com softmax sobre as saídas do modelo.
- Armazenamento dos resultados com o score contínuo e a classe dominante.

4.3.3 Exemplo de Dados Gerados

As Tabelas 2, 3 e 4 mostram exemplos do dataset final com as análises de sentimento:

Tabela 2 – Exemplo de notícias (Tesla) com scores gerados pelo FinBERT.

Data	**Título**	**Positivo**	**Neutro**	**Negativo**	**Classe**
2018-01-31	California Today: The Los Angeles Times Has (A)	0.11	0.02	0.87	negative
2018-01-30	Weekly News Quiz for Students: Grammys, Mascot	0.04	0.04	0.92	negative
2018-01-29	JPMorgan Promotes Pinto and Smith to Co-President	0.08	0.01	0.91	negative
2018-01-24	Ford Says It's a New Era. Wall Street Isn't Buying It	0.04	0.85	0.11	neutral
2018-01-04	In Norway, Electric and Hybrid Cars Outsell Conventional Models	0.94	0.01	0.05	positive

Tabela 3 – Exemplo de notícias (Inflação) com scores gerados pelo FinBERT.

Data	**Título**	**Positivo**	**Neutro**	**Negativo**	**Classe**
2018-01-31	Facebook Calms Concerns About Ad Revenue	0.14	0.10	0.76	negative
2018-01-30	4 Questions as Amazon, Berkshire and JPMorgan Disrupt Health Care	0.08	0.01	0.91	negative
2018-01-29	Janet Yellen's Fitting Finale: Fed Plans to Stand Pat	0.05	0.09	0.85	negative
2018-01-26	Consumer Confidence Is Lifting the Economy. But for How Long?	0.49	0.03	0.49	positive
2018-01-25	Theresa May Arrives in Davos as U.K.'s Post-'Brexit' Future Looms	0.34	0.31	0.35	negative

Tabela 4 – Exemplo de notícias (Taxa de Juros) com scores gerados pelo FinBERT.

Data	**Título**	**Positivo**	**Neutro**	**Negativo**	**Classe**
2018-01-31	Facebook Calms Concerns About Ad Revenue	0.14	0.10	0.76	negative
2018-01-31	Federal Reserve Leaves Rates Unchanged	0.54	0.22	0.24	positive
2018-01-30	As Deficit Soars Toward $1 Trillion, Congress Shrugs	0.51	0.40	0.10	positive
2018-01-30	Exxon Mobil Tripling Its Bet on the Hottest U.S. Shale Field	0.94	0.01	0.05	positive
2018-01-30	4 Questions as Amazon, Berkshire and JPMorgan Disrupt Health Care	0.08	0.01	0.91	negative

4.3.4 Resumo das Classificações

Das notícias do New York Times relacionadas à palavra-chave **Tesla**, o modelo FinBERT gerou as seguintes classificações:

- 406 notícias foram classificadas como neutras.
- 816 notícias foram classificadas como negativas.
- 139 notícias foram classificadas como positivas.

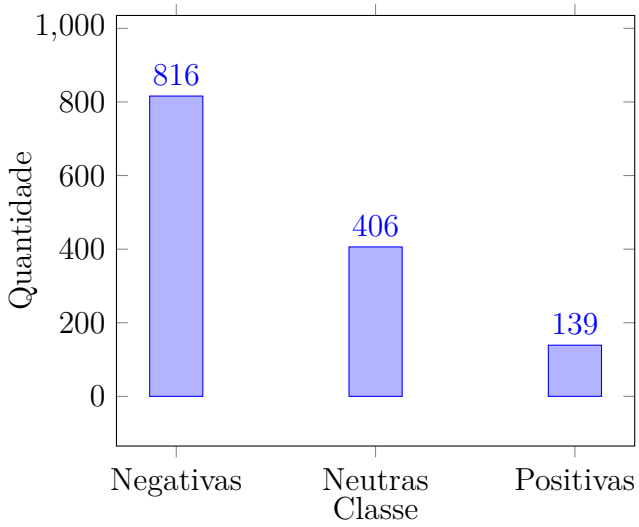

Figura 5 – Distribuição das classificações de sentimento para notícias relacionadas à Tesla.

Das notícias do New York Times relacionadas à palavra-chave **inflação**, o modelo FinBERT gerou as seguintes classificações:

- 1196 notícias foram classificadas como neutras.
- 718 notícias foram classificadas como negativas.
- 455 notícias foram classificadas como positivas.

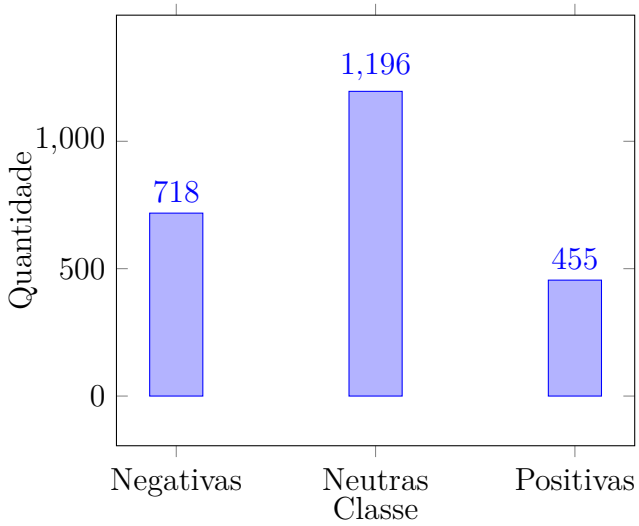

Figura 6 – Distribuição das classificações de sentimento para notícias relacionadas à inflação.

Das notícias do New York Times relacionadas à palavra-chave **taxa de juros**, o modelo FinBERT gerou as seguintes classificações:

- 1766 notícias foram classificadas como neutras.
- 1676 notícias foram classificadas como negativas.
- 562 notícias foram classificadas como positivas.

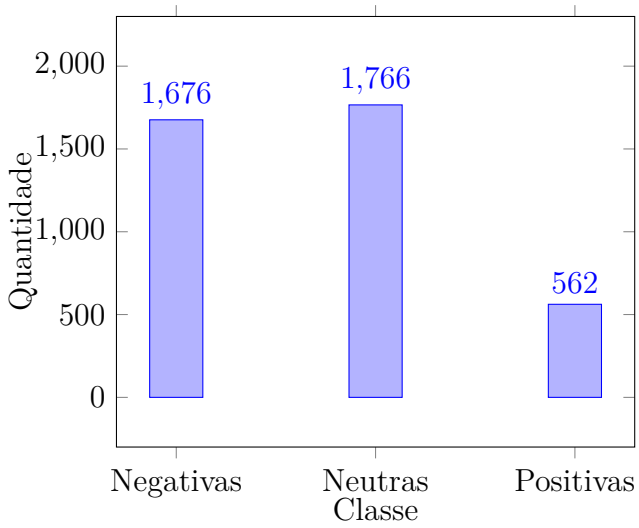

Figura 7 – Distribuição das classificações de sentimento para notícias relacionadas à taxa de juros.

A Figura 8 apresenta a distribuição das classificações de sentimento para os três conjuntos distintos de notícias coletadas do New York Times: Tesla, inflação e taxa de juros.

Observa-se que:

- O conjunto **Tesla** apresentou uma predominância marcante de classificações negativas (816 notícias), refletindo a cobertura crítica frequente em relação à empresa, suas polêmicas e desafios no mercado automotivo.
- O conjunto **inflação** teve como maioria as classificações neutras (1196 notícias), sugerindo uma cobertura mais técnica e informativa sobre o fenômeno econômico, com menor polarização emocional.
- O conjunto **taxa de juros** mostrou uma distribuição relativamente equilibrada entre negativas (1676) e neutras (1766), indicando que mudanças nas taxas são frequentemente relatadas de forma factual, mas também carregam peso negativo quando associadas a pressões sobre consumo e investimentos.

Esse panorama reforça a importância de analisar não apenas o volume de notícias, mas também a sua tonalidade emocional, que pode impactar diferentemente modelos de previsão financeira. Os resultados obtidos aqui servirão como base para a etapa seguinte, de integração dessas informações aos modelos preditivos.

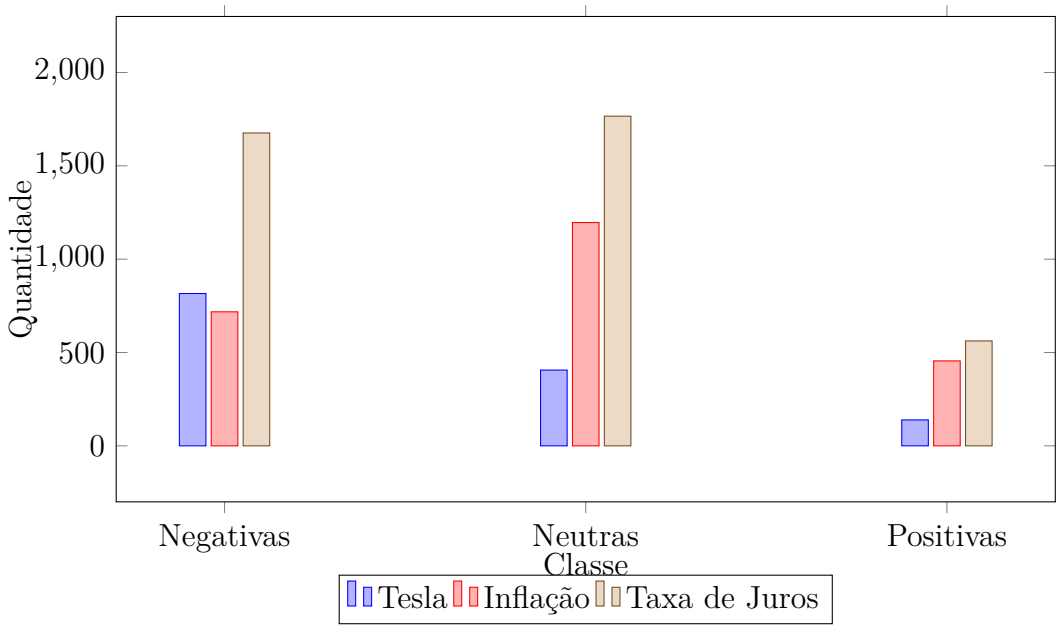

Figura 8 – Comparação das classificações de sentimento entre os diferentes conjuntos de notícias do New York Times.

4.4 Modelagem

4.4.1 Modelagem com Notícias Brasileiras

Para a modelagem das notícias brasileiras, optou-se pelo uso de redes neurais LSTM (*Long Short-Term Memory*). Essa escolha foi motivada pelo fato de o conjunto de dados conter pouco mais de 2000 notícias, coletadas entre janeiro e março, cobrindo um intervalo de tempo relativamente curto. Em séries temporais com baixa quantidade de dados, modelos como o LSTM apresentam vantagens, pois conseguem capturar dependências temporais mesmo em contextos com amostras limitadas.

Dados de Entrada:

Os dados de entrada utilizados consistiram em preços das ações brasileiras VALE3, ITUB4, PETR4 e BOVA11 a cada 5 minutos, resultando em 2504 entradas para o modelo. Como as notícias não estavam disponíveis em granularidade de 5 minutos, aplicou-se uma técnica chamada **fill forward**, que consiste em preencher os intervalos vazios com o último valor conhecido, garantindo que o sentimento mais recente permanecesse disponível até a chegada de uma nova notícia.

As **features** (variáveis independentes) selecionadas foram:

- **SMA_10** (Simple Moving Average): média móvel simples de 10 períodos.
- **EMA_10** (Exponential Moving Average): média móvel exponencial de 10 períodos.
- **RSI_14** (Relative Strength Index): índice de força relativa de 14 períodos.
- **Bollinger Upper / Lower**: bandas de Bollinger superior e inferior, que indicam volatilidade.
- **MACD** (Moving Average Convergence Divergence): indicador de convergência/divergência de médias móveis.

- **FinBERT_score**: score de sentimento obtido pelo modelo FinBERT-PT-BR.

O **target** (variável dependente) escolhido foi o preço de fechamento da ação (`Close`), que foi removido das variáveis de entrada para evitar *overfitting*.

Pré-processamento:

Os dados financeiros foram normalizados usando *MinMaxScaler*, preservando a escala entre 0 e 1, enquanto os scores de sentimento foram deixados sem normalização para manter sua contribuição direta ao modelo. Uma janela temporal de 30 períodos foi usada para gerar as sequências de entrada para o LSTM.

O conjunto foi dividido em **85% treino** e **15% teste**, garantindo que as previsões fossem avaliadas em dados nunca vistos pelo modelo.

Arquitetura do Modelo:

A arquitetura do modelo LSTM contou com:

- Três camadas LSTM sequenciais (**128**, **64** e **32** neurônios), cada uma com ativação `tanh`, *dropout* de 10% e regularização L2 para evitar *overfitting*.
- Camadas densas finais (**32** e **16** neurônios) com ativação `relu`.
- Saída final com ativação `linear`, adequada para regressão.

O modelo foi treinado com **150 épocas** e *batch size* de **32**, usando o otimizador Adam com taxa de aprendizado de 0.001. Foi aplicado **early stopping** com paciência de 10 épocas para interromper o treinamento quando não houvesse mais ganhos na validação.

Validação e Métricas:

Para avaliar o desempenho, foram calculadas as métricas:

- R^2 (coeficiente de determinação): mede a proporção da variância explicada.
- MAE (Mean Absolute Error): erro médio absoluto.
- MAPE (Mean Absolute Percentage Error): erro percentual médio absoluto.
- RMSE (Root Mean Squared Error): raiz do erro quadrático médio.

Comparação dos Modelos:

Foram treinados dois modelos:

1. LSTM utilizando apenas as variáveis técnicas (SMA, EMA, RSI, Bollinger, MACD).
2. LSTM utilizando as variáveis técnicas + os scores de sentimento do FinBERT-PT-BR.

Os resultados foram comparados por meio das métricas acima e por gráficos que mostraram o preço real versus os preços previstos por cada modelo. O objetivo foi verificar se a inclusão do sentimento das notícias melhorava a performance preditiva em relação ao modelo puramente técnico.

4.4.2 Modelagem com Notícias do New York Times

Para as notícias extraídas do New York Times, foi adotada uma abordagem comparativa entre dois modelos distintos: uma rede neural do tipo LSTM (*Long Short-Term Memory*) e um modelo de árvore de decisão baseado em LightGBM (*Light Gradient Boosting Machine*). Essa escolha foi possível graças ao volume consideravelmente maior de dados disponíveis: 1362 notícias relacionadas à Tesla, 4005 sobre taxa de juros e 2372 sobre inflação, abrangendo o período de 2018 a 2022.

Dados de Entrada:

O conjunto de dados incluiu:

- Scores de sentimento específicos para Tesla, inflação e taxa de juros (`score_tesla`, `score_inflation`, `score_interest_rate`).
- Variáveis técnicas extraídas da série temporal do preço da ação, incluindo:
  - **SMA10** (Simple Moving Average): média móvel simples de 10 períodos.
  - **EMA10** (Exponential Moving Average): média móvel exponencial de 10 períodos.
  - **RSI14** (Relative Strength Index): índice de força relativa de 14 períodos.
  - **MACD** (Moving Average Convergence Divergence): indicador de convergência/divergência de médias móveis.
  - **Volatilidade**: medida de dispersão do preço.

O **target** (variável dependente) foi definido como o preço de fechamento (`Close`).

Modelagem com LSTM:

A estrutura geral do modelo LSTM seguiu os mesmos princípios aplicados às notícias brasileiras:

- Criação de janelas temporais de 5 períodos para capturar a dinâmica sequencial.
- Normalização das variáveis de entrada usando *MinMaxScaler*.
- Treinamento em três **folds** utilizando *TimeSeriesSplit*, garantindo validação robusta para séries temporais.
- Arquitetura composta por:
  - Uma camada LSTM com 64 neurônios.
  - Camada de *Dropout* de 20% para reduzir *overfitting*.
  - Camada densa final para saída do valor predito.
- Otimização realizada com o otimizador Adam e função de perda `mean_squared_error`.
- Aplicação de **early stopping** para interromper o treinamento quando não houvesse mais ganho em validação.

As métricas calculadas para avaliação foram:

- MSE (Mean Squared Error).

- RMSE (Root Mean Squared Error).
- MAE (Mean Absolute Error).
- MAPE (Mean Absolute Percentage Error).
- R^2 (Coeficiente de determinação).

Os resultados foram gerados para cada fold, e no final, o último fold foi utilizado para comparação gráfica entre valores reais e previstos.

Modelagem com LightGBM:

O modelo LightGBM foi escolhido por sua eficiência em lidar com grandes conjuntos de dados tabulares e pela capacidade de interpretar a importância das variáveis. Ao contrário do LSTM, o LightGBM não captura automaticamente dependências temporais, sendo necessário adicionar explicitamente variáveis defasadas (*lags*) para os scores e volatilidade (`t-1`).

As etapas incluíram:

- Inclusão dos lags para capturar a informação temporal.
- Divisão do conjunto em treino (85%) e teste (15%).
- Treinamento do modelo LightGBM com:
  - 100 estimadores.
  - Taxa de aprendizado de 0.05.
- Avaliação do desempenho utilizando as mesmas métricas do LSTM.
- Extração da **importância das features**, permitindo identificar quais variáveis mais impactaram o modelo.

Comparação Final:

Ao término da modelagem, foram comparados:

- Preço real versus preço previsto pelo modelo LSTM apenas com variáveis técnicas.
- Preço real versus preço previsto pelo modelo LSTM com variáveis técnicas + scores de sentimento.
- Preço real versus preço previsto pelo modelo LightGBM com variáveis técnicas + scores de sentimento.

Essas comparações permitiram avaliar não apenas a performance dos modelos, mas também o valor agregado pela incorporação de dados qualitativos (sentimento) às abordagens quantitativas tradicionais. Além disso, a análise das importâncias das *features* no LightGBM forneceu insights sobre quais fontes de informação (Tesla, inflação, taxa de juros) apresentaram maior influência na previsão dos preços.

5 Resultados

5.1 Ativos Brasileiros

Nesta seção, analisamos o desempenho dos modelos preditivos aplicados a quatro ativos do mercado brasileiro: VALE3 (Vale), PETR4 (Petrobras), ITUB4 (Itaú) e BOVA11 (Ibovespa). Foram testados dois modelos principais:

- Um modelo LSTM alimentado apenas por variáveis técnicas (como SMA, EMA, RSI, MACD etc.);
- Um modelo LSTM aprimorado, incluindo variáveis de sentimento geradas a partir das notícias (scores derivados do FinBERT-PT-BR).

As figuras a seguir mostram, para cada ativo, os gráficos de comparação entre os preços reais e os preços previstos pelos modelos.

A seguir, têm-se todos os gráficos gerados durante os experimentos, separando os modelos LSTM com features técnicas, LSTM + Score de Sentimento, e as comparações finais. Os dias exibidos na horizontal correspondem ao conjunto de teste, que representa os 15% finais da série temporal, conforme a divisão 85/15 adotada na modelagem.

5.1.1 Gráficos Comparativos - Vale

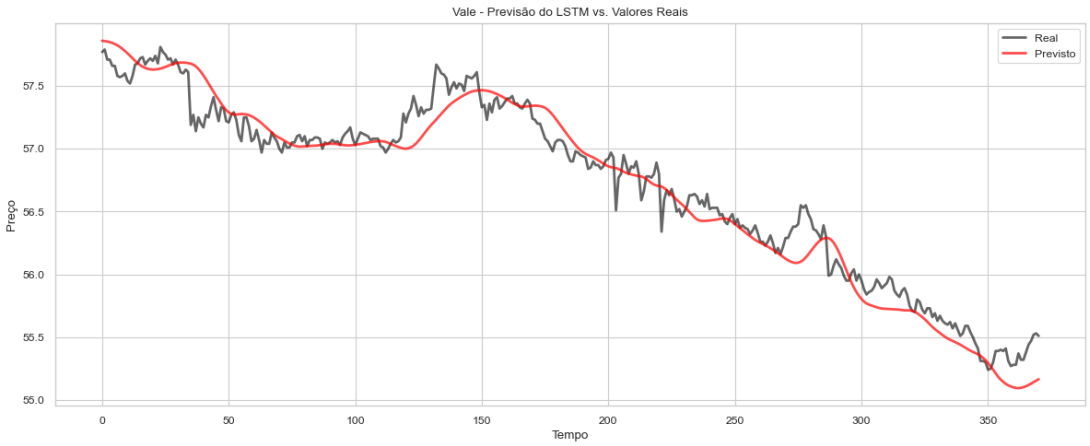

Figura 9 – VALE3 — LSTM com variáveis técnicas

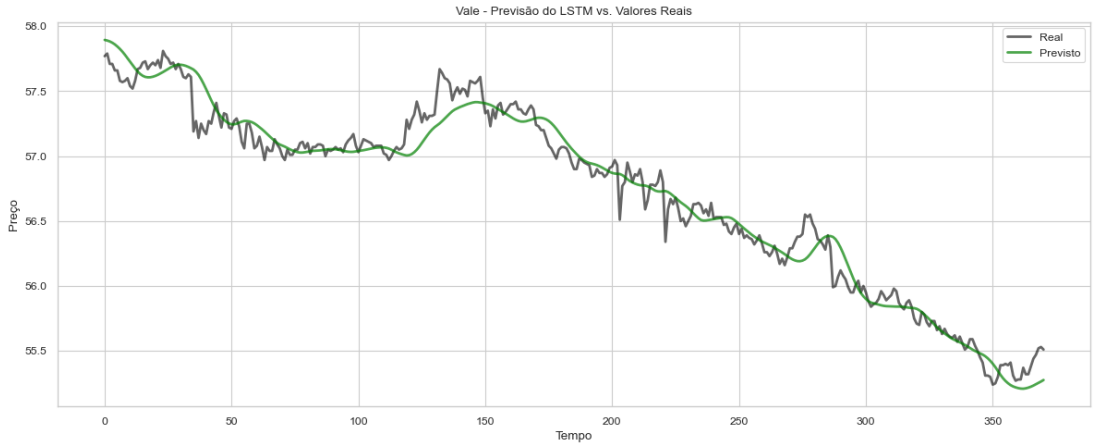

Figura 10 – VALE3 — LSTM + Score de Sentimento

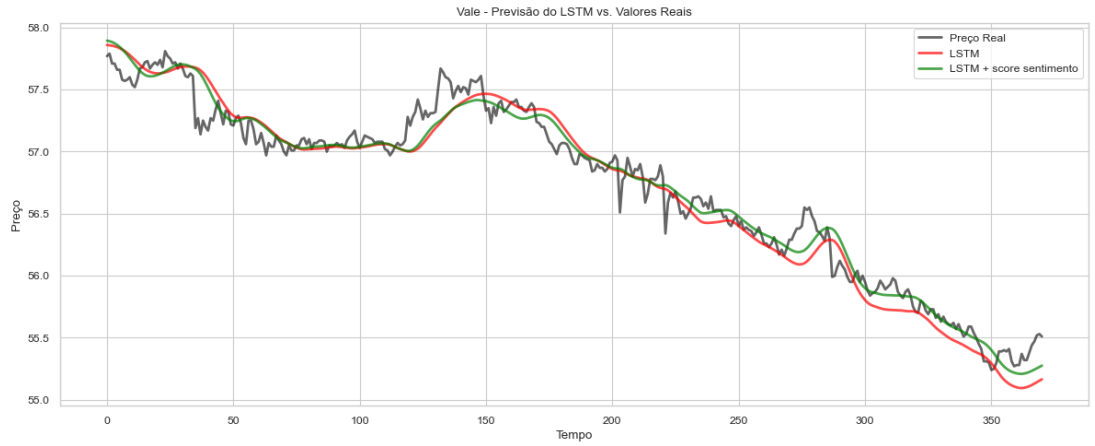

Figura 11 – VALE3 — Comparação geral

5.1.2 Gráficos Comparativos - Petrobrás

Observa-se, nas Figuras 12 e 13, uma queda abrupta no preço real da ação da Petrobras durante o período de teste, possivelmente relacionada a algum evento ou notícia de impacto relevante no mercado. Embora a introdução do *score* de sentimento no modelo LSTM tenha proporcionado um leve ganho na resposta do modelo diante da movimentação de preço, a queda real ainda não foi completamente capturada. Isso ocorre porque, apesar de os sentimentos adicionarem informações qualitativas ao modelo, sua eficácia depende da correta representação do impacto da notícia no *score* e da sua temporalidade em relação ao evento. Assim como nas previsões com variáveis técnicas, o modelo continua a suavizar variações bruscas, evidenciando a dificuldade de antecipar choques exógenos de forma precisa — mesmo com a incorporação de informações textuais.

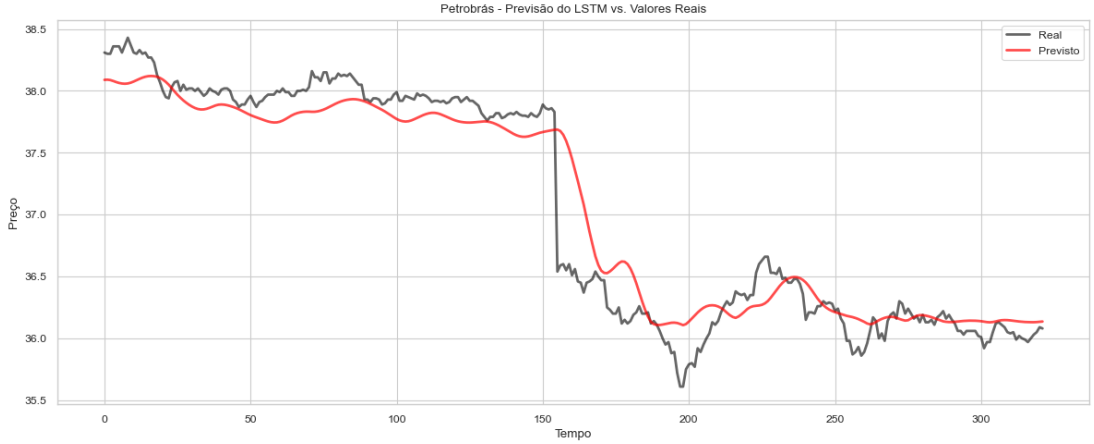

Figura 12 – PETR4 — LSTM com variáveis técnicas

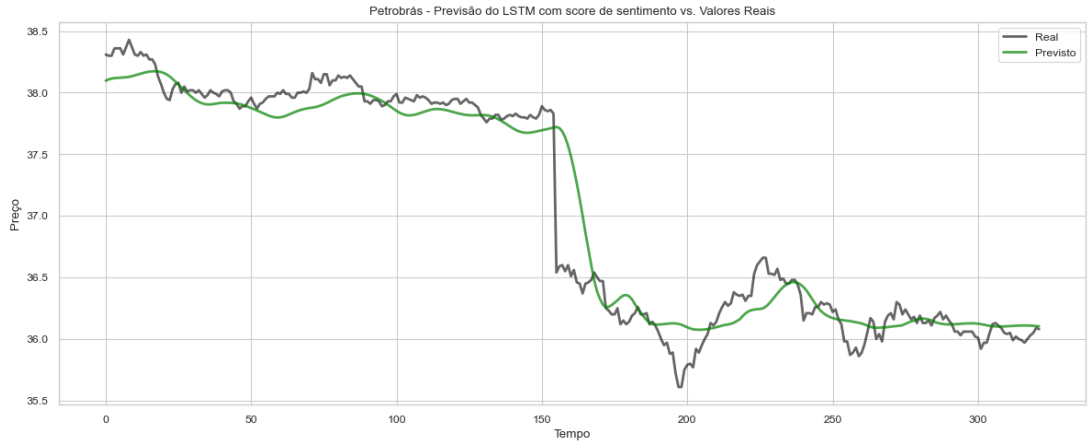

Figura 13 – PETR4 — LSTM + Score de Sentimento

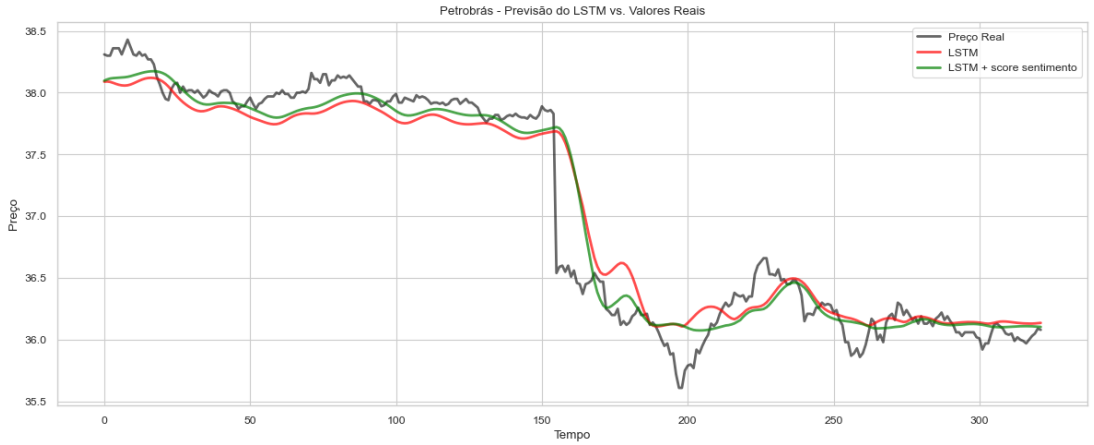

Figura 14 – PETR4 — Comparação geral

5.1.3 Gráficos Comparativos - Itaú

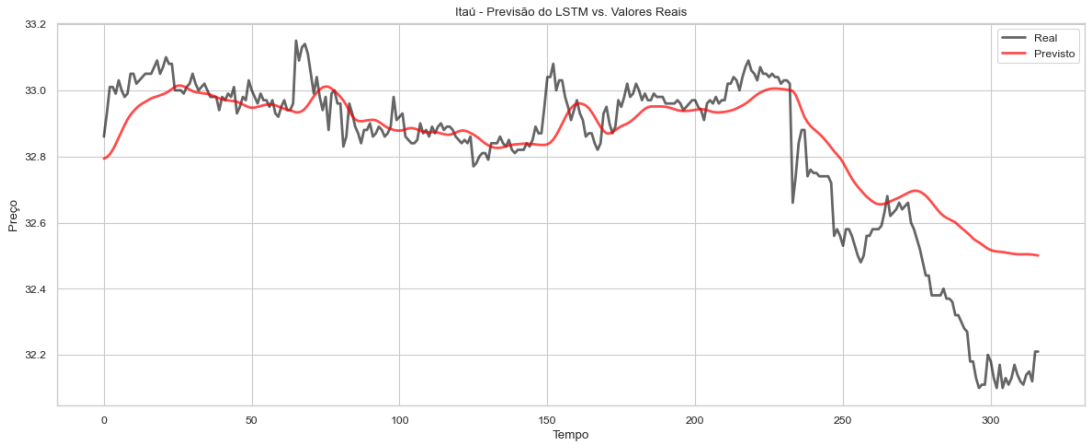

Figura 15 – ITUB4 — LSTM com variáveis técnicas

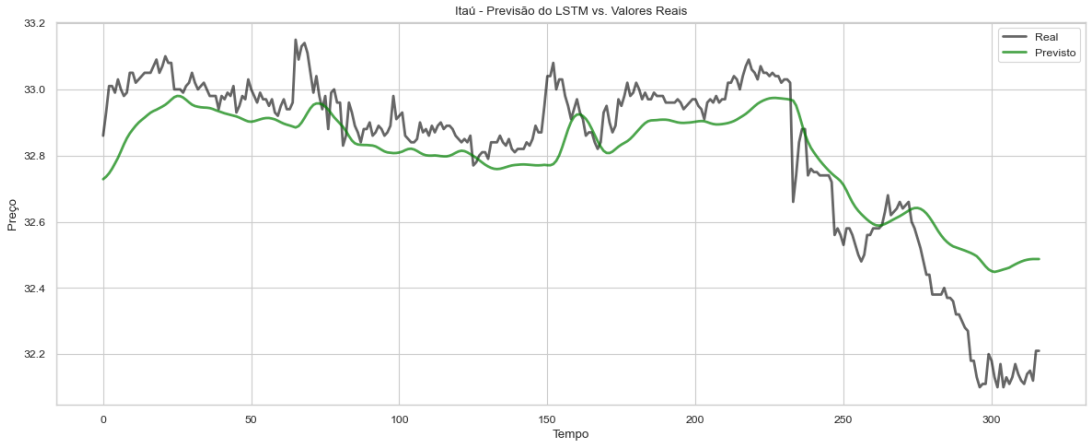

Figura 16 – ITUB4 — LSTM + Score de Sentimento

Figura 17 – ITUB4 — Comparação geral

5.1.4 Gráficos Comparativos - Ibovespa

Figura 18 – BOVA11 — LSTM com variáveis técnicas

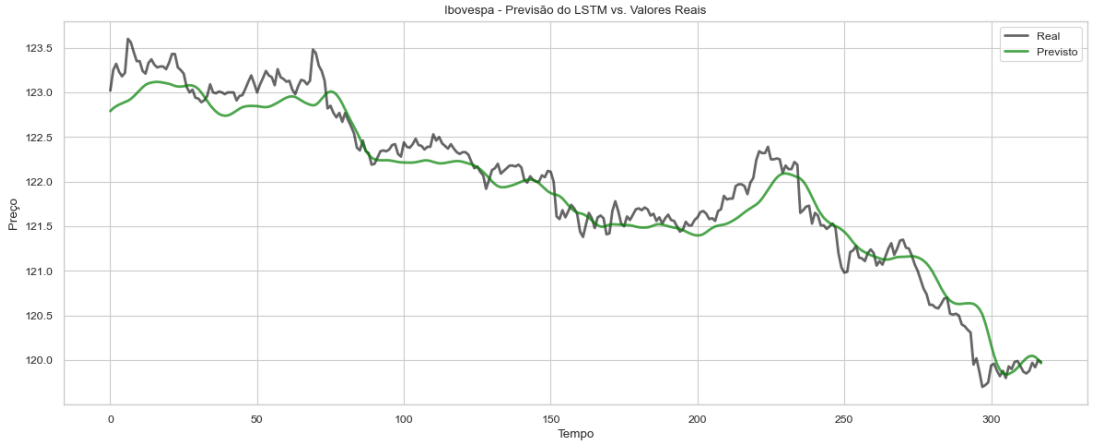

Figura 19 – BOVA11 — LSTM + Score de Sentimento

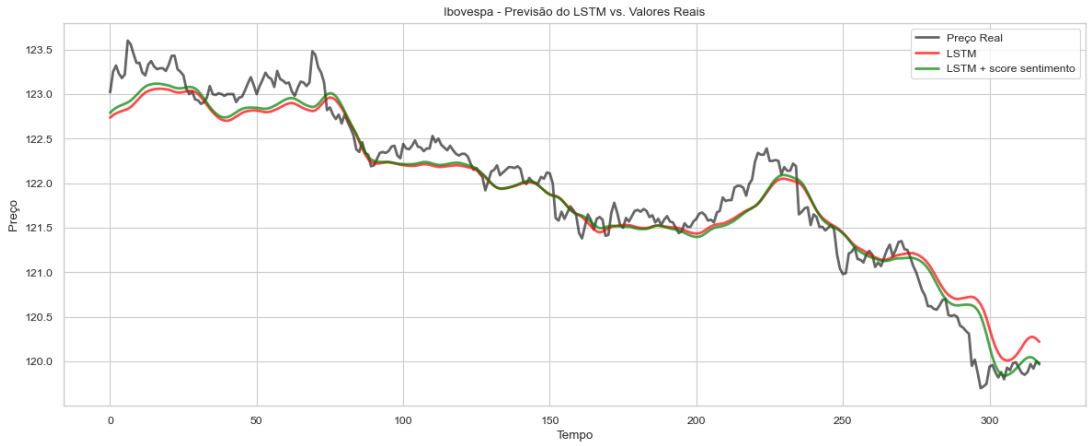

Figura 20 – BOVA11 — Comparação geral

5.1.5 Desempenho dos Modelos

A Tabela 5 resume as métricas de desempenho para os dois modelos em todos os ativos brasileiros.

Tabela 5 – Métricas do LSTM puro e LSTM + Score de Sentimento para ativos brasileiros

Ativo	**Modelo**	R^2_{Treino}	R^2_{Teste}	MAE	MAPE (%)	RMSE
VALE3	LSTM	0.9619	0.9494	0.1189	0.21	0.1559
	LSTM + Score	0.9730	0.9642	0.0968	0.17	0.1311
PETR4	LSTM	0.9374	0.9295	0.1758	0.48	0.2453
	LSTM + Score	0.9227	0.9427	0.1414	0.38	0.2210
ITUB4	LSTM	0.8907	0.7015	0.0952	0.29	0.1432
	LSTM + Score	0.9453	0.7214	0.1065	0.33	0.1384
BOVA11	LSTM	0.9481	0.9242	0.2081	0.17	0.2624
	LSTM + Score	0.9561	0.9408	0.1833	0.15	0.2318

5.1.6 Análise Geral

De modo geral, observa-se que a inclusão das variáveis de sentimento trouxe ganhos relevantes em todos os casos, principalmente no conjunto de teste, conforme mostra a Tabela 5. Por exemplo:

- Para VALE3, houve um ganho de **1,48%** no R^2_{teste}.
- Para PETR4, o ganho foi de **1,42%**.
- Para ITUB4, o aumento foi mais modesto, cerca de **2%**.
- Para BOVA11, a melhora alcançou **1,66%**.

Além dos ganhos no coeficiente de determinação (R^2), os modelos com sentimento apresentaram reduções consistentes em métricas de erro, como MAE, MAPE e RMSE, reforçando que a introdução de variáveis derivadas de notícias não apenas melhora a capacidade preditiva, mas também torna as previsões mais precisas e robustas.

5.2 Ativo Americano — Tesla (TSLA)

A análise do ativo Tesla foi a mais robusta deste trabalho, pois contou com a maior base de dados temporal (de 2018 a 2022), o maior volume de notícias coletadas e uma diversidade maior de variáveis econômicas incluídas, como scores de sentimento não apenas da Tesla, mas também de inflação e taxa de juros, além de indicadores técnicos clássicos.

Os modelos testados incluíram:

- Um modelo LSTM utilizando apenas variáveis técnicas, sem qualquer input textual.
- Um modelo LSTM utilizando variáveis técnicas + scores de sentimento derivados das notícias.
- Um modelo LightGBM utilizando variáveis técnicas + scores de sentimento + defasagens temporais (*lags*), já que o LightGBM não captura dependências temporais como o LSTM e precisa desses lags explicitamente incluídos.

5.2.1 Importância das Variáveis

A Figura 21 apresenta a importância relativa das variáveis no modelo LightGBM. Observa-se que os indicadores técnicos tradicionais, como EMA10, RSI14 e SMA10, aparecem com forte peso, mas variáveis relacionadas a sentimento (como os *scores* de sentimento da Tesla e dos indicadores macroeconômicos) também aparecem, mesmo que em menor proporção. Esse resultado reforça que, apesar dos indicadores técnicos ainda serem dominantes, o sentimento extraído das notícias acrescenta uma dimensão adicional relevante para o modelo.

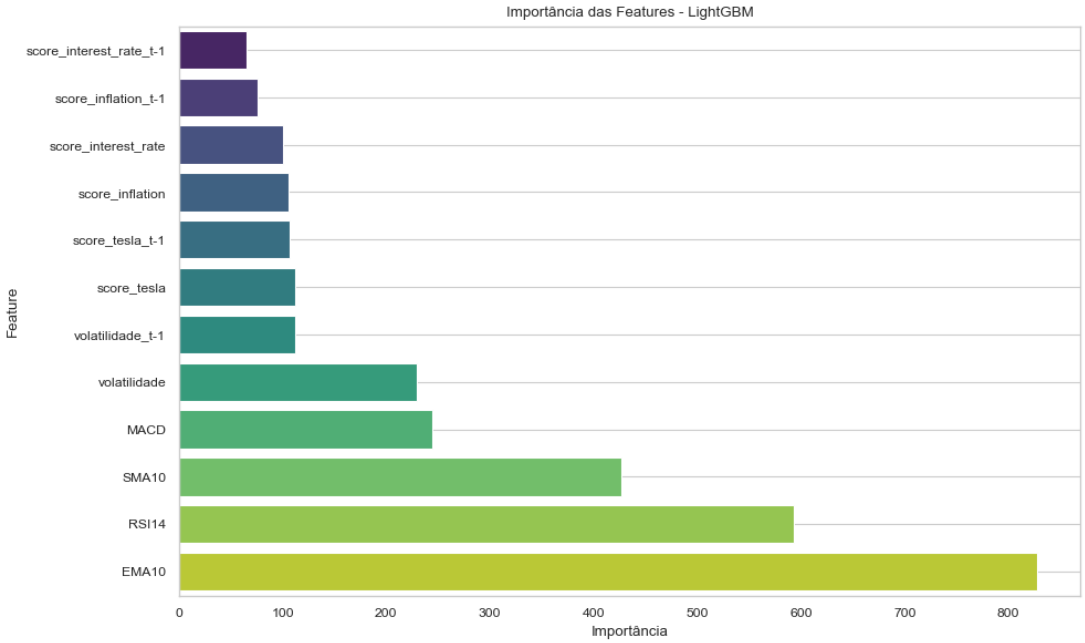

Figura 21 – Importância das features no modelo LightGBM para Tesla

5.2.2 Comparação Visual dos Modelos

As Figuras 22, 23, 24 e 25 ilustram o desempenho preditivo dos modelos testados. Inicialmente, avalia-se o LSTM utilizando apenas variáveis técnicas, seguido do LSTM com a inclusão dos scores de sentimento extraídos das notícias. Em seguida, apresenta-se os resultados do LightGBM, também comparando as versões com e sem a incorporação das variáveis de sentimento. Por fim, a Figura 26 sobrepõe os quatro cenários — LSTM técnico, LSTM com sentimento, LightGBM técnico e LightGBM com sentimento — permitindo uma comparação direta entre os modelos quanto à sua capacidade preditiva em relação aos preços reais.

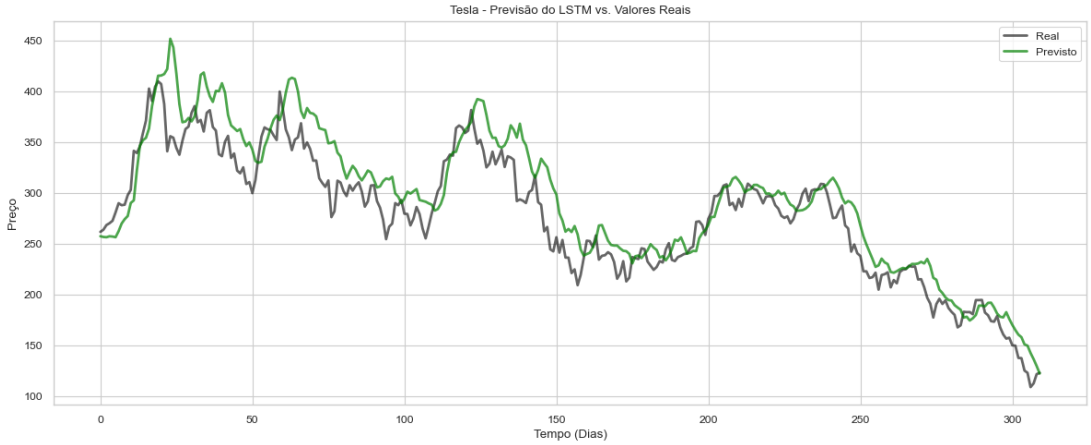

Figura 22 – Tesla — Previsão do LSTM vs. Valores Reais

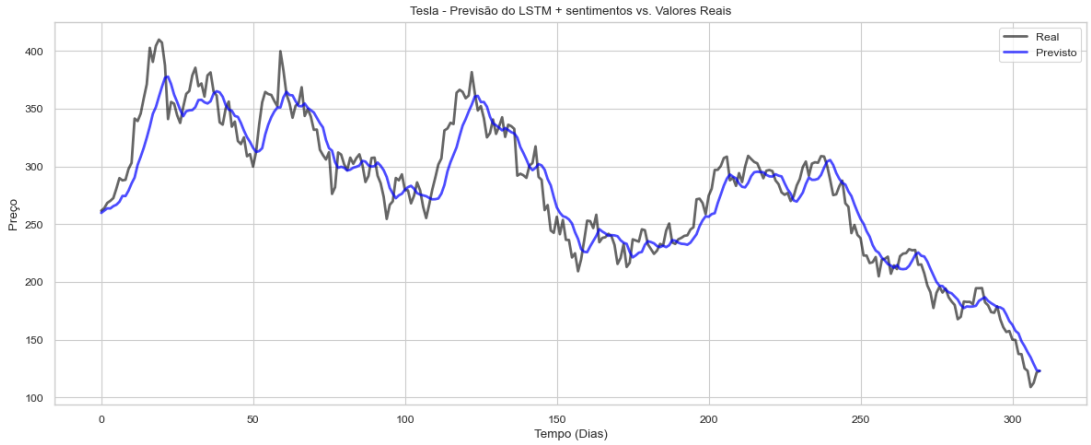

Figura 23 – Tesla — Previsão do LSTM com sentimentos vs. Valores Reais

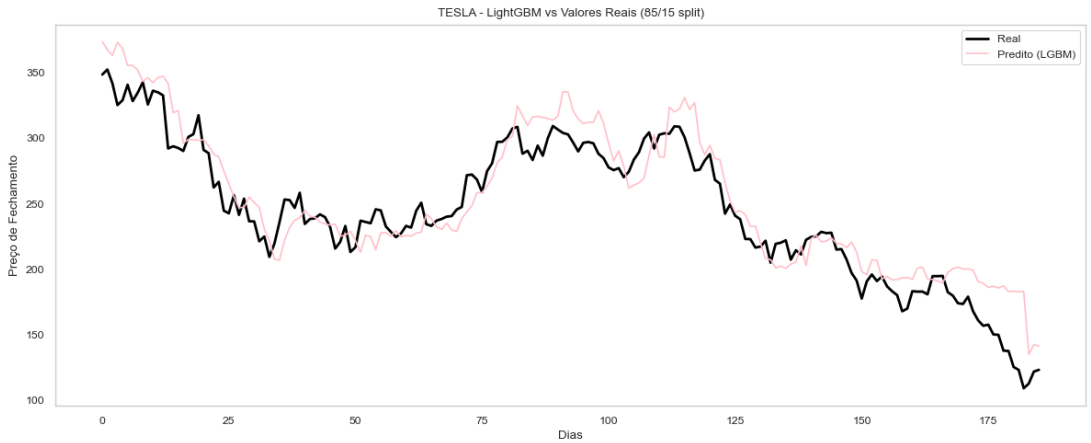

Figura 24 – Tesla — Previsão do LightGBM vs. Valores Reais

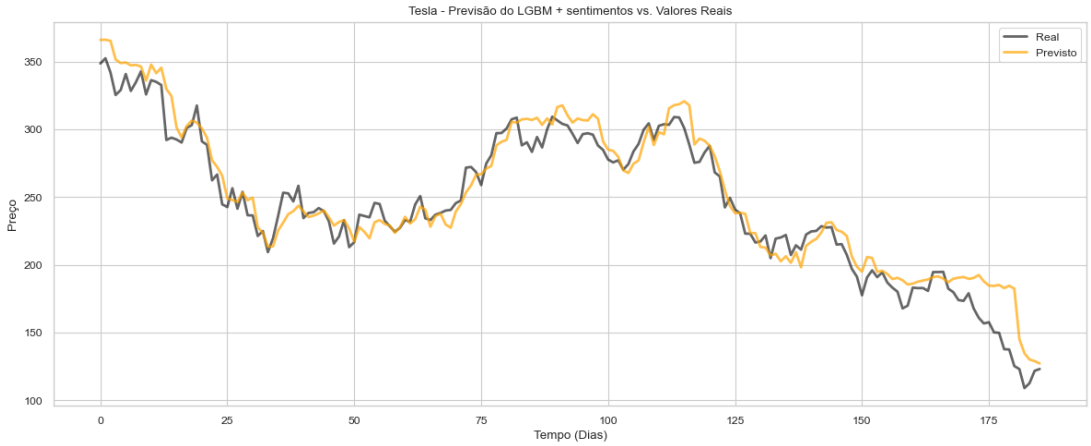

Figura 25 – Tesla — Previsão do LightGBM com sentimentos vs. Valores Reais

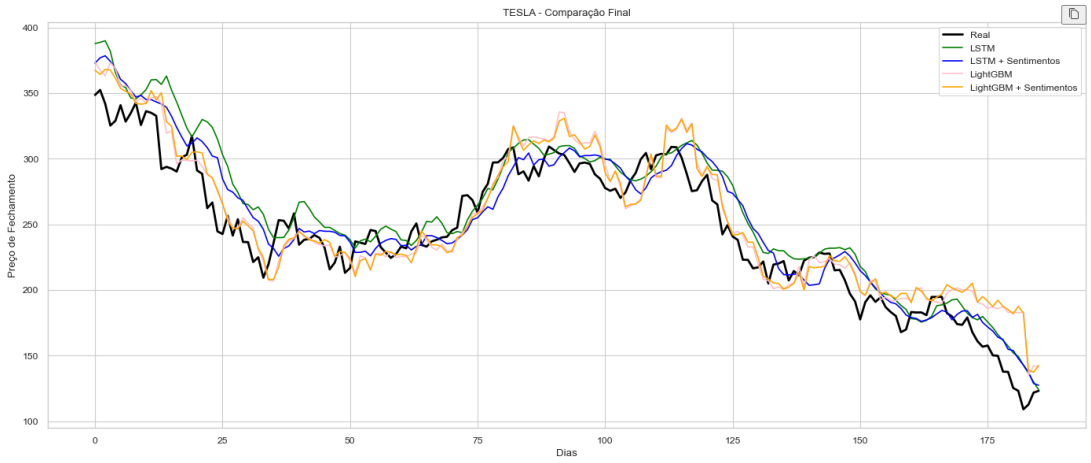

Figura 26 – Tesla — Comparação final: LSTM puro, LSTM com sentimentos e LightGBM

5.2.3 Desempenho Quantitativo

Os resultados quantitativos finais dos modelos testados para a Tesla são apresentados na Tabela 6. O modelo LSTM puro obteve um R^2 de 0.7903, com erros médios (MAE, MAPE e RMSE) relativamente elevados. A inclusão dos *scores* de sentimento no LSTM gerou uma melhora significativa: o R^2 saltou para 0.9214, e os erros médios reduziram substancialmente. Por fim, o modelo LightGBM, combinando variáveis técnicas, scores de sentimento e defasagens temporais, obteve o melhor desempenho geral, alcançando um R^2 de 0.9302 e os menores valores absolutos de erro médio.

Tabela 6 – Métricas comparativas dos modelos para Tesla

Modelo	**MSE**	**RMSE**	**MAE**	**MAPE (%)**	R^2
LSTM	591.35	24.32	18.70	7.98	0.7903
LSTM + Sentimentos	221.64	14.89	12.19	5.22	0.9214
LightGBM	430.18	20.74	16.64	7.63	0.8475
LightGBM + Sentimentos	196.97	14.03	10.72	4.98	0.9302

5.2.4 Interpretação dos Resultados

Os resultados obtidos para a Tesla confirmam a hipótese central deste trabalho: a inclusão de informações provenientes de notícias — capturadas por meio de modelos de NLP e traduzidas em scores de sentimento — melhora significativamente o desempenho preditivo de modelos de séries temporais financeiras.

Enquanto o LSTM puro se baseia apenas em padrões históricos de preços e indicadores técnicos, ele falha em capturar informações exógenas relevantes, como mudanças regulatórias, eventos corporativos ou choques macroeconômicos. Ao adicionar variáveis de sentimento, o modelo passa a incorporar essas nuances, resultando em ganhos expressivos de desempenho, como evidenciado pelo salto de mais de 13 pontos percentuais no R^2 (de 0.79 para 0.92).

Além disso, o LightGBM com os *scores* de sentimento se destacou como um competidor relevante: mesmo sendo um modelo não sequencial, sua habilidade de lidar com variáveis defasadas e importância relativa de features permitiu que ele superasse ligeiramente o LSTM com sentimentos, alcançando o maior R^2 (0.93) e os menores valores de erro médio. Esse resultado sugere que, em certos contextos, modelos baseados em árvores podem competir de forma muito eficiente com redes neurais, desde que alimentados com os inputs corretos.

Outro ponto relevante é a importância relativa das variáveis, mostrada na Figura 21, que destaca a dominância dos indicadores técnicos, mas também a presença consistente das variáveis de sentimento — indicando que, mesmo com menor peso, elas contribuem significativamente para os ganhos de performance.

Em resumo, a análise da Tesla demonstra que:

- A inclusão de sentimentos melhora substancialmente a capacidade preditiva dos modelos.
- Modelos como o LightGBM, mesmo não projetados para sequências, podem superar redes neurais quando configurados corretamente.
- O uso combinado de inputs técnicos e textuais representa um caminho promissor para aprimorar predições no mercado financeiro.

Estes achados reforçam a relevância da fusão entre dados numéricos e textuais no contexto de finanças quantitativas e abrem portas para estudos futuros que explorem novas fontes de informação e combinações de modelos.

6 Conclusão

6.1 Análise Crítica dos Resultados

Os resultados obtidos ao longo desta dissertação indicam que a inclusão de variáveis de sentimento, extraídas de notícias de mercado, teve impacto positivo no desempenho preditivo dos modelos aplicados. Comparando as abordagens puramente técnicas com aquelas que incorporaram os *scores* de sentimento, foi observada uma melhoria consistente nas métricas de avaliação, incluindo redução nos valores de RMSE e MAPE e aumento no coeficiente de determinação R^2. Em termos quantitativos, os modelos enriquecidos com variáveis textuais apresentaram ganhos de desempenho que variaram entre 1.5% e 13%, a depender do ativo analisado e do modelo utilizado.

A análise dos diferentes ativos utilizados demonstrou que a quantidade e a frequência das notícias relacionadas ao ativo são fatores determinantes para a relevância da variável de sentimento na previsão de preços. Em ativos com alta cobertura midiática, como ações de grandes empresas globais, o modelo apresentou maior ganho de desempenho. Por outro lado, em ativos com menor volume de notícias, os resultados mostraram-se menos sensíveis à inclusão dessa variável.

Além disso, a comparação entre os algoritmos utilizados mostrou diferenças relevantes em função da natureza dos dados. O modelo LSTM apresentou melhor capacidade de capturar dependências temporais, sendo mais eficaz em séries com maior autocorrelação. Por outro lado, o LightGBM com o sentimento das notícias demonstrou robustez na manipulação de um conjunto maior de variáveis, especialmente quando múltiplos *lags* foram incorporados ao conjunto de entrada.

As análises de desempenho nos períodos de validação e teste indicam que, apesar das limitações do tamanho da base de dados disponível, os modelos foram capazes de aprender padrões significativos entre as variáveis de entrada e o comportamento futuro dos preços dos ativos.

6.2 Considerações Finais e Trabalhos Futuros

Esta dissertação contribui para o avanço da literatura ao integrar variáveis textuais derivadas de notícias com técnicas tradicionais de previsão baseadas em séries temporais financeiras. A metodologia desenvolvida demonstra o potencial de uso de análise de sentimento em conjunto com modelos de aprendizado de máquina para aprimorar a previsão de preços de ativos.

Para trabalhos futuros, recomenda-se a ampliação da base de dados, tanto em termos de horizonte temporal quanto de variedade de ativos analisados, a fim de melhorar a capacidade de generalização dos modelos. Outra possibilidade é o uso de técnicas mais sofisticadas de Processamento de Linguagem Natural, como o RoBERTa ou o GPT, visando uma extração mais precisa e contextualizada dos sentimentos presentes nas notícias.

Além disso, por um lado, futuros estudos podem explorar o uso de arquiteturas de redes neurais mais avançadas, como LSTM com mecanismos de atenção, ou ainda a combinação de diferentes modelos preditivos através de técnicas de ensemble. Por outro lado, a aplicação da metodologia proposta em outros mercados, como câmbio, commodities ou títulos de renda fixa, pode oferecer novas evidências empíricas sobre a eficácia da abordagem.

Estas linhas de investigações até aqui aventadas visam aperfeiçoar a técnica e a metodologia desenvolvidas no presente trabalho. Uma linha de investigação complementar seria explorar a integração desses modelos preditivos com algoritmos de otimização de portfólio, permitindo que as previsões de sentimento influenciem diretamente as decisões de alocação de ativos em estratégias quantitativas de investimento. Uma possibilidade preliminar neste sentido seria construir um mapeamento para medir a sensibilidade de um dado ativo (em última análise, empresa i) ou índice de mercado à atuação das notícias na dinâmica de seu retorno (total ou em excesso). Tal medida de sensibilidade poderia ser desenhada nestas linhas:

(i) Para dar conta das realizações dos retornos de um ativo, considerar o espaço normado $C[0,h]$ de funções contínuas definidas num intervalo (horizonte) $[0,h]$ com a norma

$$||f|| := \max_{t\in[0,h]} |f(t)| \quad , \quad \forall f \in C[0,h].$$

(ii) Para cada ativo i, seja A_i uma realização de seu retorno obtida através das técnicas tradicionais de previsão de séries temporais (i.e., que não usam análise de sentimento) e seja A_i^{news} uma realização de seu retorno obtida pela metodologia do presente trabalho. Então, podemos definir o número $\gamma_i = ||A_i - A_i^{\text{news}}||$ como uma medida do grau de sensibilidade do ativo i à atuação das notícias na dinâmica dos retornos do índice de mercado.

A disponibilidade de tal fator γ_i poderia ser incorporada de alguma forma determinada à equação de regressão do modelo de índice único (que é em termos dos retornos de algum índice de mercado) para aperfeiçoar os resultados destes modelos na otimização de carteiras.

Referências

ALBERG, D.; LIPTON, Z. et al. Predicting financial markets using deep learning: A survey. *arXiv preprint arXiv:2107.10483*, 2021.

ARACI, D. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

BOLLERSLEV, T. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, Elsevier, v. 31, n. 3, p. 307–327, 1986.

BOX, G. E.; JENKINS, G. M.; REINSEL, G. C. Time series analysis: forecasting and control. Holden-Day, 1976.

CHEN, S.; PELGER, M.; ZHU, J. Deep learning in asset pricing. *The Review of Financial Studies*, Oxford University Press, v. 35, n. 5, p. 2184–2226, 2022.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

DIEBOLD, F. X. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold-mariano tests. *Journal of Business & Economic Statistics*, Taylor & Francis, v. 30, n. 1, p. 1–11, 2012.

ENGLE, R. F. Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, v. 50, n. 4, p. 987–1007, 1982.

FAMA, E. F. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, Wiley, v. 25, n. 2, p. 383–417, 1970.

FISCHER, T.; KRAUSS, C. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, Elsevier, v. 270, n. 2, p. 654–669, 2018.

HOCHREITER, S.; SCHMIDHUBER, J. *Long short-term memory*. [S.l.]: MIT Press, 1997. v. 9. 1735–1780 p.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2017. p. 3146–3154.

LOUGHRAN, T.; MCDONALD, B. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, v. 66, n. 1, p. 35–65, 2011.

PRADO, M. L. de. *Advances in Financial Machine Learning*. Hoboken, NJ: Wiley, 2018.

TETLOCK, P. C. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, v. 62, n. 3, p. 1437–1467, 2007.

WOOLDRIDGE, J. M. *Introdução à Econometria: Uma Abordagem Moderna*. São Paulo: Cengage Learning, 2006.