






# Brazilian Stock Market Forecast with Heterogeneous Data Integration for a Set of Stocks

Michele J. A. Rosa , Marcos R. Souza , Carlos L. S. Machado , Sandro J. Rigo , and Jorge L. V. Barbosa 

**Abstract**—The significant growth of the Brazilian stock market, coupled with the increase in investors in riskier assets, has generated a demand for automated forecasting tools. This research investigated the behavior and movement of stocks in the Brazilian market by integrating historical price series and textual data extracted from sentiments in X old Twitter messages and news collected from Google News. The analysis used natural language processing techniques for sentiment analysis, enabling an efficient fusion between numerical and textual information. Experiments were carried out with the assets PETR4, VALE3, BBDC4, and ITUB4, applying the Long Short-Term Memory, Deep Neural Network, and Linear Regression models to predict the behavior of these assets. The results indicated that the LSTM models, especially Model 2, presented the best performance in terms of predictive capacity, with the lowest values of RMSE 0.0171 and high values of coefficient of determination ranging from 0.9707 to 0.9873. The study concludes that integrating numerical and textual data, combined with deep learning techniques, offers a promising approach to stock market forecasting, increasing forecasting gains.

Link to graphical and video abstracts, and to code: <https://latam.ieee9.org/index.php/transactions/article/view/9184>

**Index Terms** — Financial market, Heterogeneous data, Natural Language Processing, Stock Exchange.

## I. INTRODUÇÃO

O mercado de capitais desempenha um papel fundamental no desenvolvimento econômico de um país e oferece a oportunidade de participação coletiva nos resultados da economia. Isso ocorre principalmente no mercado acionário, onde as empresas de capital aberto disponibilizam suas ações, possibilitando a negociação de compra e venda desses ativos financeiros [1], [2].

O risco na tomada de decisão para a compra de ativos no mercado acionário está associado à possibilidade de variações nos retornos esperados, sendo mensurável por meio de uma medida estatística que refletem a dispersão dos resultados em relação ao valor médio esperado [3], [4]. Já a incerteza, conforme definido por [5], refere-se à impossibilidade de

determinar as probabilidades de eventos futuros.

As decisões dos investidores baseiam-se em cenários econômicos, que podem ser sistemáticos, influenciados por fatores amplos de mercado, ou não sistemáticos, específicos de empresas ou setores [3], [4]. Além disso, a previsão dos preços dos ativos no mercado enfrenta desafios como a alta linear dos dados, que tornam o processo mais complexo e volatilidade e a natureza não exigem métodos avançados de modelagem [6].

Como forma de superar estes desafios, abordagens para a previsão de tendências futuras dos preços têm buscado analisar os movimentos e as flutuações do mercado de ações por meio de recursos de Inteligência artificial e aprendizado de máquina [7], [8], [9], [10], [11], [12], [13], [14].

Tradicionalmente, essas abordagens utilizam fontes de dados numéricos provenientes dos valores das ações. No entanto, observa-se uma oportunidade de ampliar o conjunto de dados utilizado nas previsões, considerando a abundância de informações geradas com o uso da internet, o aumento dos bancos de dados abertos e a publicação em tempo real nas mídias sociais, fontes de dados que vêm crescendo de forma exponencial [15], [16], [17], [18].

Este trabalho tem como objetivo explorar a ampliação do campo de atuação das técnicas tradicionais baseadas na análise gráfica, integrando-as com dados provenientes da abordagem fundamentalista. Esses dados, tradicionalmente obtidos por meio de relatórios técnicos e demonstrações contábeis, agora são complementados por uma vasta quantidade de informações provenientes de notícias, comentários em redes sociais e portais de notícias online, oferecendo uma perspectiva mais abrangente para a análise e tomada de decisão [19].

Estudos anteriores investigaram a integração do movimento diário dos preços das ações por meio de redes neurais profundas, treinadas com dados textuais de notícias financeiras e dados numéricos de preços e volume das ações [15], [16], [20], [21], [22]. No entanto, essas abordagens demonstraram uma diversidade limitada nas fontes de dados complementares e necessitam de uma análise mais aprofundada dos fatores específicos relacionados aos ativos avaliados. Essas lacunas evidenciam oportunidades relevantes para a ampliação da pesquisa e o desenvolvimento de métodos mais robustos nesta área de estudo.

Dessa forma, este artigo realiza uma análise da previsão do comportamento de determinados ativos do mercado de ações brasileiro, integrando dados de numéricos e dados textuais diversificados por meio da aplicação de técnicas avançadas de *Deep Learning* e Processamento de Linguagem Natural.

The associate editor coordinating the review of this manuscript and approving it for publication was Gladston Moreira (*Corresponding author: Michele Rosa*).

Michele Rosa, M. R. Souza, C. L. S. Machado, S. J. Rigo, and J. L. V. Barbosa are with Universidade do Vale do Rio dos Sinos, São Leopoldo, Rio Grande do Sul, Brazil (e-mails: michele.rosa@unemat.br, marcosrsouza@edu.unisinos.br, carlossmachado@edu.unisinos.br, rigo@unisinos.br, and jbarbosa@unisinos.br).

Este estudo utiliza dados diários para análise, permitindo capturar com maior precisão as variações e dinâmicas do mercado ao longo do tempo. Para a avaliação desse movimento de preços, foram selecionados quatro ativos listados na Bolsa de Valores Brasileira (B3). Adicionalmente, dados textuais provenientes de redes sociais, como o X/Twitter, e de portais de notícias, como o *Google News*, foram explorados com o uso de técnicas de análise de sentimentos.

Essa abordagem visa contribuir para o desenvolvimento de um modelo proposto mais avançadas, capazes de oferecer ganhos na capacidade preditiva. Nesse contexto, a combinação de dados numéricos e textuais visa aprimorar o desempenho das previsões e apoiar uma tomada de decisão mais informada e eficaz, por meio da redução dos riscos associados, ao incorporar informações adicionais sobre o mercado e seu contexto.

Este artigo está estruturado em cinco seções. Após esta introdução, a segunda seção apresenta os trabalhos relacionados, a terceira apresenta a modelo proposto utilizada na pesquisa e a quarta seção apresenta os resultados e as análises dos experimentos realizados. Por fim, a quinta seção apresenta as considerações finais do trabalho.

## II. TRABALHOS RELACIONADOS

Foi realizada uma revisão não sistemática sobre abordagens para previsão no mercado acionário envolvendo a integração de dados numéricos e dados textuais oriundos de notícias e outras fontes de dados. Observa-se a existência de trabalhos cujo objetivo foi explorar integrações mais simples de dados a partir de métodos tradicionais de predição [15], [16], [17], [18], [23], [24] demonstrando o potencial desta abordagem a partir de uso de análise de sentimentos em texto como base de integração dos dados.

Estes trabalhos destacam como lacunas de pesquisa a necessidade de tratamento diferenciado para as diversas fontes de dados e sua integração, aspectos que são aprofundados em outros trabalhos com abordagens considerando a ampliação de fontes de dados [15], [16], [17], [18], [23], [24], [25], [26], [27].

O estudo [15] analisaram a associação de sentimentos dos usuários e previsibilidade do movimento futuro do preço de ações, utilizando Rede Neural Artificial (RNA). Os dados foram extraídos da Bolsa de Valores de Gana (GSE) e abrangem o período de janeiro de 2010 a setembro de 2019. A previsão foi realizada para janelas de tempo de 1 dia, 7 dias, 30 dias, 60 dias e 90 dias, com as seguintes precisões: 49.4 a 52.95% com base no índice do *Google trends*, 55.5 a 60.05% com base no *Twitter*, 41.52 a 41.77% com base nas postagens de fóruns, 50.43 a 55.81% com base em notícias da web e 70.66 a 77.12% com a combinação dos conjuntos de dados.

O estudo [17] apresentou uma nova estrutura de previsão de valores de ações com fusão de múltiplas fontes de dados, usando uma arquitetura de rede neural híbrida integrando uma Rede Neural convolucional e uma rede *Long Short-Term Memory* (LSTM), chamada IKN-ConvLSTM.

Os dados utilizados no estudo foram divididos em três conjuntos quantitativos e três qualitativos. Os conjuntos quantitativos incluíram: (i) dados históricos de ações, coletados

no site oficial da Bolsa de Valores Gana (GSE) (<https://gse.com.gh/>), com 10 variáveis e 744 dias de negociação; (ii) dados macroeconômicos, obtidos no site oficial do Banco de Gana, com 44 indicadores econômicos ao longo de 744 dias de negociação; e (iii) o índice de tendências do Google, que incluiu 221 registros [17].

O estudo de [27] apresentou a previsão do preço das ações da BGI *Genomics*, com uma estrutura com fontes de dados de negociações diárias, notícias on-line e indicadores técnicos. Foi aplicada uma rede de *Long Short-Term Memory* (LSTM) e os resultados demonstraram que o modelo melhorou o desempenho da previsão por meio de fusão de informações heterogêneas.

O modelo proposto foi comparado o desempenho com métodos tradicionais, como regressão logística, máquina de vetores de suporte (SVM), árvores de decisão com aumento de gradiente e o LSTM original, em tarefas específicas, como a previsão da direção diária do preço e do preço de fechamento, além do desenvolvimento de estratégias de negociação. Os resultados experimentais demonstraram que o modelo LSTM com atenção supera significativamente os métodos comparativos, tanto em precisão estatística quanto em desempenho de negociação, evidenciando a eficácia da fusão de informações heterogêneas provenientes de diversas fontes [27].

Em [16] o objetivo foi prever os preços das ações e integrar a análise de sentimentos das postagens do *Twitter*, através de uma Rede Neural de *Long Short-Term Memory* (LSTM). Os resultados demonstram que a análise de sentimento ajudou a alcançar um *Root Mean Square Error* (RMSE) de 0.021 para a empresa Vale com a fusão do conjunto de dados de cotações de preços e análise de sentimentos.

No estudo [44] realizaram um comparativo sobre a previsão de preços das ações utilizando métodos de aprendizagem profunda em um conjunto de dados integrados. O estudo destaca que a combinação de preços das ações, indicadores técnicos, dados macroeconômicos e fundamentais apresenta um potencial significativo para aprimorar a compreensão do uso de conjuntos de dados na previsão de preços.

Portanto, para aumentar a precisão da previsão do mercado acionário, alguns trabalhos propuseram uma estrutura de fusão de dados numéricos e textuais, utilizando técnicas de aprendizado de máquina e processamento de linguagem natural. Observa-se que existem algumas limitações para realizar esta abordagem, tais como a dificuldade de integração dos dados em formatos distintos e a precisão das técnicas de análise de sentimentos em textos.

Na seção a seguir apresenta uma abordagem para tratar do contexto de fontes adicionais. Assim, este estudo visa contribuir na exploração e quantificação de melhorias marginais na precisão da previsão por meio da integração de dados numéricos e textuais.

## III. MATERIAIS E MÉTODOS

Esta seção apresenta o modelo proposto, detalhando as fontes de dados, o pré-processamento, a fusão e os modelos de predição. A Fig. 1 ilustra o fluxo completo. Foram utilizados dados numéricos de preços e volume das ações, além de dados textuais do *Google News* e X/Twitter. Após o pré-processamento individual, os dados foram integrados em uma base única para o treinamento e

avaliação da rede neural, visando a previsão dos valores das ações. Em seguida, foi realizada a simulação do investimento com os modelos selecionados.

O conjunto de dados utilizado neste estudo inclui, portanto, dados de numéricos (preços e volume das ações) e dados

textuais (*Google News* e *X/Twitter*), integrados com o objetivo de fornecer a base de treinamento para uma rede neural e desta forma prever o comportamento de ativos do mercado acionário brasileiro.

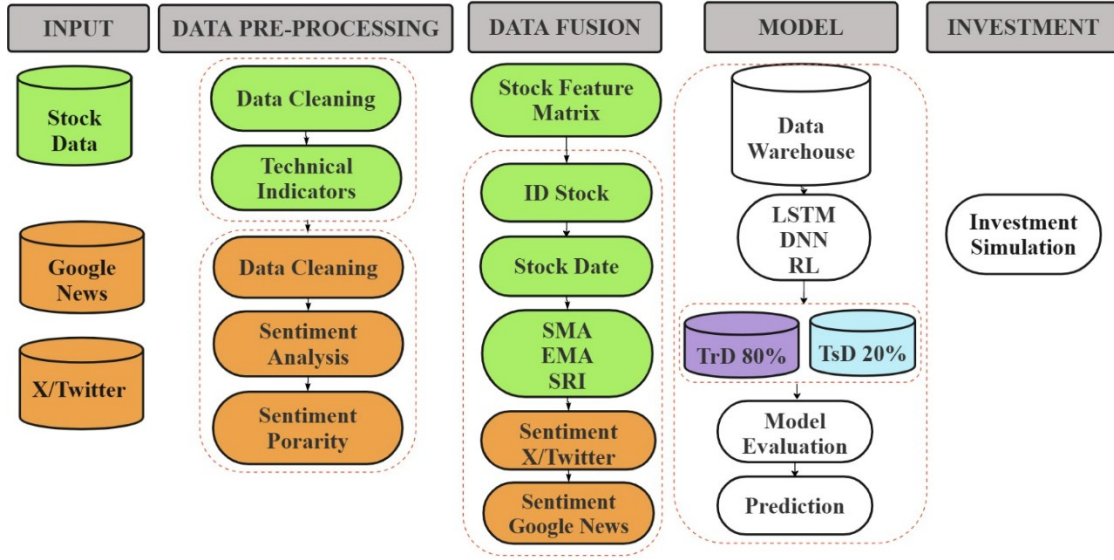


Fig. 1. Detalhamento do fluxo completo do pré-processamento e integração dos dados numéricos e textuais para o modelo proposto.

A seguir são descritos os procedimentos para tratar cada conjunto de dados, seu processamento, integração e os modelos de previsão.

#### A. Dados Numéricos

As empresas selecionadas para o experimento foram a Petrobras (PETR4), a Vale (VALE3), o Itaú Unibanco (ITUB4) e o Bradesco (BBDC4), pela sua grande relevância no mercado brasileiro, por fazerem parte do índice Ibovespa e serem as maiores empresas negociadas na B3. A Vale S.A. possui a maior participação no índice, com 13.73%, seguida pela Petrobras com 7.76%, o Itaú Unibanco com 7.14%, e o Bradesco com 3.74%, de acordo com informações referentes ao mês de janeiro de 2024 [3].

Desta forma, a escolha dos ativos para previsão considerou a liquidez, volatilidade, setores, sensibilidade a eventos, histórico de dados, participação no índice, e a disponibilidade de dados relevantes.

Os dados numéricos utilizados foram as séries históricas dos preços das ações, volume, fechamento, abertura, mínimo e máximo, coletados no Yahoo *finance*<sup>1</sup>. As séries foram analisadas a partir do preço de fechamento ajustado para considerar as alterações de preço devido à distribuição de proventos, sendo observado o comportamento diário das ações em dias úteis. O período analisado vai de 01/01/2008 a 20/09/2022, totalizando 3652 observações. Esse período foi marcado por vários eventos econômicos e políticos que geraram impactos significativos nas ações das empresas analisadas no estudo.

Os dados de preços das ações foram utilizados para calcular indicadores técnicos que ajudam a identificar a tendência no mercado de ações. Os indicadores técnicos utilizados foram a média móvel simples (MMS), média móvel exponencial (MME) e o índice de força relativa (IFR) [15], [29].

Estes indicadores são descritos a seguir [5], sendo a média móvel simples (MMS) descrita na equação (1),

$$MMS = \frac{\sum_{i=1}^n x_i}{n} \quad (1)$$

$x_i$  é o preço de fechamento em um determinado dia;  $n$  – é o número total de dias, que foi de 5 e 30 dias.

A média móvel exponencial (MME) na equação (2),

$$MME_d = MME_{d-1} + \frac{2}{n+1}x(P_d - MME_{d-1}) \quad (2)$$

$MME_{d-1}$  – é média móvel exponencial do dia anterior;  $P_d$  – é preço de fechamento no dia  $d$ ;  $n$  o número de períodos para a MME (5 e 30 dias usados no estudo).

O Índice de Força Relativa (IFR) mede a velocidade e as mudanças nos movimentos dos preços, variando entre 0 e 100. É tradicionalmente considerado sobrecomprado quando acima de 70 e sobrevendido quando abaixo de 30. No estudo, foi utilizado um período de 14 dias, no conforme sugerido por J.Welles Wilder, seu criador, conforme apresentado na equação (3).

$$IFR = 100 - \left( \frac{100}{1 + FR} \right) \quad (3)$$

<sup>1</sup><https://pypi.org/project/yahoofinancials/>

*FR* - (fator de força Relativa) é média dos ganhos em dias de alta dividida pela média das perdas em dias de baixa, ao longo do período considerado.

Para padronizar os dados numéricos, foi aplicada uma normalização nos valores de volume, fechamento, abertura, mínimo e máximo. Esse processo ajustou os valores para o intervalo entre 0 e 1, facilitando a análise de padrões de variação de preço ao longo do tempo [28], descrito na equação (4),

$$\text{normalized\_value} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (4)$$

$x$  é o valor original na série;  $x_{\min}$  é o menor valor na série;  $x_{\max}$  é o maior valor na série.

### B. Dados Textuais

O conjunto de dados textuais utilizado nos experimentos foi composto por notícias do *Google News* e postagens do X (antigo *Twitter*). Foram pesquisadas mensagens no período entre 01/01/2008 a 20/09/2022, mantendo-se apenas os dias úteis e descartando-se feriados e finais de semana. Foram encontradas 2869 observações do *Google News* e 2748 observações do *Twitter* para Petrobras (PETR4). Para Vale (VALE3) foram coletadas 1833 notícias no *Google News* e 1915 mensagens do *Twitter*. Para o Bradesco (BBDC4) foram obtidas 1782 notícias do *Google News* e 1585 mensagens do *Twitter*. Para o Itaú Unibanco (ITUB4), obteve-se 1843 notícias do *Google News* e 1853 menções do *Twitter*. Após esta coleta foram realizados os processos de limpeza e filtragem dos textos, descritos a seguir, no item "Filtragem e Limpeza".

A Tabela I apresenta exemplos dos dados textuais obtidos da empresa Petrobras (PETR4), como forma de ilustrar os desafios de identificação de aspectos positivos ou negativos neste contexto textual.

Os dados de notícias foram obtidos diretamente do portal de notícias *Google News* utilizando a biblioteca *pygooglenews*<sup>2</sup> no período predeterminado. Seguindo uma abordagem semelhante à proposta [30], onde as notícias foram coletadas e integradas aos dados de preços e volume das ações, de modo que os dados já estivessem estruturados e processados para classificação.

TABELA I

EXEMPLOS DE MENSAGENS OBTIDAS PARA PETROBRAS

Data	Notícias extraída do Google News
15/07/2010	Brasil começa a produção comercial de petróleo do pré-sal - Site Inovagro Tecnológica
30/07/2010	Entenda os riscos da exploração do Petróleo   Economia e Negócios   G1 — Globo.com
18/08/2010	Petrobras é condenada a indenizar família por acidente com explosivo — Consultor Jurídico
30/08/2010	MP investigará venda de ativos de Petrobras a Braskem - Veja
15/09/2010	MP vai investigar se policiais de SP quebraram sigilos para Petrobras — Globo.com
<b>Postagens extraída do X (antigo Twitter) tweets</b>	
23/10/2009	Petrobras, TAM e NET captam US\$ 5.6 bilhdes
26/10/2009	Petrobras e Vale descolam Bovespa da cena externa
06/11/2009	Petrobras confirma descoberta de gás no Peru
10/11/2009	Petrobras desbanca quadrilha que atuava em projetos
13/11/2009	Lucro da Petrobras cai para R\$ 7,303 bilhdes

A extração dos dados do *Google News* foi realizada utilizando como parâmetros de busca o nome do ativo em conjunto com o nome da empresa, como por exemplo, "PETR4" e "Petrobras". Em seguida os dados foram unificados para datas iguais, a fim de evitar datas em duplicidade e garantir uma análise de sentimentos para as notícias relacionadas ao dia corrente. Após a extração, obteve-se como resultado um *dataset* com as notícias unificadas para cada dia.

O conjunto de dados do *Twitter* foi obtido através da leitura das postagens dos *tweets* contendo o código dos ativos ou o nome das empresas selecionadas. Por exemplo, foram utilizadas as *hashtags* "PETR4" e "Petrobras", "VALE3", "Vale do Rio Doce" e "minério", "ITUB4" e "Itaú", "BBDC4" e "Bradesco". Os *tweets* foram coletados utilizando a API fornecida pelo *Twitter*<sup>3</sup>, conforme estudo [31]. Foram empregadas quatro fontes de notícias do mercado financeiro sendo elas, InvestingBrasil<sup>4</sup>, Money Times<sup>5</sup>, InfoMoney<sup>6</sup> e Valor Econômico<sup>7</sup>.

### C. Filtragem e Limpeza

Para a preparação dos dados foram efetuados processos para garantir a limpeza dos dados e maior coerência nas análises. Algumas das técnicas utilizadas [32], basearam-se nos processos de vetorização e indexação, nos quais as palavras foram transformadas em *tokens*. Foram removidos textos duplicados; em seguida os textos foram convertidos para letras minúsculas garantindo uma padronização; após isso foram removidas as pontuações, números e símbolos que não seriam utilizáveis mantendo-se apenas espaços em branco e alguns caracteres especiais que estão ligados a língua portuguesa. Logo após, foi efetuada a remoção dos nomes dos ativos utilizados nos testes (Petrobras, Vale, Itaú, Bradesco) para garantir que o modelo proposto seja robusto, generalizável e livre de vieses específicos.

O próximo passo foi efetuar a tokenização dos textos e realizar a remoção de todas as palavras (*stop words*) da lista de *tokens*. Neste processo foi utilizado um dicionário de *stop words* contido na biblioteca "spacy" [33]. Por fim, utilizamos um processo de lematização das palavras para reduzir os *tokens* às suas formas básicas ou infinitivas. Após esses processos, foi criado um *dataframe* com os dados textuais já processados e unificados pela data de ocorrência.

### D. Análise de Sentimentos

De modo a realizar a fusão de dados, os elementos textuais obtidos no *Twitter* e no *Google News* foram processados de modo a gerar uma análise de sentimentos em texto dos títulos das notícias. Para isso foi utilizada a biblioteca *Leia* [34], que é uma variação da biblioteca *Vader Sentiments* [35], na qual é utilizado um dicionário de termos voltado para a língua portuguesa. Para a aplicação dos processos de polarização dos sentimentos, foram analisados os dados obtidos para cada dia especificamente. Os sentimentos são classificados em três categorias, sendo elas

<sup>2</sup><https://pypi.org/project/pygooglenews/>

<sup>3</sup><https://developer.x.com/en/docs/twitter-api/>

<sup>4</sup><https://twitter.com/InvestingBrasil>

<sup>5</sup><https://twitter.com/leiamoneytimes>

<sup>6</sup><https://twitter.com/infomoney>

<sup>7</sup><https://twitter.com/valoreconomico>

positiva, negativa e neutra. A Tabela II apresenta alguns exemplos de textos analisados e sua classificação.

TABELA II  
EXEMPLOS DE ANÁLISE DE SENTIMENTOS DAS NOTÍCIAS

Data	Sentimento	Textos	Compound	Positive	Negative	Neutral
18/08/2010	Positive	Brasil tem sete entre 500 maiores empresas do mundo, aponta "Fortune"	0.8689	0.495	0	0.505
7/7/2014	Negative	Petrobras é condenada a indenizar família por acidente com explosivo	-0.7184	0	0.5	0.5
25/12/2015	Neutral	O que foi 2015 para os brasileiros segundo as pesquisas no Google	0	0	0	1

Esta é definida como a soma dos valores da polarização, que ao final terá como resultante um valor normalizado entre  $-1$  e  $+1$ . Neste caso, o valor  $-1$  representa um resultado extremamente negativo e o valor  $+1$  indica um resultado extremamente positivo.

#### E. Fusão do Conjunto de Dados

Para a fusão dos dados numéricos e textuais, três conjuntos de dados diferentes foram combinados: dados numéricos normalizados de séries históricas de preços e volume das ações, dados polarizados por sentimentos de análises do *Google News* e dados polarizados por sentimentos de postagens no X/Twitter.

Para as séries históricas de preços e volume das ações foram utilizados os dados "data", "fechamento", "abertura", "mínimo", "máximo" e "volume". No caso dos dados do X/Twitter e *Google News* utilizou-se os dados "data", "compound", "negativo", "neutro", "positivo", gerados pelo processo de análise de sentimentos em textos. Na etapa seguinte todos os conjuntos de dados textuais foram integrados aos dados de preços e volume de acordo com a sua data.

Para evitar a ocorrência de dias sem valores numéricos devido a não haver postagens ou notícias no dia em questão, definiu-se um método para preencher os dados evitando as polarizações sem valores. Caso o primeiro dia não possua valores, este é definido como neutro. Para os dias posteriores, caso não haja valores polarizados, utiliza-se um processo de redução em 50% para o próximo dia em relação ao anterior. Um exemplo deste processo pode ser observado na Tabela III, no qual as ocorrências de valores "NaN" são substituídas de acordo com este método.

TABELA III  
DIAS SEM POSTAGENS E MÉTODO DE PREENCHIMENTO

Data	Fechamento	Compound Twitter	Compound Twitter 50% do valor anterior
20/05/2022	0.025	0.153	0.153
21/05/2022	0.05	NaN	0.0765
22/05/2022	0.033	NaN	0.03825
23/05/2022	0.04	-1	-1
24/05/2022	0.045	NaN	-0.5

Esse processo foi realizado para que os dias sem postagens ou notícias diminuam seu valor de sentimento atual, até que se aproximem de zero, ou seja, tendam à neutralidade com o tempo.

#### F. Composição dos Dados para Treinamento

Para a composição dos dados de treinamento dos modelos, foi adotado um conjunto de procedimentos para preparar todas as informações necessárias. Do total de dados disponíveis, 80% foram utilizados para o treinamento do modelo e 20% para o teste.

Como exemplifica a Tabela II, os valores relativos aos percentuais entre 0 e 1 são compostos com uma métrica de composição dos sentimentos chamada *compound*.

Após essa etapa, foi definida a variável *close\_next\_day*, utilizada para armazenar os valores de fechamento do dia seguinte. As demais variáveis consideradas no modelo foram:

- Indicadores de preço e volume: *open*, *high*, *low*, *close*, *volume*.
- Métricas de sentimento: *compound\_gn*, *negative\_gn*, *neutral\_gn*, *positive\_gn*, *compound\_tw*, *negative\_tw*, *neutral\_tw*, *positive\_tw*.
- Indicadores técnicos: *mms\_30*, *mms\_5*, *mme\_5*, *mme\_30*, *ifr\_14*.

#### G. Estrutura dos Modelos de Predição

Os modelos de previsão utilizados foram: *Long Short-Term Memory* (LSTM) sem retroalimentação, LSTM com retroalimentação, *Deep Neural Network* (DNN) e Regressão Linear (RL).

Será realizada uma comparação entre os modelos para otimizar a previsão de preço no mercado de ações. Cada modelo possui vantagens distintas, e compreender essas diferenças possibilita uma aplicação mais eficaz em diferentes condições de mercado, aprimorando a precisão e a relevância das previsões para estudo proposto, pode ser observado na Tabela IV.

O primeiro, LSTM sem retroalimentação, possui duas camadas: a primeira camada é uma LSTM com 16 neurônios, e a camada de saída é uma *Dense* com 1 neurônio. Utilizou-se o otimizador *Adam*, com *batch size* de 64 e 100 épocas para treinamento. Esse modelo capta dependências de longo prazo em séries temporais de forma simples, como em dados de ações, sem considerar influências adicionais dos resultados passados.

O segundo modelo, uma LSTM com retroalimentação, tem 4 camadas com LSTM com 200 neurônios, 300, e 400 neurônios, seguidas por uma camada *Dense* com 1 neurônio. As camadas LSTM usam ativação de tangente hiperbólica (*tanh*) e, função recorrente *sigmoid* e *dropout* de 3% nas duas últimas camadas. Treinado com otimizador *Adam*, *batch size* 32 e 25 épocas, o modelo captura padrões complexos ao usar a saída anterior com entrada para próxima previsão, essencial em mercados financeiros, onde eventos recentes impactam o futuro [46].

O terceiro modelo tem 3 camadas: a camada de entrada é uma *Dense* com 32 neurônios e ativação *relu*, seguida por uma *Dense* de 8 neurônios e ativação *relu* e uma camada de saída *Dense* com 1 neurônio. Embora as DNNs geralmente sejam menos eficazes em capturar dependências temporais diretamente, elas são

TABELA IV  
DESCRIÇÃO DA ARQUITETURA DOS MODELOS DE APRENDIZAGEM

Modelo	Camadas e Arquitetura	Ativação	Parâmetros de Treinamento
<b>Modelo 1 - LSTM sem retroalimentação</b>	1 camada: LSTM (16 neurônios) Camada de saída: Dense (1 neurônio)	LSTM: padrão Dense: linear	Otimizador: Adam Batch size: 64 Épocas: 100
<b>Modelo 2 -LSTM com retroalimentação</b>	camada: (200 neurônios) camada: (300 neurônios)38 camada: (400 neurônios) Camada de saída: Dense (1 neurônio)	LSTM: tanh Recurrent: sigmoid Dense: linear	Dropout: 3% nas 2 últimas LSTM Otimizador: Adam Batch size: 32 - Épocas 25
<b>Modelo 3 - Deep Neural Network (DINN)</b>	1ª camada: Dense (32 neurônios, ativação ReLU) - 2ª camada: Dense (8 neurônios, ativação ReLU) Saída: Dense (1 neurônio)	ReLU nas camadas ocultas, linear na saída	
<b>Modelo 4 - Regressão Linear (RL)</b>	Linear Regression do scikit-learn	Linear ( $\alpha + \beta x + \varepsilon$ )	

poderosas para trabalhar com dados tabulares ou combinados, explorando relações não lineares complexas e oferecendo maior flexibilidade.

O quarto modelo, de regressão linear (RL), utiliza a função *LinearRegression* do *scikit-learn* para replicar matematicamente a equação (5). Ele estima os coeficientes minimizando o erro quadrático médio (MSE), garantindo que a saída seja uma combinação linear ponderada das entradas. Esse método captura padrões lineares nos dados, mas possui limitações diante de relações mais complexas [45].

$$Y = \alpha + \beta x + \varepsilon \quad (5)$$

sendo  $\alpha + \beta x$  a equação da reta, e  $\varepsilon$  o termo de erro.

Portanto, regressão linear é uma técnica de aprendizado de máquina de algoritmo supervisionado, utiliza equação linear que usa os valores de entradas para prever as saídas, trabalhando apenas com valores numéricos e os pesos são atualizados conforme a função que minimiza o erro [45].

Os LSTMs são mais adequados para modelar sequências temporais, capturando dependências de longo prazo, enquanto as DNN são eficazes para capturar padrões complexos em dados não sequenciais [36]. Assim, o estudo propôs uma comparação para prever o comportamento das ações, observando qual modelo se adequa melhor à captura e dinâmica dos ativos e fonte de dados utilizadas.

#### H. Métricas de Desempenho

As métricas de avaliação escolhidas foram: *Root Mean Square Error* (RMSE), e Coeficiente de Determinação ( $R^2$ ). Essas métricas oferecem uma avaliação equilibrada e detalhada dos modelos de previsão [15], [24], [37].

A Raiz do Erro Quadrático Médio (*Root Mean Squared Error* - RMSE) mede o erro médio entre os valores previstos e os reais. Diferentemente do MAPE, o RMSE não expressa o erro em percentual, mas como um valor numérico que representa a magnitude do erro. Valores mais próximos de zero indicam maior precisão no modelo. O RMSE é calculado pela equação (6),

$$RMSE = \sqrt{\frac{1}{n} \sum (previsto_i - real_i)^2} \quad (6)$$

Na equação 6,  $N$  é o número de dias da série temporal sendo analisada,  $previsto_i$  é o valor previsto para a série no dia  $i$  e  $real_i$  é o valor real da série também no dia  $i$ .

O Coeficiente de Determinação ( $R^2$ ) avalia a proporção da variância dos dados que é explicada pelo modelo. Seus valores variam entre 0 e 1, sendo que valores mais próximos de 1 indicam maior capacidade explicativa do modelo em relação aos dados.

$$R^2 = 1 - \frac{\sum_{i=1}^N (real_i - previsto_i)^2}{\sum_{i=1}^N (real_i - \overline{real})^2} \quad (7)$$

A abordagem utilizada da Acurácia, foi uma métrica de precisão adaptada para problemas de regressão, complementando métricas tradicionais como MSE, MAE e MAPE. Assim, essas métricas permitem uma análise quantitativa abrangente do desempenho dos modelos.

#### IV. RESULTADOS

A combinação dos dados numéricos e textuais foi composta por preços e volume das ações, sentimentos *Google News*, sentimentos *X/Twitter*, cálculo do índice de força relativa (IFR), média móvel simples (MMS) e a média móvel exponencial (MME), totalizando sete combinações. Com essa abordagem, foram realizados experimentos em todos os modelos e para todos os ativos PETR4 (Petróleo Brasileiro S.A), VALE3 (Vale S.A), BBDC4 (Banco Bradesco S.A) e ITUB4 (Itaú Unibanco S.A), no período entre 01/01/2008 e 20/09/2022.

Essas empresas, pertencentes a setores econômicos distintos, como petróleo, mineração e bancos, desempenham um papel significativo no mercado da B3 e no índice Ibovespa. Sua ampla cobertura em notícias e discussões em mídias sociais proporciona um grande volume de dados textuais e numéricos, viabilizando uma análise abrangente das dinâmicas de mercado e das reações a eventos econômicos e financeiros.

##### A. Resultado dos Experimentos

As Figs. 2 e 3 apresentam os resultados da métrica de desempenho Raiz do Erro Quadrático Médio (RMSE) para cada modelo, considerando as sete combinações de dados numéricos e textuais utilizadas na análise.



Os resultados dos indicadores foram obtidos a partir da amostra de testes e com as variáveis transformadas no intervalo  $[0,1]$ . A comparação entre as informações textuais do *Google News* e do *X/Twitter*, quando combinadas com dados numéricos, revela que a diferença de desempenho entre as duas fontes não é expressiva.

No entanto, o *X/Twitter* apresentou um RMSE ligeiramente maior em alguns modelos, possivelmente devido ao ruído e à superficialidade de suas publicações em comparação com as notícias mais elaboradas do *Google News*, como observado no estudo [15].

A inclusão de indicadores técnicos (IFR, MMS e MME) junto às informações textuais e numéricas geralmente melhora o desempenho do modelo, reduzindo o RMSE em alguns casos.

Como mostrado nas Figs. 2 e 3, a eficácia das combinações varia conforme o modelo e o ativo analisado. No geral, as combinações 4 (*Stock Data* + IT) e 5 (*Google News* + *Twitter* + IT) apresentaram os menores erros para a maioria dos modelos e ativos.

Na Fig. 3 observa-se que os modelos 1, 2 e 3 tiveram um RMSE baixo e consistente, indicando que conseguiram captar padrões de mercado de forma eficiente. Em contrapartida, a Fig. 2 mostra que o modelo 4 (Regressão Linear) apresentou RMSE erros mais elevados, especialmente para VALE3, sugerindo maior dificuldade desse modelo em prever esse ativo.



Fig. 2. Desempenho comparativo dos modelos de predição, com base na Raiz do Erro Quadrático Médio (RMSE), considerando diferentes combinações de integração de dados para os ativos analisados.

No geral, VALE3 teve os maiores erros, enquanto BBDC4 e ITUB4 foram os ativos mais simples de prever. Na Tabela V, podemos observar os resultados para as métricas de desempenho para a Combinação 5 (*Google News* + *Twitter* + *IFR* + *MMS*), para os modelos. Os modelos de aprendizagem profunda (Modelos 1, 2 e 3) apresentaram os melhores desempenhos em termos de RMSE e  $R^2$ . Os resultados obtidos que o Modelo 2 – LSTM foi o mais consistente e obteve o melhor desempenho geral entre os quatro modelos testados.

Esse modelo apresentou os valores menores de RMSE, como 0.0171 para PETR4, e manteve baixos erros de previsão para todos os ativos. Destacou-se especialmente em ativos como VALE3, com um  $R^2$  de 0.9873, e ITUB4, com  $R^2$  de 0.9730.

O Modelo 1 – LSTM também apresentou um bom desempenho, com  $R^2$  elevados, variando entre 0.9446 e 0.9679, e valores de RMSE relativamente baixos. No entanto, para VALE3, exibiu um erro elevado de 0.04532, maior dificuldade na previsão desse ativo.

O Modelo 3 – DNN também apresentou um desempenho sólido, com RMSEs baixos e  $R^2$  elevados, embora ligeiramente inferiores ao Modelo 2 – LSTM. Por outro lado, o Modelo 4 – Regressão Linear apresentou o pior desempenho, com RMSEs significativamente mais altos e  $R^2$  negativo para alguns ativos, como VALE3 e BBDC4, revelando que esse modelo não conseguiu capturar bem os padrões dos dados.



Fig. 3. Comparação entre diferentes formas de integração de dados aplicadas a três modelos de previsão: (1) LSTM sem retroalimentação, (2) LSTM com retroalimentação e (3) *Deep Neural Network* (DNN).

TABLE V  
RESULTADOS DAS MÉTRICAS DE DESEMPENHO RMSE E  $R^2$  PARA OS  
MODELOS DE PREDIÇÃO

Ativo	Modelo 1 - LSTM		Modelo 2 - LSTM	
	RMSE	$R^2$	RMSE	$R^2$
PETR4	0.01802	0.9677	0.0171	0.9707
VALE3	0.04532	0.9446	0.0216	0.9873
BBDC4	0.02358	0.9679	0.0224	0.9709
ITUB4	0.02285	0.9658	0.0202	0.9730
Ativo	Modelo 3 - DNN		Modelo 4 - RL	
	RMSE	$R^2$	RMSE	$R^2$
PETR4	0.0180	0.9675	0.0742	0.4539
VALE3	0.0277	0.9792	0.2409	-0.5641
BBDC4	0.0223	0.9711	0.1809	-0.8869
ITUB4	0.0198	0.9742	0.0641	0.7309

Esses resultados reforçam a superioridade dos modelos de aprendizado profundo na previsão dos preços das ações quando combinados com dados numéricos, indicadores técnicos e informações textuais.

No período analisado, os ativos selecionados foram impactados por diversos eventos econômicos e políticos que afetaram os mercados de *commodities* e o setor financeiro, tais como: crise financeira global (2008-2009); boom das *commodities* (2009-2011); a descoberta do Pré-sal e os investimentos na Petrobras (2007-2014); crise política e econômica no Brasil (2014-2016); a tragédia de Brumadinho (2019); a pandemia de COVID-19 (2020-2022); entre outros eventos que refletiram nos resultados dos modelos para cada ativo.

A empresa Vale apresentou maior volatilidade no preço de suas ações devido ao desastre da barragem de Mariana -MG [16]. Podemos observar que os ativos enfrentam desafios

significativos devido à Hipótese de Eficiência dos Mercados (HME). De acordo com a HME [38], [39], os preços das ações refletem todas as informações disponíveis e seu ajuste à novas informações é instantâneo. Porém suas limitações e as ineficiências observadas no mercado sugerem que, em prática, os mercados podem não ser completamente eficientes. Assim, possibilitando a previsão de preços de ativos, pois tem informações que não refletem imediatamente.

A teoria clássica da economia (HME) assume que os agentes econômicos são racionais e que desvios dessa racionalidade são anomalias. Em contraste, a economia comportamental argumenta que a racionalidade humana é limitada e que as decisões econômicas são influenciadas por uma variedade de fatores, incluindo aspectos emocionais, sociais, econômicos, cognitivos e culturais[40], [41], [42], [43].

Desta forma, as séries históricas e os indicadores técnicos são potenciais recursos para previsão do comportamento e movimento dos preços das ações, mas as informações de notícias (dados textuais), apresentam resultados eficientes, com potenciais ganhos nas previsões.

### B. Simulação do Investimento

Com base nos experimentos realizados, selecionamos o Modelo 2 – LSTM e duas combinações de dados: Combinação 4 (Stock Data + IT: IFR, MMS, MME) e Combinação 5 (Google News + Twitter + IFR + MMS), que apresentaram o melhor desempenho entre as opções analisadas. Também foi incluída uma comparação das simulações utilizando os modelos 1 e 3 com a combinação 5.

Para avaliar a aplicabilidade no mercado real, desenvolvemos uma estratégia de investimento baseada nas previsões do modelo. Sempre que o modelo indicasse alta para o próximo dia, uma compra era realizada na abertura do pregão seguinte, desde que não houvesse uma posição já comprada. Caso a



posição já estivesse em carteira, ela era mantida. Se o modelo previsse queda e houvesse uma posição comprada, a venda era efetuada na abertura do pregão seguinte; caso contrário, nenhuma operação era realizada.

Consideramos um capital inicial de R\$ 10.000.00 e realizamos a simulação apenas nos dados de teste dos modelos, ou seja, utilizando exclusivamente os 20% dos dados não empregados no treinamento. Esse conjunto cobre o período de 23/10/2019 a 16/09/2022, incluindo a fase de maior volatilidade do mercado devido à pandemia de Covid-19. Os custos operacionais não foram considerados nas simulações.

Para comparação, adotamos a estratégia de investimento *Buy & Hold*, na qual o investidor adquire o ativo no início do período e mantém a posição até o final [13]. Os resultados obtidos com as diferentes estratégias são apresentados na Tabela VI.

TABLE VI

RESULTADO DA SIMULAÇÃO DE INVESTIMENTO, EM REAIS (R\$), COMPARANDO OS ATIVOS ANALISADOS COM DUAS ESTRATÉGIAS: *BUY & HOLD* E UMA ESTRATÉGIA BASEADA EM MODELOS DE PREDIÇÃO

Ativo	Buy & Hold (R\$)	Retorno Buy & Hold (%)	Estratégia (Modelo 2) (R\$)	Retorno Estratégia (%)	Estratégia (Modelo 2) (R\$)	Retorno Estratégia (%)
Investimento direto na ação R\$			Combinação 4		Combinação 5	
PETR4	10,739.39	7.39%	17,504.55	75.05%	13,426.86	34.27%
VALE3	15,489.48	54.89%	15,442.85	54.43%	13,114.02	31.14%
BBDC4	6,553.20	-34.47%	12,441.31	24.41%	14,383.19	43.83%
ITUB4	6,668.49	-33.32%	7,050.50	-29.50%	12,249.94	22.50%
Ativo	Buy & Hold (R\$)	Retorno Buy & Hold (%)	Estratégia (Modelo 1) (R\$)	Retorno Estratégia (%)	Estratégia (Modelo 3) (R\$)	Retorno Estratégia (%)
Investimento direto na ação R\$			Combinação 5		Combinação 5	
PETR4	10,739.39	7.39%	11,475.02	14.75%	11,965.57	19.66%
VALE3	15,489.48	54.89%	12,456.36	24.56%	14,403.46	44.03%
BBDC4	6,553.20	-34.47%	8,607.81	-13.92%	7,550.06	-24.50%
ITUB4	6,668.49	-33.32%	6,080.85	-39.19%	11,147.27	11.47%

O Modelo 2 - LSTM com retroalimentação apresentou, de forma geral, o melhor desempenho nas simulações. Podemos observar na Tabela VI a estratégia baseada na combinação 5 para os modelos 1 e 3. O Modelo 1 - LSTM sem retroalimentação teve retornos mais modestos, especialmente para o ativo ITUB4 com um retorno negativo de -39,19%, com um bom desempenho para PETR4 de 14,75%, mas limitou-se na captura de dependências de longo prazo. Já o Modelo 3 - DNN demonstrou boa capacidade de modelar relações não lineares, especialmente em VALE3 44,03% e ITUB4 11,47%, ainda que seus resultados não tenham superado consistentemente os do Modelo 2.

Os resultados obtidos reforçam o potencial da utilização de modelos baseados em aprendizado profundo, como o Modelo 2 - LSTM com retroalimentação, para estratégias de previsão e tomada de decisão no mercado financeiro, especialmente quando combinados com indicadores técnicos e dados textuais. Embora o desempenho financeiro varie conforme o ativo e as condições de mercado, a abordagem proposta demonstrou capacidade de superar estratégias tradicionais como o *Buy & Hold* em diversos cenários, sobretudo em ativos com tendência negativa. Dessa forma, o estudo contribui para a compreensão das aplicações práticas de modelos preditivos no contexto de investimentos, destacando a importância da escolha adequada de modelos e da seleção das variáveis utilizadas na previsão, além de reforça o

Utilizamos duas combinações de dados: a Combinação 4 (*Stock Data + IT (IFR + MMS + MME)*) e a Combinação 5 (*Google News + Twitter + IFR + MMS*). Os resultados para o modelo 2 indicam que, para a ação PETR4, a estratégia baseada na Combinação 4 apresentou o melhor retorno, atingindo 75,05%, enquanto a Combinação 5 obteve 34,27%, ambos superando o *Buy & Hold*, que teve um retorno de 7,39%. No caso de VALE3, o desempenho da estratégia foi semelhante ao investimento passivo, com retornos de 54,43% e 31,14%, respectivamente, para as Combinações 4 e 5, em comparação aos 54,89% do *Buy & Hold*. Para os ativos BBDC4 e ITUB4, a estratégia mostrou-se mais eficiente na reversão de perdas. Enquanto o *Buy & Hold* resultou em retornos negativos de -34,47% e -33,32%, a estratégia baseada na Combinação 5 conseguiu reverter esses prejuízos, alcançando retornos positivos de 43,83% para BBDC4 e 22,50% para ITUB4.

papel das informações textuais na melhoria das decisões no mercado financeiro, conforme o ativo analisado.

## V. CONCLUSÃO

Esta pesquisa investigou o comportamento de ativos do mercado acionário brasileiro por meio da integração de dados numéricos e dados textuais, para prever o preço de fechamento das ações no dia seguinte, utilizando técnicas de processamento de linguagem natural para análise de sentimentos. Foram realizados experimentos com os ativos PETR4, VALE3, BBDC4 e ITUB4, empregando os modelos *Long Short Term Memory* (LSTM), *Deep Neural Network* (DNN) e Regressão Linear (RL), a fim de avaliar a eficiência dessa abordagem.

Os modelos baseados em LSTM se mostraram mais adequados para a previsão dos ativos, com ênfase para o Modelo 2 que obteve os valores menores de RMSE (de 0,0171 PETR4) e altos valores de  $R^2$  (entre 0,9707 e 0,9873), enquanto o Modelo 4, baseado em regressão linear, teve desempenho limitado e é menos indicado para esse contexto. O Modelo 3, baseado em DNN, mostrou potencial em casos específicos, mas cuidado de consistência geral.

O mercado acionário brasileiro, possui características distintas de mercados mais maduros. Isso limita as possibilidades de generalização dos modelos para outros contextos e ativos, especialmente para aqueles que exibem comportamentos diferentes dos analisados. Como observado nos resultados dos

modelos propostos, os comportamentos dos ativos variaram de acordo com a situações que afetaram os setores analisados e os impactos internos e externos da economia.

Os resultados da simulação mostraram que a estratégia baseada em aprendizado profundo obteve retornos superiores ao investimento passivo para a maioria dos ativos analisados, evidenciando o potencial do modelo para otimização de decisões no mercado financeiro.

O estudo apresenta limitações que podem impactar a generalização dos resultados, como a influência do período analisado (2008-2022), inclui a pandemia de Covid-19, um evento atípico que pode ter influenciado os padrões do mercado. Além disso, possíveis ruídos nos dados textuais, ausência de custos operacionais, sensibilidade dos modelos a eventos extremos de cada ativo.

Os experimentos futuros podem incluir variações macroeconômicas, como inflação e PIB, para avaliar seu impacto na previsão de ativos, além de aprimorar o Modelo 2 (LSTM) com novos parâmetros e explorar ativos de baixa rentabilidade.

#### REFERÊNCIAS

- [1] F. S. Alzazah e X. Cheng, “Chapter Recent Advances in Stock Market Prediction Using Text Mining: A Survey”, em *E-Business - Higher Education and Intelligence Applications*, London, United Kingdom: IntechOpen, 2020 [Online], 2021. doi: 10.5772/intechopen.92253.
- [2] F. Lemos, *Technical Analysis of Financial Markets: A Complete and Definitive Guide to Asset Trading Methods*, 3<sup>o</sup> ed. São Paulo, 2022.
- [3] Assaf Neto, *financial market*, 15<sup>o</sup> ed, vol. 15. Barueri: SP, 2021.
- [4] J. L. Pinheiro, *Capital markets*, 9<sup>o</sup> ed. São Paulo: Atlas, 2019.
- [5] H. F. Knight, *Risk, uncertainty and profit*. Hart, Schaffner and Marx, 1921.
- [6] P. da J. Silva, *financial analysis of companies*, 13<sup>o</sup> ed. São Paulo: Atlas, 2016.
- [7] G. Sismanoglu, M. A. Onde, F. Kocer, e O. K. Sahingoz, “Deep Learning Based Forecasting in Stock Market with Big Data Analytics”, em *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, IEEE, abr. 2019, p. 1–4. doi: 10.1109/EBBT.2019.8741818.
- [8] L. Werner, C. Bisognin, e C. W. Araujo, “Analysis of forecasting techniques: a case study for the volume of Petrobras shares”, *Brazilian Journal of Development*, vol. 6, n<sup>o</sup> 1, p. 1103–1115, 2020, doi: 10.34117/bjdv6n1-078.
- [9] B. Kaczorowski, M. Kleina, M. Augusto Mendes Marques, e W. de Assis Silva, “Artificial Intelligence and The Multivariate Approach In Predictive Analysis Of The Small Cap Index Of The Brazilian Stock Exchange”, *IEEE Latin America Transactions*, vol. 19, n<sup>o</sup> 11, p. 1924–1932, nov. 2021, doi: 10.1109/TLA.2021.9475626.
- [10] S. Du, D. Hao, e X. Li, “Research on stock forecasting based on random forest”, em *2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA)*, IEEE, out. 2022, p. 301–305. doi: 10.1109/ICDSCA56264.2022.9987903.
- [11] B. Albaoth, “The Role of Artificial Intelligence Prediction in Stock Market Investors Decisions”, em *2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, IEEE, dez. 2023, p. 1–5. doi: 10.1109/CSDE59766.2023.10487719.
- [12] M. Korablyov, O. Fomichov, D. Antonov, S. Dykyi, I. Ivanisenko, e S. Lutsky, “Hybrid stock analysis model for financial market forecasting”, em *2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT)*, IEEE, out. 2023, p. 1–4. doi: 10.1109/CSIT61576.2023.10324069.
- [13] Carosia, “Using Machine Learning to Prevent Losses in the Brazilian Stock Market During the Covid-19 Pandemic”, *IEEE Latin America Transactions*, vol. 21, n<sup>o</sup> 8, p. 867–873, ago. 2023, doi: 10.1109/TLA.2023.10246342.
- [14] Ruke, S. Gaikwad, G. Yadav, A. Buchade, S. Nimbarkar, e A. Sonawane, “Predictive Analysis of Stock Market Trends: A Machine Learning Approach”, em *2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)*, IEEE, mar. 2024, p. 1–6. doi: 10.1109/ICDECS59733.2023.10503557.
- [15] K. Nti, A. F. Adekoya, e B. A. Weyori, “Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence From Ghana”, *Applied Computer Systems*, vol. 25, n<sup>o</sup> 1, p. 33–42, maio 2020, doi: 10.2478/acss-2020-0004.
- [16] G. Vargas, L. Silvestre, L. Rigo Júnior, e H. Rocha, “B3 Stock Price Prediction Using LSTM Neural Networks and Sentiment Analysis”, *IEEE Latin America Transactions*, vol. 20, n<sup>o</sup> 7, p. 1067–1074, jul. 2022, doi: 10.1109/TLA.2021.9827469.
- [17] K. Nti, A. F. Adekoya, e B. A. Weyori, “A novel multi-source information-fusion predictive framework based on deep neural networks for accuracy enhancement in stock market prediction”, *J Big Data*, vol. 8, n<sup>o</sup> 1, p. 17, dez. 2021, doi: 10.1186/s40537-020-00400-y.
- [18] Maqbool, P. Aggarwal, R. Kaur, A. Mittal, e I. A. Ganaie, “Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor: A Machine Learning Approach”, *Procedia Comput Sci*, vol. 218, p. 1067–1078, 2023, doi: 10.1016/j.procs.2023.01.086.
- [19] Shi, Z. Teng, L. Wang, Y. Zhang, e A. Binder, “DeepClue: Visual interpretation of text-based deep stock prediction”, *IEEE Trans Knowl Data Eng*, vol. 31, n<sup>o</sup> 6, p. 1094–1108, jun. 2019, doi: 10.1109/TKDE.2018.2854193.
- [20] X. Ding, Y. Zhang, T. Liu, e J. Duan, “Using Structured Events to Predict Stock Price Movement: An Empirical Investigation”, em *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, p. 1415–1425. doi: 10.3115/v1/D14-1148.
- [21] X. Ding, Y. Zhang, T. Liu, e J. Duan, “Deep Learning for Event-Driven Stock Prediction”, *24th. International Joint Conferences on Artificial Intelligence Organization - IJCAI*, p. 2327–2333, 2015.
- [22] Y. Peng e H. Jiang, “Leverage Financial News to Predict Stock Price Movements Using Word Embeddings and Deep Neural Networks”, em *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, p. 374–379. doi: 10.18653/v1/N16-1041.
- [23] X. Zhang, Y. Zhang, S. Wang, Y. Yao, B. Fang, e P. S. Yu, “Improving stock market prediction via heterogeneous information fusion”, *Knowl Based Syst*, vol. 143, p. 236–247, mar. 2018, doi: 10.1016/j.knosys.2017.12.025.
- [24] K. Nti, A. F. Adekoya, e B. A. Weyori, “Efficient Stock-Market Prediction Using Ensemble Support Vector Machine”, *Open Computer Science*, vol. 10, n<sup>o</sup> 1, p. 153–163, jul. 2020, doi: 10.1515/comp-2020-0199.
- [25] X. Zhang, S. Qu, J. Huang, B. Fang, e P. Yu, “Stock Market Prediction via Multi-Source Multiple Instance Learning”, *IEEE Access*, vol. 6, p. 50720–50728, 2018, doi: 10.1109/ACCESS.2018.2869735.
- [26] X. Zhang, Y. Li, S. Wang, B. Fang, e P. S. Yu, “Enhancing stock market prediction with extended coupled hidden Markov model

- over multi-sourced data”, *Knowl Inf Syst*, vol. 61, n° 2, p. 1071–1090, nov. 2019, doi: 10.1007/s10115-018-1315-6.
- [27] Q. Zhang, L. Yang, e F. Zhou, “Attention enhanced long short-term memory network with multi-source heterogeneous information fusion: An application to BGI Genomics”, *Inf Sci (N Y)*, vol. 553, p. 305–330, abr. 2021, doi: 10.1016/J.INS.2020.10.023.
- [28] M. Bishop, *Pattern recognition and machine learning*. 2006.
- [29] Nabipour, P. Nayyeri, H. Jabani, S. S., e A. Mosavi, “Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis”, *IEEE Access*, vol. 8, p. 150199–150212, 2020, doi: 10.1109/ACCESS.2020.3015966.
- [30] K. Joshi, B. H. N, e J. Rao, “Stock Trend Prediction Using News Sentiment Analysis”, *International Journal of Computer Science and Information Technology*, vol. 8, n° 3, p. 67–76, jun. 2016, doi: 10.5121/ijcsit.2016.8306.
- [31] J. Briggs, “Sentiment Analysis for Stock Price Prediction in Python”, Towards Data Science. Accessed: June 10, 2024. [Online]. Available at: <https://towardsdatascience.com/sentiment-analysis-for-stock-price-prediction-in-python-bed40c65d178>
- [32] R. P. Schumaker e H. Chen, “Textual analysis of stock market prediction using breaking financial news”, *ACM Trans Inf Syst*, vol. 27, n° 2, p. 1–19, fev. 2009, doi: 10.1145/1462198.1462204.
- [33] Honnibal e I. Montani, “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing”. Accessed: August 10, 2024. [Online]. Available at: <https://spacy.io/>
- [34] R. J. A. Almeida, “LeIA - Léxico para Inferência Adaptada”, GitHub. Accessed: August 10, 2024. [Online]. Available at: <https://github.com/rafjaa/LeIA>
- [35] Hutto e E. Gilbert, “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text”, *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, n° 1, p. 216–225, maio 2014, doi: 10.1609/icwsm.v8i1.14550.
- [36] K. G. Kim, “Book Review: Deep Learning”, *Healthc Inform Res*, vol. 22, n° 4, p. 351, 2016, doi: 10.4258/hir.2016.22.4.351.
- [37] H. Liu e Z. Long, “An improved deep learning model for predicting stock market price time series”, *Digit Signal Process*, vol. 102, p. 102741, jul. 2020, doi: 10.1016/j.dsp.2020.102741.
- [38] F. Fama, “Efficient Capital Markets: A Review of Theory and Empirical Work”, *J Finance*, vol. 25, n° 2, p. 383, maio 1970, doi: 10.2307/2325486.
- [39] F. Fama, “Random Walks in Stock Market Prices”, *Financial Analysts Journal*, p. 55–59, 1965, Acessado: 5 de março de 2022. [Online]. Disponível em: <https://www.jstor.org/stable/4469865>
- [40] Statman, “Behavioral Finance versus Standard Finance”, *AIMR Conference Proceedings*, vol. 1995, n° 7, p. 14–22, dez. 1995, doi: 10.2469/cp.v1995.n7.4.
- [41] R. H. Thaler, “Behavioral Economics: Past, Present, and Future”, *American Economic Review*, vol. 106, n° 7, p. 1577–1600, jul. 2016, doi: 10.1257/aer.106.7.1577.
- [42] B. G. Malkiel e E. F. Fama, “Efficient Capital Markets: A Review of Theory and Empirical Work”, *J Finance*, vol. 25, n° 2, p. 383–417, maio 1970, doi: 10.1111/j.1540-6261.1970.tb00518.x.
- [43] F. Fama, “Efficient Capital Markets: II”, *J Finance*, vol. 46, n° 5, p. 1575–1617, dez. 1991, doi: 10.1111/j.1540-6261.1991.tb04636.x.
- [44] Y. Juwono, R. Sarno, R. N. E. Anggraini, A. T. Haryono and A. F. Septiyanto, “Comparative Study on Stock Price Forecasting Using Deep Learning Method Based on Combination Dataset,” *IEEE International Conference on Artificial Intelligence*, 2024, doi: 10.1109/AIMS61812.2024.10513288.
- [45] S. Russell, P. Norvig. *Artificial intelligence. Translation: Regina Célia Simille*. Rio de Janeiro: Elsevier, 2013.
- [46] Hochreiter, S., Schmidhuber, J., Long Short-Term Memory, *Neural Computation* (1997) 9 (8): 1735–1780. 1997. doi: 10.1162/neco.1997.9.8.1735



**Michele Jackeline Andressa Rosa** received a PhD in Applied Computing from the University of Vale do Rio dos Sinos. Bachelor of Economic Sciences from the State University of Mato Grosso. Master's degree in Regional Development and Agribusiness from the Federal University of Mato Grosso. MBA in Finance with emphasis on Capital Markets, Pitágoras University Unopar Anhanguera. Professor at State University of Mato Grosso and Economist.



**Marcos Roberto Souza** Bachelor in Computer Science from the University of Vale do Rio dos Sinos.



**Carlos Leandro Silva Machado** Bachelor in Computer Science from the University of Vale do Rio dos Sinos.



**Sandro José Rigo** bachelor's in computer science from the Pontifical Catholic University of Rio Grande do Sul. Master's degree in computer science from the Federal University of Rio Grande do Sul. PhD in Computer Science from the Federal University of Rio Grande do Sul. Postdoctoral fellow at the Friedrich Alexander University in Erlangen-Nuremberg Germany. Professor at the University of Vale do Rio dos Sinos and researcher at the Graduate Program in Applied Computing.



**Jorge Luis Victória Barbosa** received MSc and a PhD in computer science from the Federal University of Rio Grande do Sul, Porto Alegre, Brazil. He conducted post-doctoral studies at Sungkyunkwan University, Suwon, South Korea. Jorge is a full professor at the Applied Computing Graduate

Program of the University of Vale do Rio dos Sinos, head of the university's Mobile Computing Lab, and a researcher at the Brazilian Council for Scientific and Technological Development. His main research interests are mobile and ubiquitous computing, context prediction using context histories, mainly through similarity and pattern analysis, and ubiquitous computing applications mainly in health, accessibility, and learning.