Análise de Sentimento na Precificação de Ativos na Bolsa Brasileira

Felipe da Silva Morishita Garbi [1] Matheus Barbosa Ferrari [1]
Rogério de Oliveira [1]

[1]Faculdade de Computação e Informática (FCI)
Universidade Presbiteriana Mackenzie São Paulo, SP – Brasil

<felipe.garbi,matheus.ferrari@mackenzista.com.br>
<rogerio.oliveira@mackenzie.br>

2025

Resumo

Este trabalho investiga o emprego da análise de sentimento de notícias financeiras na previsão de preços de ações da Bolsa de Valores Brasileira. O objetivo é avaliar se a incorporação de features de sentimento extraídas através do FinBERT-PT-BR melhora a capacidade preditiva de modelos de aprendizado de máquina. Para isso, foram coletadas 8.762 notícias de três portais brasileiros entre maio e setembro de 2025, classificadas quanto ao sentimento e integradas aos dados históricos de preços de quatro ativos (PETR4, VALE3, EMBR3, BOVA11). Os modelos LSTM e SVR foram treinados com diferentes janelas temporais (7 e 14 dias) e horizontes de previsão (1, 2 e 5 dias). O estudo evidencia que o sentimento midiático pode complementar modelos quantitativos, especialmente em arquiteturas que exploram relações temporais nativamente, como a LSTM.

Palavras-chave: Análise de Sentimento; FinBERT-PT-BR; Previsão de Preços; LSTM; SVR; Mercado de Ações Brasileiro.

Abstract

This study investigates the use of sentiment analysis of financial news in predicting stock prices on the Brazilian Stock Exchange. The objective is to assess whether incorporating sentiment features extracted using FinBERT-PT-BR improves the predictive capacity of machine learning models. To this end, 8,762 news articles were collected from three Brazilian websites between May and September 2025, classified according to sentiment, and integrated with historical price data for four assets (PETR4, VALE3, EMBR3, BOVA11). The LSTM and SVR models were trained with different time windows (7 and 14 days) and forecast horizons (1, 2, and 5 days). The study shows that media sentiment can complement quantitative models, especially in architectures that natively exploit temporal relationships, such as LSTM.

Keywords: Sentiment Analysis; FinBERT-PT-BR; Stock Price Prediction; LSTM; SVR; Brazilian Stock Market.

1 Introdução

A relação entre o comportamento humano e os mercados financeiros tem sido objeto de estudo desde o início das primeiras teorias econômicas. Contudo, com a digitalização das informações e o aumento exponencial do volume de dados produzidos por notícias financeiras, emergiu uma demanda crescente para compreender como aspectos qualitativos, como o sentimento manifestado em notícias, podem afetar os preços dos ativos. Esses fatores vão além das variáveis econômicas convencionais e englobam aspectos subjetivos, como a maneira como as notícias são compreendidas e divulgadas pelo mercado.

Nos últimos anos, a relação entre a mídia e o mercado financeiro tem sido amplamente estudada, especialmente em mercados globais, com foco no impacto das informações divulgadas no comportamento dos investidores (TETLOCK, 2015; BARBER; ODEAN, 2008). Esse fenômeno se torna ainda mais relevante no Brasil, cujo mercado financeiro enfrenta tanto desafios locais quanto fatores externos, amplificando a reação dos investidores a eventos diversos. A volatilidade da economia brasileira e a influência de fatores políticos e econômicos instáveis tornam a Bolsa de Valores Brasileira (B3) um ambiente propício para o estudo da relação entre sentimento midiático e variações no preço das ações. Contudo, a maioria dos estudos existentes foca em mercados estrangeiros, utilizando notícias em inglês e ativos internacionais, deixando uma lacuna significativa no contexto brasileiro.

O problema central a ser tratado nesta pesquisa é mensurar quantitativamente o impacto do sentimento expresso em notícias financeiras no desempenho dos ativos no mercado brasileiro. Para isso, avaliamos de que forma a incorporação de variáveis de sentimento pode contribuir para aprimorar a previsão de preços de ações ao serem integradas a diferentes arquiteturas de aprendizado de máquina.

São empregadas notícias de três portais de relevância nacional: Exame, Infomoney e Valor Econômico, publicadas entre 01/05/2025 e 01/09/2025. Essas notícias foram processadas e classificadas quanto ao sentimento (positivas, negativas ou neutras) por meio do modelo FinBERT-PT-BR, especializado em textos financeiros em português, e os scores resultantes foram agregados diariamente pela mediana e integrados aos dados históricos de preços de quatro ativos da Bolsa de Valores Brasileira (PETR4, VALE3, EMBR3 e BOVA11). Para a previsão dos valores, são empregados dois modelos de aprendizado de máquina: o SVR (Regressor de Vetores de Suporte) e a rede LSTM (Long Short-Term Memory), avaliados com e sem features de sentimento em diferentes janelas temporais (7 e 14 dias) e horizontes de previsão (1, 2 e 5 dias). As melhorias observadas foram validadas

estatisticamente por testes t de Student unicaudais ($\alpha = 0{,}05$) sobre erro absoluto, erro quadrático e correlação, verificando em quantos casos a inclusão do sentimento produziu ganhos significativos. O trabalho também disponibiliza um repositório público com as notícias coletadas, o código-fonte e as análises realizadas (FERRARI; GARBI, 2025).

Este trabalho contribui para complementar o estudo sobre análise de sentimento aplicada ao mercado brasileiro, fornecendo evidências empíricas sobre quando e como o sentimento midiático pode complementar modelos quantitativos de previsão. Além disso, também contribui fornecendo uma base de dados primária e aberta de notícias nacionais, que pode ser empregada em outros estudos.

Este artigo está estruturado da seguinte forma: a Seção 2 apresenta a fundamentação teórica sobre eficiência de mercado, finanças comportamentais, análise de sentimento e trabalhos relacionados; a Seção 3 descreve detalhadamente a metodologia empregada, incluindo coleta de dados, processamento de sentimento com FinBERT-PT-BR, arquiteturas dos modelos (LSTM, SVR) e protocolos de validação estatística; a Seção 4 apresenta os resultados obtidos, incluindo análises comparativas entre arquiteturas, testes de hipótese, análise por horizonte temporal, avaliação de desempenho absoluto versus melhoria relativa, e identificação de configurações ótimas; e, por fim, a Seção 5 apresenta as conclusões da pesquisa e direções para trabalhos futuros.

2 Referencial Teórico

A previsão de preços de ativos financeiros é um assunto amplamente estudado, tanto no campo das finanças quanto na ciência de dados. No começo, era um dos métodos mais utilizados para compreender a relação entre variáveis de mercado. Contudo, esses modelos têm limitações ao enfrentar a complexidade e a natureza não linear dos mercados financeiros, principalmente em momentos de alta volatilidade.

O uso de métodos baseados de aprendizado de máquina se tornou mais comum nesse tipo de análise. Esses modelos conseguem identificar padrões complexos em grandes volumes de dados e se adaptam melhor a mudanças repentinas no mercado. Com a integração de bases de dados textuais, como notícias, relatórios e postagens financeiras, foi permitido a inclusão de informações que demonstram o sentimento de investidores sobre diferentes ativos e eventos econômicos.

A combinação das duas abordagens, aprendizado de máquina e análise de sentimento, mostrou bons resultados ao oferecer previsões mais precisas sobre as condições do mercado. A tabela apresenta alguns estudos recentes e suas métricas de desempenho.

Tabela 1 – Melhor modelo utilizado em cada estudo e suas métricas de desempenho.

Autor (Ano)	**Melhor Modelo**	**MAE**	**MAPE**
Vieira (2025)	LightGBM + Sentimento	10.72	4.98%
Halder (2022)	FinBERT-LSTM	174.94	1.41%
Gu et al. (2024)	FinBERT-LSTM	173.67	4.5%
Kasture et al. (2024)	RNN-LSTM + Sentimento	0.036	

2.1 Eficiência de Mercado e Finanças Comportamentais

A Hipótese de Mercado Eficiente (HME) diz que os preços dos ativos refletem de forma rápida toda a informação disponível, tornando ineficaz a obtenção sistemática de retornos anormais a partir de dados públicos (FAMA, 1970). Contudo, pesquisas em Finanças Comportamentais mostram que viéses cognitivos e a atenção midiática podem gerar anomalias de curto prazo. Trabalhos clássicos mostram que a cobertura da mídia e o tom das notícias influenciam volume e direção de negociações, criando reações que nem sempre se alinham imediatamente aos fundamentos econômicos (TETLOCK, 2007; BARBER; ODEAN, 2008). Esses estudos contradizem a visão da HME e sugerem que a tonalidade da mídia pode gerar oscilações temporárias nos preços dos ativos, possibilitando que modelos que integrem esse tipo de informação obtenham oportunidades diante do mercado financeiro.

2.2 Análise de Sentimento e Processamento de Linguagem Natural

A análise de sentimento busca determinar a polaridade de textos, classificando-os como positivos, negativos ou neutros em variados níveis de detalhamento, como documento, sentença ou aspecto (LIU, 2012).

Com o desenvolvimento dos modelos de *deep learning*, particularmente os *transformers*, como o BERT (DEVLIN et al., 2019), passou-se a representar palavras de maneira contextualizada, possibilitando a captura de nuances linguísticas. Nesse campo, o vocabulário técnico e as expressões especializadas demandam modelos que consigam entender o contexto em que são utilizadas.

No domínio financeiro, o vocabulário, as siglas e as expressões apresentam sentidos dependentes do contexto, o que torna a aplicação direta de modelos genéricos relativamente ineficaz. Para superar isso, surgiram modelos especializados, como o FinBERT (ARACI, 2019), que ampliam essa metodologia ao calcular as probabilidades ligadas às classes de sentimento (positiva, negativa e neutra). Isso possibilita a obtenção de scores contínuos de polaridade, resultantes da diferença entre as probabilidades de positividade e negatividade. Essa representação, utilizada também neste estudo, permite captar nuances de tom e incorporar o sentimento como uma variável numérica em modelos preditivos. Essa representação numérica possibilita incorporar o sentimento como variável em modelos quantitativos, contribuindo em previsões previamente baseadas apenas em dados de preço e volume.

No contexto brasileiro, o modelo FinBERT-PT-BR desenvolvido por Santos, Bianchi e Costa (2023) adapta essa arquitetura para o português, sendo pré-treinado em um grande conjunto de textos econômicos e financeiros nacionais. O português apresenta construções sintáticas que diferem significativamente do inglês, e os termos econômicos brasileiros também possuem características próprias. O modelo, portanto, consegue capturar com precisão as nuances e contextos das notícias locais. Por esse motivo, este estudo utiliza as probabilidades de sentimento extraídas pelo FinBERT-PT-BR como base para a construção dos indicadores de sentimento diário empregados nos modelos preditivos subsequentes.

2.3 Modelos de Aprendizado de Máquina e Séries Temporais

O aprendizado de máquina tem sido cada vez mais utilizado em finanças por conseguir identificar padrões complexos e não lineares em séries de preços, algo que métodos tradicionais nem sempre conseguem fazer. Modelos como LSTM e combinações

com análise de sentimento, como o FinBERT-LSTM, têm mostrado bons resultados na previsão de preços de ativos (HALDER, 2022; GU et al., 2024). No Brasil, estudos indicam que incluir informações de sentimento extraídas de notícias pode melhorar a precisão das previsões, trazendo melhorias, dependendo do ativo e da cobertura midiática (VIEIRA, 2025). Isso mostra como o aprendizado de máquina é útil para integrar informações não estruturadas aos modelos quantitativos no mercado financeiro.

O aprendizado de máquina é aplicado a séries temporais de forma supervisionada, ou seja, para prever valores futuros com base nos padrões observados anteriormente. Para isso, a série pode ser dividida em janelas deslizantes, conhecidas como *sliding windows*, com um tamanho w. Essas janelas refletem o histórico recente que será utilizado como entrada para prever o próximo valor, levando em conta um horizonte h.

Além das variáveis intrínsecas da série (como preços ou volume), a inclusão da análise de sentimento em modelos de séries temporais é realizada pela incorporação de variáveis exógenas, ou seja, informações externas ao histórico de preços que podem influenciar o comportamento do ativo. Nesse caso, os indicadores de sentimento extraídos de notícias financeiras são agregados diariamente e adicionados como variáveis adicionais às janelas de entrada. Dessa forma, o modelo pode aprender a ponderar o impacto emocional do noticiário sobre o movimento dos preços.

O estudo utiliza os modelos LSTM e SVR por serem métodos que possuem abordagens muito diferentes entre si. A rede Long Short-Term Memory (LSTM), desenvolvida por Hochreiter e Schmidhuber (1997), possui a habilidade de captar dependências de longo prazo em séries temporais através de unidades de memória que mantêm informações importantes ao longo do tempo. Isso a torna eficiente em situações com alta dinâmica sequencial e volatilidade, como o mercado de ações. Já modelo Support Vector Regression (SVR), criado por Drucker et al. (1996), tenta ajustar uma função que minimize os erros dentro de uma margem de tolerância, usando diferentes funções de kernel para captar relações não lineares. Enquanto a LSTM aprende padrões diretamente da sequência ao longo do tempo, o SVR trabalha com janelas fixas de observações, sendo mais adequado para conjuntos menores.

3 Metodologia

Esta pesquisa caracteriza-se como quantitativa e experimental, utilizando métodos computacionais para análise de dados financeiros e textuais.

3.1 Coleta de Dados

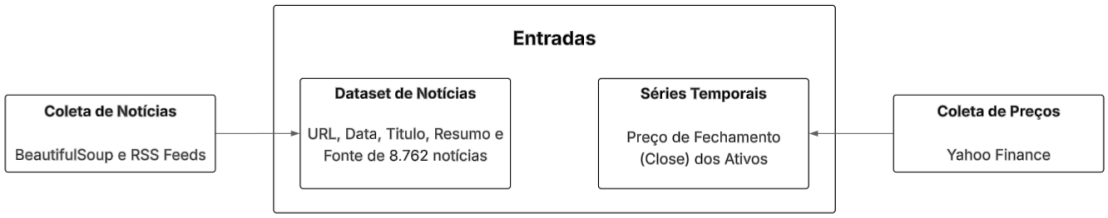

Figura 1 – Fluxo geral de captura inicial dos dados.

As notícias foram coletadas de três portais brasileiros: *Exame* (<https://exame.com>), *Infomoney* (<https://www.infomoney.com.br>) e *Valor Econômico* (<https://valor.globo.com>) A escolha desses três portais se deve à sua ampla cobertura do mercado financeiro brasileiro, credibilidade editorial e frequência de atualização. O Valor Econômico é reconhecido um dos principais veículos de imprensa econômica nacional. Já o Infomoney destaca-se pela agilidade na divulgação de notícias de mercado e análises de curto prazo, enquanto a Exame fornece cobertura mais ampla, abordando economia, negócios, política e acontecimentos com relevância para o comportamento dos ativos. Além disso, esses portais disponibilizam estruturas HTML consistentes, o que viabiliza uma coleta automatizada estável.

A coleta foi realizada entre 01/05/2025 e 01/09/2025 (4 meses) através de web scraping (BeautifulSoup) e RSS Feeds. O BeautifulSoup é uma biblioteca Python utilizada para processar o HTML das páginas (RICHARDSON, 2007), e os RSS Feeds funcionam como uma interface simplificada disponibilizada pelos próprios portais, permitindo o acesso às notícias via requisições HTTP.

O processo de coleta foi automatizado por meio de uma rotina diária executada via GitHub Actions, utilizando um arquivo `.yml` configurado no repositório do projeto. O fluxo era acionado automaticamente às 23h30 (UTC-3) e executava o script Python responsável pela extração das notícias mais recentes dos três portais. A cada execução, o script realizava: (1) requisições aos RSS feeds dos portais Exame e Infomoney; (2) raspagem direta da página de "Últimas Notícias" do Valor Econômico utilizando a biblioteca BeautifulSoup4; e (3) extração e limpeza dos dados de cada artigo, incluindo URL, data, título, corpo do texto e fonte.

Cada execução gerava um arquivo CSV diário, nomeado de acordo com a data da coleta, contendo apenas as notícias publicadas no próprio dia. Essa restrição foi implementada para reduzir duplicações de conteúdo entre coletas sucessivas e garantir que o conjunto refletisse fielmente as notícias de cada data de pregão. Após a consolidação dos arquivos individuais, o conjunto final resultou em 8.762 notícias, armazenadas publicamente no repositório GitHub do projeto (FERRARI; GARBI, 2025).

Tabela 2 – Amostra do conjunto de notícias coletadas.

url	data	titulo	texto	fonte
<https://exame.com/mundo/pacotes-vindos-da-china-aos-eua-sofrerao-tarifa-punitiva-a-partir-de-sexta-feira/>	2025-05-01	Pacotes vindos da China aos EUA sofrerão tarifa punitiva a partir de sexta-feira	Pacotes provenientes da China, especialmente aqueles contendo roupas baratas dos gigantes do comércio eletrônico Shein e Temu, serão afetados por tarifas punitivas nos Estados Unidos ...	Exame
<https://www.infomoney.com.br/mundo/trump-compra-de-petroleo-iraniano-precisa-parar-e-fala-em-sancoes-a-compradores/>	2025-05-01	Trump: Compra de petróleo iraniano precisa parar e fala em sanções a compradores	O presidente dos Estados Unidos, Donald Trump, disse que todas as compras de petróleo ou produtos petroquímicos iranianos devem parar e qualquer país ...	Infomoney
<https://valor.globo.com/empresas/noticia/2025/05/01/moderna-registra-prejuizo-no-trimestre-citando-sazonalidade-dos-negocios-respiratorios.ghtml>	2025-05-01	Moderna registra prejuízo no trimestre, citando sazonalidade dos negócios respiratórios	A Moderna registrou prejuízo no primeiro trimestre de 2025 devido à queda na receita, influenciada pela queda nas vendas de produtos, embora a empresa ...	Valor Econômico

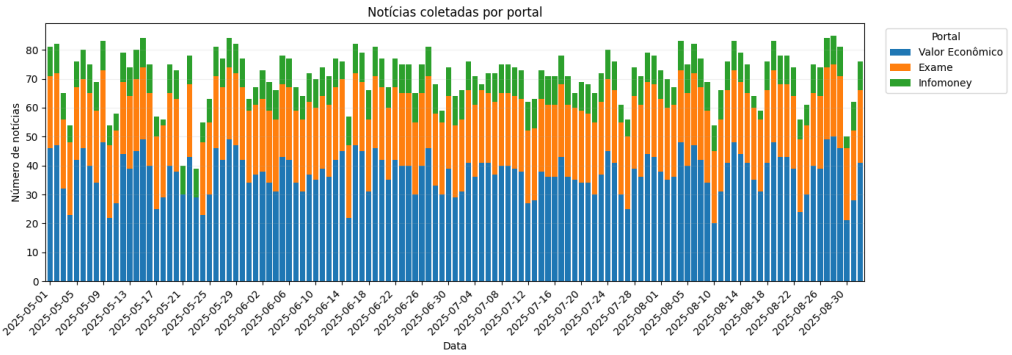

Figura 2 – Quantidade total de notícias coletadas por portal.

Já os dados históricos (open, high, low, close, volume) dos quatro ativos (PETR4.SA, VALE3.SA, EMBR3.SA, BOVA11.SA) foram obtidos através da biblioteca Python yfinance, totalizando 85 dias úteis de pregão entre o mesmo período da coleta das notícias. A Tabela 3 compõe uma amostra dos dados obtidos, tendo como exemplo a PETR4.SA.

Tabela 3 – Amostra dos dados históricos de preços para PETR4.SA.

date	open	high	low	close	volume
2025-05-02	28.67	29.25	28.45	29.25	33,146,600
2025-05-05	29.02	29.06	28.15	28.15	59,011,700
2025-05-06	28.58	28.85	28.44	28.62	52,751,300
2025-05-07	28.76	28.76	28.38	28.75	35,050,400

3.2 Processamento de Sentimento

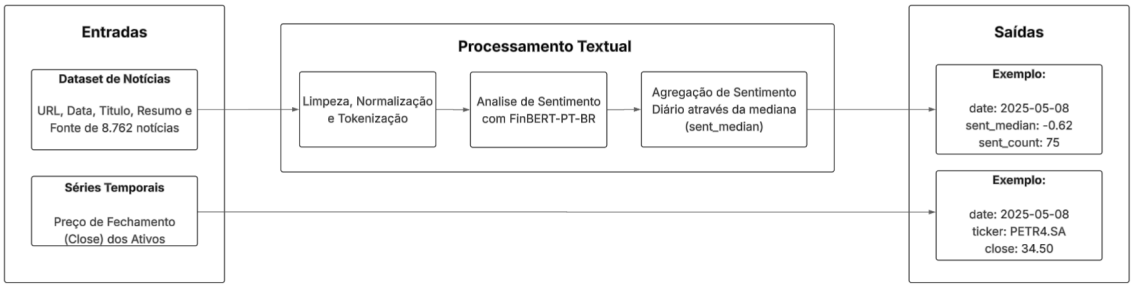

Figura 3 – Diagrama de processamento dos dados.

O conjunto de notícias passou por um processamento textual leve, visando padronizar o conteúdo e remover elementos indesejados. O processo foi: (1) Limpeza de tags HTML, caracteres especiais e URLs; (2) Normalização para letras minúsculas; (3) Tokenização (com AutoTokenizer), que divide o texto em unidades linguísticas básicas (tokens), permitindo sua posterior interpretação.

O modelo FinBERT-PT-BR foi então empregado para realizar a inferência, gerando assim, para cada notícia, as probabilidades das classes de sentimento Positiva, Negativa e Neutra (p_{pos}, p_{neg} e p_{neu}). A partir destas probabilidades, foi calculado o score contínuo de sentimento definido como score $= p_{pos} - p_{neg}$, onde score pertence ao intervalo de [-1, +1]. Esse valor representa o grau de polaridade da notícia, variando de fortemente negativa (-1) a fortemente positiva (+1), com valores próximos de zero ($-0,05 < p < +0,05$) indicando neutralidade.

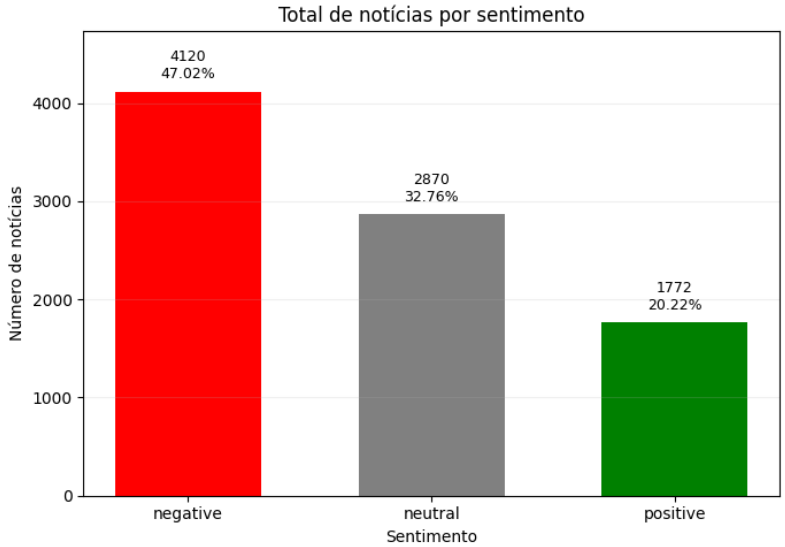

Figura 4 – Distribuição das notícias segundo a classificação de sentimento

Considerando que várias notícias são publicadas no mesmo dia, os resultados de sentimento são consolidados em uma única medida para cada data de pregão t (agregação temporal de sentimento). Para isso, calculou-se a mediana dos scores de todas as notícias do dia, com a intenção de reduzir a influência de valores extremos:

$$\text{sent\_median}_t = \text{mediana}(\{\text{score}_1, \ldots, \text{score}_{N_t}\}), \quad (1)$$

em que N_t corresponde ao número de notícias disponíveis na data t. O resultado desse processo é uma série temporal diária de sentimento (sent_median$_t$).

3.3 Integração e Preparação

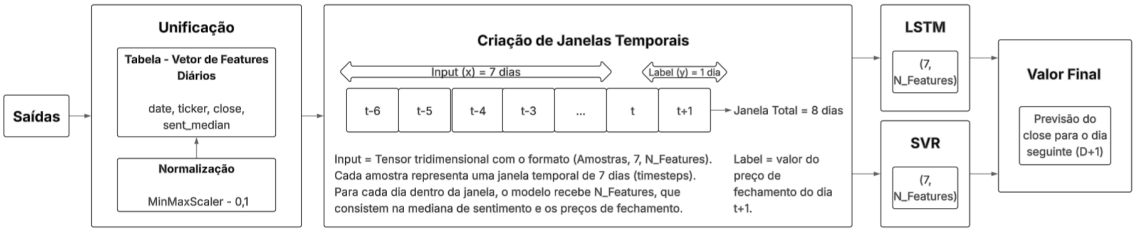

Figura 5 – Diagrama descrevendo um exemplo de integração com janela de 7 dias, e horizonte de previsão de 1 dia.

3.3.1 Alinhamento Temporal

Os conjuntos de sentimento (sent_median$_t$) e de preços (close) por ticker foram sincronizados conforme o calendário de pregões da B3, considerando apenas dias úteis. Notícias publicadas em finais de semana e feriados foram excluídas do processo de agregação diária, de forma a manter correspondência entre cada observação de sentimento e a data do pregão.

Tabela 4 – Amostra dos dados finais para utilização nos modelos.

date	**ticker**	**close**	**sent_median**
2025-05-02	BOVA11.SA	131.85	-0.41
2025-05-02	EMBR3.SA	66.05	-0.41
2025-05-02	PETR4.SA	29.24	-0.41
2025-05-02	VALE3.SA	51.01	-0.41
2025-05-05	BOVA11.SA	130.33	0.05

3.3.2 Normalização

Na etapa seguinte, realizou-se a normalização, para garantir que todas as variáveis numéricas estejam em uma mesma escala de valores, de modo a evitar que diferenças de magnitude prejudiquem o aprendizado do modelo. Para tal, foi aplicada a normalização do tipo MinMaxScaler, que transforma cada variável para o intervalo [0,1] segundo a expressão:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{2}$$

com $x_{\min}$ e $x_{\max}$ representando, respectivamente, os valores mínimo e máximo observados da variável no conjunto de dados. Assim, todos os atributos passam a ter a mesma faixa de variação, o que impede que variáveis com valores mais altos (como o preço de fechamento) dominem aquelas em escalas menores (score de sentimento). A normalização foi realizada sobre todo o conjunto de dados, em caráter exploratório. Essa

abordagem é adequada para nosso objetivo de comparar o desempenho relativo entre modelos e configurações, ao invés de construir um sistema de trading em produção.

3.3.3 Janelamento Temporal

Modelos de previsão de séries temporais requerem que os dados sejam organizados de modo a capturar a dependência temporal entre eventos sucessivos. Para isso, foram criadas janelas temporais, que consistem na segmentação da série contínua de observações em subconjuntos (janelas). Cada janela representa um intervalo de n dias consecutivos de informações passadas, utilizados como entrada do modelo para prever o valor futuro dos preços em um horizonte h.

O janelamento cria pares de entrada–saída (X_i, y_i), em que cada entrada X_i contém as observações das w datas anteriores e a saída y_i corresponde ao valor que se deseja prever após um deslocamento temporal h. Isso pode ser descrito da seguinte forma:

$$X_i = [x_i, x_{i+1}, \ldots, x_{i+w-1}] \in R^{w \times d}, \qquad y_i = \text{close}_{i+w+h-1}, \tag{3}$$

onde d representa o número de variáveis consideradas por dia:

- $d = 1$ no modo sem sentimento, utilizando apenas o preço de fechamento (close);
- $d = 2$ no modo com sentimento, combinando o preço de fechamento e o sentimento agregado diário (sent_median).

3.4 Arquiteturas de Modelos

Foram empregados dois tipos de modelos para previsão da variação de preços das ações: uma rede recorrente do tipo LSTM (Long Short-Term Memory), e um modelo de Regressão por Vetores de Suporte (SVR).

3.4.1 LSTM

Neste trabalho, empregou-se uma arquitetura compacta, composta por:

- Uma camada LSTM composta por 64 neurônios, responsável por capturar padrões temporais e relações de dependência entre os valores anteriores da série;
- Uma camada de dropout de 20%, empregada para reduzir o overfitting e aumentar a capacidade de generalização do modelo;
- Uma camada densa intermediária contendo 32 neurônios e função de ativação ReLU, que possibilita a modelagem de relações não lineares entre as representações aprendidas;
- Uma camada de saída formada por um único neurônio e ativação linear, correspondente à estimativa do preço de fechamento futuro.

O treinamento foi realizado com o otimizador Adam, taxa de aprendizado $\alpha = 0{,}001$, função de perda Mean Absolute Error (MAE), 100 épocas e tamanho de lote (*batch size*) igual a 8. Essa configuração buscou equilibrar o desempenho e a estabilidade do modelo, considerando o número limitado de amostras disponíveis.

3.4.2 SVR

Os principais hiperparâmetros utilizados foram:

- parâmetro de penalização $C = 1{,}0$, que controla o equilíbrio entre erro de treinamento e suavidade do modelo;
- parâmetro gamma = 'scale', que define a amplitude do kernel RBF de forma adaptativa ao conjunto de dados.

Como o SVR precisa de uma matriz de entrada bidimensional, as sequências temporais resultantes do janelamento foram achatadas em vetores de tamanho $w \times d$, onde w representa a janela temporal e d o número de variáveis por dia. Essa estrutura permite a compatibilidade entre os dados temporais e o formato exigido pelo modelo de regressão vetorial.

3.5 Experimentos e Validação

Os experimentos foram realizados com o objetivo de avaliar se a incorporação do sentimento das notícias financeiras contribui com as predições das séries de preços dos ativos.

A metodologia adotada se baseou em janelas deslizantes (*sliding windows*), conforme ilustrado na Figura 6. Cada janela de tamanho w (7 ou 14 dias) foi utilizada para prever o preço de fechamento h dias à frente ($h = 1, 2, 5$). Após cada previsão, a janela é deslocada um dia à frente, garantindo que todas as observações disponíveis sejam aproveitadas. O procedimento é caracterizado como uma análise *in-sample*, isto é, as previsões são geradas dentro do próprio conjunto de dados observado. Essa abordagem tende a produzir resultados mais otimistas em relação à acurácia preditiva, contudo, ainda é adequada ao foco do trabalho, que é a análise exploratória.

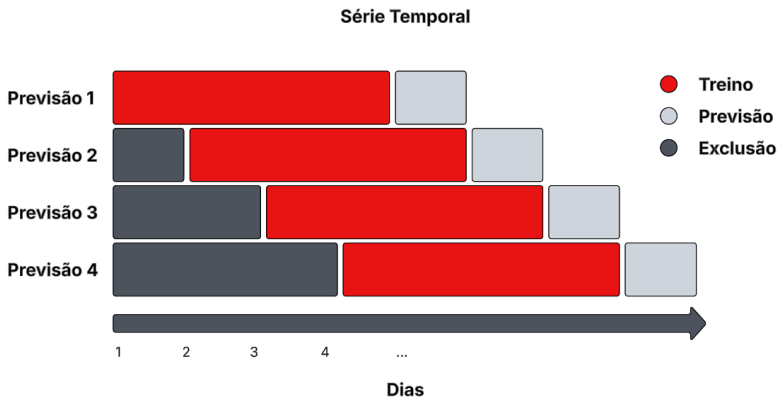

Figura 6 – Janela deslizante aplicada nas previsões. Cada bloco representa o conjunto de dias usados para treino (cinza) e o ponto previsto (dourado).

As janelas de 7 e 14 dias foram selecionadas por representarem períodos curtos e médios de retenção de informação, coerentes com a dinâmica de reação do mercado a notícias recentes. Já os horizontes refletem diferentes prazos de reação: imediata (1 dia), curta (2 dias) e semanal (5 dias).

Tabela 5 – Configurações experimentais testadas.

Janela (w)	**Horizonte** (h)	**Descrição**
7 dias	1 dia	Reação rápida a notícias recentes
7 dias	2 dias	Efeito de curto prazo
7 dias	5 dias	Efeito semanal médio
14 dias	1 dia	Reação rápida com janela temporal mais longa
14 dias	2 dias	Curto prazo com memória ampliada
14 dias	5 dias	Influência média estendida no tempo

O desempenho dos modelos foi avaliado por meio de 4 métricas:

- **MAE (Erro Médio Absoluto)** - indica, em média, o quanto as previsões do modelo se afastam dos valores reais.
- **MAPE (Erro Percentual Médio Absoluto)** - apresenta o erro médio em termos percentuais, possibilitando a comparação do desempenho do modelo entre ativos com diferentes faixas de preço.
- **Correlação de Pearson (rolling)** - mede o grau de alinhamento entre as previsões e os valores reais ao longo do tempo.
- **MAPE Agregado** - Média aritmética simples dos MAPEs individuais dos ativos. Dada uma configuração (janela w, horizonte h e modo), definimos o MAPE agregado como:

$$\text{MAPE}_{\text{agregado}}(w, h, \text{modo}) = \frac{1}{K} \sum_{k=1}^{K} \text{MAPE}_k,$$

onde cada MAPE_k é o erro percentual do ativo k.

Para cada ativo, foram testadas 12 configurações distintas: 2 janelas temporais (7 e 14 dias) $\times$ 3 horizontes de previsão (1, 2 e 5 dias) $\times$ 2 modos (com/sem sentimento), gerando 48 experimentos por modelo (4 tickers $\times$ 12 configurações).

Para cada combinação (ticker, janela, horizonte), foram realizados testes t de Student. O teste avalia se a diferença entre as médias é estatisticamente significativa (rejeita-se a hipótese nula $H_0 : \mu_{\text{sem}} = \mu_{\text{com}}$ quando $p < 0{,}05$) ou pode ser atribuída ao acaso. As 24 combinações de janelas, tickers e horizontes ($2 \times 4 \times 3$) foram submetidas a dois testes de erro (um com erro absoluto e outro com erro quadrático), e um teste de correlação totalizando 72 testes estatísticos ($24 \times 2 + 24$).

4 Resultados e Discussão

A seguir são apresentados os resultados dos experimentos conduzidos e sua interpretação. Além de reportar métricas, buscamos explicar o que esses números significam em termos de utilidade preditiva e comportamento de mercado.

4.1 LSTM

Tabela 6 – Desempenho agregado do LSTM, com MAPE e MAE médios entre 4 ativos (por janela e horizonte)

Janela	Horizonte	MAE Sem	MAE Com	MAPE Sem (%)	MAPE Com (%)	Melhoria MAPE (%)	Erros (Sig.)	Correlação (Sig.)
7	1	0.751	0.695	1.10	1.02	+7.09	0/8	1/4
7	2	1.001	0.819	1.49	1.25	+16.10	2/8	2/4
7	5	1.644	1.215	2.49	1.98	+20.30	4/8	3/4
14	1	0.748	0.662	1.10	0.99	+10.10	1/8	2/4
14	2	1.007	0.719	1.48	1.07	+27.82	6/8	4/4
14	5	1.372	0.827	2.14	1.33	+38.10	8/8	4/4

Com e Sem - se referem a Com Sentimento e Sem Sentimento.

Erros (Sig.) - número de testes t pareados com diferença significativa ($p < 0{,}05$) entre os modelos *com* e *sem* sentimento, considerando métricas de erro quadrático e absoluto (4 ativos × 2 métricas).

Correlação (Sig.) - número de correlações de Pearson (rolling) com diferença significativa ($p < 0{,}05$) entre as previsões e preços reais (4 ativos × 1 por janela).

O modelo LSTM mostrou ganhos notáveis quando o sentimento foi incluído, especialmente em horizontes de previsão maiores. Para resumir o desempenho entre os quatro ativos (PETR4, VALE3, EMBR3 e BOVA11) usamos o MAPE médio agregado. A Tabela acima sintetiza esses resultados: o LSTM apresentou uma melhoria média global de +19,91% ao incorporar sentimento, e o efeito cresce com o horizonte de previsão.

Os resultados demonstraram que o LSTM passa a prever melhor quando é permitido algum tempo para que o efeito do sentimento apareça nos preços. Para previsões de curtíssimo prazo (próximo dia útil, $h = 1$), o ganho é estatisticamente insignificante, para 2 dias já há melhora, e no horizonte de 5 dias os ganhos chegam a reduzir o erro em quase 40% na melhor configuração (janela $w = 14$, $h = 5$).

Uma outra perspectiva interessante para esse resultado é a observação do impacto por ativo na Tabela 7:

4.2 Impacto por Ativo

Tabela 7 – Taxa de sucesso estatístico por ativo, por horizonte.

Ativo	Tipo	Horizonte = 1	Horizonte = 2	Horizonte = 5	Total
BOVA11.SA	ETF (Índice)	50% (3/6)	83.3% (5/6)	100% (6/6)	77.7% (14/18)
VALE3.SA	Ação	0% (0/6)	33.3% (2/6)	66.6% (4/6)	33.3% (6/18)
EMBR3.SA	Ação	16.6% (1/6)	50% (3/6)	100% (6/6)	55.5% (10/18)
PETR4.SA	Ação	0% (0/6)	50% (3/6)	50% (3/6)	33.3% (6/18)

Horizonte: É definido pela combinação de 2 janelas (7 e 14 dias), onde é feito 3 testes estatísticos para cada uma (2 para erros e 1 para correlação).

É perceptível que a BOVA11.SA obteve a maior taxa de melhoria estatisticamente significante, indicando a possibilidade de que indicadores macroeconômicos são mais propensos a influência do sentimento quando estes estão refletindo o mercado inteiro.

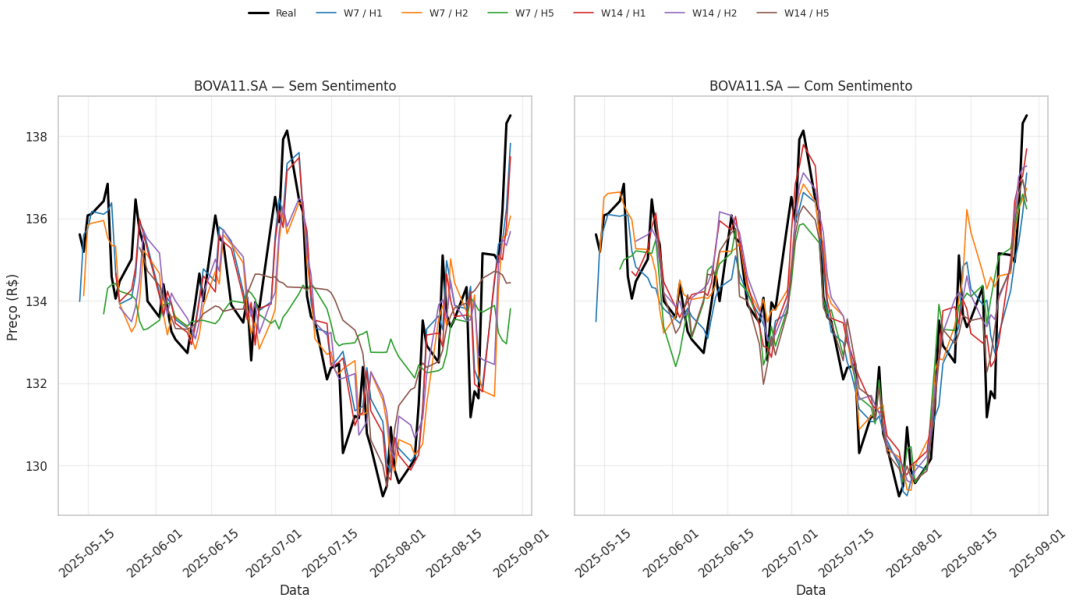

Figura 7 – Previsões do modelo LSTM para BOVA11.SA

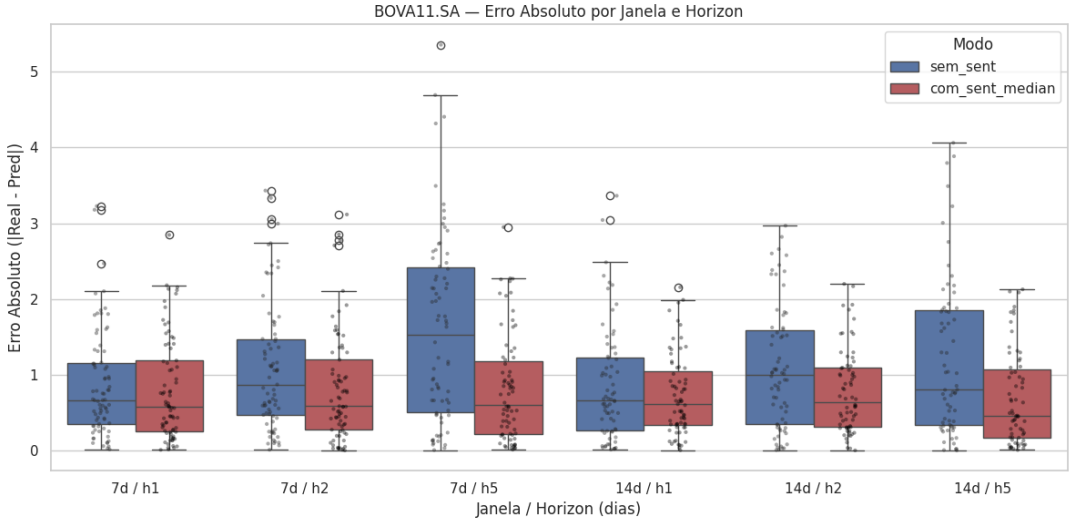

Figura 8 – Distribuição de erros absolutos do modelo LSTM para BOVA11.SA.

As figuras mostradas para BOVA11.SA revelam: as previsões do LSTM sem sentimento tendem a ser mais dispersas para $h = 5$, enquanto as previsões com sentimento apresentam melhor aderência à série real. Os boxplots de erro confirmam uma redução consistente da mediana e menos outliers quando o sentimento é incluído para $h \geq 2$ (ver Figuras 7 e 8).

É possível perceber um padrão: o sentimento do mercado costuma levar algum tempo para ser incorporado integralmente aos preços. Redes com memória temporal conseguem modelar essa defasagem, e por isso o ganho cresce com o horizonte. Para previsões de próximo dia útil, o comportamento recente de preços já carrega a maior parte do sinal disponível, tornando o sentimento redundante.

4.3 SVR - Support Vector Regression

Tabela 8 – Desempenho agregado do SVR, com MAPE e MAE médios entre 4 ativos (por janela e horizonte)

Janela	Horizonte	MAE Sem	MAE Com	MAPE Sem (%)	MAPE Com (%)	Melhoria MAPE (%)	Erros (Sig.)	Correlação (Sig.)
7	1	0.707	0.691	1.05	1.06	−0.99	0/8	3/4
7	2	0.817	0.755	1.23	1.16	+5.59	1/8	2/4
7	5	1.117	0.852	1.64	1.32	+19.76	6/8	3/4
14	1	0.683	0.717	1.01	1.12	−10.90	0/8	3/4
14	2	0.708	0.745	1.07	1.16	−8.79	0/8	3/4
14	5	0.728	0.799	1.15	1.24	−7.07	0/8	3/4

Erros (Sig.) — número de testes t pareados com diferença significativa ($p < 0{,}05$) entre os modelos com e sem sentimento, considerando métricas de erro quadrático e absoluto (4 ativos × 2 métricas).

Correlação (Sig.) — número de correlações de Pearson (rolling) com diferença significativa ($p < 0{,}05$) entre as previsões e os preços reais (4 ativos × 1 por janela).

Diferentemente do LSTM, o modelo SVR apresentou comportamento inconsistente com a adição de features de sentimento. Embora tenha mostrado ganhos pontuais, a maioria das configurações resultou em perdas de desempenho, com uma média geral de apenas +1,41%.

Figura 9 – Previsões do modelo SVR para BOVA11.SA.

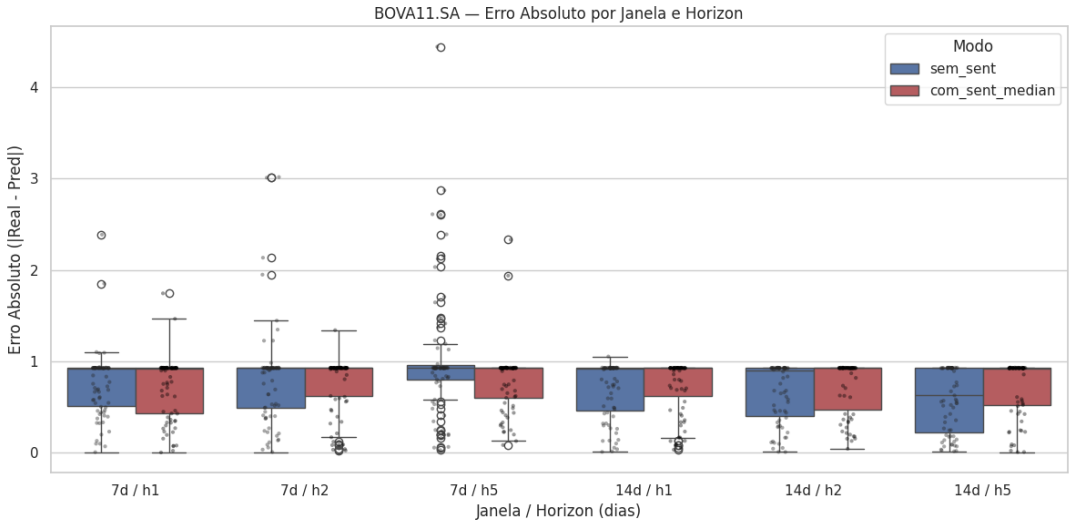

Figura 10 – Distribuição de erros absolutos do modelo SVR para BOVA11.SA.

4.4 Comparação entre Modelos

A Tabela 9 consolida o desempenho agregado final dos dois modelos:

Tabela 9 – Comparação geral entre modelos - desempenho agregado e taxa de sucesso estatístico

Modelo	MAPE Sem (%)	MAPE Com (%)	Melhoria (%)	Taxa Sig. (Erro / Corr.)
LSTM	1.63	1.27	+22.06	43.8% / 66.7%
SVR	1.19	1.18	+1.41	14.6% / 70.8%

Observação: os valores de MAPE agregados foram calculados como média aritmética das seis configurações exibidas nas tabelas por janela/horizonte (2 janelas × 3 horizontes).

Aqui há um ponto importante para entender: o LSTM ganha mais ao incorporar sentimento (melhoria relativa), mas o SVR apresenta menor erro absoluto sem sentimento. Em outras palavras, modelos mais simples (SVR) generalizam melhor com poucas amostras, enquanto modelos sequenciais (LSTM) beneficiam-se mais do sinal de sentimento explorar relações temporais.

Paradoxalmente, o SVR apresentou maior taxa de sucesso em testes de correlação (70,8%) do que o LSTM (66,7%), mas menor taxa em testes de erro (14,6% vs 43,8%). Este resultado indica que o SVR captura a direção correta das variações de preço, entretanto com uma magnitude de erro maior, evidenciando que correlação sozinha não garante utilidade preditiva prática.

5 Conclusão

Este estudo conclui que a utilização de dados de sentimento, aliados a modelos como o LSTM, tem grande potencial para a previsão de preços de ativos na Bolsa de Valores Brasileira. A combinação entre análise de sentimentos e modelagem temporal mostrou-se promissora, ainda que não tenha superado integralmente o desempenho obtido

por modelos tradicionais, como o SVR, demonstrando que modelos mais simples podem superar arquiteturas complexas em bases de dados pequenas.

A inclusão do modelo de linguagem especializado, FinBERT-PT-BR, trouxe para este estudo novas perspectivas sobre a utilização de informações subjetivas para a área financeira, e preenche lacuna significativa na literatura brasileira. Esse enfoque amplia a fronteira da pesquisa em finanças computacionais no Brasil, demonstrando a viabilidade do uso de modelos de PLN para capturar aspectos emocionais e cognitivos do investidor.

Para trabalhos futuros, podem ser utilizadas bases de dados maiores, reconhecimento de entidades para aplicar sentimento por ativo, fontes de dados alternativas, outros modelos preditivos complementares, como XGBoost ou diferentes RNN's, e realizar validações fora da amostra. Essas extensões podem transformar esta prova de conceito em um sistema de previsão com aplicabilidade real, contribuindo para o avanço do campo de análise de sentimento e precificação de ativos no mercado financeiro brasileiro.

Referências

ARACI, D. *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. 2019. Disponível em: <https://arxiv.org/abs/1908.10063>.

BARBER, B. M.; ODEAN, T. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The Review of Financial Studies*, Oxford University Press, v. 21, n. 2, p. 785–818, 2008.

DEVLIN, J. et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. Disponível em: <https://arxiv.org/abs/1810.04805>.

DRUCKER, H. et al. Support vector regression machines. In: MOZER, M.; JORDAN, M.; PETSCHE, T. (Ed.). *Advances in Neural Information Processing Systems*. MIT Press, 1996. v. 9. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf>.

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

FERRARI, M. B.; GARBI, F. d. S. M. *Análise de Sentimento na Precificação de Ativos na Bolsa Brasileira*. 2025. Repositório GitHub. Disponível em: <https://github.com/Matheus-B-Ferrari/AnaliseDeSentimentosB3>. Acesso em: 10 nov. 2025. Disponível em: <https://github.com/Matheus-B-Ferrari/AnaliseDeSentimentosB3>.

GU, W. jun et al. Predicting stock prices with finbert-lstm: Integrating news sentiment analysis. In: *Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing*. ACM, 2024. (ICCBDC 2024), p. 67–72. Disponível em: <http://dx.doi.org/10.1145/3694860.3694870>.

HALDER, S. *FinBERT-LSTM: Deep Learning based stock price prediction using News Sentiment Analysis*. 2022. Disponível em: <https://arxiv.org/abs/2211.07392>.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Computation*, v. 9, p. 1735–1780, 11 1997.

KASTURE, P.; SHIRSATH, K. Enhancing stock market prediction: A hybrid rnn-lstm framework with sentiment analysis. *Indian Journal Of Science And Technology*, v. 17, p. 1880–1888, 04 2024.

LIU, B. *Sentiment Analysis and Opinion Mining*. San Rafael, CA: Morgan & Claypool Publishers, 2012. (Synthesis Lectures on Human Language Technologies).

RICHARDSON, L. *Beautiful Soup: Python Library for HTML and XML Parsing*. 2007. <https://www.crummy.com/software/BeautifulSoup/>. Acesso em: 8 nov. 2025. Versão utilizada: BeautifulSoup4.

SANTOS, L.; BIANCHI, R.; COSTA, A. Finbert-pt-br: Análise de sentimentos de textos em português do mercado financeiro. In: *Anais do II Brazilian Workshop on Artificial Intelligence in Finance*. Porto Alegre, RS, Brasil: SBC, 2023. p. 144–155. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/bwaif/article/view/24960>.

TETLOCK, P. C. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, Wiley Online Library, v. 62, n. 3, p. 1139–1168, 2007.

TETLOCK, P. C. Chapter 18 - the role of media in finance. In: ANDERSON, S. P.; WALDFOGEL, J.; STRöMBERG, D. (Ed.). *Handbook of Media Economics*. North-Holland, 2015, (Handbook of Media Economics, v. 1). p. 701–721. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780444636850000188>.

UNIVERSIDADE PRESBITERIANA MACKENZIE. *Guia do TCC: Orientações gerais para a elaboração do trabalho de conclusão dos cursos de graduação*. São Paulo: UPM, 2021.

VIEIRA, J. E. A. L. *Modelo preditivo para precificação de ativos com integração de notícias do mercado financeiro e técnicas de machine learning*. Dissertação (Dissertação de Mestrado) — Fundação Getúlio Vargas (FGV), São Paulo, 2025. Disponível em: <https://repositorio.fgv.br/items/ca7c23d2-302c-4477-b9fe-f4102bdc523e>.